BIBLIOTHEQUE
Institut de Phonétique
AIX-EN-PROVENCE

Groupement des Acousticiens de Langue Française



8èmes JOURNEES D'ETUDES du GROUPE de la " COMMUNICATION PARLEE "

Avec la participation
du CNRS et du GALF

VOLUME I
Textes des exposés

...es d'Etudes organisées par ... Aix-En-Provence
...TITUT DE PHONETIQUE ... 25-27 Mai 1977
... Aix-En-Provence

Groupement des Acousticiens de Langue Française

8èmes JOURNEES D'ETUDES du GROUPE de la " COMMUNICATION PARLEE "

Avec la participation
du CNRS et du GALF

VOLUME I
Textes des exposés

BIBLIOTHÈQUE
Institut de Phonétique
AIX-EN-PROVENCE

Journées d'Etudes organisées par
l'INSTITUT DE PHONETIQUE
d'Aix-En-Provence

Aix - En - Provence
25 - 27 Mai 1977.

Institut de Phonétique
Inventaire n° 1683
Cote n° A/JEP 8/1A

TABLE DES MATIERES

THEME I

VARIATIONS DE LA PERIODICITE

- Programme pour le traitement automatique des données obtenues par extraction du fondamental de la parole (application à la pathologie de la prosodie en psychiatrie et en neurologie).

  C. CHEVRIE-MULLER, P. DECANTE 5

- Incidence des variables situationnelles sur quelques paramètres de l'intonation.

  I. FONAGY, V. LUCCI, L.J. BOE 13

- Evaluation de l'état pathologique des cordes vocales d'après l'analyse des variations du fondamental.

  R. GUBRYNOWICZ, W. MIKIEL, P. ZARNECKI 21

- Variations à court terme de la fonction de source vocale.

  A. LANDERCY 29

REGULATION DE LA VIBRATION DES CORDES VOCALES

- La régulation de la vibration des cordes vocales : simulation à l'aide d'un modèle à deux masses.

  B. GUERIN, L.J. BOE 37

- Sur les corrélatifs physiologiques de la fréquence du fondamental de la parole.

  S. MAEDA 43

DETECTION DE Fo ET DU VOISEMENT

- Mesure du fondamental par filtrage variable.

  J. CAELEN, P. CAZENAVE 53

- Automate de segmentation et détection du fondamental.

  D. DOURS, R. FACCA, G. PERENNOU 61

- Détection de mélodie par un modèle fonctionnel du système auditif périphérique.

  L. BASTET, J.M. DOLMAZON, B. GUERIN, V. SCHUPLIAKOV (Léningrad) 71

- Détermination de la période fondamentale du signal vocal avec un filtre digital.

  W. HESS 79

- Détecteur numérique de mélodie.

  I. EL MALLAWANY, J.F. ZURCHER 87

- Un analyseur-visualisateur de mélodie à microprocesseur.
Ph. MARTIN 95

- Etude et réalisation d'un vocoder à codage direct de la mélodie.
S. EL SAMKARI, F. HAMELIN 103

- Un système de détection automatique du fondamental et de l'intensité.
B. TESTON, M. ROSSI 111

- La mesure du fondamental par la détection de crêtes:techniques employées, résultats.
J.F. ZURCHER 119

THEME II

PERCEPTION DE LA HAUTEUR ET DE SES VARIATIONS (MICROPROSODIE)

- Validation d'indices perceptifs de l'intonation par analyse multidimensionnelle.
F. CARTON, F. LONCHAMP 133

- La perception de l'intonation anglaise par des anglophones et neerlandophones.
R. COLLIER 139

- Les faits microprosodiques du français : voyelles, consonnes, coarticulation.
A. DI CRISTO, M. CHAFCOULOFF 147

- Le rôle de certains traits prosodiques dans la perception du message.
G. KONOPCZYNSKI 159

- Vers une base psychophonétique de la stylisation intonative.
J. 't HART 167

- On the interpretation of Raw fundamental frequency tracings.
N. THORSEN 175

- Quelques expériences d'analyse perceptuelle en français.
J. VAISSIERE 183

CORRELATIONS ENTRE L'INTENSITE, LA DUREE ET LA FREQUENCE FONDAMENTALE

- Relation entre l'évolution de la fréquence laryngienne et de l'intensité pour la phrase énonciative du français, premiers résultats.
P. GRAILLOT 193

ROLE DES TRAITS PROSODIQUES DANS LA TRANSMISSION ET LA SYNTHESE DE LA PAROLE

- Sur la commande par la glotte d'une prothèse des membres supérieurs.
P. ANDRE, F. LHOTE 203

- Propositions pour un modèle d'analyse de l'intonation.
  M. ROSSI, A. DI CRISTO 323
- Prise en compte d'informations prosodiques dans un modèle de compréhension du discours continu.
  J.P. HATON, J.F. MARI, G. MESSENET, J.M. PIERREL, C. SANCHEZ 332
- Les pauses et les groupes rythmiques : leur durée et distribution en fonction de la vitesse d'élocution.
  M. SAINT-BONNET, L.J. BOE 337
- Premiers essais d'utilisation de la durée pour la segmentation en mots dans un système de reconnaissance.
  J. VAISSIERE 345
- Utilisation, pour la reconnaissance de la parole continue, de marqueurs prosodiques extraits de la fréquence du fondamental.
  R. VIVES, C. LE CORRE, G. MERCIER, J. VAISSIERE 353

INTONATION ET COMMUNICATION PARLEE

- Analyse et rééducation des paramètres prosodiques chez l'enfant sourd.
  M.C. HATON, J.P. HATON 364
- De l'analyse phonologique à l'analyse phonétique des énonciatives en français.
  L. THERRIEN-ROY, D. JACQUEMIN, L.J. BOE 371

- Prosodie automatique pour la synthèse par diphonèmes.
  C. CHOPPY, J.S. LIENARD 211
- Analyses des redondances de systèmes symboliques et degrés de voisement.
  J. DREYFUS-GRAF 219
- Analyse de la structure intonative de la phrase en français.
  D. LARREUR, F. EMERARD 227
- Traits prosodiques distinctifs de certaines attitudes intellectuelles et émotives.
  I. FONAGY, J. SAP 237
- Utilisation d'un modèle prosodique pour la synthèse par concaténation de mots en français et en anglais.
  Ph. MARTIN 247
- La place des éléments voisés dans l'acquisition du langage.
  N. TIRANDAZ, C. BERGER-VACHON 255

THEME III

INTONATION ET SYNTAXE

- Traits, indices et propriétés du voisement : vers une phonologie des unités infra-phonémiques.
  C. ABRY, L.J. BOE 267
- Ligne mélodique et "que" complétif en français de Montréal.
  L. DAGENAIS, A. MARCHAL 275
- Expériences perceptives sur l'intonation.
  B. DELGUTTE 281
- Pour une détermination des fonctions grammaticales de l'intonation : analyse de quelques tests d'encodage et de décodage.
  M. C. HAZAEL-MASSIEUX 289
- L'intonation générative.
  D. HIRST 297
- Problèmes de neutralisation des marques prosodiques, application à la reconnaissance automatique.
  Ph. MARTIN 305

ROLE DE L'INTONATION DANS LA SEGMENTATION DE L'ENONCE ET LA RECONNAISSANCE DE LA PAROLE

- Corrélations entre mélodie, intensité et durée dans la phrase française énonciative.
  G. CAELEN, G. MAURAND 315

THEME 1

Variations de la périodicité

VIIIèmes JOURNEES D'ETUDE SUR LA PAROLE

Aix-en-Provence, 25-27 Mai 1977

PROGRAMME POUR LE TRAITEMENT AUTOMATIQUE DES DONNEES OBTENUES PAR EXTRACTION DU FONDAMENTAL DE LA PAROLE.(APPLICATION A LA PATHOLOGIE DE LA PROSODIE EN PSYCHIATRIE ET EN NEUROLOGIE).

C. CHEVRIE-MULLER

Maître de Recherches à l'INSERM

P. DECANTE Ingénieur à l'INSERM

• Résumé :

Une méthode d'analyse automatisée de deux paramètres acoustiques de la parole est présentée (fréquence fondamentale - débit). Le principe retenu est celui d'un prétraitement du signal de parole par le Melograph M G P2 (C.N.E.T.) suivi d'un échantillonnage de la courbe analogique par un analysateur multicanaux et d'un traitement statistique programmé sur ordinateur. Les variables suivantes sont obtenues : moyenne, écart-type, statistique de VON NEUMAN pour la fréquence fondamentale ; durée de parole et durée des pauses pour le débit.

• Summary :

A data processing method is shown here for two speech parameters (fundamental frequency - speech rate). Speech is preprocessed by a MG P2 Melograph (manufactured by C.N.E.T.). The analog output is sampled into a multichannel analyzer connected to a tape punch.
This tape is processed by a general purpose computer by a statistical program. The following variables are computed : mean, standard deviation, VON NEUMAN statistics of the pitch ; speech and pauses durations.

PROGRAMME POUR LE TRAITEMENT AUTOMATIQUE DES DONNEES OBTENUES PAR EXTRACTION DU FONDAMENTAL DE LA PAROLE (APPLICATION A LA PATHOLOGIE DE LA PROSODIE EN PSYCHIATRIE ET EN NEUROLOGIE).

C. CHEVRIE-MULLER

Maître de Recherches à l'INSERM

P. DECANTE

Ingénieur à l'INSERM

L'analyse des paramètres acoustiques support des informations prosodiques ou "supra-segmentales" a connu une période historique au cours de laquelle les mesures étaient effectuées manuellement sur des tracés (oscillographie, spectrographie...) (SNIDECOR 1943, BONNER 1943, HOLLIEN et COPPELAND 1965, MAHL et SCHULZ 1964, MAC GLONE 1966, CHEVRIE-MULLER 1967, JAFFE et RIEBER 1973).

Plus récemment l'analyse de la parole, normale et pathologique, a bénéficié des méthodes automatisées (compteurs et ordinateurs) permettant des études statistiques étendues (GOLDMAN-EISLER 1968, JAFFE et RIEBER 1973, HOLLIEN et JACKSON 1973, CHEVRIE-MULLER et DECANTE 1973, BOE et RAKOTOFIRINGA 1975).

On présentera ici la méthode utilisée au Laboratoire d'Etude de la Voix et de la Parole (U3-INSERM, Hôpital de la Salpétrière - PARIS) accompagnée d'applications à l'étude de malades neurologiques et psychiatriques.

1 - MATERIEL ANALYSÉ

- Il s'agit d'une phrase de 16 mots (21 syllabes) qui constitue le corpus.
- L'enregistrement est effectué en cabine insonorisée.

2 - METHODE

2.1. Contraintes imposées par le travail sur la parole pathologique.

- L'analyse porte sur des sujets variés : enfants, adultes, sujets âgés, des deux sexes.
- L'intensité de la voix est extrêmement variable d'un sujet à l'autre et parfois aussi chez un même sujet. Il en est de même de la fréquence fondamentale.
- Le signal de parole, du fait de troubles laryngés, peut être surchargé de souffle. La périodicité des ondes acoustiques peut disparaître transitoirement.

2.2. Dispositif adopté pour le pré-traitement du signal de parole : Melograph M G P2 (C.N.E.T. - Lannion).

Sa description ne sera pas reprise ici (cf. CHEVRIE-Muller 1973).

Il permet l'extraction du fondamental.

2.3. Echantillonnage de la courbe analogique par analyseur multi-canaux (INTERTECHNIQUE - HC 25 + BM 25).

Il a été effectué sur une base de temps de 50 millisecondes.

Les valeurs échantillonnées à la sortie analogique "pitchée" du Melograph sont stockées dans la mémoire. Elles peuvent-être obtenues sur imprimante et sur ruban perforé pour traitement ultérieur par ordinateur.

Cette procédure est une solution transitoire liée à des problèmes d'interfaçage. Un interfaçage est en cours qui relie la sortie binaire pure du Melograph à une entrée calculateur : Accès Direct Mémoire sur des mots de 16 bits ; le bit de poids fort indique la présence ou l'absence de pitch.

2.4. Traitement par programme des données recueillies.

2.4.1. Elimination des valeurs "non pertinentes"

Certaines valeurs des courbes analogiques correspondent à des fréquences très éloignées de la moyenne pour le corpus considéré. Il s'agit de phénomènes très brefs (<100ms), non périodiques, n'intervenant pas dans la perception d'une ligne mélodique.

Le programme élimine une valeur éloignée de 1 octave ou plus de la Fréquence fondamentale moyenne du corpus, dans le cas où l'échantillon correspondant à cette valeur est précédé et suivi d'un échantillon "non-pitché".

2.4.2. Calculs statistiques pour les paramètres Fréquence fondamentale (F.f.) et Débit (D.).

Les calculs sont effectués d'une part sur les valeurs correspondant aux tensions de la sortie logarithmique, d'autre part sur la transformation en Hz de ces valeurs. Un fichier-étalon est introduit pour permettre la transformation programmée.

2.4.2.1. Statistique sur F.f.

Le programme donne pour un corpus la moyenne, la variance, l'écart-type et la statistique de VON NEUMAN (1941 a et b).

VON NEUMAN a proposé d'utiliser à la place de la variance (S^2), la moyenne des carrés des différences successives

$$\Delta^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x})^2$$

Le rapport $\frac{\Delta^2}{S^2}$ (coefficient êta) permet d'apprécier l'influence respective d'un mouvement de tendance centrale (ici la ligne mélodique) et celle des mouvements aléatoires.

2.4.2.2. Statistique sur le débit

Le programme donne la durée "voisée" (d.v.), la durée "non voisée" (d.n.v.), la durée totale du corpus (d.t.). d.t. = d.v. + d.n.v.

Cependant d.n.v. comporte d'une part les pauses vraies (d. pauses) d'autre par les consonnes non voisées ou sourdes (d.c.s.).

Un histogramme des pauses est réalisé avec une classe $\leq$ 150 millisecondes et une classe $>$ 150 millisecondes. La somme des pauses $<$ 150 ms est considérée comme une approximation de d.c.s., la somme des pauses $>$ 150 ms comme une approximation de d. pauses.

On calcule enfin : d. parole =d.v. - d.c.s. , puis d.v/d.t. et d. parole /d.t..

3 - RESULTATS

3.1. Corrélations entre les mesures des paramètres acoustiques et la perception des caractères prosodiques.

3.1.1. Corrélation de la variance de F.f. avec l'appréciation subjective de la richesse intonative.

Un seul et même auditeur a apprécié pour chaque phrase enregistrée la "quantité de mélodie" selon une échelle à 9 pas (1 : absence de variations de hauteur ; 9 : variations de hauteur nettement exagérées ; 5 : variations moyennes, normales).

La corrélation est très significative : pour un nombre de phrases n = 75, r = . 60 - ce qui satisfait au seuil de . 01).

3.1.2. Corrélation entre la mesure de débit et l'appréciation subjective de la vitesse.

L'auditeur note la rapidité du débit de 1 : extrêmement lent à 9 : extrêmement rapide. Pour n_p = 51.

d.t. /percept. vitesse	d. parole/percept.vitesse	d. pauses/percept. vitesse
r= - . 53 **	r = - . 61 **	r = - . 28 **
* signif. au seuil . 05	** signif. au seuil . 01	

3.2. Analyse de paroles pathologiques.

3.2.1. Affections neurologiques : maladie de Parkinson

L'analyse de la parole de 11 cas de Parkinson est présentée à titre d'application de cette méthode. La parole des malades était diversement altérée et même normale pour un malade.

Pour chaque sujet on possède une appréciation (semi) quantitative des symptômes de la maladie et un jugement subjectif sur les paramètres prosodiques qui peuvent-être confrontés aux données acoustiques.

3.2.1.1. Analyse de la variabilité de F.f.

Variance de F.f. relativement élevée (pour les valeurs arbitraires en échelle log : S = 16 à 23) : il s'agit des 5 malades les moins gravement atteints par la maladie sur le plan général, leur parole étant également relativement préservée.

Variance basse (S = 7 à 12) : il s'agit de 6 malades dont la maladie est beaucoup plus sérieuse, la parole étant très gravement altérée dans 3 cas.

3.2.1.2. Analyse de la statistique de VON NEUMAN.

Parmi les 5 cas à variance basse le coefficient "êta" peut-être bas, traduisant la prédominance dans la variabilité de la "tendance centrale" c'est-à-dire de la ligne mélodique ; il s'agit de 3 cas comprenant le malade dont la parole est normale. Dans les 2 autres cas le coefficient est élevé par prédominance des mouvements aléatoires.

Parmi les 6 cas à variance basse le coefficient de VON NEUMAN a une valeur faible pour le sujet le moins atteint du groupe. Dans tous les autres cas il est élevé.

3.2.1.3. Analyse du Débit

On objective, par la mesure de d.T,d. parole et d. pause, trois cas de grande accélération de la parole (tachylalie parkinsonienne) avec réduction de durée des pauses et des mots

3.2.2. Affections psychiatriques : Schizophrénie.

Deux observations de malade à comportement mélodique opposé sont proposées :

	AGE	SEXE	MOY. en Hz	Ec.Type en Hz
1. S O L ...	20 ans	F	260	22
2. B R A ...	16 ans	F	250	8

Obs. SOL. : Variabilité importante de F.f. cliniquement : logorrhée, maniérisme, humeur et attention instables, "coqs à l'âne".

Obs. BRA. : Monotonie de F.f. cliniquement : contact superficiel, langage pauvre mimique inexpressive, inadaptée, syndrome de dépersonnalisation.

La variabilité ou la monotonie de l'intonation sont ici le reflet du comportement général du malade.

DISCUSSION - CONCLUSION

La Longueur restreinte du corpus mérite d'être discutée ainsi que le choix de la base de temps d'échantillonnage. La durée moyenne de nos corpus est de 4 sec, pour de telles durées la distribution des fréquences laryngiennes risque de ne pas être gaussienne (BOE et RAKOTOFIRINGA 1975). On remarquera cependant que la forte corrélation trouvée entre la variance de F.f. et la perception de la variabilité de hauteur par un auditeur apporte une validation à notre méthode, au moins pour une utilisation clinique. Il est de toute façon dans l'intention des auteurs d'allonger les corpus et de raccourcir la base de temps de l'échantillonnage.

Le seuil choisi pour attribuer la valeur de consonne non voisée (sourde) à une pause : 150 millisecondes, peut également être critiqué. Il a été fixé d'après des mesures réalisées sur des tracés oscillographiques de sujets normaux et en référence aux travaux de la littérature (GOLDMAN-EISLER 1958- BOOMER et DITTMAN 1962). La grande variabilité du débit en pathologie supposerait que le seuil de sélection des pauses soit pondéré en fonction de la vitesse d'élocution moyenne du sujet.

La méthode proposée devra donc être encore affinée, mais elle semble, dans son principe, apporter une première solution à l'analyse objective de la parole pathologique. Elle laisse entrevoir des possibilités non seulement pour une nosologie de la parole en neurologie et en psychiatrie, mais également pour la mesure de la gravité d'une atteinte, l'action ultérieure d'un traitement médicamenteux, rééducatif ou chirurgical.

Les tableaux de résultats, les figures et la bibliographie n'ont pu prendre place dans cette publication restreinte.

BOE (L.J.), CONTINI (M), RAKOTOFIRINGA (H).

Etude statistique de la fréquence laryngienne.
- Application à l'analyse et à la synthèse des faits prosodiques du français.

Phonetica 1975, 32, 1-23.

BOE (L.J.), RAKOTOFIRINGA (H.).

A statistical analysis of laryng-eal frequency : its relationship to intensity level and duration.

Language and Speech 1975, 18, 1-13.

BONNER (M.R.)

Changes in speech pattern under emotional tension.

Amer. J. Psychol. 1943, 56, 262-273.

CHEVRIE-MULLER (C.), GREMY (F.).

Contribution à l'établissement de quelques constantes physiologiques de la voix parlée.

J. Franç.O.R.L. 1967, 16, 149-154.

CHEVRIE-MULLER (C.), DECANTE (P.).

Etude de la fréquence fondamentale en pathologie (A. Méthode oscillographique, B. Extraction du fondamental).

Bull. Audiophonologie 1973, 3, 147-194.

GOLDMAN-EISLER (F.).

Psycholinguistics Experiments in Spontaneous Speech.

(Academic Press, London 1968).

HOLLIEN (H.), COPPELAND (R.H.).

Speakin g Fundamental Frequency Characteristics of Mongoloid Girls.

J. Speach Hear. Disord. 1965, 30, 344-349.

HOLLIEN (H.), JACKSON (B.).

Normative data on the speaking fundamental frequency characteristics of young adult males.

J. of Phonetics 1973, 1, 117-120.

JAFFE (J.); ANDERSON (S.W.); RIEBER (R.W.).

Research and clinical approaches to disorders of speech rate.

J. of Commun. Disord.1973, 6,225-246.

MAC GLONE (R.E.)

Vocal pitch characteristics of children aged one and two years.

Speech Monograph. 1966, 33, 178-181.

MAHL (F.F.), SCHULZ (G.).

Psychological Research in the Extralinguistic area, in

Approach to Semiotics (Mouton edit. La Hogue 1964).

SNIDECOR (J.C.).

A comparative study of the pitch and duration characteristics of impromptu speaking and oral reading.

Speech Monograph 1943, 10, 50-56.

VON NEUMAN (J.).

Distribution of the ratio of the mean square successive difference to the variance.

Ann. Math. Statist. 1941, 12, 367-395.

VON NEUMAN (J.), KENT (R.H.), BELLINSON (N.R.), HART (B.I.).

The mean square successive difference.

Ann. Math. Statist. 1941, 12, 153-162.

TABLEAU N° 1

ANALYSE DE LA FREQUENCE FONDAMENTALE — ANALYSE CLINIQUE

Sujets N°	F.f. HZ	F.f. log.	S log.	Δ log.	$\frac{1}{2}\frac{\Delta^2}{S^2}$ brut	$\frac{1}{2}\frac{\Delta^2}{S^2}$ cor	Mélodie (écoute) 1 = nulle 4 = import.	SEXE	Tremblement	Masque facial	Akinésie	Rigidité	Piétinement
GROUPE A(1)													
141.1	128.	121.	18.65	15.79	0.36	0.19	4	M	*	0	0	(*)	0
082.1	138.	132.	17.55	15.13	0.37	0.16	3	M	*	***	**	***	*
130.2	135.	129.	17.71	18.45	0.54	0.39	3	M	*(*)	**	*(*)	**	**
092.2	181.	159.	18.94	25.83	0.92	0.90	3	F	0	**	*	**	0
156.2	272.	199.	17.75	23.44	0.87	0.67	4	M	0	**	*	(*)	*
GROUPE B(2)													
059.1	116.	112.	8.66	6.49	0.28	0.28	2	M	**	***	**	**	**
063.1	242.	186.	10.24	11.22	0.60	0.31	2	F	***	***	***	**	***
078.4	169.	155.	8.96	10.38	0.67	0.70	2	F	*	****	****	(*)	-
097.2	180.	161.	7.17	9.39	0.86	0.70	2	M	0	***	***	***	***
158.1	204.	172.	7.97	10.06	0.80	0.60	2	F	**	***	***	***	****
129.1	190.	166.	6.71	8.50	0.80	0.87	2	F	0	***	***	**	***

(1) Groupe A = Sujets avec variance élevée

(2) Groupe B = Sujets avec variance basse

VIIIèmes JOURNEES D'ETUDE SUR LA PAROLE

Aix-en-Provence, 25-27 Mai 1977

INCIDENCE DES VARIABLES SITUATIONNELLES SUR QUELQUES PARAMETRES DE L'INTONATION

I. FONAGY & V. LUCCI - L.J. BOË

Institut d'Etudes Linguistiques et Phonétiques de la Sorbonne Nouvelle

Institut de Phonétique de Grenoble

Résumé :

Les auteurs ont étudié l'incidence de quelques variantes linguistiques situationnelles, telles que l'exposé, la lecture, l'interview, sur la variabilité intonative du message (valeur moyenne de la fréquence laryngienne, registre, intervalles inter et intrasyllabiques, indices configuratifs). Une interprétation de cette variabilité est proposée à partir des caractéristiques situationnelles ou spécifiques de chacune des variantes.

Summary :

The authors have studied the incidence of a few linguistic variants (such as the account, the reading and the interview) on the intonative variability of the message (mean value of the laryngeal frequency, register, frequency range inter or intrasyllabic, intonation patterns). An interpretation of this variability is proposed from situational or specific characteristics of each item.

INFLUENCE DES VARIABLES SITUATIONNELLES SUR QUELQUES PARAMETRES DE L'INTONATION

I. FONAGY V. LUCCI L.J. BOË

Nous nous proposons d'examiner l'incidence de la nature situationnelle du message (exposé, lecture, interview) sur les réalisations intonatives dans la chaîne parlée. Les quatre locuteurs (tous universitaires) retenus, ont fourni un échantillon différent (90 secondes en moyenne) pour chacun de ces trois types de messages (soit 12 extraits d'une longueur totale de 18 minutes).

1) L'analyse automatique

Dans une première étape nous avons, pour chaque extrait, analysé automatiquement[1] la fréquence laryngienne moyenne ainsi que la fonction de répartition, le registre et les écarts types (tableau I). Le registre fixé à $\pm 2\sigma$ soit 98% du temps total. La dernière colonne du tableau I indique les écarts (en tons) entre les moyennes du fondamental a) lecture - exposé, b) lecture - interview.

Remarque

$\overline{F_0}$ décroit régulièrement de l'exposé à la lecture et de la lecture à l'interview. Par ailleurs, le registre varie dans le même sens, unanimement, avec les trois variantes. Les variations de la moyenne du fondamental, comme celles du registre intonatif, paraissent corrélées aux types situationnels examinés, puisqu'elles ont toujours lieu de façon identique, pour l'ensemble des 12 extraits.

Cette analyse automatique ne nous renseigne que globalement sur la variabilité intonative. Pour une meilleure connaissance du mécanisme étudié, nous avons effectué, complémentairement, une analyse quantitative manuelle, ainsi qu'une analyse de quelques indices configuratifs les plus représentatifs.

2) Analyse quantitative manuelle

a) Fréquence moyenne des points "idiosyncratiques"

Ces mesures ont été effectuées à partir de six échantillons du corpus[2] (2 sujets). Nous avons déterminé les moyennes des fréquences maximales à l'intérieur d'une syllabe, et celles de "noyaux syllabiques", coïncidant avec le

---

1. Mesures effectuées à l'Institut de Phonétique de Grenoble (cf. L.J. BOË & H. RAKOTOFIRINGA, 1975).

2. Mesures effectuées à l'Institut d'Etudes Linguistiques et Phonétiques de la Sorbonne Nouvelle, avec le détecteur de fréquence laryngée construit par B. GAUTHERON.

sommet d'intensité de la syllabe[1]. Ces fréquences, que nous nommerons "idiosyncratiques", varient dans le même sens que la distribution globale des fréquences fondamentales, puisqu'elles vont croissant de l'interview à la lecture, de la lecture à l'exposé (cf. tableau II).

b) Intervalles intersyllabiques et intrasyllabiques

- Intervalles intersyllabiques.

Ces intervalles ont été déterminés par paires de syllabes, en prenant comme niveau de référence, soit la fréquence maximale à l'intérieur des syllabes, soit la fréquence mesurée au point de l'intensité maximale. Dans ces cas, les variations de la gamme moyenne ne vont pas dans le même sens que celles du registre global puisqu'il y a accroissement de l'interview à l'exposé, de l'exposé à la lecture (tableau III). La gamme totale de l'exposé dépasse largement celle de la lecture. Ceci revient à dire que la gamme plus restreinte de la lecture a été mieux exploitée par nos deux sujets que la gamme plus large de l'exposé.

- Intervalles intrasyllabiques.

Ils sont plus réduits dans l'interview (tableau III) par rapport à l'exposé et la lecture.

3) Indices configuratifs

La lecture.

Se distingue, avant tout, par sa haute régularité qui se manifeste à différents niveaux. Les schémas qui dominent la lecture sont peu nombreux, et par conséquent récurrents. L'arc mélodique reliant deux syllabes accentuées est particulièrement fréquent (VAISSIERE, 1971; FONAGY, 1956, pp. 196-202). Les syllabes accentuées flottent autour du même niveau de hauteur, et l'écart type des fréquences maximales est par conséquent moins élevé que celui de l'interview ou de l'exposé. A ces récurrences s'ajoutent des redondances d'ordre grammatical (fins de phrases marquées par une chute tonale à un niveau constant, pré-indication des propositions, reproduction de phrases enchâssées dans l'espace tonal).

L'exposé.

Les syntagmes sont rarement mis en relief dans l'exposé, où les stéréotypes mélodiques priment sur le contenu du message et neutralisent les divergences syntaxiques. Par ailleurs, la dislocation des unités syntaxiques et sémantiques par des moyens prosodiques (pauses, etc.) est de règle dans l'exposé. Il faut signaler la fréquence élevée du cliché prosodique que constitue la clausule, caractérisée par un accent très marqué frappant l'antépénultième syllabe et par une montée précipitée dans cette syllabe suivie d'une chute brusque.

---

[1]. Des tests perceptifs indiquent que ces deux points sont particulièrement importants pour la perception de la hauteur d'une syllabe (cf. I. FONAGY, 1974).

L'interview.

La réduction de la gamme tonale, la baisse de la fréquence moyenne, l'absence de structure montante-descendante à l'intérieur de la phrase, les rares apparitions d'arc tonal confèrent à l'interview un aspect plutôt monotone que vient parfois troubler une recherche de changements peu prévisibles à partir de la structure syntaxique (ton pouvant s'élever brusquement en fin de phrase assertive après une longue séquence autour d'un même niveau tonal). Les montées ne présument pas une continuation (DELATTRE, 1966). On relève par ailleurs les clichés caractéristiques suivants : fusion des deux phrases indépendantes à l'aide d'une intonation montante-descendante, montée du ton dans deux syllabes successives accentuées (FONAGY & BERARD, 1973), montée de 4 demi-tons suivie d'une descente de 3 demi-tons.

4) Tentative d'interprétation

La variabilité intonative quantitative et qualitative relevée dans les trois types de messages examinés, peut être expliquée en prenant en compte les caractéristiques situationnelles des types de messages, ainsi que d'autres caractéristiques spécifiques de chacun d'entre eux.

Les paramètres situationnels.

La nature même de l'exposé étant de s'adresser à un groupe d'individus, dans des conditions qui ne sont jamais optimales, on rencontre, en même temps qu'une intensité générale plus marquée, une élévation de la moyenne du fondamental liée à une augmentation de la pression d'air sous-glottique. Par ailleurs, pout être perçues, les variations doivent être plus marquées, à syntaxe égale (l'interview est caractérisée, de ce fait, par un registre moins large). La proximité physique des interlocuteurs permet au locuteur, dans l'interview, de se contenter d'une gamme relativement restreinte de variations mélodiques. Le conférencier, par contre, est amené, du fait de l'éloignement physique et psychologique plus marqué, à se servir de moyens vocaux propres à tenir l'attention en éveil, utilisant pour cela un registre intonatif plus large.

Si la lecture possède un registre intonatif plus large par rapport à l'interview (les deux types de messages étant produits dans des situations proches, puisque la lecture n'était pas "publique"), c'est qu'il y a conditionnement à partir d'autres caractéristiques propres à chacun des types de messages.

Les caractéres spécifiques.

Le message lu est généralement plus complexe au niveau informatif et syntaxique. Une phrase plus longue et plus structurée exige plus d'espace sonore, une plus large gamme de fréquence (FONAGY & MAGDICS, 1963). D'autre part, les pauses étant sensiblement plus importantes dans l'interview (LUCCI, 1974), contrairement à la lecture, il est prévisible que l'intonation se voie attribuer un rôle accru dans cette dernière variante. Par ailleurs, dans la lecture, les accents didactiques sont plus fréquents que dans l'interview (LUCCI, 1975) et contribuent à augmenter les contrastes intonatifs.

Rappelons enfin que l'interview est caractérisée par une fréquence élevée de "pauses sonores" (hésitations, allongements de E muets), contrai-

rement à la lecture (LUCCI, 1974, 1976). Ces phénomènes de "ralentissement du débit" agissent sur l'intonation en "bloquant" la variabilité de celle-ci puisque le niveau fréquentiel, dans ce cas, est au repos, dans une zone basse.

REFERENCES

BOË, L.J. & RAKOTOFIRINGA, H. (1975). A Statistical Analysis of Laryngeal Frequency : its Relationship to Intensity Level and Duration. Language and Speech 18, pp. 1-13.

DELATTRE, P. (1966). Les dix intonations de base du français. French Review 40, pp. 1-14.

FONAGY, I. (1956). Uber den Verlauf des Lautwandels. Acta Linguistica Hung - 6, pp. 173-278.

FONAGY, I. (1974). La voix du poète. Budapest : Akademiai Kiado.

FONAGY, I & BERARD, E. (1974). Questions totales simples et implicatives. Studia Phonetica 8, pp. 53-97.

FONAGY, I. & MAGDICS, K. (1963). Paradoxon der Sprechmelodie. Ural Altaische Jahbücher 35, pp. 1-55.

LUCCI, V. (1974). Rythme et longueur du message parlé. Bulletin de l'Institut de Phonétique de Grenoble 4, pp. 139-152.

LUCCI, V. (1975). L'accent didactique en français parlé et français lu. Cahiers de Linguistique et d'Orientalisme et de Slavistique, Université de Provence.

LUCCI, V. (1976). Le mécanisme du E muet dans différentes formes de français parlé. La Linguistique Vol. 12, Fasc. 2.

VAISSIERE, J. (1971). Contribution à la synthèse par règles du français. Thèse de 3° Cycle. Université des Langues et Lettres de Grenoble.

	$\overline{F_\ell}$ Hz	$\sigma(F_\ell)$ Hz	$\frac{\sigma(F_\ell)}{\overline{F_\ell}}$ %	$\pm 2\sigma$ tons	Ecart tons $\overline{F_\ell}$
Loc. H exposé	143	29	20,3	7,4	0,3
Loc. H lecture	138	24	17,4	6,3	1,0
Loc. H interview	123	20	16,3	5,8	
Loc. J exposé	252	65	25,8	9,9	0,57
Loc. J lecture	236	52,5	22,2	8,3	0,41
Loc. J interview	225	42,5	18,9	6,9	
Loc. XI exposé	257	62,5	24,3	9,2	1,07
Loc. XI lecture	227	47,5	20,9	7,7	0,47
Loc. XI interview	215	45	20,9	7,7	
Loc. X2 exposé	240	50	20,8	7,7	0,75
Loc. X2 lecture	220	45	20,4	7,5	0,40
Loc. X3 interview	210	39	18,6	6,7	

Tableau I

	Moyenne des hauteurs maximales (1/2 tons)	Moyenne des hauteurs aux points des sommets d'intensité (1/2 tons)
Locuteur A.B.		
Conversation	14	13
Exposé	16	15
Cours lu	14	14
Locuteur F.R.		
Conversation	10	9,5
Exposé	13	12
Cour lu	11	10

Tableau II

Moyennes de la hauteur syllabique dans trois types de discours chez deux sujets féminins, déterminées en demi-tons à partir du niveau de fréquence inférieur des deux sujets

	Ecart de fréquence moyen entre les sommets de hauteur des syllabes consécutives (1/2 tons)	Ecart de fréquence moyen entre les points d'intensité maximale des syllabes consécutives (1/2 tons)	Ecart de fréquence intrasyllabique
Locuteur A.B.			
Conversation	2,2	2,1	1,7
Exposé	2,4	2,6	3
Cours lu	3,1	2,8	2,3
Locuteur F.R.			
Conversation	1,8	1,7	1,6
Exposé	3,2	2,7	2,3
Cours lu	3,4	3,2	2,3

Tableau III

Ecart de fréquence intersyllabique et intrasyllabique en demi-tons dans trois types de discours chez deux locuteurs (1759 syllabes).

VIII èmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

EVALUATION DE L'ETAT PATHOLOGIQUE DES CORDES VOCALES D'APRES L'ANALYSE DES VARIATIONS DU FONDAMENTAL

R. GUBRYNOWICZ, W. MIKIEL et P. ŻARNECKI

Laboratoire de l'Acoustique de la Parole
Institut de Recherches Fondamentales de Technologie
Académie Polonaise des Sciences
rue Świętokrzyska 21, 00-049 VARSOVIE

Résumé :

Un système d'analyse des variations de la fréquence fondamentale a été élaboré. A l'aide de ce système nous avons étudié les variations de la fréquence fondamentale pour des sujets ayant les cordes vocales partiellement paralysées. A part d'une analyse qualitative de l'influence de l'état pathologique des cordes vocales sur la mélodie de la parole, nous avons effectué une analyse statistique de variations de la fréquence fondamentale à long terme. L'objectif de notre étude est d'élaborer une méthode objective d'évaluation de l'état des cordes vocales après les intervention chirurgicales et au cours du traitement de réadaptation.

Summary :

A mini-computer processing system has been developped for studies of the fundamental frequency variations of pathological voices. A qualitative analysis of the influence of vocal cords diseases /especially partial paralysis/ on FO contours and also on long-term statistical distribution of FO variations were carried out. The aim of our work is to elaborate an objective method of analysis which allows on appreciation of the state of vocal cords after surgical treatement and during the process of the rehabilitation.

EVALUATION DE L'ETAT PATHOLOGIQUE DES CORDES VOCALES D'APRES L'ANALYSE DES VARIATIONS DU FONDAMENTAL

R. GUBRYNOWICZ, W. MIKIEL, P. ŻARNECKI
Laboratoire de l'Acoustique de la Parole
Institut de Recherches Fondamentales de Technologie
Académie Polonaise des Sciences
rue Świętokrzyska 21, 00-049 VARSOVIE

1. INTRODUCTION

Cette étude fait partie du projet de réalisation d'un système d'analyse acoustique de la parole aidant le diagnostic médical de l'appareil phonatoire /1, 2/. L'un des premiers objectifs que nous nous étions fixés, était de fournir aux phoniatres un outil suffisamment simple et souple, travaillant en temps réel ou avec un petit décalage dans le temps. D'autre part ce système doit donner au médecin des moyens d'analyse objectives et plus puissants que les traditionnels, et permettre d'évaluer avec précision les changements de la voix après un traitement médical et au cours du traitement de réadaptation.

Les avantages de la méthode acoustique d'analyse de l'appareil phonatoire sont considérables. En premier lieu, on évite l'introduction dans l'organe phonatoire des sondes, capteurs etc. qui troublent les conditions naturelles de la phonation. En second, on peut étudier les fragments de la parole continue choisis à volonté et d'une durée convenable pour des analyses à court, soit à longue terme.

Les travaux ont débutés vers la fin du 1972 par la construction du "FONOSPEKTROSKOP" conçu spécialement pour les recherches phoniatriques et qui comportait un analyseur à canaux avec un système de visualisation, un circuit de mesure de niveaux de signaux analysés et un détecteur de mélodie. Quelques années d'expériences effectuées en coopération étroite avec les phoniatres de l'Académie de Médecine de Varsovie, nous ont permis de préciser les buts et les méthodes d'analyse du diagnostic acoustic de l'appareil phonatoire.

Actuellement des recherches sont en cours sur deux thèmes principaux:

a/ analyse du signal de la parole pour des conduits vocals pathologiques /2/,
b/ évaluation de l'état du larynx par l'analyse des variations de la fréquence fondamentale.

Les recherches concernant le laryhx ont commencé par l'étude systématique des rélations entre les variations de la fréquence fondamentale et de l'état pathologique des cordes vocales. Outre le système de mesure utilisé à ces recherches nous présenterons dans cette communication les résultats préliminaires d'analyse des variations du fondamental, obtenus pour des sujets ayant une ou les deux cordes vocales atteintes d'une paralysie partielle, ou une totalement immobilisée /latérofixation/. Nous avons répété les mêmes

---

Les recherches décrites dans cette communication sont menées dans le cadre du contrat M.R. 10.4.3.02

analyses pour certains de ces sujets qui ont subi un traitement médical.

2. METHODE ET SYSTEME DE MESURE DU FONDAMENTAL

Afin d'effectuer les recherches nous avons construit un système de mesure du fondamental et de traitement de données représenté sur la fig. 1.

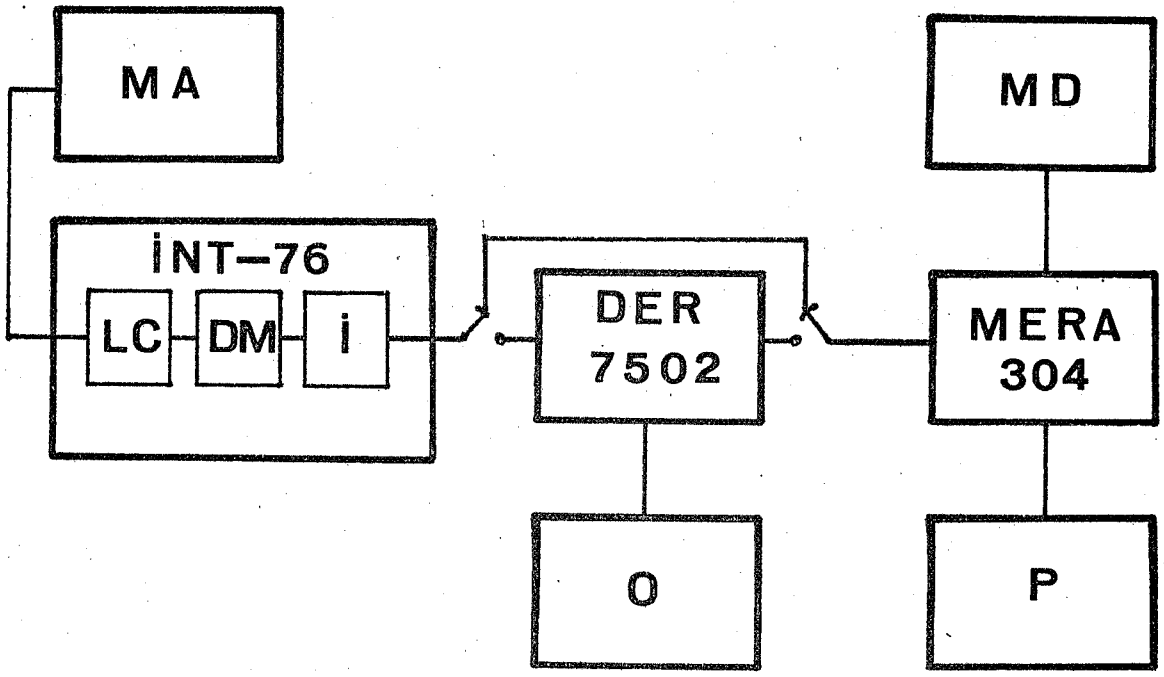


FIGURE 1.

LEGENDE:

MA - Magnétophone
INT - Intonographe
LC - Limiteur-compresseur
I - Interface
O - Oscilloscope
DER - Digital Event Recorder B&K
MERA - Mini-ordinateur
P - Imprimante alpha-numérique
MD - Mémoire à disque
DM - Détecteur de mélodie

Il se compose d'un détecteur de mélodie /DM/ équipé d'un bloc de numérisation et d'un interface pour la transmission de données en temps réel au miniordinateur muni d'une mémoire à disque.

Le détecteur de mélodie qui est la troisième version améliorée de celui-là élaboré pour "FONOSPEKTROSKOP", relève la valeur de la période du fondamental toutes les 12.5 ms et fonctionne pour une gamme de fréquences fondamentales entre 80 et 500 Hz. L'erreur de mesure est en moyenne 1.4% et pour les fréquences entre 80 et 200 Hz ne dépasse pas 1%. Le principe de fonctionnement du détecteur de mélodie est suivant.

Le signal reçu par le micro ou enregistré sur bande magnétique est amplifié et appliqué sur l'étage de préfiltrage qui est un filtre passe-bas avec la fréquence limite 80 Hz et la pente

environ 12 dB/oct. Un compresseur d'amplitude réduit de 30 à 3 dB la dynamique du signal d'entrée du filtre d'asservissement qui poursuit le fondamental. Le signal de sortie pratiquement dépourvu ainsi des harmoniques supérieures est appliqué sur l'entrée du circuit de mesure de l'intervalle entre deux passages à zéro successifs. Les résultats de mesure sont envoyés au miniordinateur. Le détecteur de mélodie a également une sortie analogique qui est pourvue d'un indicateur et d'un système d'enregistrement sur papier.

Un programme de saisie de données conversationnel permet d'enregistrer sur disque en temps réel les variations de la fréquence fondamentale. Au cours du traitement qui suit immédiatement la saisie de données on peut utiliser un ensemble de procédures qui permettent: d'enregistrer à l'aide d'une imprimante alpha-numérique les évolutions du fondamental et de ses accroissements déterminés pour chaque intervalle de quantisation; de calculer et de visualiser les histogrammes pour ces deux variables /FO et ΔFO/ ; et de calculer leurs paramètres de distribution - moyennes, écarts types, moments centrals, asymétrie, excès etc.

3. LE MATÉRIEL ANALYSÉ

Chaque sujet, après l'examen médical à la suite duquel le type d'altération des cordes vocales était établi sans équivoque, a enregistré un texte composé: a/ d'une série de six voyelles isolées, b/ d'un ensemble de phrases à la forme affirmative et interrogative, c/ et d'un extrait d'un journal d'une durée de 2 minutes.

Tous les enregistrements ont été effectués à l'aide d'un microphone de bonne qualitée, dans une cabine amortie et bien isolée de bruits extérieurs. Pour les enregistrements on a utilisé le magnétophone NAGRA IV et les bandes magnétiques BASF. On a recueilli un ensemble d'enregistrements pour 15 voix pathologiques et 3 voix normales.

4. RESULTAT OBTENUS

A titre d'example on a représenté sur la fig. 2 l'évolution de la fréquence fondamentale relevée pendant la prononciation des deux dernières syllabes d'une phrase interrogative par une femme chez laquelle la corde vocale gauche était paralysée à cause de la strumectomie /2a/. La fig. 2b présente l'évolution de la fréquence fondamentale pendant la prononciation des mêmes syllabes par la même femme après l'operation supplémentaire et le traitement de réadaptation. On constate que la malade a eu une grande difficulté de reproduire pour l'interrogation le schema mélodique montant. Cette constation a une valeur générale et nous avons observé chez d'autres sujets malades que la montée de la fréquence fondamentale est substituée par un allongement de la durée de dernières syllabes de la phrase interrogative.

Sur la fig. 3 nous avons présenté les résultats de l'analyse statistique à long terme des variations de la fréquence fondamental obtenus pour le même sujet. On voit que après le traitement médical il y a un abaissement net de la valeur moyenne de la fréquence fondamentale /env. de 16% /. L'amplitude des variations de la fréquence fondamentale a aussi diminué mais dans une proportion plus faible.

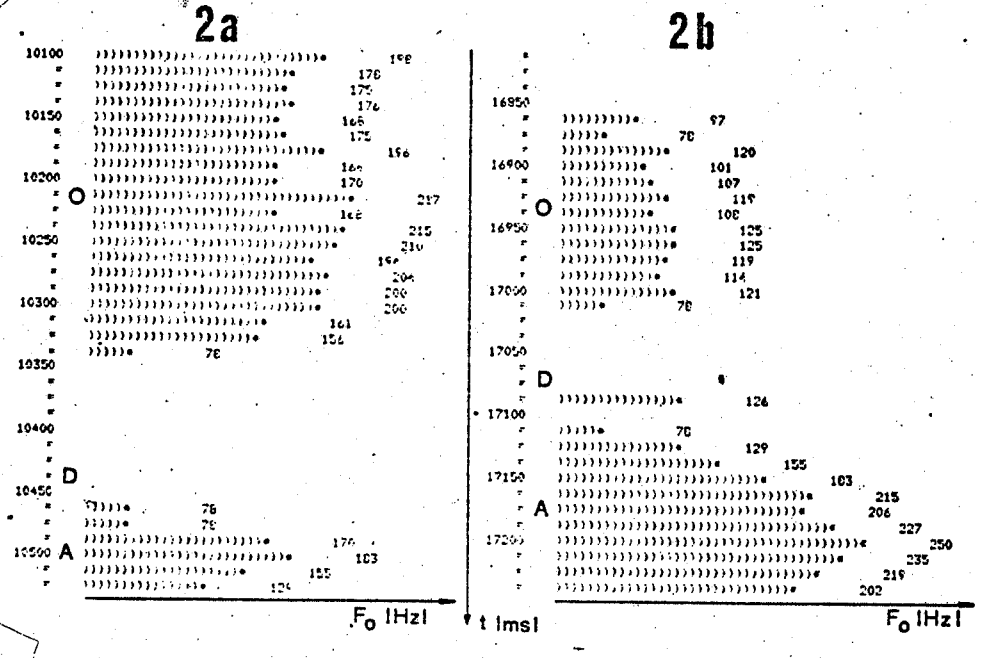
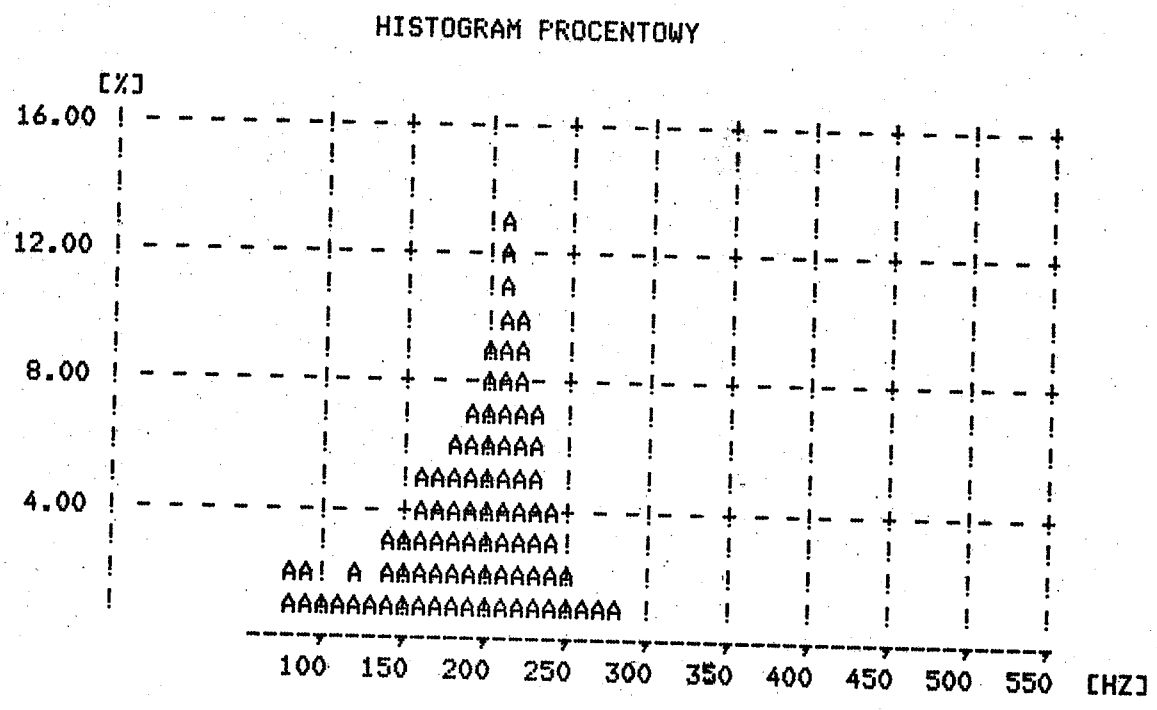


FIGURE 2.

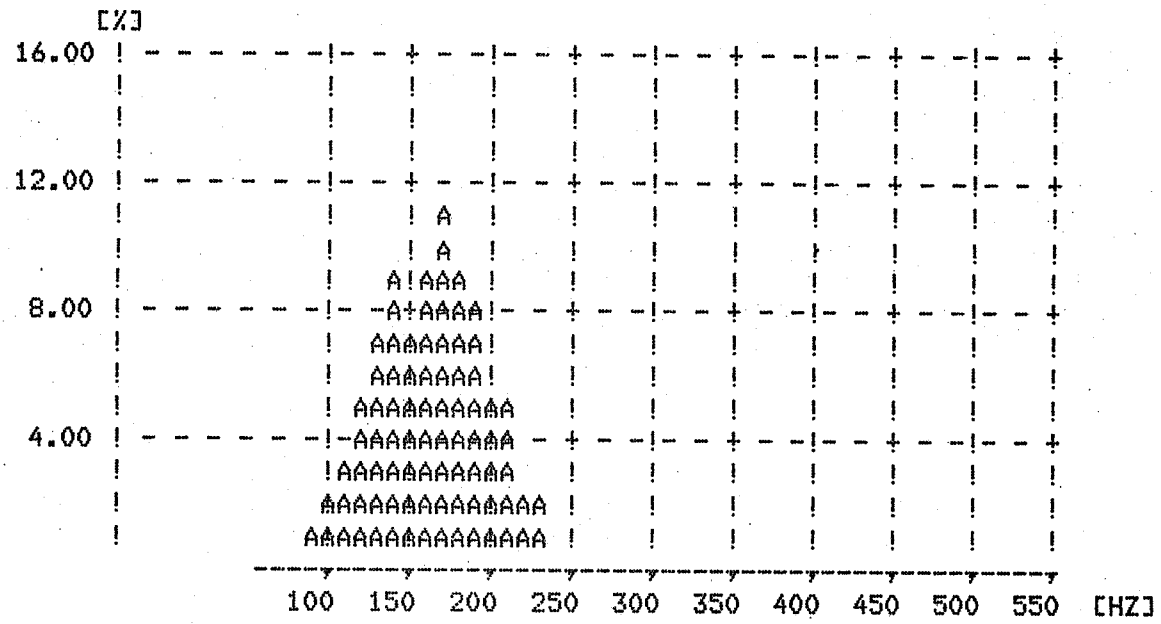


FO moyen = 199 Hz , Ecart type = 57 Hz

APRES LA STRUMECTOMIE

FIGURE 3A.

HISTOGRAM PROCENTOWY



FO moyen = 171 Hz, Ecart type = 51 Hz

APRES LE TRAITEMENT MEDICAL ET DE READAPTATION

FIGURE 3B.

Pour d'autres cas pathologiques nous avons constaté une forte corrélation entre la forme, le paramètres de distribution statistique des variations de FO et le type de paralysie des cordes vocales et leur position réciproque.

Le cas de maladie cité plus tôt était rélativement léger. Pour les cas plus graves nous avons observé une dispersion des valeurs de FO beaucoup plus importante, surtout pour les fragments initiaux et finals de phonation. En plus, s'est produit simultanément une réduction du temps d'articulation de phonèmes et un allongement des pauses aspiratoires. En général, le nombre des pauses augmente. En conclusion, on peut affirmer que la méthode d'analyse des variations de la fréquence fondamentale s'est relevée particulièrement utile pour l'évaluation des progrès au cours de traitement médical et de réadaptation de la voix.

Tous nos remerciements aux Docteur W. Tłuchowski et ses collaborateurs de l'Académie de Médecine de Varsovie pour leur coopération à la réalisation des enregistrements et à la discussion des problèmes médicaux.

BIBLIOGRAPHIE

/1/ TŁUCHOWSKI W., MIKIEL W., NIEDŹWIECKI A., KOMOROWSKA A.: Une méthode nouvelle d'analyse phonospectroscopique utilisée dans la phoniatrie clinique - Conférence de l'Union Européenne de Phoniatrie, Interlaken 1974.

/ 2/ KACPROWSKI J., MIKIEL W., SZEWCZYK A.: Acoustical modelling of cleft palate, Archives of Acoustics N° 2, v. 2 /1976/, 137 - 158.

VIIIèmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

VARIATIONS A COURT TERME DE LA FONCTION DE SOURCE VOCALE

A. LANDERCY

UNIVERSITE DE MONS

Résumé :

Une méthode numérique d'analyse par synthèse est utilisée pour étudier les variations dans le fonctionnement de la source vocale. Des variations de la fréquence fondamentale et de forme de l'impulsion glottique sont mises en évidence au sein de vocoïdes choisis pour leur extrême stabilité. L'importance de ces variations sur l'audition a été vérifiée par des tests perceptifs.

Summary :

A numerical analysis-by-synthesis method is used to study the variations in the functioning of the vocal source. Variations in the fundamental frequency and in the shape of the glottal pulse are shown up within each really stable vowel sound. The importance of these variations on audition has been shown by perception tests.

Variations à court terme de la fonction de source vocale

1. Introduction

Le pourcentage d'intelligibilité des sons de parole (synthétiques ou naturels) dépendant essentiellement de la fonction de transfert des cavités supraglottiques, l'étude de la source vocale a été négligée pendant de nombreuses années. Certains travaux ont cependant démontré que la fonction de source joue un rôle important pour le naturel et les caractéristiques individuelles de la voix (7, 9, 11). Par analogie, on a pu mettre en évidence l'influence de la pression et des dimensions des cordes vocales sur la forme et la périodicité de l'onde glottique (1, 10). Ces variations de forme et de fréquence de l'onde glottique se traduisent, au niveau spectral, par des variations d'amplitude, de fréquence et de phase qui doivent jouer un rôle au niveau perceptif. Aussi, nous sommes-nous proposé, à partir de sons de parole naturelle, d'étudier les variations à court terme de l'onde laryngienne et de mettre en évidence l'importance de ces variations sur la perception des sons.

2. Méthode expérimentale

Nous avons analysé des parties de vocoïdes oraux prononcés de manière soutenue - séquences sonores habituellement considérées comme extrêmement stables - afin de poser l'hypothèse de la constance de la fonction de transfert des cavités supraglottiques. Ces sons ont été échantillonnés période par période (3) et les éléments ainsi obtenus ont été traités numériquement de manière à en extraire les coefficients de Fourier ainsi que les valeurs de la fréquence fondamentale et de l'intensité.

L'examen des résultats permet effectivement de constater - de période à période - des variations de l'amplitude, de la fréquence et de la phase des harmoniques, variations dont l'importance augmente avec la fréquence (5).

Afin d'obtenir la répartition spectrale de l'impulsion glottique, nous avons employé une méthode numérique d'analyse par synthèse qui nous a permis d'obtenir le spectre de la fonction de source S(f) en extrayant la fonction de transfert des cavités T(f) de la fonction de parole aux lèvres P(f). Pour visualiser simplement les variations d'une période à l'autre, nous avons reconstitué, par transformée inverse, la forme temporelle de l'onde glottique (4).

3. Analyse des résultats

3.1. Variations de la fréquence fondamentale

Malgré l'extrême stabilité des échantillons traités, les analyses ont mis en évidence de légères variations de la fréquence fondamentale. Sur dix périodes consécutives, elles atteignent généralement $\pm$ 1,5% de la valeur de la fondamentale moyenne et paraissent aléatoires (fig. 1). Si nous tenons compte de la précision de la mesure, due à la période d'échantillonnage du signal (60 μs), et que nous analysons les variations qui interviennent au sein de trois périodes consécutives, nous constatons que, dans

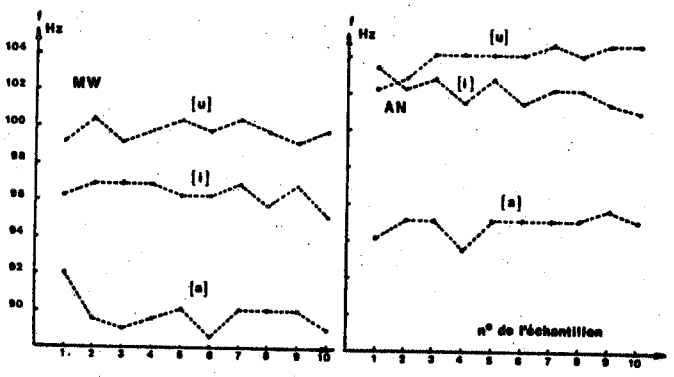


Fig. 1.

45% des cas, les variations sont supérieures à 120 µs. Lieberman (7), sur un corpus de voix parlée, avait trouvé que dans 86% des cas envisagés, la période varie de plus de 100 µs dans ces conditions. Si nous appliquons à notre corpus les conclusions tirées par cet auteur, nous pouvons supposer que les variations de la fréquence fondamentale décelées sur la fonction de parole aux lèvres P(f) sont essentiellement attribuables à des variations dans la périodicité des vibrations des cordes vocales et que, par conséquent, nous ne devrions pas constater de modifications importantes de la forme de l'impulsion glottique au sein d'un même échantillon.

3.2. Variations de l'enveloppe spectrale de l'onde de source

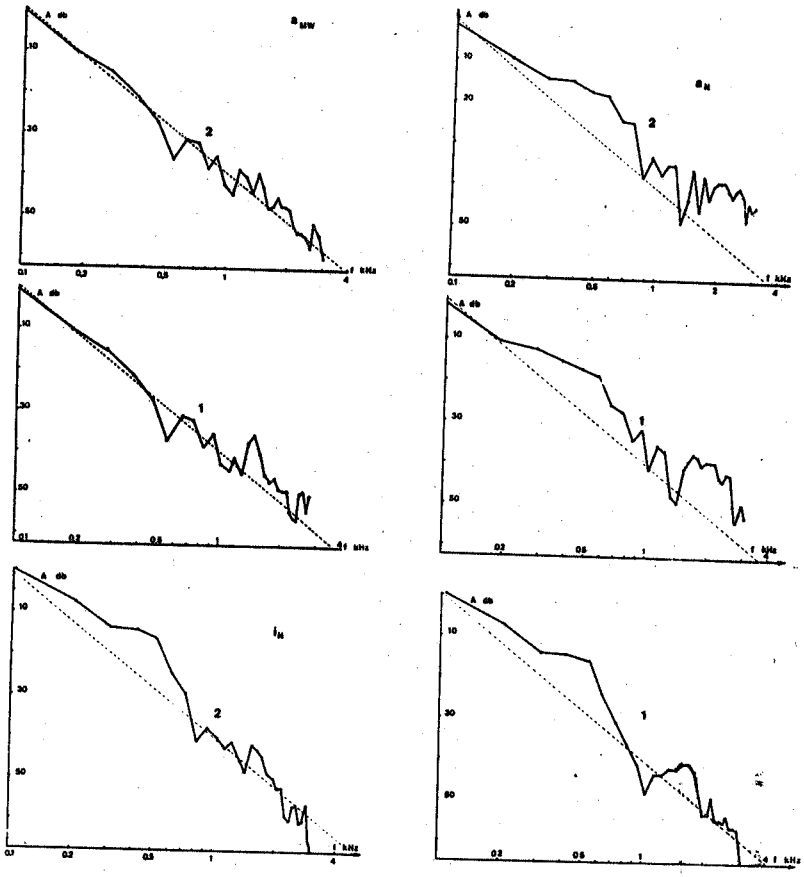


Fig. 2.

De manière générale, les enveloppes spectrales (fig. 2) présentent l'aspect d'une ligne brisée, dont la pente moyenne est de 12 dB/octave. Si nous tenons compte des informations de phase, nos résultats confirment les recherches antérieures (2, 10) indiquant que le spectre de source est à non minimum de phase et présente de nombreux minima spectraux.

Si nous comparons les enveloppes spectrales d'une période à l'autre, dans la partie du spectre, - en deça de 1000 Hz, - où se concentre la majorité de l'énergie, on observe très peu de variations; les minima qui se présentent dans cette gamme de fréquences sont stables, aussi bien du point de vue de leur fréquence que de leur acuïté. Pour les fréquences supérieures, des différences importantes d'intensité et de phase peuvent se manifester pour des harmoniques de même rang, l'ampleur de ces différences variant selon le locuteur et le son prononcé. Enfin, au sein d'un même vocoïde, les zones fréquentielles des principaux minima varient peu d'une période à l'autre.

En deçà de 1000 Hz, nous pouvons donc supposer que le spectre

de source est caractéristique du locuteur et relativement indépendant du vocoïde.

3.3. Variations de forme de l'impulsion glottique

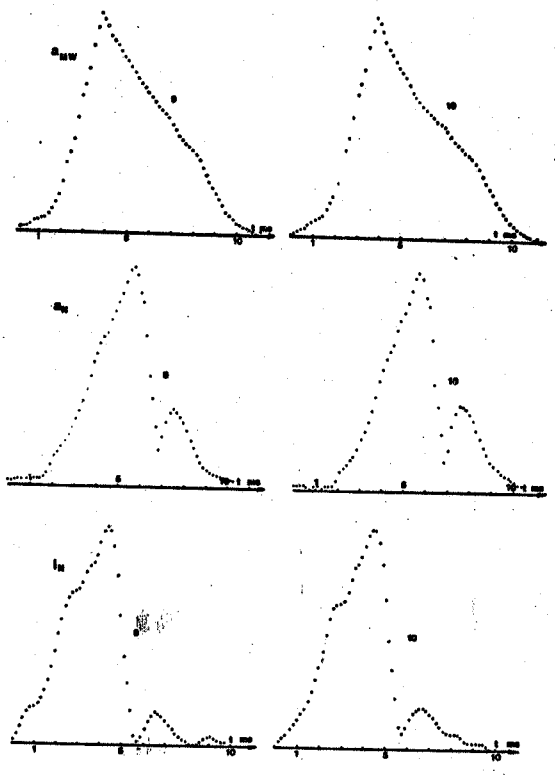


Fig. 3.

De manière générale, et conformément aux résultats obtenus antérieurement (2, 8, 9, 10), nous constatons que (fig. 3) :
a) les impulsions glottiques présentent toutes une dissymétrie entre les fronts d'ouverture et de fermeture;
b) pendant l'occlusion des cordes vocales, des rebondissements peuvent apparaître;
c) la forme générale semble être typique du locuteur et indépendante du vocoïde.

Afin de mettre en évidence les différences qui interviennent de période à période, nous avons, à partir d'une approximation triangulaire, mesuré les temps de montée t1 et de descente t2, la durée de l'impulsion to ainsi que la période du phénomène T; nous avons également calculé les rapports k = t2/t1 entre les temps de descente t2 et de montée t1 ainsi que les rapports R = T/to entre la période et la durée de l'impulsion.

D'une période à l'autre, les variations de forme sont relativement peu importantes. Cependant, sur dix périodes consécutives d'un même vocoïde, les variations deviennent plus conséquentes :

	Variations autour de la moyenne	
	Minimales	Maximales
t1 temps de montée	4 %	7 %
t2 temps de descente	2 %	13 %
to durée de l'impulsion	2 %	7 %
k = t2/t1	5 %	20 %
R = T/to	2 %	15 %

Nous constatons donc que, même sur un corpus choisi pour son extrême stabilité, des variations de la période et de la forme apparaissent dans l'onde glottique. Nous avons tenté de mettre en évidence leur influence sur la perception.

4. Analyse perceptive

Diverses expériences psycho-acoustiques (9, 11) ont prouvé que des changements fondamentaux de la forme de l'onde glottique étaient nettement perçus par les sujets. Nous avons pu montrer que de tels changements modifiaient considérablement l'allure générale du spectre de sonie (5), ce qui explique la différenciation par

les sujets. Par contre, les variations à court terme ne perturbent pas, - ou très peu, - le spectre de sonie et ces légères modifications semblent être imputables davantage à la fréquence fondamentale qu'à la forme. Afin de vérifier séparément l'influence des modulations de forme et de fréquence fondamentale sur la perception, nous avons construit un test auditif à partir de matériels naturels et synthétiques.

Les stimuli synthétiques ont été construits à partir d'une fonction de source triangulaire dont les paramètres (t1, t2, T) représentaient la valeur moyenne de l'impulsion naturelle. Couplés avec des fonctions de transfert correctes, les spectres de sonie de ces sons synthétiques reproduisaient fidèlement ceux des sons naturels (des vérifications au niveau du spectre et de la forme d'onde temporelle étaient également effectuées).

Nous avons demandé aux sujets de comparer entre eux trois types de stimuli :

a) naturels (t1, t2 et T variables);
b) synthétiques, présentés à fréquence fondamentale fixe (t1, t2 et T fixes);
c) synthétiques, pilotés en fréquence fondamentale par les stimuli naturels correspondants (t1, t2 fixes, T variable).

En ce qui concerne les comparaisons entre les stimuli b et c, 91% des sujets sont capables de les différencier. Par contre, pour les paires de stimuli a et c, 57% seulement peuvent les différencier, ce résultat n'étant pas significatif au seuil .05.

En d'autres termes, à partir de l'échantillon envisagé, nous observons que les sujets sont capables de détecter la faible modulation de la fréquence fondamentale (T), mais ils ne perçoivent pas les modulations de la forme de l'onde glottique (t1, t2).

Conclusions

La méthode d'analyse utilisée permet de visualiser simplement la forme de l'onde glottique et de mesurer les variations qui s'y produisent. Du point de vue perceptif, il semble que si la forme générale reste la même, les modifications des temps de montée et de descente ne sont pas détectées. Par contre , des modulations de 1,5% de la fréquence fondamentale permettent aux sujets de différencier les stimuli.

BIBLIOGRAPHIE

(1) FLANAGAN, J.L., Some Properties of the Glottal Sound Source, J. Speech Hear. Res., 1, 1958, 99-116.

(2) FLANAGAN, J.L., LANDGRAFT, L.L., Self-Oscillating Source for Vocal Tract Synthesisers, IEEE Trans. on Audio and Electroacoustics, AU 16, 1, 1968.

(3) LANDERCY, A., Temporal Segmentation. - Influence on the Envelope Function on Perception, J. Speech Hear. Res. 14, 1971, 47-57.

(4) LANDERCY, A., Contribution à l'étude de la source vocale et de ses variations à court terme, Thèse de doctorat, Univ. de Bruxelles, 1973.

(5) LANDERCY, A., Les variations à court terme dans les vocoïdes, R.A. Inst. Phon. de Bruxelles, 7/1, 1973, 73-91.

(6) LANDERCY, A., Les paramètres acoustiques de l'audition et la perception des sons de la parole, Revue de Phonétique Appliquée, 25, 1973, 3-23.

(7) LIEBERMAN, Ph., Perturbations in Vocal Pitch, J. Acoust. Soc. Am., 33, 1961, 597-603.

(8) MATHEWS, M.V., MILLER, J.E., DAVID, E.E., Pitch Synchronous Analysis of Voiced Sounds, J. Acoust. Soc. Am. 33, 1961, 179-186.

(9) NAKATSUI, M., SUZUKI, J., Method of Observation of Glottal Source Wave Using Digital Inverse Filtering in Time Domain, J. Acoust. Soc. Am., 47, 1970, 664-665.

(10) PAILLE, J., BEAUVIOLA, J.P., CARRE, R., Synthèse de la parole: description et utilisation d'un synthétiseur du type "à formants", Revue de Physique Appliquée, 5, 1970, 785-793.

(11) ROSENBERG, A.E., Effect of Glottal Pulse Shape on the Quality of Natural Vowels, J. Acoust. Soc. Am., 49, 1971, 583-590.

Régulation de la vibration des cordes vocales

TESTON Bernard
10, Butte des 3 Moulins
13 - AIX-EN-PROVENCE

VIII èmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

LA REGULATION DE LA VIBRATION DES CORDES VOCALES : SIMULATION A L'AIDE D'UN MODELE A DEUX MASSES

Bernard GUERIN
Laboratoire de la Communication parlée*
E.N.S.E.R. - GRENOBLE

et

Louis-Jean BOË
Institut de Phonétique
GRENOBLE

• Résumé :

Les modifications du régime de fonctionnement du larynx peuvent être imputées à des variations de la pression subglottique et/ou de la tension des cordes vocales. A l'aide d'un modèle à deux masses (simulé numériquement), les auteurs précisent les influences respectives de ces deux paramètres. Les résultats permettent :
. de proposer un modèle de commande physiologique utilisable en synthèse ;
. d'induire les commandes physiologiques qui sous-tendent les paramètres supra-segmentaux.

• Summary :

Subglottal pressure and/or stiffness of vocal cords are the two main physiological control parameters of laryngeal frequency. With a two-mass model (digital simulation) the respective influence of these commands has been investigated. Results allow to :
. submit a physiologic model of command to be used in synthesis ;
. predict the part which plays each physiologic parameter in the prosody of speech.

* Equipe de recherche associée au C.N.R.S.

LA REGULATION DE LA VIBRATION DES CORDES VOCALES : SIMULATION A L'AIDE D'UN MODELE A DEUX MASSES

1/ INTRODUCTION

L'évolution de la fréquence laryngienne peut être imputable pour l'essentiel à des variations de la pression subglottique P_S, relativement commode à mesurer, ou à des modifications de la tension des cordes vocales, assez difficiles à induire de l'E.P.G. (par ailleurs peu aisée à mettre en oeuvre).

Dans l'étude de l'intonation, la part respective de ces deux commandes a été largement discutée (LIEBERMAN, 1967 ; LADEFOGED, 1967) ; pour essayer de la mieux préciser, nous avons utilisé le modèle à deux masses (ISHIZAKA et MATSUDAIRA, 1968) qui a l'avantage, par rapport à d'autres modèles proposés, d'être à la fois relativement simple et de simuler le fonctionnement des cordes vocales dans des limites relativement satisfaisantes.

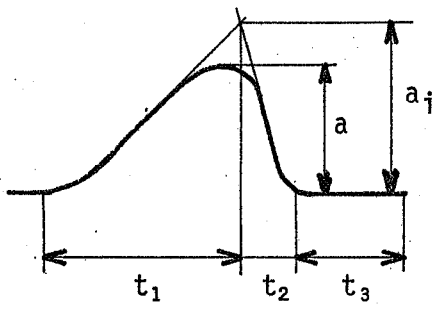
2/ PROCEDURE

2.1. LES DONNEES DE DEPART

Nous avons rassemblé un certain nombre de résultats concernant le fonctionnement du larynx ; ils constituent une base de départ satisfaisante dans la mesure où, obtenus par des méthodes souvent très différentes, ils se recoupent relativement bien.

2.1.1. L'ONDE DE DEBIT

Sa connaissance est essentielle pour pouvoir juger du bien-fondé de la modélisation. Seules des méthodes indirectes permettent de l'induire à partir :
. du signal acoustique : filtrage inverse,
. des variations de l'aire de la glotte : mesure sur un film laryngoscopique ou utilisation de la glottographie photo-électrique.



$$T.O. = \frac{t_1 + t_2}{t_1 + t_2 + t_3} \quad \frac{t_1}{t_2} > 2$$

$$0{,}40 \leq T.O. \leq 0{,}85$$

$$\frac{a}{a_1} \simeq 0{,}8$$

$$F_\ell = \frac{1}{t_1 + t_2 + t_3}$$ (on en connaît très bien les limites).

2.1.2. LA PRESSION SUBGLOTTIQUE

La mesure de P_S a été faite par de multiples procédés (canule de trachéotomie, trocart piqué dans la trachée, ballonnet oesophagien, sonde introduite sous la glotte par le nez, préparation de larynx excisés...) et sa relation avec F_ℓ est assez bien connue. Selon les auteurs :

$$F_\ell = k_1 P_S + k_2 \qquad k_1 = 5 \text{ Hz/cm d'}H_2O$$

ou $$\log F_\ell = k_1' P_S + k_2' \qquad k_1' = \frac{1}{2} \text{ octave/7,5 cm d'}H_2O \text{ .}$$

2.1.3. LA TENSION DES CORDES VOCALES

Il faut bien voir qu'il existe deux sortes de tension :

. l'une est passive : elle résulte de l'étirement du ligament thyro-aryténoïdien sous l'effet de la contraction des muscles crico-thyroïdiens (C.T.) ;

. l'autre est active : elle correspond à la contraction de la partie interne des muscles thyro-aryténoïdiens (T.A.).

Les études E.M.G. montrent bien que l'action des C.T. et T.A. sont pour l'essentiel simultanées (SHIPP et al., 1971) ; d'autres renseignements complémentaires peuvent être apportés par l'étude des variations de la longueur et de l'épaisseur des cordes vocales (HOLLIEN, 1974). En effet, la tension passive a pour conséquence de les allonger en les étirant alors que seule la tension active se traduirait par une diminution de leur longueur et une augmentation de leur épaisseur dans la partie centrale.

2.2. SIMULATION DU MODELE

2.2.1. RESOLUTION

Nous sommes partis de l'étude d'ISHIZAKA et FLANAGAN (1972). Le modèle a été simulé numériquement et chargé par un circuit figurant l'impédance d'entrée du conduit vocal correspondant à la voyelle neutre (F_1 = 500, F_2 = 1500). Le comportement de la glotte peut être représenté par un circuit électrique analogue (figure 1). On aboutit ainsi à deux systèmes d'équations différentielles non linéaires à coefficients non constants. Pour résoudre chacun d'eux, nous avons utilisé la méthode de RUNGE KUTTA à quatre approximations, le système des relations mécaniques et aérodynamiques étant transformé respectivement en 4 et 3 équations du premier ordre. L'ensemble a été implanté sur un PDP 11/20 ; autour du programme principal de simulation des cordes vocales, on a organisé un ensemble de sous-programmes utilitaires permettant de modifier les paramètres de la source, de réaliser les entrées/sorties nécessaires à cette simulation (tracé des déplacements des masses et de l'onde de débit, figure 2), de calculer F_ℓ, le T.O. et la puissance du débit.

FIGURE 1.
Circuit électrique équivalent au modèle de la source vocale chargé par l'impédance d'entrée du conduit vocal.

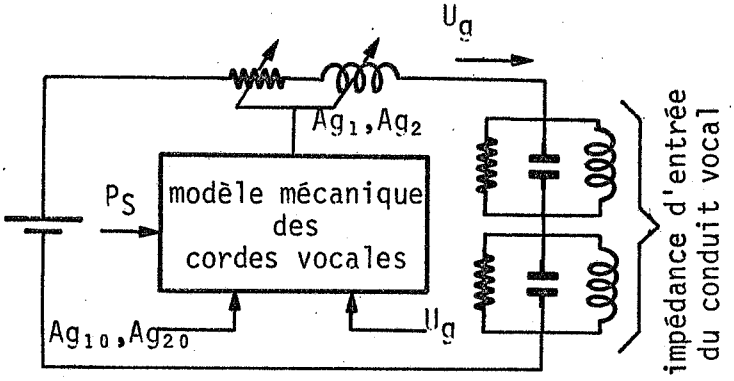
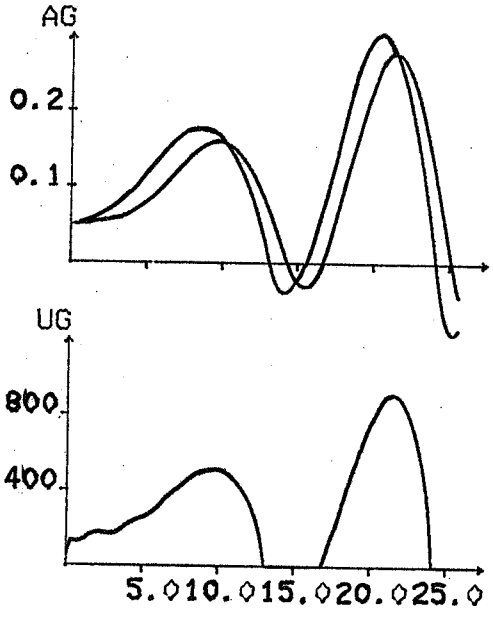


FIGURE 2.
Aire d'ouverture des cordes vocales et onde de débit, obtenues par simulation.



2.2.2. DETERMINATION DES PARAMETRES DE COMMANDE

Dans un premier temps, il a fallu préciser quantitativement les valeurs des paramètres de commande qui permettent d'obtenir, pour un domaine de fréquences correspondant à une voix d'homme (70 à 200 Hz), une onde débit conforme à celle qui a été relevée à partir de signal acoustique ou de la fonction aire de la glotte. Tout d'abord, nous avons fait varier simultanément *tous* les éléments de l'ensemble de commande $\{C\} = \{ P_S , m_1 , m_2 , k_1 , k_2 , k_{12} \}$ dans les gammes suivantes :

P_S	m_1	m_2	k_1	k_2	k_{12}
2 à 19	100 à 380	10 à 70	10 à 80	2 à 10	6,5 à 30
cm d'H_2O	mg		kdyn / cm		

A l'inverse du point de vue adopté jusque-là, nous avons fait évoluer m_1 et m_2 sans que leur somme ne reste constante. En effet, le modèle simule pour l'essentiel le comportement de la partie médiane de cordes vocales et celle-ci évolue en épaisseur (et donc en masse) avec la tension. Parmi toutes les simulations, nous en avons retenu 167 : *celles qui correspondent à une onde de débit de forme satisfaisante*. Nous avons opéré sur ces $\{C\}$ un traitement statistique : l'analyse de régressions linéaires nous permet de constater l'existence de relations entre les commandes, à savoir :

$m_2 = 0,14\ m_1 + 5,9$ $\qquad$ $k_1 = 1,8.10^{-2}\ m_1 + 9,9$

$k_1 = -\ 1,0.10^{-2}\ m_1 + 57$ $\qquad$ $k_{12} = -\ 3,7.10^{-2}\ m_1 + 27$

et plus particulièrement, on relève en moyenne la relation : $\frac{m_2}{m_1} = \frac{k_2}{k_1} = 0,17$.

A partir de ces corrélations qui reflètent les contraintes mécaniques du modèle, il est possible de proposer un paramètres de contrôle Q, qui donne une bonne approximation des liaisons constatées (GUERIN et BOË, 1977) :

$$m_1 = \frac{M}{Q} \quad ; \quad \frac{m_2}{m_1} = \frac{k_2}{k_1} = 0,15 \quad ; \quad k_1 = Q \,.\, K \quad ; \quad k_{12} = 5\ Q + 10$$

$$M = 380 \text{ mg} \quad ; \quad K = 57 \text{ kdyn/cm}$$

Nous avons opéré 45 simulations avec : P_S = 2 à 19 cm d'H_2O

$Q = 1 - 1,52 - 2,71$.

Q	$F_\ell = f(P_S)$	T.O. $= f(P_S)$	N.I. $= f(P_S)$
1	$F_\ell = 2,9\ P_S + 62$	T.O. $= 0,002\ P_S + 52$	N.I. $= 0,16\ P_S + 19$
1,52	$F_\ell = 2,9\ P_S + 86$	T.O. $= -\ 0,17\ P_S + 54$	N.I. $= 0,53\ P_S + 27$
2,71	$F_\ell = 3,9\ P_S + 137$	T.O. $= -\ 0,33\ P_S + 59$	N.I. $= 0,57\ P_S + 26$

3/ CONCLUSIONS

De l'étude attentive des résultats indiqués ci-dessus, on peut tirer un certain nombre de remarques en comparaison de mesures sur la parole réelle.

En effet, on peut constater que :
. la relation de F_ℓ avec P_S varie entre 2,5 et 3,9 Hz/cm d'H_2O ;
. le T.O. évolue entre 0,50 et 0,60 : il dépend d'autant moins de P_S que Q est faible, c'est à dire que la masse des cordes vocales est élevée ;
. P_S et le niveau d'intensité (N.I.) sont liés à peu près identiquement pour Q = 2,71 et 1,52, puis plus faiblement.

Ces résultats concordent assez bien avec ceux qui ont été relevés par observation du larynx et témoignent en faveur de la validité quantitative du modèle.

En ce qui concerne les influences respectives de P_S et Q sur F_ℓ, on peut remarquer sur la régression globale :

$$F_\ell = 3{,}13\ P_S + 50{,}5\ Q + 6{,}86$$

- que si P_S évolue de 2 à 19 cm d'H_2O, F_ℓ varie de 50 Hz ;
- que si Q évolue de 1 à 2,71, F_ℓ varie de 86 Hz .

Bien que nous n'ayons pas beaucoup de renseignements chiffrés sur la dynamique de Q, on peut noter que F_ℓ est plus sensible à des variations de ce qui représente une tension qu'à la pression subglottique. C'est d'ailleurs ce qui semble ressortir des travaux de SHIPP et al.

Quant au niveau d'intensité :

$$N.I. = 0{,}57\ P_S - 1{,}06\ Q + 28{,}7$$

il est essentiellement lié aux variations de P_S.

L'observation simultanée des évolutions temporelles de la fréquence laryngienne et du niveau d'intensité de la parole naturelle devrait permettre de préciser la part respective de ces deux types de commande.

BIBLIOGRAPHIE

GUERIN B., BOË L.J. (1977)
A two-model of the vocal cords : determination of control parameters and their respective consequences
I.E.E.E. Int. Conf. on Acoustics, Speech and Signal process, Hartford Conn. (à paraître)

HOLLIEN H. (1974)
On vocal registers - Journal of Phonetics, 2, 125-143

ISHIZAKA K., FLANAGAN J.L. (1972)
Synthesis of voiced sounds from a two-mass model of the vocal cords - B.S.T.J., 54, 485-506

ISHIZAKA K., MATSUDAIRA M. (1968)
What makes the vocal cords vibrate - 6th Int.Congr.Acoust., Tokyo - B1-3

LADEFOGED P. (1967)
Three areas of experimental phonetics, Chap.1 - Oxford Univ.Press, London

LIEBERMAN Ph. (1967)
Intonation, perception and language - M.I.T. Press, Cambridge Mass.

SHIPP T., McGLONE R., MORISSEY P. (1971)
Some physiologic correlates of voice frequency change - 7th Int. Congr. Phonetic Sci., Montréal, 407-411.

VIIIèmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

SUR LES CORRELATIFS PHYSIOLOGIQUES DE LA FREQUENCE DU FONDAMENTAL DE LA PAROLE

SOME PHYSIOLOGICAL CORRELATES OF FUNDAMENTAL FREQUENCY OF SPEECH

S. MAEDA C.N.E.T. LANNION

• Résumé :

Une analyse de contours de fréquence de fondamental (Fo) a montré que ces contours peuvent être analysés comme la superposition de deux composants : (1) des mouvements de Fo locaux : une séquence de mouvements de Fo vers le haut et vers le bas, qui composent des sortes de "hat-patterns", et (2) un mouvement non local, constitué par une diminution progressive de Fo le long des phrases et des constituants : cette pente est appelée "la ligne de déclinaison" (base-line).

Dans cet article, nous discuterons surtout du mécanisme physiologique possible sous-jacent à la ligne de base. Une expérience de cinéradiographie semble montrer qu'un écourtement progressif des cordes vocales est le correlat physiologique de la ligne de déclinaison.

• Summary :

An analysis of the fundamental frequency (Fo) contours of speech indicates that the contours may be analyzed into two components : (1) localized contours, i. e., a sequence of up-and-down Fo movements composing the so-called "hat-pattern", and (2) a non-localized one, we way call it a "baseline", which is a gradual Fo falling along a sentence. In this paper, we shall discuss primarily on a possible physiological mechanism that generates the baseline. A Cineradiographic experiment indicates that a gradual shortening of the vocal folds seems to be the physiological correlate of the baseline.

SUR LES CORRELATIFS PHYSIOLOGIQUES DE LA FREQUENCE DU FONDAMENTAL DE LA PAROLE

S. MAEDA C.N.E.T. LANNION

I - INTRODUCTION

Une étude de contours de la fréquence du fondamental (Fo) dans des phrases déclaratives lues par des locuteurs Anglo-américains a montré que ceux-ci pouvaient être schématisés par deux classes d'attributs = (1) la première classe désigne des mouvements de Fo locaux, tel que une "montée" (attribut R - de l'anglais "Raise"), une"descente" (attribut L - de l'anglais "Lowering"), un "pic" (attribut P - de l'anglais "Peak") qui est souvent associé avec l'attribut R, - et une pente ascendante (attribut R_1 - de l'anglais "Rise") sur un plateau de Fo entre R et L, et (2) la seconde classe désigne un mouvement non local, une ligne de déclinaison graduelle le long de la phrase, l'attribut BL (de l'anglais "Baseline") (Maeda, 1974). La figure 1 illustre la schématisation idéalisée d'un contour de base. La combinaison de R (Montée) et de L (Descente) forme un plateau supérieur de Fo (appelé un "hat-pattern" (Cohen et t'Hart, 1967)). Les symboles représentant les attributs, tels que BL, R et L, sont utilisés pour décrire les contours à l'endroit même où les mouvements qui leur correspondent se produisent dans la phrase (voir Fig. 1). La structure de contours peut être décrit en termes d'attributs par l'utilisation d'un simple réseau de transition (voir Fig. 2). Le réseau permet de rendre compte de toutes les séquences d'attributs attestées dans notre corpus.

On peut cependant se demander si une telle description de Fo correspond en quelque manière à un processus de parole existant dans l'être humain, ou si elle n'est rien d'autre qu'une des approximations géométriques possibles pour décrire les courbes de Fo. Par exemple, Olive (1973) a proposé une tout autre façon de décrire les contours de Fo. C'est cette question qui nous a amené à investiger les corrélations physiologiques des attributs. Si ces attributs sont une manifestation d'un certain processus de parole, nous devrions alors trouver, à un niveau physiologique, des corrélations de ces attributs. Dans cette communication nous allons parler en particulier des corrélations physiologiques de la ligne de déclinaison, qui semble être un fait commun à toutes les langues.

II - PROPRIETES ACOUSTIQUES DE LA LIGNE DE DECLINAISON

Le contour de Fo de 30 phrases de longueurs diverses, prononcées par 3 locuteurs ont été analysées, puis divisées en 5 à 6 groupes selon leur longueur, pour chaque locuteur. Les contours des phrases appartenant au même groupe ont éré superposés, afin de visualiser la pente de déclinaison générale. Un résultat de cette superposition est montré à la Fig. 3. On peut représenter la pente générale par une ligne droite, comme le montre la figure, cette ligne étant déterminée par un examen visuel du résultat de la superposition. La valeur de la pente de déclinaison, r (Hz/Sec), est mesurée pour chacune des droites ainsi déterminée. Les résultats sont présentés à la Fig. 4, sur laquelle chaque point indique la valeur mesurée de r versus la longueur moyenne des phrases dans le groupe concerné. Les lignes incurvées sont définies par l'équation : $r = \overline{\Delta F}\ t$, dans laquelle $\overline{\Delta F}$ est l'amplitude moyenne de la pente de Fo sur tous les groupes, et t la durée d'une phrase. Le fait que les valeurs r mesurées et les valeurs

estimées coincident de façon satisfaisante suggère que l'amplitude du déclin de Fo, ΔF, doit être considérée comme constante pour un locuteur donné. L'amplitude de ce déclin a été trouvé égale à 32,2 Hz, 28,2 Hz et 21,0 Hz, pour les trois locuteurs, respectivement.

De plus, les valeurs de Fo à la fin des phrases sont également approximativement constantes. En utilisant ces deux propriétés, la ligne de déclinaison peut être déterminée à partir de la longueur de la phrase.

III. CORRELATIONS PHYSIOLOGIQUES DE LA LIGNE DE DECLINAISON : UNE EXPERIENCE CINERADIOGRAPHIQUE SUR LES MOUVEMENTS DU LARYNX

Des valeurs de ΔF relativement élevées ont été trouvées lors de l'analyse acoustique. Le mécanisme qui génère un tel déclin de Fo représente un grand intérêt d'étude pour une expérience physiologique.

La valeur de Fo de la parole est déterminée principalement par deux facteurs : la pression de l'air subglottale, Ps, et les états des cordes vocales. En général, l'influence de Ps sur les valeurs de Fo est estimée à 5 Hz/cm H_2O ou moins. La diminution de Ps dans une phrase est inférieure à 3 cm/H_2O, bien que Ps diminue rapidement à la fin de la phrase (par exemple, voir Atkinson 1973). En conséquence, l'influence de Ps sur ΔF est sans doute inférieure à 15 Hz, ce qui est beaucoup moindre que les ΔF attestées dans les phrases. Nous avons donc fait l'hypothèse, partant de ce fait, que non seulement Ps, mais aussi quelque (s) autre (s) facteur (s), devrait être impliqué dans la génération de la ligne de déclinaison.

Les valeurs de Fo sont dépendantes de la longueur des cordes vocales. (Hollien, Brown et Hollien, 1971). D'autres auteurs, par exemple Vanderslice (1967) et Ohara (1970), ont fait l'hypothèse que la hauteur verticale du larynx est également en relation avec les valeurs de Fo.

Afin d'obtenir quelque lumière sur ce problème, nous avons effectué une expérience cinéradiographique pour mesurer la dynamique du larynx durant l'acte de parole pour l'un des trois locuteurs. Dasn cette expérience, un film latéral de Rayons-X (35 mm) de 35 prises/seconde et un enregistrement du signal acoustique ont été réalisés en parallèle. Seules 4 phrases de longueur courte ont été enregistrées, de façon à ce que la dose totale de Rayons-X absorbée par le sujet reste dans les limites de sécurité.

A la Fig. 5, nous montrons un exemple de la variation de hauteur verticale mesurée sur la portion antérieure du cartilage thyroïde (TYD), sur l'os hyoïde (HYD), et sur le mandibule (JAW). Le contour de Fo correspondant et l'enveloppe d'énergie (A.E.) sont illustrés en parallèle.

Nous avons trouvé que les mouvements verticaux de cartilage thyroïde et les contours de Fo, en particulier la ligne de déclinaison (indiquée par la ligne en pointillés sur la Fig. 5 pour cette phrase) ne sont pas corrélés de façon consistante dans chacune des 4 phrases. Néanmoins, nous avons pu faire les deux remarques suivantes.

1/ La position du larynx durant l'acte de parole est relativement stable. Cette remarque est illustrée dans notre exemple. L'os hyoïde bouge considérablement, peut-être en corrélation avec les mouvements de la langue. Par contre les cartilages thyroïdes, qui sont reliés à l'os hyoïde par les muscles, ont une position fixe. Les muscles laryngiens extrinsèques doivent participer à la stabilisation du larynx.

2/ Le larynx s'élève au début de chaque phrase, et s'abaisse à la fin. Ce mouvement peut être considéré comme une transition entre l'état de repos et la production de la parole.

En ce qui concerne les 4 phrases, et dans la limite de nos expériences, la hauteur du larynx n'est pas corrélée de façon consistante à la ligne de déclinaison. Et comme la longueur des cordes vocales est directement corrélée avec les valeurs de Fo, nous avons essayé de mesurer l'évolution de leur logueur. Il n'est pas possible, cependant, de mesurer cette longueur directement sur les films de Rayons-X. Nous avons alors mesuré les variations de longueur des ventricules, image par image. (Les ventricules sont les espaces entre les vraies cordes vocales et les fausses). Malheureusement, la portion postérieure des ventricules est souvent masquée par les cartilages aryténoïdes, et elle ne peut pas être vue clairement sur le film. Ce fait se retrouve dans la dispersion des points sur la Fig. 6. Comme les points montrent une dispersion considérable entre deux images successives, une comparaison détaillée de la variation de longueur et du contour de Fo ne peut être fait. Mais la tendance générale des variations de la longueur du ventricule peut néanmoins être étudiée avec profit.

La droite représentant les variations de longueur image par image est déterminée par un algorithme utilisant les moindres carrés. L'algorithme a été appliqué sur les seuls points qui sont situés à l'intérieur de la phrase, délimitée par les deux traits verticaux sur la Fig. 6. Pour chacune des 4 phrases, la droite montre un gradient négatif, correspondant à un écourtement graduel de la longueur des cordes vocales le long de la phrase. L'amplitude moyenne de l'écourtement est de l'ordre de 3,8 mm.

Selon Hollien, Brown et Hollien (1971), l'influence de la longueur des cordes vocales sur les valeurs de Fo est de l'ordre de 8 Hz/mm, dans l'échelle des valeurs de Fo qui nous concernent ici. La chute de Fo due à l'écourtement des cordes vocales est alors estimée à 30 Hz, valeur qui peut être positivement comparée à la valeur de F mesurée dans l'analyse acoustique. En ce qui concerne les 4 phrases analysées, la ligne de déclinaison est corrélée de façon plus consistante avec la longueur du ventricule, donc avec la longueur des cordes vocales, qu'avec la hauteur du larynx.

IV - DISCUSSION

Quel mécanisme est sous-jacent à l'écourtement des cordes vocales ? Le mécanisme que nous proposons ici est simple. La longueur des cordes vocales est déterminée essentiellement par la relation géométrique entre le cartilage thyroïde et les cartilages cricoïdes. Les deux cartilages sont connectés par le joint cricothyroïde et leur angle par rapport au joint définit la longueur des cordes vocales (en assumant que le joint est un joint de rotation). Si nous supposons que la position du cartilage thyroïde est fixe, par exemple, comme sur la Fig. 5, le mouvement du cricoïde est directement en rapport avec la longueur des cordes vocales, et donc avec les valeurs de Fo.

En nous basant sur les mesures de la dimension du cartilage laryngien (Maue et Dickson, 1971), nous estimons que le rapport de levier entre le mouvement du cricoïde et de la longueur des cordes vocales est d'environ de 3 ou plus. En conséquence, pour effectuer l'écourtement observé sur les cordes vocales, c'est-à-dire 3,8 mm, un mouvement vers le bas de seulement 1,3 mm du cartilage cricoïde est nécessaire.

Deux forces principales peuvent être considérées comme ayant une action sur le cartilage cricoïde : une force générée par les muscles cricothyroïdes, et une traction trachéale, qui est une force agissant vers le bas. Dans notre hypothèse, une diminution de volume des poumons durant l'acte de parole peut causer une augmentation de la traction trachéale, résultant en une rotation graduelle du cartilage cricoïde, et en conséquence en un écourtement des cordes vocales. Nous n'avons pas de preuves directes d'une augmentation de la traction trachéale le long des phrases. Cependant, une interprétation des fonctions respiratoires semble apporter un élément positif à la théorie décrite ci-dessus.

En résumé, un écourtement progressif de la longueur des cordes vocales durant la parole semble rendre compte d'une part importante de la chute de Fo durant la ligne de déclinaison. Nous avons fait l'hypothèse qu'une augmentation dans la traction trachéale due à une compression des poumons durant la parole, pousse le cartilage cricoïde vers l'avant, et en conséquence diminue la longueur des cordes vocales. La contribution de la pression de l'air subglottale Ps, à la descente de Fo dans la ligne de déclinaison, est probablement inférieure à 30 % de l'amplitude de la baisse de Fo. Pour conclure, nous devons insister sur le fait que notre recherche ne porte que sur un nombre très réduit de phrases, et qu'en conséquence, les résultats décrits ici ne doivent pas être considérés comme définitifs.

REMERCIEMENTS

Nous remercions Monsieur le Professeur Kenneth M. Stevens pour l'aide et les encouragements qu'il nous a apportés tout au long de ce travail. Je remercie également M. Robert Donahue du Cardiac Catheterization Laboratory de l'Hopital Général du Massachusetts pour le temps qu'il a passé pour l'expérience de cinéradiographie. Ce travail a été effectué au Laboratoire de Recherche en Electronique de M.I.T., partiellement grâce à un contrat de la National Institute of Health (contrat NS 04332). Cette communication est basée sur un des travaux réalisés pour l'obtention d'une partie de PhD.

REFERENCES

Atkinson, J., (1973), "Aspects of Intonation in Speech : Imprications from an Experimental Study of Fundamental Frequency," Ph. D. Thesis, The University of Connecticut.

Cohen, A. and t'Hart, J., (1967), "On the Anatomy of Intonation," Lingua, 19, 117-192.

Hollien, H., Brown, W., and Hollien, K., (1971), "Vocal Fold Length Associated with Modal, Falsetto and Varying Intensity Phonations," Folia Phoniatr., 23, 66-78.

Maeda, S., (1974), "A Characterization of Fundamental Frequency Contrours of Speech," Quarterly Progress Report, N° 114, Research Laboratory of Electronics, M.I.T., 198-213.

Maue, W. and Dickson, D., (1971), "Cartilages and Ligaments of the Adults Human Larynx," Arch Otolaryng., 94, 432-439.

Ohara, J., (1970), "Aspects of the Control and Production of Speech," Working Papers in Phonetics, University of California, Los Angeles, 15, 1 - 192.

Olive, J., (1975), "Fundamental Frequency Rules for the Synthesis of Simple Declarative English Sentences," J. Acoust. Soc. Am., 57, 476-482.

Vanderslice, r., (1967), "Larynx vs. Lungs : Cricothyrometer Data Refuting Some Recent Claims Concerning Intonation and Archetypality," Working Papers in Phonetics, University of California, Los Angeles, 7, 69-79.

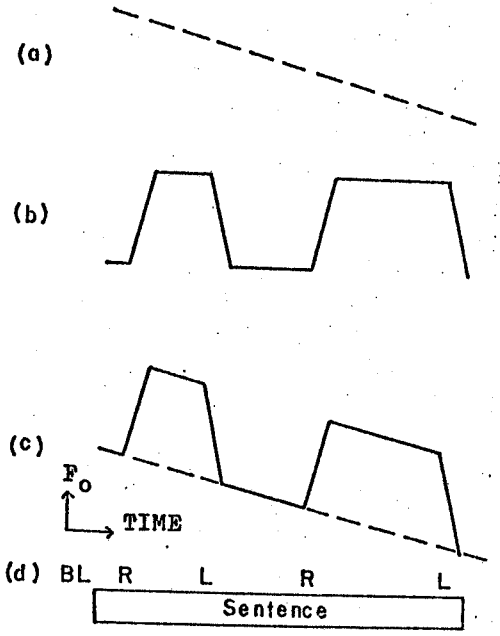


Fig. 1.

Le contour de Fo dans (c) est le résultat de l'addition de la ligne de déclinaison dans (a) et des mouvements locaux de Fo (combinaison de montées et de descentes) dans (b). La séquence d'attributs correspondante associée avec la phrase est représentée dans (d).

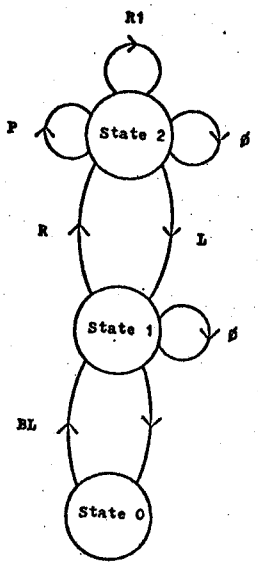


Fig. 2.

Un réseau de transition d'états décrivant la structure des contours de Fo en termes d'attributs.

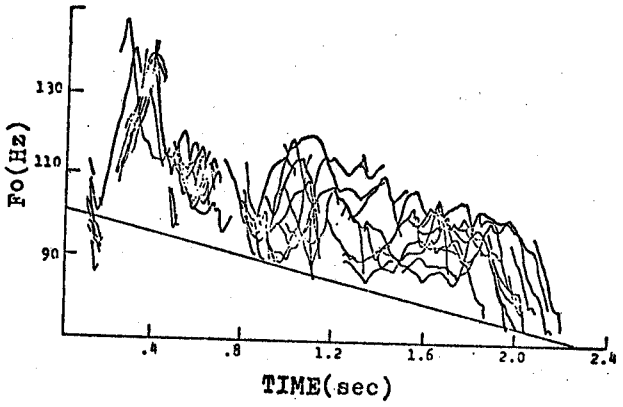


Fig. 3.

Superposition des contours de Fo de 11 phrases de longueur similaire. La droite représente une chute graduelle des contours, droite dont la pente est déterminée par inspection visuelle.

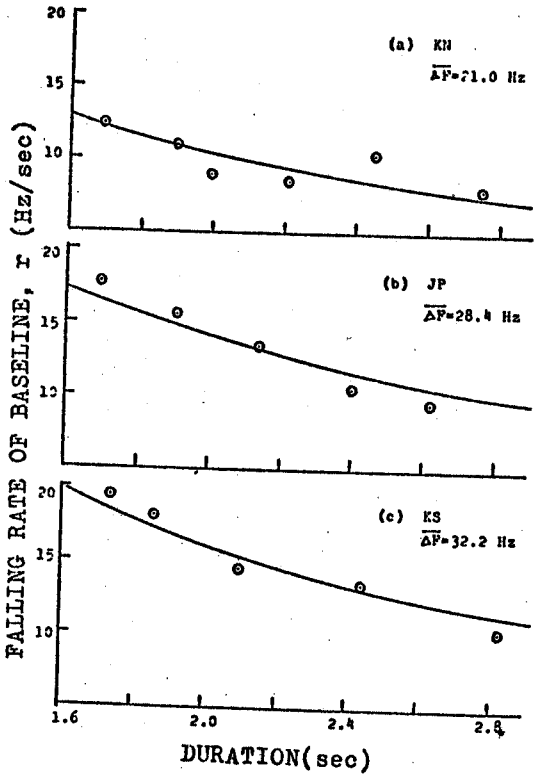


Fig. 4.

Correspondance entre le taux de baisse, r (Hz/sec) et la durée en sec., pour chacun des trois locuteurs.

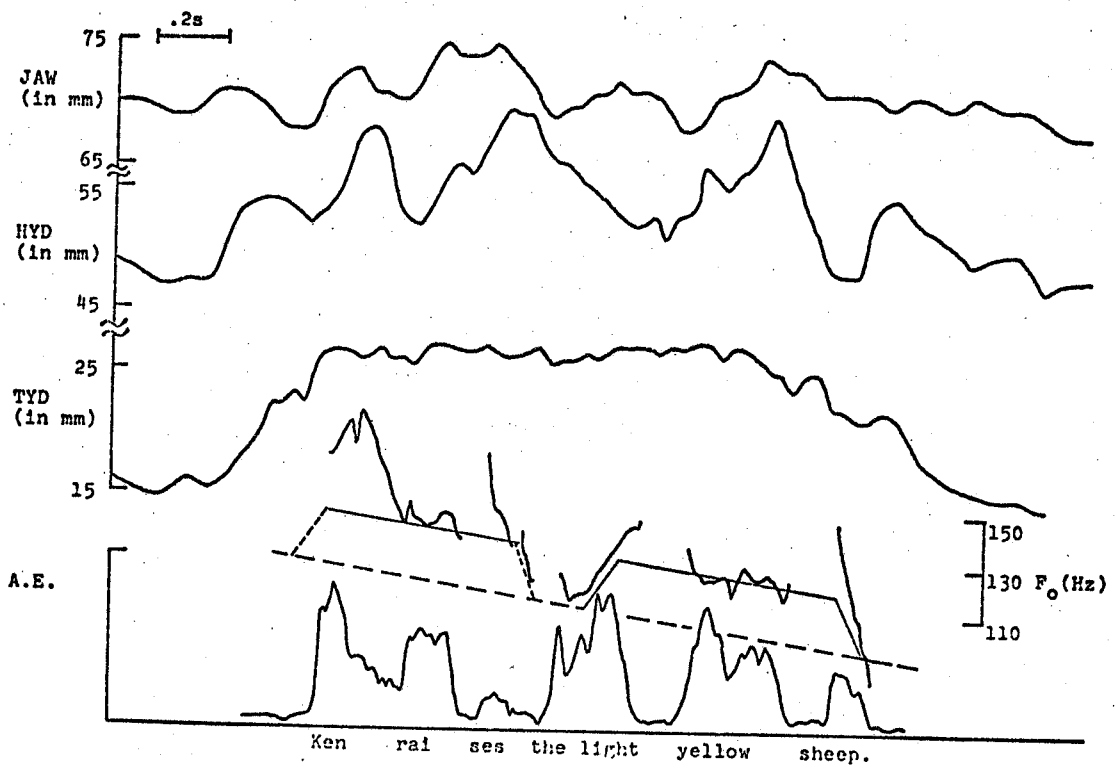


Fig. 5.

Les mouvements verticaux du mandibule (JAW), de l'os hyoïde (HYD) et du cartilage thyroïde (TYD), le contour de Fo et sa schématisation, ainsi que l'enveloppe (A.E.) de l'énergie sont représentés sur cette figure.

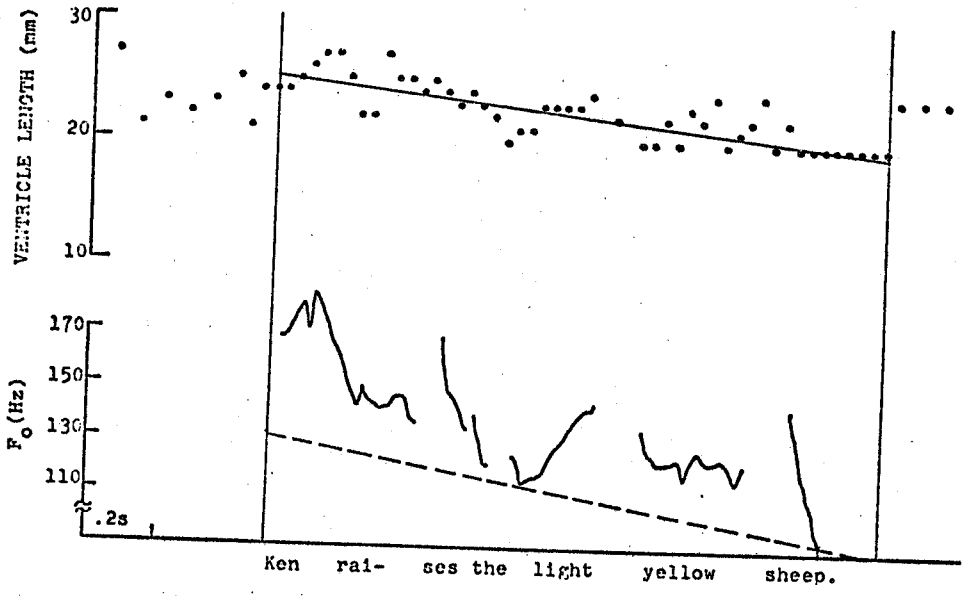


Fig. 6.

Les variations de la longueur du ventricule laryngien, qui correspondent à celles de la longueur des cordes vocales, calculées à partir des données cinéradiographiques pour la phrase représentée à la Fig. 5.
La droite superposée aux points est une approximation par les moindres carrés.

Détection de Fo et du voisement

VIII èmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

MESURE DU FONDAMENTAL PAR FILTRAGE VARIABLE

J. CAELEN et P. CAZENAVE

Laboratoire C.E.R.F.I.A.
Université P. Sabatier TOULOUSE

• Résumé :

Parmi les méthodes de mesure du fondamental, celle du filtrage suivi suivi de la détection des passages par zéro du signal filtré est bien connue. La méthode présentée ici rend le filtrage adaptable à la parole, c'est à dire à divers paramètres acoustiques tels que la stabilité, l'énergie, et à l'erreur de mesure. L'organe de décision tient compte également du contexte phonétique ce qui rend tout l'ensemble du système de décision solidaire de la structure acoustique de la parole.

• Summary :

Among the methods which permit to measure the pitch, the are of filtering followed by the detection of zero-crossing of output sihnal is well-know. The method presented here, renders filtering adaptative to speech, that is to say, to acoustical parameters such a stability, energy and errot of measure. The system of decision takes into account as well phonetic context, that makes the whole detection dependent of acoustical structure of speech.

MESURE DU FONDAMENTAL PAR FILTRAGE VARIABLE

J. CAELEN et P. CAZENAVE
Laboratoire CERFIA
Université P. Sabatier TOULOUSE

RESUME

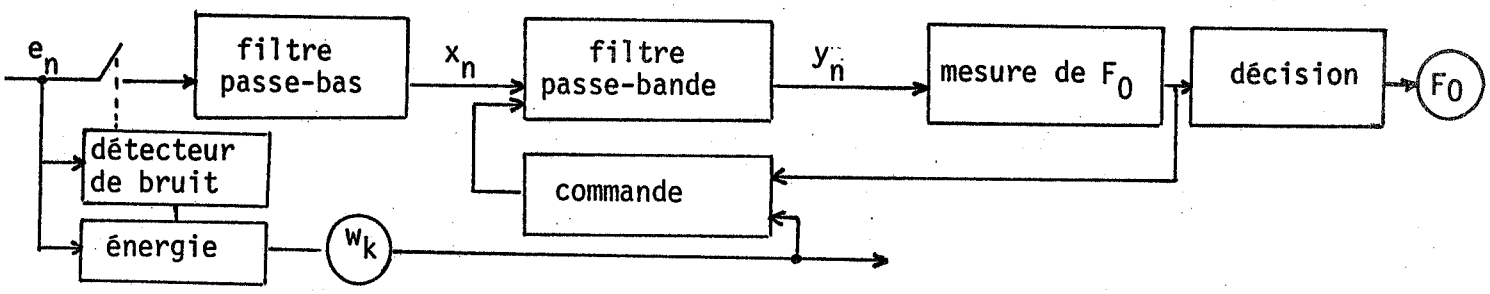
Parmi les méthodes de mesure du fondamental, celle du filtrage suivi de la détection des passages par zéro du signal filtré est bien connue. La méthode présentée ici, rend le filtrage adaptable à la parole, c'est-à-dire à divers paramètres acoustiques tels que la stabilité, l'énergie et à l'erreur de mesure. L'organe de décision tient également compte du contexte phonétique ce qui rend tout l'ensemble du système de détection solidaire de la structure acoustique de la parole.

SUMMARY

Among the methods which permit to measure the pitch, the are of filtering followed by the detection of zero-crossing of output signal is well-know. The method presented here, renders filtering adaptive to speech, that is to say, to acoustical parameters such a stability, energy and error of measure. The system of decision takes into account as well phonetic context, that makes the whole detection dependent of acoustical structure of speech.

1. PRINCIPE

Le signal est échantillonné à 15 khz et attaque un système de filtres commandés en fonction de l'entrée et de la sortie. Sur le schéma ci-dessous on peut reconnaître tout d'abord un détecteur de bruit (logique à seuil) qui inhibe les procédures de filtrage, le système de filtres commandables, la logique de mesure du fondamental, celle de l'énergie et enfin la logique de décision qui affecte ou non la valeur trouvée à la mélodie.



1.1. Le filtre passe-bas

Est un filtre du 1er ordre, récursif, de fréquence de coupure f_0. Son rôle est d'effectuer un pré-filtrage passe-bas pour atténuer les composantes haute fréquence de e_n.

$$\begin{cases} x_n = a x_{n-1} + e_n \\ x_0 = 0 \end{cases} \qquad \text{avec} \quad a = \exp(-2\pi f_0/f_e)$$

x_n $n^{ème}$ échantillon du signal de sortie

e_n $n^{ème}$ échantillon du signal d'entrée

f_e fréquence d'échantillonnage.

1.2. Le filtre passe-bande variable

Est un filtre récursif de 2ème ordre de fréquence centrale f_c et de cœfficient de surtension Q. Ce sont ces deux paramètres qui font l'objet du contrôle et de l'adaptation.

$$\begin{cases} y_n = b_1 y_{n-1} + b_2 y_{n-2} + x_n \\ y_0 = y_1 = 0 \end{cases} \qquad \text{avec} \quad b_1 = 2\exp(-\pi f_c/f_e Q)\cos(2\pi f_c/f_e)$$
$$b_2 = \exp(-2\pi f_c/f_e Q)$$

y_n $n^{ième}$ échantillon du signal de sortie.

1.3. La mesure du fondamental

Se fait par comptage des passages par zéro du signal de sortie y_n. Sur une fenêtre d'observation T on évalue la fréquence moyenne f_k ($k^{ième}$ observation) et le cœfficient de variation v_k des passages par zéro.

Soit :

$$\theta_k = \frac{1}{m} \sum_{i=1}^{m}{}' \tau_i$$

$$f_k = f_e/2\theta_k$$

$$V_k = \left(\frac{A}{m} \sum_{i=1}^{m}{}' (\tau_i - \theta_k)^2\right)^{1/2} / \theta_k$$

τ_i écart entre deux passages par zéro successifs θ_n et θ_{n-1}

m nombre de p.a.z sur T

θ_n est un passage par zéro (p.a.z) si $y(\theta_n) = 0$ ou $y(\theta_{n-1})y(\theta_n) < 0$

Dans ces conditions f_k sera une évaluation de la fréquence du signal d'entrée qui, s'il est stable $\Longrightarrow$ v_k faible. Inversement v_k sera l'image de la stabilité du signal e_n et de l'erreur de détection de F_0. En zone de voisement il faut donc chercher à minimiser v_k.

1.4. La commande

Les conditions optimales de filtrage seront obtenues pour $f_c = F_0$ et Q max, or F_0 n'est pas connue et l'instabilité de e_n impose Q borné. On doit donc rechercher des conditions sous-optimales et tenir compte de l'instabilité tant fréquentielle qu'énergétique de e_n.

pour v_k petit, f_k sera une bonne approximation de F_0 d'où une commande possible de f_c :

$$f_c = f_0 + (f_k - f_0)/(1 + v_k)$$

si v_k est grand le filtre reste bloqué à sa position initiale f_0 .

pour $v_k \to 0$ $\lim f_c = f_k$ avec $\lim f_k = F_0$

La commande de Q doit tenir compte de l'instabilité d'amplitude de e_n puisque c'est ce paramètre qui joue directement sur le temps de montée du filtre. Le principe est de choisir un filtre sélectif lorsque e_n est stable et un filtre peu sélectif dans le cas contraire, afin de suivre au mieux ces instabilités.

$$Q = Q_0 + \Delta Q/(1+v_k)(1+w_k-w_{k-1})$$

ΔQ incrément de Q ;

w_k énergie de e_n sur la $k^{ième}$ observation de durée T.

Cette commande confère au filtre une sélectivité variable en fonction de la stabilité de e_n.

1.5. La logique de décision

Tout en restant la plus simple possible (temps de calcul), cette logique doit intégrer le contexte phonétique de la parole. C'est pourquoi ici encore, on a tenté d'adapter les seuils de décision aux évènements acoustiques changeants.

Condition 1 :

$v_k > v_s \implies f_k$ non significatif, pas de voisement

v_s cœfficient seuil qui dépend du contexte acoustique. En effet, on sait (1) que l'on peut classer certains phonèmes d'après leur énergie moyenne et les variations de cette énergie dans le temps. Par expérience puis apprentissage on a donc été amené à chercher v_s comme fonction de w_k et $w'_k = (w_k - w_{k-1})/T$ d'où :

$$v_s = h(w_k, w_{k-1}, T)$$

et par souci de simplification $v_s = A_k w_k + A_{k-1} w_{k-1} + BT$, les cœfficients A et B déterminés d'abord par expérience puis affinés par apprentissage.

Condition 2 :

$f_k > 2f_{k-1} \implies f_k$ non significatif

On suppose en effet ici, qu'entre deux observations et en élocution normale il est impossible que le fondamental se soit élevé d'une octave. Cette condition sert donc à la correction de f_k dans le cas où le filtre se serait accordé sur un harmonique de F_0.

Ces deux conditions permettent alors de décider :

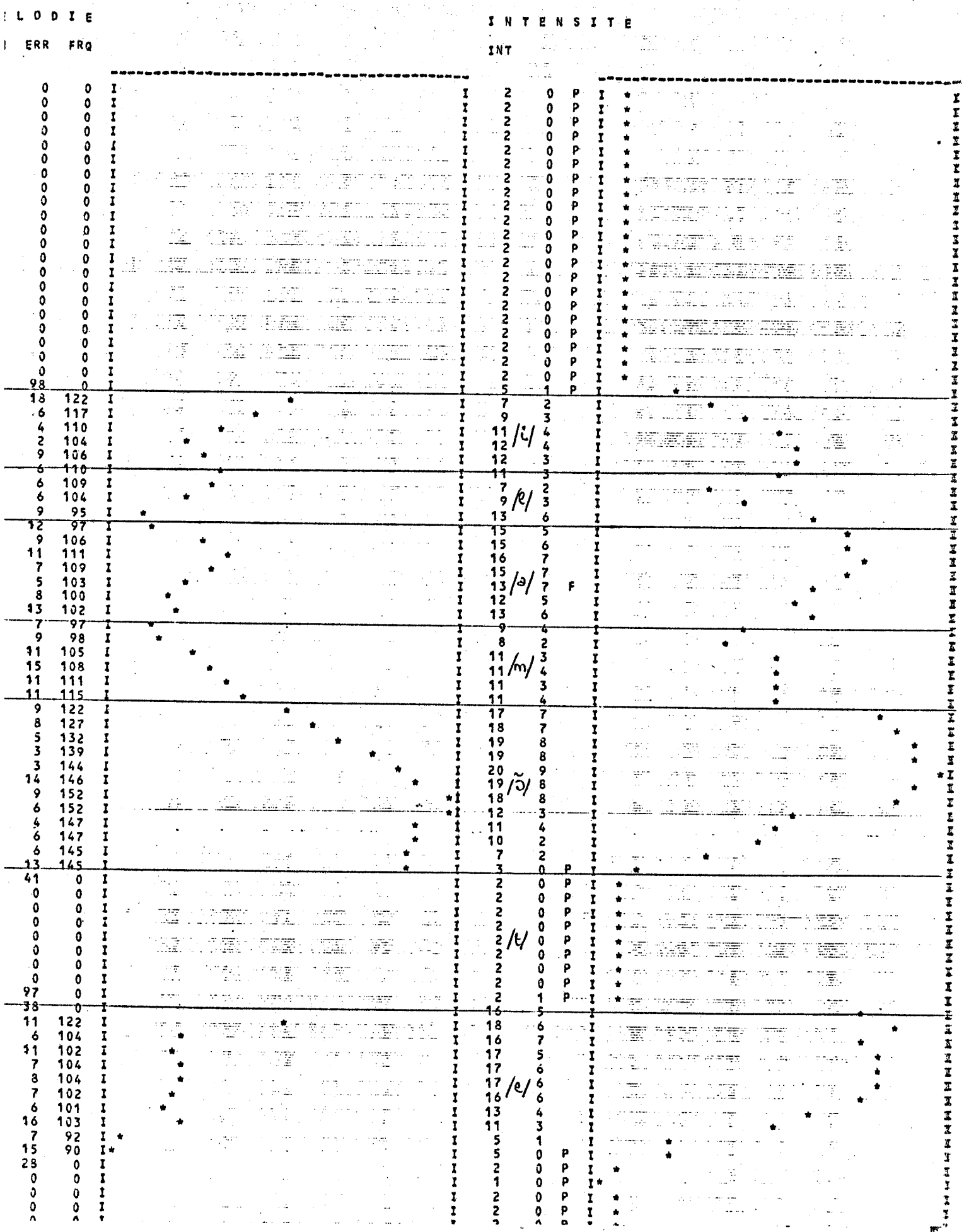
f_k est significatif,

f_k n'est pas significatif $\implies f_k = 0$

$f_k/2$ est significatif.

2. RESULTATS

Ce détecteur de mélodie fonctionne en temps réel sur IRIS 80 et est exploité églament sur B3500. Les résultats seront largement commentés dans (2) on ne constate


MELODIE
ERR FRQ
INTENSITE
INT
/i/
/ʁ/
/ə/
/m/
/ɔ̃/
/t/
/e/

pas d'erreur de détection mais uniquement des "manques" notamment lors des trops brusques variations d'intensité du signal (explosions), ou lorsque certaines plosives voisées sont noyées dans le bruit. On savait déjà (3) que la détection par passages à zéro sur le signal filtré donnait de bons résultats, ici, l'adaptation et le contrôle adjoints au filtrage augmentent encore la fiabilité et rend la méthode plus apte à s'adapter à un grand nombre de voix.

CONCLUSION

Cette méthode complète les techniques habituelles de filtrage et l'utilisation de filtres "suiveurs" convenablement guidés en fonction de la structure propre de la parole ne peut qu'améliorer les résultats obtenus. Et c'est bien ce que nous obtenons. Il n'est pas exclu que des procédures de commande liées à une reconnaissance sommaire des phonèmes ne lèvent pas les dernières imperfections du système. On obtiendrait alors un analyseur de mélodie totalement adapté à la parole, sans pour autant s'éloigner du temps réel.

REFERENCES

(1) A propos de marqueurs lexicosyntaxiques : quelques exemples commentés de phrases issues de l'analyseur du projet A.R.I.A.
D. DOURS, R. FACCA, Y. LAURENTIE, G.MAURAND, G. PERENNOU, TOULOUSE 1975

(2) Corrélations entre mélodie, intensité et durée dans la phrase française énonciative.
G. CAELEN, G. MAURAND 8° JEP AIX-EN-PROVENCE 1977

(3) La détection du voisement par les propriétés physiques résultant de l'excitation périodique du conduit vocal : comparaison statistique de trois procédés.
C. ABRY, L.J. BOE, J.F. ZURCKER 6 JEP TOULOUSE 1975

(4) Etude de filtres de 2^e ordre variables et couplés. Application à l'analyse de signaux non stationnaires.
J. CAELEN Note CERFIA TOULOUSE 1976

VIII èmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

AUTOMATE DE SEGMENTATION ET DETECTION DU FONDAMENTAL

D. DOURS, R. FACCA, G. PERENNOU

C. E. R. F. I. A., TOULOUSE

• Résumé :

La détection du voisement s'effectue lors de la décomposition du signal vocal en une suite de segments tels que chacun d'eux corresponde à la réalisation d'un phonème. Pour assurer cette décomposition, l'automate de segmentation tient compte des réalisations particulières de certains phonèmes.

L'algorithme de détection de l'onde glottique est fondé sur une recherche de périodicité sur le signal pré-traité par un filtre numérique. A partir des pics détectés, on effectue une correction des pics aberrants en s'aidant des zones où la pseudo-périodicité est stable.

• Summary :

The detection of voicing takes places during the vocal signal fragmentation (splitting), into a series of segments so that each of them corresponds to the realization of one phoneme. To ensure this splitting, the segmentation automaton takes into account the peculiar realization of certain phonemes.

The glottal vawe detection algorithm is founded on a recurrence research by a numerical filter on the processed signal. From the peaks which have been spotted, the erratic peaks are corrected through the zones where the pseudo periodicity is steady.

AUTOMATE DE SEGMENTATION ET DETECTION DU FONDAMENTAL

D. DOURS, R. FACCA, G. PERENNOU

C. E. R. F. I. A., TOULOUSE

1. INTRODUCTION

L'automate de segmentation développé dans le projet A.R.I.A. vise à satisfaire à deux objectifs : d'une part décomposer le signal vocal en une suite de segments tels que chacun d'eux corresponde à la réalisation d'un phonème, d'autre part, détecter l'onde glottique dans les segments voisés en vue d'une analyse période par période.

Pour effectuer la détection du voisement, l'automate de segmentation tient compte des réalisations particulières de certains phonèmes voisés. Cette méthode permet de prendre des décisions sur le voisement ou le non voisement des phonèmes.

2. SEGMENTATION EN PHONEMES

2.1. Automate de segmentation

Il est composé de quatre automates distincts.

L'automate a_1, découpe le signal en intervalles réguliers de 256 pts et affecte une classe $C_i \in C$ à chaque intervalle.

C = {Silence, Buzz, voisé, voisé faible, fricative, fricative voisée}.

L'automate a_2, constitue des segments homogènes en regroupant tous les intervalles ayant été affectés de la même classe et en leur affectant une classe qui dépend de C_i et de leur longueur.

L'automate a_3 découpe les segments homogènes voisés en consonne voisée, voyelle, consonne voisée courte.

L'automate a_4 tient compte des réalisations particulières de certains phonèmes pour effectuer la segmentation définitive.

2.1.1. Automate a_1

Il découpe le signal échantillonné $s(t) \rightarrow (s_1, s_2, \ldots, s_n)$ en intervalles $I_k = (s_{(k-1)N+1}, s_{(k-1)N+2}, \ldots, s_{(k-1)2N})$.

Pour un échantillonnage à 15 KHz nous avons pris N = 256.

A chaque intervalle I_k est associé le quadruple (V_k, Z_k, F_k, B_k), dans lequel

V_k est l'énergie

Z_k un indice de voisement

F_k un indice de friction

B_k un indice de Buzz

Le quadruple (V_k, Z_k, F_k, B_k) étant obtenu, on procède à l'affectation d'une classe à un intervalle I_k de la façon suivante :

2.1.1.1. Détection des fricatives

Si $F_k \neq 0$, I_k est considéré comme un intervalle fricatif. On effectue alors la séparation fricatives voisées, fricatives sourdes.

Si $Z_k \neq 0$, I_k est considéré comme un intervalle fricatif sourd et affecté de la classe (FRI).

Sinon I_k est considéré comme un intervalle fricatif voisé et affecté de la classe (FRIV).

2.1.1.2. Détection des silences

Si $Z_k = 0$ et $F_k = 0$, I_k est considéré comme un silence et affecté de la classe (SI).

2.1.1.3. Détection du buzz

Si $B_k = 1$, $Z_k \neq 0$ et $F_k = 0$, I_k est considéré comme appartenant à une plosive sonore et affecté de la classe (BUZZ).

2.1.1.4. Distinction voisé fort - voisé faible

Pour que l'on puisse effectuer une recherche du fondamental, il est indispensable que l'énergie des sons voisés atteigne un certain seuil. C'est pour cette raison que l'on fait la distinction voisé fort - voisé faible.

Si $V_k > \theta_v$, $B_k \neq 1$, $Z_k \neq 0$ et $F_k = 0$, I_k est considéré comme un intervalle voisé fort et affecté de la classe (VOI).

Si $V_k < \theta_v$, $B_k \neq 1$, $Z_k \neq 0$ et $F_k = 0$, I_k est considéré comme un intervalle voisé faible et affecté de la classe (VOIF).

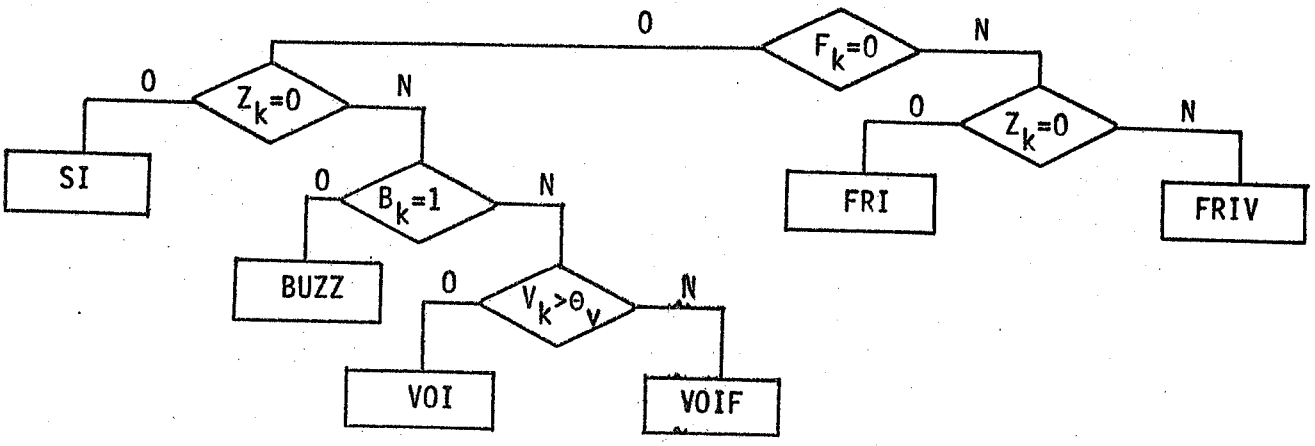


Fig. 1 : Affectation d'une classe à un intervalle I_k

2.1.2. Automate a_2

Les segments homogènes sont constitués par concaténation d'intervalles I_k ayant été affectés de la même classe. On leur affecte une classe notée [CL] s'ils ont une longueur suffisante sinon on leur affecte une classe notée CL .

Les règles d'affectation sont les suivantes :

$$(VOI)^n = \begin{cases} [VOI] & \text{si } n \geq 2 \\ (VOI) & \text{si } n = 1 \end{cases}$$

$$(SI)^n = \begin{cases} [SI] & \text{si } n \geq 16 \\ [PL] & \text{si } 2 \leq n \leq 15 \\ (SI) & \text{si } n = 1 \end{cases}$$

$$(FRI)^n = \begin{cases} [FRI] & \text{si } n \geq 4 \\ (FRI) & \text{si } n \leq 3 \end{cases}$$

$$(BUZZ)^n = \begin{cases} [BUZZ] & \text{si } n \geq 3 \\ (BUZZ) & \text{si } n \leq 2 \end{cases}$$

$$(FRIV)^n = \begin{cases} [FRIV] & \text{si } n \geq 3 \\ (FRIV) & \text{si } n \leq 2 \end{cases}$$

$$(VOIF)^n = \begin{cases} [VOIF] & \text{si } n \geq 2 \\ (VOIF) & \text{si } n = 1 \end{cases}$$

Nous voyons apparaître ici, une nouvelle classe [PL] qui est la classe des plosives sourdes.

2.1.3. Automates a_3

Le découpage en phonèmes d'un segment classé voisé [VOI] s'effectue à partir des variations de la courbe d'énergie du signal.

On peut en effet constater que la courbe d'énergie (Fig. 2) présente une très grande variation d'énergie lors des transitions Consonne-voyelle et voyelle-consonne. Nous avons donc développé un algorithme fondé sur cette remarque.

A l'issue de cette segmentation, un segment classé [VOI] se décompose en une suite de segments classés [VO] (voyelle), [CV] (consonne) et [CVC] (consonne voisée courte si moins de 512 points).

A ce niveau, il n'est pas possible de distinguer toutes les semi-voyelles. Elles ne pourront être détectées qu'au niveau de l'analyse, elles sont pour l'instant classées soit comme voyelle soit comme consonne.

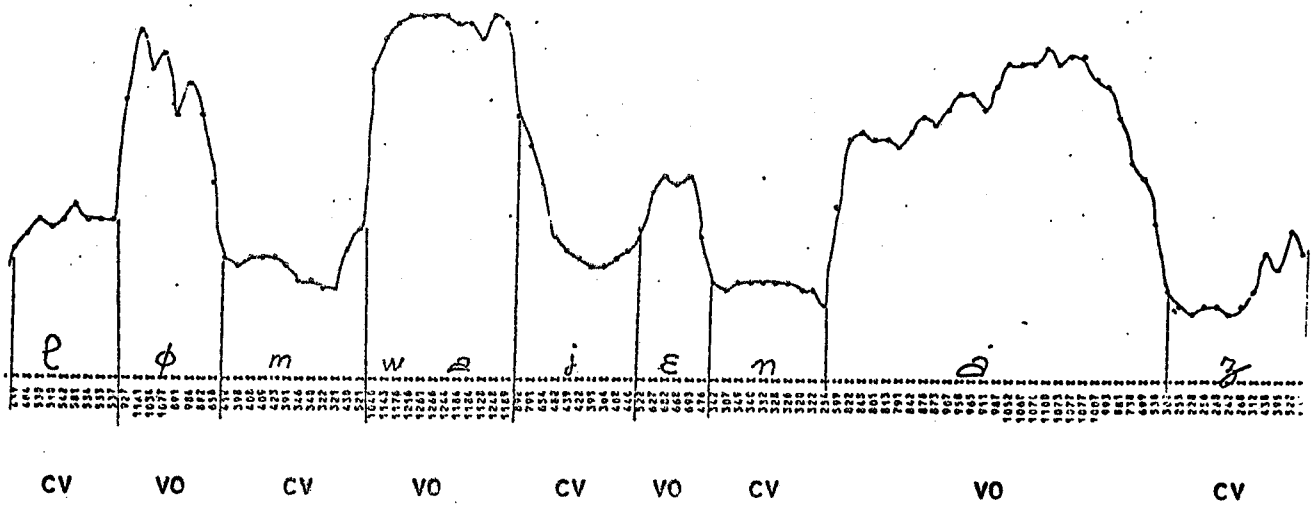


Fig. 2 : Courbe d'énergie de la séquence "LE MOYEN AGE"

2.1.4. Automate a_4

La segmentation en phonème a pour but de découper le signal en segments tels que chacun d'eux corresponde à la réalisation d'un phonème et de leur affecter une des classes : voyelle, consonne voisée, fricative sourde, fricative sonore, plosive sonore, plosive sourde.

On pourrait penser que le découpage en phonèmes se borne à la constitution de segments homogènes. Il n'en est rien car la réalisation de certains phonèmes n'est pas aussi simple.

On remarquera par exemple qu'une fricative voisée telle que [Z] ou [ʒ] peut être composée d'une partie classée [FRIV] suivie d'une partie classée (FRI). Il faudra donc concaténer ces deux segments homogènes pour constituer le segment correspondant à ce phonème.

Il en est de même pour d'autres phonèmes (voir tableau fig. 3). De plus les transitions entre phonèmes ne sont pas toujours très nettes, et des segments parasites peuvent s'insérer à la frontière des phonèmes. Par exemple, une plosive sourde (p, t, k), classée [PL] et précédée d'un segment voisé, peut être constituée de la concaténation de deux segments (Buzz) et [PL]. D'autres exemples pourront être trouvés dans le tableau de la fig. 3.

Il est bien connu des phonéticiens que les plosives sourdes (t, k) sont souvent constituées d'un segment [PL] suivi d'une friction de courte durée (FRI). Ces deux segments seront concaténés pour constituer une nouvelle classe [PLFRI].

Nous concaténons également les deux segments [PL] et (VOI) pour créer la classe [PLT] qui correspond à une plosive sourde en fin de mot comme dans "Cap".

De même la classe [PLFRIT] est constituée des segments [PLFRIT] et [VOI]. Elle correspond à une plosive sourde telle que (t,k) en fin de mot comme comme dans "sac".

Cet automate est caractérisé par un alphabet d'entrée I, un ensemble d'états internes identifié à I_2 (dans le cas présent ceci permet d'alléger les notations)

$$I = I_1 \cup I_2$$

avec $I_1 = \{(SI),(BUZZ),(FRI),(FRIV),(VOIF),(VOI)\}$

$I_2 = \{[SI],[BUZZ],[PLT],[PLFRI],[PLFRIT],[FRI],[FRIV],[VOIF],[CV],[CVC],$
$[VO]\}.$

A ce stade, il fonctionne comme un automate fini dont les transitions sont régies par la table fig. 3.

Toute case blanche dans une ligne (X) ou [X] doit être interprétée comme [X]. La distinction avec celui-ci ne s'opérant qu'au moment des sorties.

L'alphabet de sortie est un triplet (S, t_1, t_2) ou $S \in I_2$, t_1 l'instant de début, $t_2 \geq t_1$ l'instant de fin de segment.

A un instant donné, la sortie prend en compte l'état interne défini précédemment, la durée de l'entrée suivant les règles ci-dessous :

Soit $(S,E) \rightarrow T$ la dernière transition avec : $S, T \in I_2$ $E \in I$.

Soit L la durée de E.

Soit (S, t_1, t_2) le dernier triplet déjà écrit.

états internes I_1 I_2 / entrées	[SI]	[BUZZ]	[PL]	[PLT]	[PLFRI]	[PLFRIT]	[FRI]	[FRIV]	[VOIF]	[CV]	[CVC]	[VO]
(SI)	[SI]	[BUZZ]	[PL]	[PLT]	[PLFRI]	[PLFRIT]	[FRI]	[FRIV]	[VOIF]	[CV]	[CV]	[VO]
(BUZZ)	[SI]	[BUZZ]		[PLT]		[PLFRIT]			[VOIF]			
(FRI)	[PLFRI]		[PLFRI]				[FRI]	[FRIV]			[FRIV]	[FRIV]
(FRIV)							[FRI]	[FRIV]			[FRIV]	
(VOIF)	[SI]		[PL]	[PLT]	[PLFRI]				[VOIF]		[CV]	
(VOI)	[SI]	[BUZZ]	[PLT]	[PLT]	[PLFRIT]	[PLFRIT]	[FRI]	[FRIV]	[VOIF]	[CV]	[CV]	[VO]
[SI]	[SI]								[SI]			
[BUZZ]									[VOIF]			
[PL]		[PL]							[PL]			
[PLT]												
[PLFRI]												
[PLFRIT]												
[FRI]		[FRI]					[FRI]	[FRI]	[FRI]			
[FRIV]												
[VOIF]									[VOIF]		[CV]	
[CV]									[CV]	[CV]		
[CVC]									[CV]			
[VO]									[VO]			

Fig. 3 - Table de transition

1er cas : (S,E) correspond à une case blanche. On écrit un nouveau triplet $(T, t_2 + 1, t_2 + L)$.

2ème cas : (S,E) ne correspond pas à une case blanche, on modifie uniquement t_2 en $t_2 + L$.

2.2. Résultats et commentaires sur la segmentation

Nous avons testé le module de segmentation sur les onze phrases suivantes :

- Le moyen-âge est une époque très intéressante
- Le samedi les petits enfants sont en vacance
- Il a monté
- Il joue du tuba
- Y-a-t-il une personne s'appelant Dupont
- C'est selon ce que vous choisirez
- La leçon est longue
- Il a mis un nouveau toit
- Il a trouvé le savon
- Quelle volonté ! C'est étonnant
- La passion a des limites.

On remarquera que les principales règles de production sont utilisées, en particulier les règles concernant les fricatives sourdes, les fricatives sonores, les plosives sourdes et sonores. La friction dans les plosives sourdes (t, k) n'a pas été toujours détectée.

La plosive sourde [P] dans la séquence "les petits" a été confondue avec une plosive sonore. Il apparaît également que deux [œ̃] sur huit attestés dans les phrases traitées présentent une partie classée [CV] en fin de phonème. Ceci correspond à la fin de la nasale et certains indices de nasalité que nous avons pu mettre en évidence nous permettrons certainement de supprimer cette cause d'erreur.

Ces résultats sont très encourageants car ils ont été obtenus sur des phrases relativement complexes et dont la vitesse d'élocution est grande. Ils demandent à être confirmés sur un plus grand nombre de phrases et pour des locuteurs différents. Mais nous pensons qu'après quelques retouches il sera opérationnel pour un nombre quelconque de phrases et de locuteurs et ce, avec un très grand pourcentage de réussite.

3. DETECTION DE L'ONDE GLOTTIQUE

La détection est effectuée sur les segments classés [VOI] et [FRIV].Le filtrage du signal temporel fait apparaître un pic significatif pour chaque période de fondamental.

1. Filtrage

Soient S_i les échantillons du signal parole, FS_i les échantillons du signal filtré.

Le filtre est défini par la relation :

$$FS_i = \sum_{j=i}^{i+T-1} |S_j| - \sum_{j=i-T+1}^{i-1} |S_j| \qquad i \geq T$$

Les principales caractéristiques de ce filtre sont :

- Calcul rapide en valeurs entières par récurrence à partir de FS_T

$$FS_{i+1} = FS_i + |S_{i+T}| + |S_{i-T+1}| - 2\ |S_i| \qquad i \geq T$$

- T doit être adapté au fondamental moyen du locuteur. En pratique on choisit T voisin d'un quart de la période moyenne du locuteur analysé.

- Les maxima de FS émergent clairement mais ne donnent pas de résultats suffisamment précis parce qu'il y a un déphasage avec le début réel de la période du fondamental qui dépend du type de son analysé.

- Dans certains cas, il y a apparition de pics parasites ou inversement un "trou".

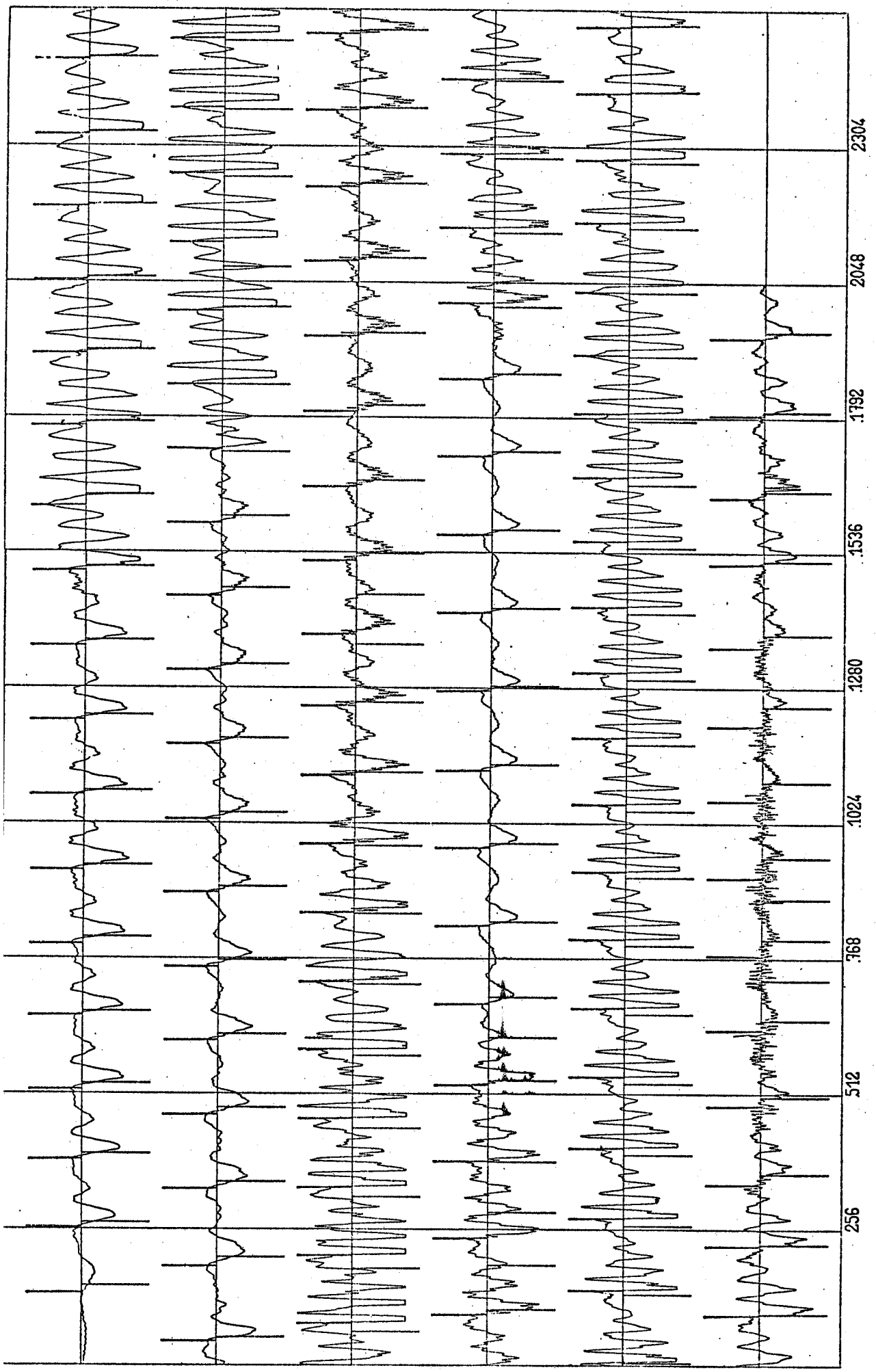
2. Recherche des maxima de FS

Sur quelques périodes consécutives les variations du signal et du fondamental d'une période à la suivante sont négligeables vis à vis du signal filtré.

Ce n'est plus le cas sur la durée d'un phonème et à fortiori sur plusieurs phonèmes .

La recherche des paxima de FS se fait donc sur des intervalles successifs de longueur L (environ 30 msec).

A partir du maximum le plus grand, on recherche récursivement tous les maxima du même ordre de grandeur que celui-ci.



Maxima du signal filtré vers le bas
Marqueurs de périodes vers le haut

3. Recherche des "marqueurs" de période

Les repères qui nous intéressent sont déphasés par rapport aux maxima de FS. En général ce déphasage est inférieur à T points.

Une analyse plus fine du signal parole lissé sur T points de part et d'autre du maximum de FS permet de déterminer avec précision le repère exact qui nous intéresse.

4. Correction des pics aberrants

L'algorithme se décompose en deux parties :

- Détermination des zones régulières. Celles-ci sont constituées par la concaténation de périodes sur lesquelles la pseudo-périodicité du fondamental détecté est stable.

- Correction des zones irrégulières. Trois cas peuvent se présenter :

. La zone irrégulière est située entre deux zones régulières. La correction consiste à ajuster les pics en fonction des valeurs du fondamental aux extrémités des zones régulières et de la longueur de la zone irrégulière à traiter (Interpolation linéaire).

. La zone irrégulière est située en fin de zone voisée. La correction est effectuée par extrapolation linéaire à partir de la zone régulière la précédant.

. La zone irrégulière est en tête de la zone voisée. La correction est effectuée par extrapolation linéaire à partir de la zone régulière la suivant.

5. Résultats et commentaires

En général la détection des pics est satisfaisante.

La correction des pics aberrants n'intervenant que dans les cas suivants : fricative voisée faible, fin de voyelle nasale et surtout début de voyelle précédée de la plosive [P].

En effet la plosion du [P] perturbe fortement la structure pseudo-périodique du signal.

BIBLIOGRAPHIE

D. DOURS, R. FACCA

Analyse temporelle du signal vocal comparée à l'analyse fréquentielle du point de vue de la reconnaissance.
Projet SESORI IV/137 n° 74-94.

D. DOURS, R. FACCA

Méthode de segmentation et d'analyse par traitement direct du signal vocal. Application à la classification et la reconnaissance des voyelles et des consonnes.
Thèses présentées à l'U.P.S. - Toulouse 1974.

VIII èmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

DETECTION DE MELODIE PAR UN MODELE FONCTIONNEL DU SYSTEME AUDITIF PERIPHERIQUE

L. BASTET, J.M. DOLMAZON, B. GUERIN, V. SCHUPLIAKOV (Léningrad

Laboratoire de la Communication Parlée et de l'Instrumentation de Mesure *
Ecole Nationale Supérieure d'ELECTRONIQUE et de RADIOELECTRICITE
23, Avenue des Martyrs, GRENOBLE

Résumé :

On propose une méthode de détection de mélodie fondée sur l'utilisation d'un modèle fonctionnel du système auditif périphérique suivi d'un détecteur tenant compte des effets de phase qui régissent le fonctionnement des neurones innervant les cellules ciliées. On obtient ainsi, en temps réel, les évolutions représentatives de la "fréquence fondamentale" telle qu'elle peut être interprétée par le système auditif pendant la perception d'un signal de parole.

Summary :

A functional model of the peripheral auditory system connected with a detector taking into account the phase analysis occurring in the first neuron layers is used as a pitch detector. Variations of the fundamental frequency are displayed and can be interpreted as a result of peripheral auditory system processing during speech perception.

* Equipe de Recherche associée au C.N.R.S. n° 366 .

-DETECTION DE MELODIE PAR UN MODELE FONCTIONNEL DU SYSTEME AUDITIF PERIPHERIQUE-

1/ INTRODUCTION

Pour analyser les informations fournies au système nerveux central par le système auditif périphérique lors de la perception d'un message sonore, nous avons réalisé un modèle fonctionnel de ce système qui reproduit au mieux les diverses données expérimentales disponibles. Nous nous intéressons plus particulièrement ici à la détection des fréquences correspondant à la mélodie d'un signal de parole.

L'étude des propriétés des réseaux de neurones connectés aux récepteurs sensitifs fait apparaître l'importance de la distribution temporelle des informations sur la membrane basilaire. Ceci nous a conduits à développer une simulation de cette sensibilité aux écarts de phase.

Cependant, notre détecteur ne prétend pas être une nouvelle méthode de mesure de la fréquence fondamentale mais une approche du comportement du système auditif périphérique vis à vis du fondamental.

2/ METHODE DE DETECTION DE LA MELODIE

LE MODELE FONCTIONNEL DE L'OREILLE INTERNE

Notre modèle ayant déjà été décrit par ailleurs ([2],[3],[4]), nous n'en rappelons que brièvement la structure.

Nous avons reproduit la fonction de transfert de l'oreille moyenne par l'association en serie de filtres du premier et du deuxième ordre. Les propriétés de transfert de l'oreille interne (cochlée) ont été simulées à l'aide d'un double banc de filtres. Le premier banc est constitué par la mise en série de 128 filtres passe-bas du second ordre proches de l'amortissement critique. Chacun de ces filtres possède un gain en continu très peu supérieur à 1. La fonction de transfert existant entre l'entrée et la sortie d'un de ces 128 filtres reproduit les variations de pression de part et d'autre de la membrane, en un point d'abscisse donnée, en fonction du signal d'entrée. Le second banc est également constitué de 128 filtres passe-bas du second ordre, mais surtendus. Chacun d'eux est connecté à la sortie d'un des filtres du premier banc. Les fonctions de transfert des filtres résonnants simulent, pour chacun des 128 points considérés de la membrane, la relation existant entre les variations de pression de part et d'autre de la membrane et les déplacements de ce point.

Dans une première approche, nous n'avons pas tenu compte des effets non linéaires qui se produisent aux intensités sonores élevées (au-dessus de 70 dB). Le modèle a été, tout d'abord, simulé sur ordinateur puis, pour un fonctionnement en temps réel, une version câblée analogique a été réalisée et utilisée pour nos études sur la mélodie.

METHODE DE DETECTION

Les résultats actuels se rapportant à l'innervation des cellules ciliées (SPOENDLIN, 1975 ; MORRISON et al., 1975) montrent que les dendrites d'un neurone sont connectées par des synapses à plusieurs cellules ciliées réparties le long de la membrane basilaire. L'excitation de chaque récepteur produit un potentiel post-synaptique excitateur ou inhibiteur qui se propage le long des dendrites vers le soma du neurone. Ces potentiels post-synaptiques ne peuvent déclencher un potentiel d'action le long de l'axone que si leur somme au niveau du soma atteint, à un instant donné, un certain seuil (DUVAL, 1974). Or, ces potentiels

d'excitation prennent naissance à des distances différentes du soma neuronique : il doit donc exister une certaine relation de temps entre les instants de déclenchement de ces potentiels d'excitation pour qu'ils concourent à une dépolarisation suffisante, à un instant donné, de la membrane du neurone. Dans ce cas seulement le seuil est atteint et un potentiel d'action prend naissance au début de l'axone.

Pour pouvoir tenir compte de cette relation de temps, nous avons analysé l'évolution du déphasage des signaux entre des sorties voisines du modèle en fonction de la fréquence du signal d'entrée (figure 1). Le déphasage passe par un maximum lorsque la fréquence du signal d'entrée se situe entre les fréquences propres (f_{r_1} et f_{r_2}) des deux filtres résonnants considérés. En prenant un seuil correspondant à une certaine fraction du déphasage maximal, on peut détecter la présence ou non d'un signal de fréquence comprise entre f_{r_1} et f_{r_2}.

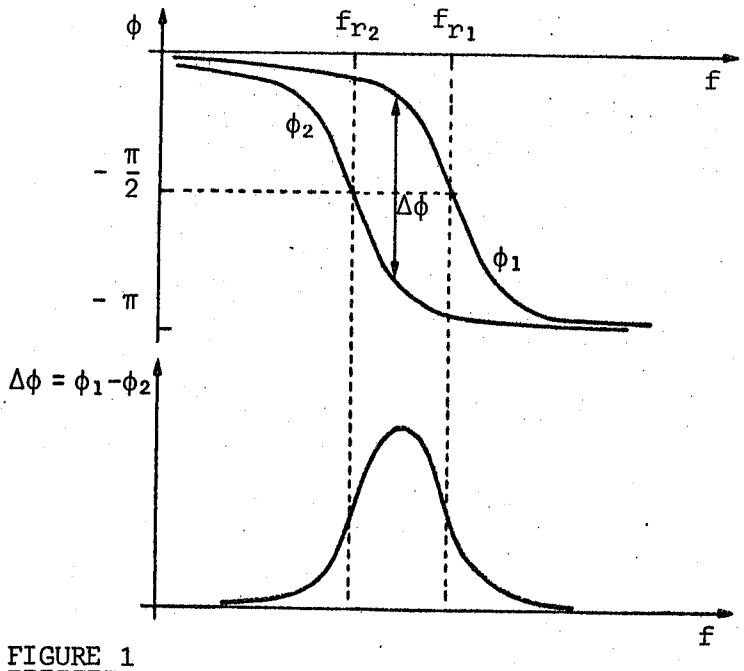


FIGURE 1
Variation du déphasage entre sorties voisines du modèle en fonction de la fréquence.
f_{r_1}, f_{r_2} : fréquences propres des deux filtres résonnants correspondant à deux points de la membrane ($x_1 < x_2$).

Pour effectuer cette détection, nous avons réalisé le montage correspondant au schéma synoptique de la figure 2. Chaque voie du détecteur est excitée par les tensions analogiques issues de deux filtres résonnants voisins. Les comparateurs transforment le signal d'entrée en signal rectangulaire et détectent les passages par zéro de chaque signal ;

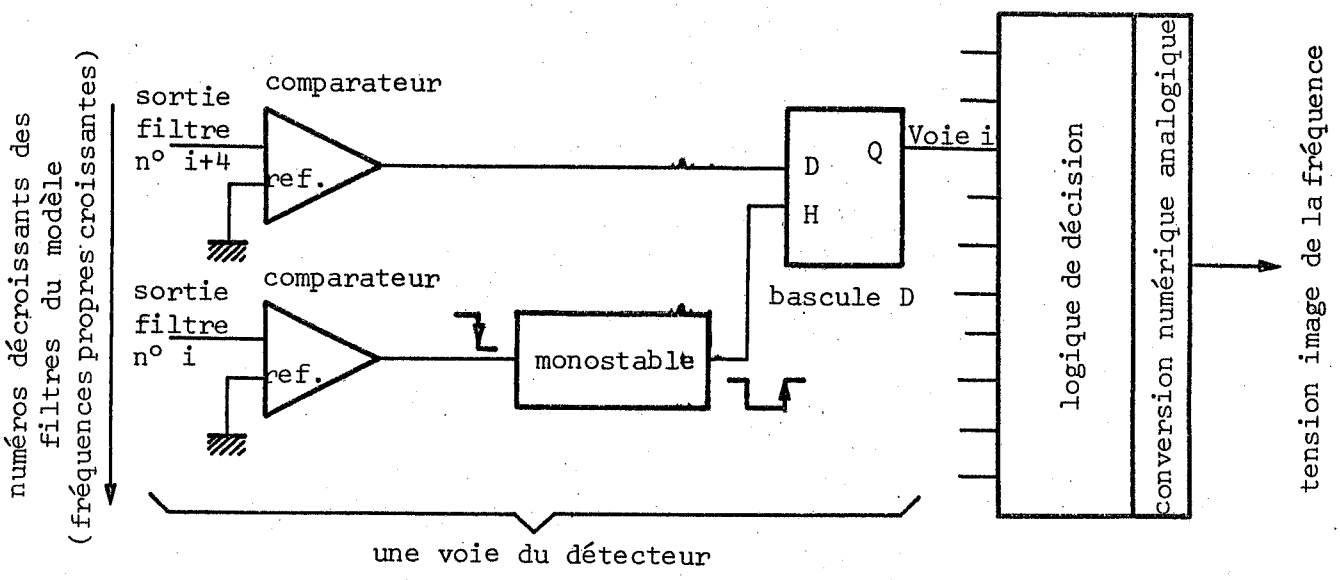


FIGURE 2
Schéma synoptique du détecteur de mélodie .

la largeur de l'impulsion délivrée par le monostable correspond au seuil de déphasage que l'on se fixe pour décider de la présence ou non de la fréquence propre à la voie considérée. La sortie de la bascule D est activée lorsque l'intervalle de temps séparant les passages à zéro des deux signaux dépasse le seuil fixé par le monostable. Les sorties de toutes les voies du détecteur sont ensuite regroupées dans un organe logique qui, en fonction des voies activées, décide de la valeur de la fréquence fondamentale détectée ou de son absence. La fréquence est obtenue en sortie sous une forme analogique qui permet sa mesure et son enregistrement. Actuellement, le détecteur se compose de vingt voies : les fréquences détectées s'étendent de 90 Hz à 200 Hz par pas de 5 Hz. La gamme choisie ne couvre certes pas toute l'étendue possible de l'évolution de la mélodie et la résolution fréquentielle pourrait être plus fine. La gamme et la résolution de notre système dépendent uniquement du nombre de filtres résonnants choisis et de l'écartement minimal de leurs fréquences propres. C'est pourquoi pour tester le bien-fondé de notre démarche nous avons limité le nombre des voies, mais une extension est possible. Pour l'instant, la stratégie de l'organe logique consiste à prendre pour fréquence détectée celle qui correspond à la voie excitée de numéro le plus faible, ce qui permet de distinguer le fondamental de ses harmoniques.

3/ COMPORTEMENT DU DETECTEUR FACE A DES SIGNAUX SIMPLES

DETECTION D'UNE FREQUENCE PURE

Le signal présenté à l'entrée du modèle est constitué de trains d'oscillations sinusoïdales de fréquence fixe. La figure 3 représente ce signal ainsi que la réponse du détecteur.

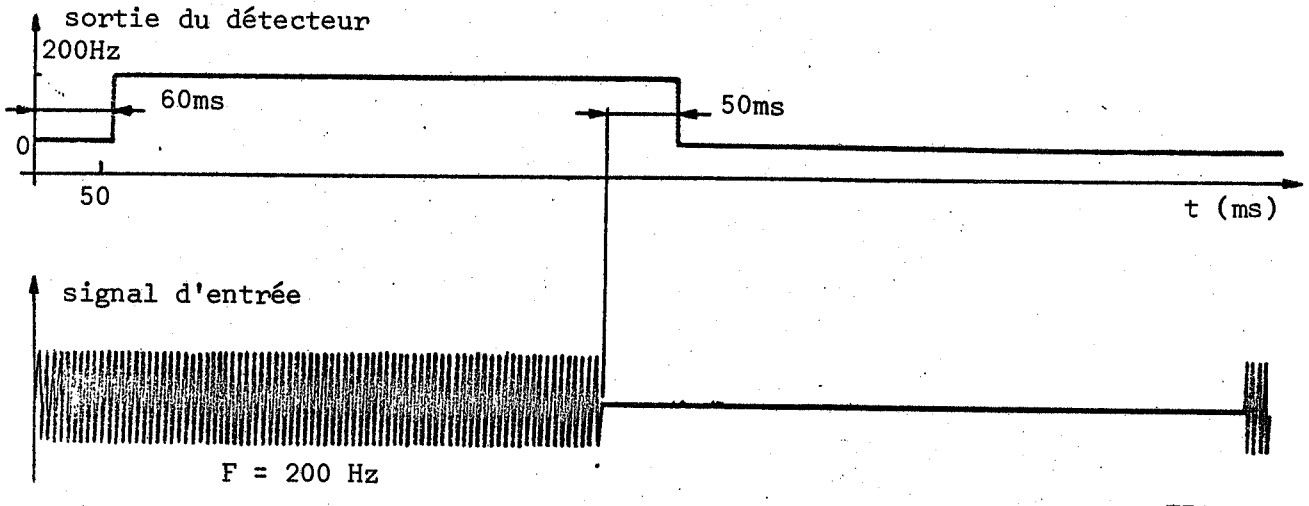


FIGURE 3
Détection d'une fréquence pure

On constate que la sortie du détecteur évolue avec un certain retard par rapport au signal d'entrée. Ce retard se décompose en deux termes. L'un est dû au temps de propagation des ondes acoustiques dans la cochlée : les fréquences de mélodie correspondant aux zones de la membrane basilaire les plus éloignées de la fenêtre ovale, ce temps de retard est maximal et il atteint 20 à 30 ms. L'autre terme est dû au temps de réponse du détecteur lui-même : la mesure de déphasage s'effectuant période par période, il faut, pour prendre une décision, attendre au moins une période correspondant à la fréquence détectable la plus faible (soit

environ 10 ms). Nous avons remarqué que ce temps de retard dépend beaucoup du facteur de surtension des filtres résonnants. Ainsi avec un facteur de surtension de 20 le retard en sortie du détecteur atteint 90 à 130 ms à la montée et 60 à 110 ms à la descente (ces temps dépendent de la voie étudiée). Un facteur de surtension de 5 permet, tout en conservant des maximums de déphasage suffisamment accusés, de réduire le retard de la mesure : nous obtenons, à la montée comme à la descente, des retards totaux de 50 à 60 ms.

REPONSE A UN SIGNAL DE FREQUENCE VARIABLE

Pour étudier la capacité de notre détecteur à suivre les variations de fréquence d'un signal, nous avons excité notre modèle par un signal sinusoïdal dont la fréquence est modulée sinusoïdalement à basse fréquence. La figure 4 montre la réponse du détecteur à ce signal. La fréquence maximale de modulation que l'on détecte avec un facteur de qualité de 20 n'est que de 2 Hz tandis qu'elle est de 6 Hz avec un facteur de qualité de 5. Au-delà de cette fréquence maximale de modulation, le détecteur ne peut plus suivre les évolutions de la fréquence du signal. La vitesse maximale d'évolution obtenue ($\Delta f/\Delta t$) atteint 2000 Hz/s environ. On peut comparer cette valeur avec celle que l'on rencontre dans la parole : la variation de fréquence peut atteindre, lors d'un glissando, cinq quarts de ton en 100 ms (L.J. BOE, 1973), ce qui représente une variation d'environ 1000 Hz/s. Il semble donc que la bande passante de détection du système auditif périphérique permette de suivre les évolutions les plus rapides de mélodie.

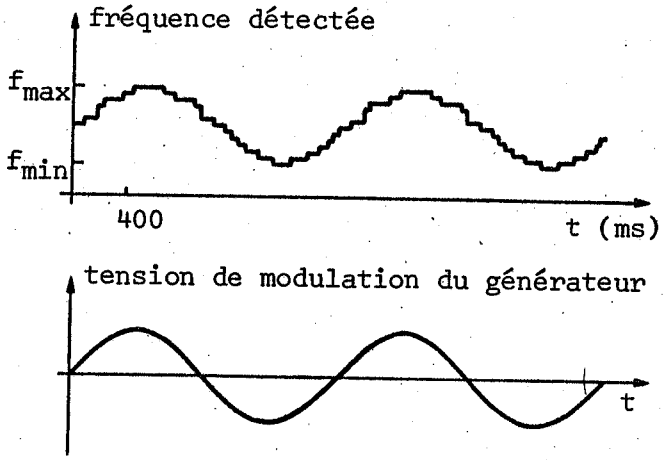


FIGURE 4
Réponse à un signal sinusoïdal de fréquence variable.

DETECTION DE LA MELODIE D'UN SIGNAL DE PAROLE

Les résultats précédents nous ont conduits à tester notre modèle avec un signal de parole naturelle enregistrée. La figure 5 donne un exemple de détection de mélodie (voir page suivante).

Le détecteur suit l'évolution de la fréquence fondamentale du signal avec un certain retard (50 à 60 ms). On constate une détection correcte du caractère voisé ou non-voisé, même pour des durées brèves sans fondamental. On peut aussi vérifier sur la figure 5 que la vitesse d'évolution de la mélodie détectée n'atteint pas la valeur maximale déterminée précédemment.

Cependant, nous avons noté que certaines portions de signal peuvent être considérées par notre détecteur comme non-voisées si la mélodie du signal atteint des zones de fréquence non couvertes par les vingt voies de mesure. Ce défaut sera facilement supprimé par une extension du nombre de voies du détecteur. Dans le cas où le rapport signal sur bruit à l'entrée diminue, la détection se dégrade car le bruit introduit des variations aléatoires dans la phase des signaux issus de la membrane basilaire.

On note que le mode de détection utilisé ne fait pas intervenir l'énergie

des différentes composantes spectrales du signal d'entrée, ce qui autorise la détection d'un fondamental de faible amplitude, même en présence d'harmoniques d'énergie très élevée.

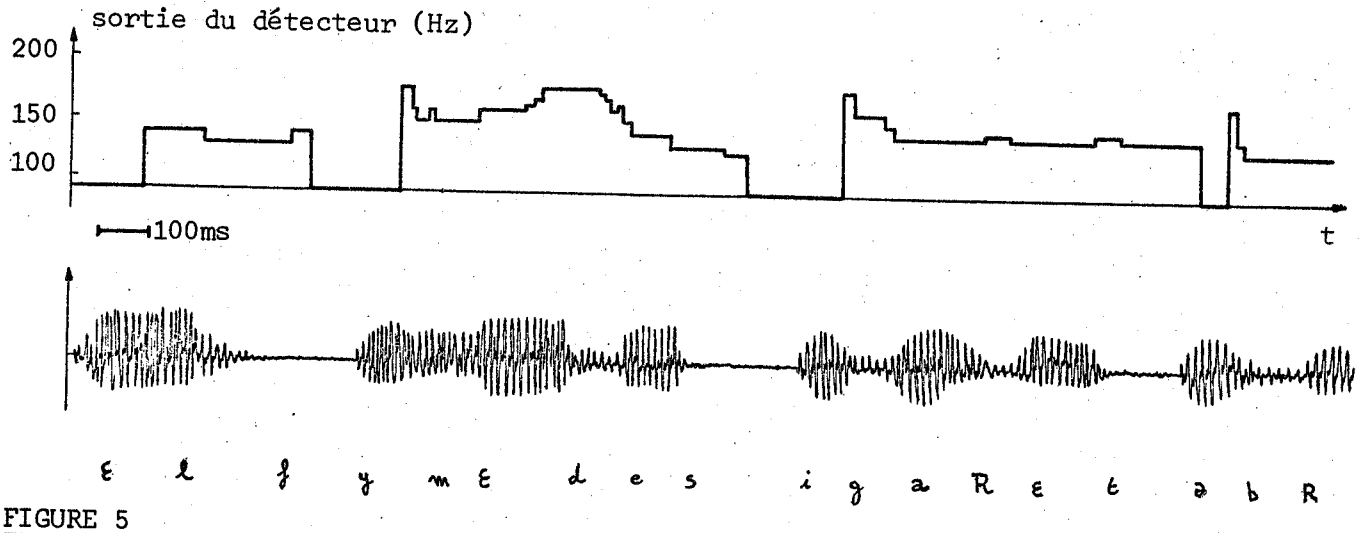


FIGURE 5
Détection de la mélodie d'un signal de parole : "Elle fumait des cigarettes brunes...".

CONCLUSIONS

Les résultats précédents montrent que l'étude des déphasages entre les informations issues de la membrane basilaire permet de situer la fréquence fondamentale d'un message acoustique. Cette analyse permet de proposer une explication vraisemblable de la détection du fondamental par le système auditif périphérique.

Compte tenu de l'innervation homogène des récepteurs sensitifs le long de la membrane basilaire, il est raisonnable de penser que ce traitement a lieu pour toute l'étendue du spectre audible. Le système nerveux central disposerait donc des informations relatives au fondamental et à tous ses harmoniques, quelle que soit la structure formantique de la parole perçue, l'énergie des différentes raies n'intervenant pas. On comprend alors comment pourrait être restitué le fondamental lorsqu'il a été artificiellement supprimé (filtrage téléphonique par exemple). Nous étudions actuellement un organe logique de décision évolué utilisant la souplesse d'un microprocesseur. Il pourrait restituer la valeur du fondamental absent lorsque la détection d'harmoniques permet de conclure à l'existence d'un fondamental dans le signal initial. Cet organe nous permettra d'utiliser une stratégie de décision beaucoup plus performante qui pourra améliorer les performances de notre détecteur.

La comparaison de nos résultats avec ceux de certaines expériences de psycho-acoustique devrait permettre de mieux préciser ce type d'analyse et d'améliorer ainsi la connaissance du fonctionnement du système auditif périphérique.

BIBLIOGRAPHIE

[1] L.J. BOË
Etude de l'interaction source laryngienne - conduit vocal dans la détermination des caractéristiques intrinsèques des consonnes du français (fréquence laryngienne) - Mesure de la durée
Bulletin de l'Institut de Phonétique de Grenoble, *II*, 1-24 (1973)

[2] L.A. CHISTOVITCH et al.
Physiologie de la parole - Perception de la parole par l'homme (en russe)
Nayka Ed., Leningrad (1976)

[3] J.M. DOLMAZON, V.S. SHUPLJAKOV, L. BASTET
Etude d'un modèle fonctionnel du système auditif périphérique
A paraître dans la Revue d'Acoustique

[4] J.M. DOLMAZON, L. BASTET, V.S. SHUPLJAKOV
A functional model of the peripheral auditory system in speech processing
Comm. au "1977 I.E.E.E. Int. Conf. on Acoustics, Speech and Signal processing", Hartford U.S.A. (9-11 mai 1977)

[5] F. DUVAL
Etude par modélisation électronique de structures à inhibition latérale et longitudinale
Thèse D.I., Grenoble (1974)

[6] D. MORRISON, R.A. SCHINDLER, J. WERSALL
A quantitative analysis of the afferent innervation of the organ of Corti in guinea pig
Acta Oto., *79*, 11-23 (1975)

[7] H. SPOENDLIN
Retrograde degeneration of the cochlear nerve
Acta Oto., *79*, 266-275 (1975).

VIII èmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

DETERMINATION DE LA PERIODE FONDAMENTALE DU SIGNAL VOCAL AVEC UN FILTRE DIGITAL

Wolfgang Hess

Université Polytechnique de Munich, Allemagne

Sommaire. Un algorithme est présenté qui détermine les cycles de voisement du signal vocal dans le domaine temporel. L'algorithme emploie une approximation simple du filtre inverse réalisé par une cascade de deux filtres digitaux non récursifs. D'abord, un filtre passe-bas coupe les formants hauts; d'après, le formant F1 est compensé individuellement par un filtre avec un zéro adaptif. La structure temporelle du signal de sortie de ce filtre est pareille à celle du signal glottique: on y trouve un extremum signifiant par cycle fondamental. Cet extremum marque préliminairement la frontière entre deux cycles successifs de voisement. Enfin, ces marques sont déplacées dans une position optimale. L'algorithme demande la présence de l'harmonique fondamentale dans le signal; pour signaux à bande limitée, il faut la restituer par une opération non linéaire en train du filtrage.

Abstract. The pitch determination algorithm presented in this paper derives the pitch periods from speech signals in time domain. It applies a crude approximation of the inverse filter which is realized as a two-step non-recursive digital filter. In the first step, a low-pass filter cuts off the higher formants; the second step compensates F1 by a time-variant zero. The output signal of the filter shows a time structure similar to the glottal waveform: one principal peak per pitch period which is detected and marked a preliminary period limitation. In a final step, these markers are shifted into an optimum phase position. The algorithm requires the presence of the first harmonic in the signal; applied to band-limited signals, the first harmonic must be reconstructed prior to marker detection by some nonlinear operation between the two filtering steps.

DETERMINATION DE LA PERIODE FONDAMENTALE DU SIGNAL VOCAL AVEC UN FILTRE DIGITAL

Wolfgang Hess

Lehrstuhl für Datenverarbeitung
Technische Universität München
Munich, Allemagne

1. Introduction.

Le signal vocal voisé est un signal acoustique approximativement périodique. Sa fréquence fondamentale f_0 ou - dans le domaine temporel - sa période fondamentale T_0 se représente par sa première harmonique. Le signal est engendré dans la glotte (la source vocale), où il reçoit sa périodicité. Le signal glottique est riche en harmoniques; sa structure temporelle est rélativement régulière. Dans le conduit vocal, le signal glottique est transformé en signal vocal avec une structure variable en temps et beaucoup plus compliquée, cette qui est contrôllée par les gestes d'articulation; ces dernières, enfin, caractérisent l'information phonétique contenue dans la parole.

La détermination du fondamental à travers le signal vocal est un problème difficile de l'analyse de la parole /3/. il y a des methodes nombreuses qui, faute de place, ne peuvent pas être discutées ici. Dans le domaine temporel, par example, on peut rechercher la structure du signal vocal et, par là, déterminer le période fondamentale par un algorithme de décision qui sera normalement assez compliqué. D'autre part, on peut prétraiter le signal vocal par un filtre afin d'obtenir une structure plus simple et régulière, dont l'on puisse dériver la période fondamentale plus facilement. Un de ces signaux plus réguliers est le signal glottique qui est reconstitué du signal vocal par le filtre inverse. La fonction de transfert de ce filtre inverse est le réciproque de celle du conduit vocal; donc, en appliquant ce filtre, on obtiendra le signal glottique reconstitué avec sa structure régulière.

Dans cet article, une methode sera décrite qui emploie une approximation simple du filtre inverse réalisée par une cascade de deux filtres digitaux non récursifs.

Le problème de trouver où il y a des sections voisées dans le signal vocal sera omis de cèt article. Chaque appareil ou algorithme à détection du fondamental doit contenir quelque discrimination entre silence et signal vocal aussi que entre des signaux voisés et non voisés. La methode appliquée ici est décrite ailleurs /1/ et, faute de place, ne peut pas être répétée dans cet article. Donc, on pourra supposer que tous les sections du signal vocal différentes de parole voisée ayont été marquées et soient exclues du traitement par le detecteur du fondamental.

2. La réalisation du filtre.

2.1 Principe (Fig. 1) et exigences. - Le signal glottique, de quoi la périodicité peut être dérivée simplemênt, est recontitué du signal vocal par le filtre invers. Pour la détection du fondamental cependant on ne réalise qu'un approximation simple. On demande que la dépense de calcul ne soit pas grande, et que le rapport signal/bruit du signal filtré soit aussi grand que possible, c'est à dire que aussi beaucoup d'harmoniques que possible doivent passer le filtre. Des rélations de phase entre le signal vocal et le signal filtré sont importantes si l'algorithme sera employé afin de détecter la phase correcte des cycles fondamentaux, c'est à dire, de détecter le debut de la réponse du conduit vocal à une nouvelle impulsion glottique. (Naturellement, cela ne se peut trouver juste que si le signal est libre de distorsions de phase.) Puisque le filtre inverse ne demande que des zéros, il peut être réalisé en façon non récursive. Mise en oeuvre avec une réponse linéaire de phase, le filtre fournira une rélation fixe de phase entre le signal directe et le signal filtré.

Comment peut-on simplifier la conception originale du filtre inverse? Evidemment, on doit compenser le formant F1 individuellement par un filtre adaptif. A supprimer les hauts formants, un filtre passe-bas à une fréquence de coupure de 1500 Hz suffira. Ce filtre supprime F3 et les formants plus hauts complêtement; F2 est éloigné en grande partie. Cette configuration est donc réalisée par une cascade de deux filtres digitaux: d'abord, le filtre passe-bas éloigne les hauts formants; le filtre suivant compense F1 par un zéro adaptif. Quant à la dépense de calcul, on gagne considérablement de la vitesse si l'on n'a besoin ni des multiplications ni des divisions pendant la course du filtrage. Afin d'éviter ces opérations, les coefficients des filtres ne doivent

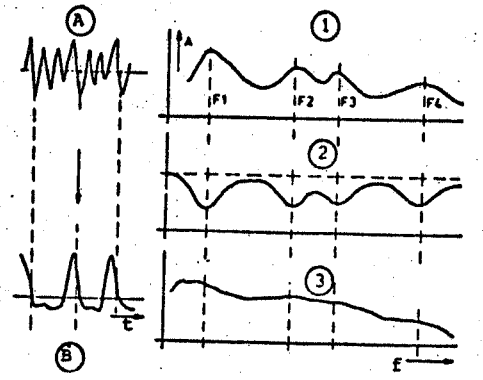


Fig.1: Principe du filtre inverse

A Signal vocal
B Signal glottique (rest.)
1 son spectre
2 fonction de transfert du filtre inverse
3 son spectre

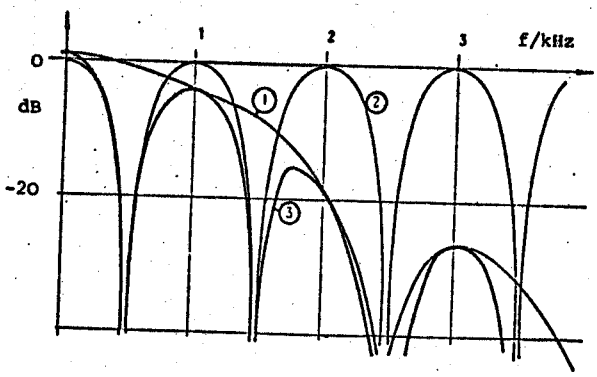


Fig.2: Fonction de transfert du filtre réalisé

1 Préfiltrage
2 Filtre à zéro adaptif
3 Fonction totale (F1=500 Hz)

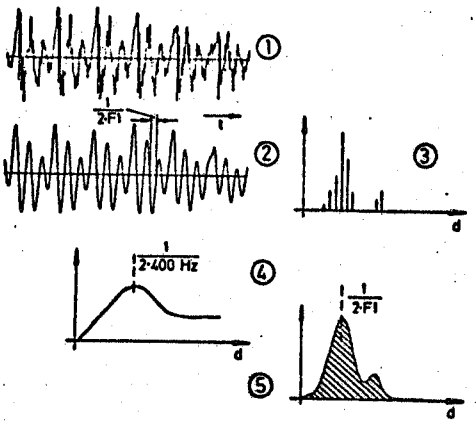


Fig.3: Détermination du formant F1.

1 Signal original (vocal)
2 Signal préfiltré
3 Histogramme directe
4 Fonction pondérée
5 Histogramme final
d Distance de passages par zero successifs

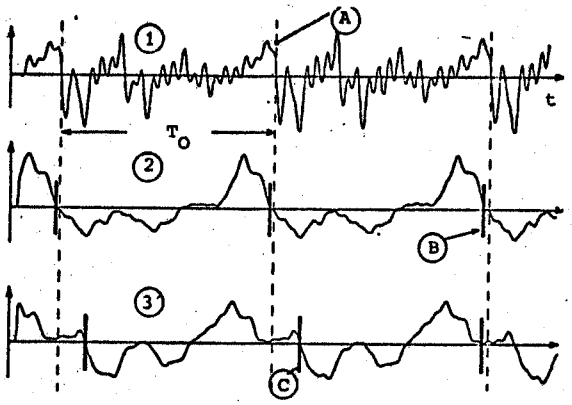


Fig.4: Fonctionnement de l'algorithme dependant de la determination de F1.

1 Signal original (F1=480 Hz)
2 Signal filtré (zéro situé à 500 Hz)
3 Signal filtré (zéro situé à 400 Hz)
A Marques en position propre
B Marques préliminaires régulières
C Marques préliminaires irrégulières

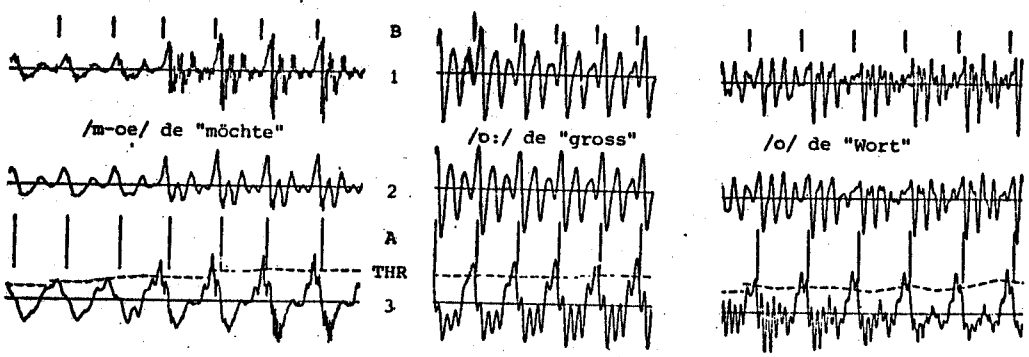


Fig.5: Examples de fonctionnement de l'algorithme à détermination des cycles fondamentaux (Locuteur: GRR)

1 Signal original (vocal)
2 Signal préfiltré (PFS)
3 Signal filtré (FS)
A Marques préliminaires
B Marques après le déplacement final
THR Seuil

prendre des valeurs que 1, 0, et -1. Naturellement, c'est une grande restriction pour le projet du filtre; en consequence, zéros parasitaires apparaissent hors de la bande désirée. Les expériments cependant ont montré que cette restriction ne degrade pas sévèrement le fonctionnement de ce filtre.

2.2 Préfiltrage et détermination du formant F1 (Fig. 3) - Le premier étape du filtrage, le "préfiltrage", demande un filtre passe-bas à fréquence de coupure de 1500 Hz. A cela, une cascade de deux filtres passe-bas à fonction de fente est appropriée. Le filtre à fonction de fente est défini par

$$y(n) = \sum_{i=n-k+1}^{n} a(i) \qquad \text{ou, récursivement} \qquad y(n) = y(n-1) + a(n) - a(n-k) \tag{1}$$

Dans ces équations, a représente le signal d'entrée, y le signal de sortie, et k la largeur de fente. Si l'on suppose une fréquence d'echantillonnage de 20 kHz, les valeurs de k seront égales à 8 et 5 pour rendre les zeros à 2,5 kHz, 4 kHz etc.

Dans le signal préfiltré (le "PFS"), les hauts formants disparaissent. Le formant F1 demeure et doit être éliminé individuellement. Avant sa élimination, l'algorithme doit avoir au moins une idée de sa valeur momentanée. Pour cette raison, F1 est déterminé approximativement. L'algorithme prend la distribution des distances des passages successifs par zéro du PFS. Après un lissage et l'application d'une fonction ponderée pour éviter des erreurs introduites par des restes du formant F2 ou par une harmonique fondamentale très forte dans le PFS, l'histogramme correspondant (Fig. 3) révèle un maximum signifiant à la valeur du réciproque de deux fois la fréquence du formant F1.

2.3 Elimination du formant F1 et marquage préliminaire des cycles fondamentales (Fig. 5) - Pour l'élimination du formant F1, on demande un filtre avec un zéro voisin de la fréquence momentanée de F1. Afin d'obtenir une bonne valeur du rapport signal/bruit du signal filtré, beaucoup d'harmoniques doivent passer le filtre, et l'on préfère un filtre à élimination de bande. Pour des raisons de vitesse, ce filtre est aussi réalisé sans multiplications. On prend un filtre à deux impulsions (FDI) defini par

$$y(n+m) = a(n+\tfrac{N}{2}) + a(n-\tfrac{N}{2}) \qquad \text{avec le premier zéro situé à} \qquad f_{01} = \frac{1}{2 \cdot N \cdot T} \tag{2} \quad (3)$$

Ici, a represente le signal d'entrée, y le signal de sortie. N dénote la distance entre les deux impulsions, T l'intervalle d'échantillonnage; enfin, m est introduit en raison de causalité. f_{01} représente le premier zéro de ce filtre. Aux fréquences de $3 \cdot f_{01}$, $5 \cdot f_{01}$ etc., on trouve des zéros parasitaires dûs à la restriction du choix des coefficients du filtre (Fig. 2).

La réalisation de ce filtre est très simple: la distance N entre les deux impulsions, pour éliminer F1, doit être égale au réciproque de deux fois cette fréquence, c'est-à-dire, égale à la location du maximum dans le histogramme utilisé par la détermination de la valeur momentanée de F1. Le signal de sortie de ce filtre, le "signal filtré" FS, est l'approximation désirée du signal glottique. Puisque ce signal présente une structure temporelle beaucoup plus régulière que le signal vocal original, on peut y employer une analyse simple de seuil (Fig. 5) pour le marquage préliminaire des cycles fondamentals. Cette analyse de seuil est déjà préparée dans le PFS qui est normalisé à l'égard de la valeur moyenne absolue à court terme définie par

$$L_S = \sum_{i=n}^{n+N-1} | a(i) | \tag{4}$$

a denote le signal vocal, et N représente la longueur de la fenêtre appliquée dans cette operation qui était nécessaire pour la discrimination entre les sections de silence et celles de parole /1/. Au signal filtré, on emploie un seuil local THR (fig. 5) qui est adapté aux valeurs des extrema du FS de temps en temps. Chaque maximum du FS dépassant le seuil est marqué en limite préliminaire de cycle fondamental ("marque préliminaire"). La location exacte de ce maximum est sous une dépendance certaine des parametres du filtre et de la fréquence du formant F1. Donc, les marques préliminaires sont immédiatement déplacées du maximum en passage par zéro suivant (voir fig. 5 num. 4 et fig. 6). Ce point possède la meilleure correspondance entre le FS et le signal original. Il est sûr mais il n'est pas le début désiré du cycle fondamental (voir sect. 2.1). On ne peut guère extrahir le point du début par des règles automatiques du signal filtré; il faut l'extrapoler de points plus sûrs. Pour cette raison, le point de début du cycle fondamental est defini en façon suivante: Puisque le signal glottique commence en direction positive, et puisque la longueur du conduit vocal est petite comparée par la longueur de l'onde acoustique du formant F1, le cycle fondamental individuel, qui est dominé par le formant F1, commencera en direction positive. Donc, le point de début sera situé la moitié de la durée d'un cycle du formant F1 avant le point de la marque preliminaire (Figs. 5 et 6).

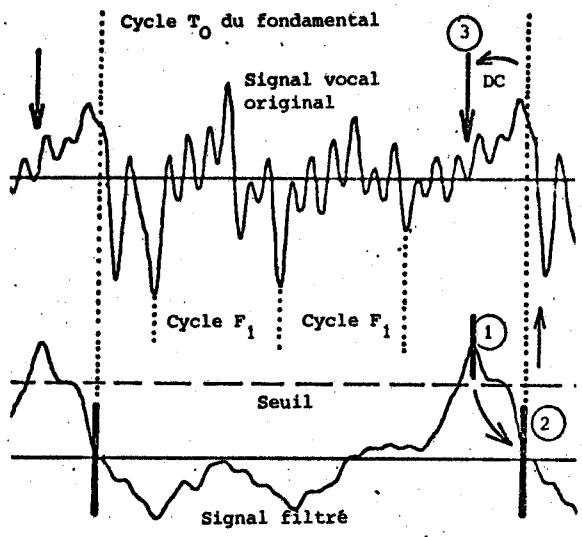


Fig.6: Evaluation des cycles fondamentaux

1 détection primaire par analyse de seuil
2 déplacement en passage par zéro suivant
3 déplacement final en position optimale

DC demi-cycle de l'oscillation de F1

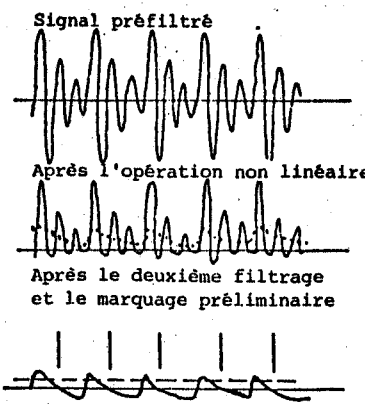


Fig.7: Principe de l'opération non linéaire

PFS Signal préfiltré
FS Signal filtré

1 Caractéristique de l'opération non linéaire
2 rectification monophasée
3 rectification biphasée avec le facteur b qui peut varier entre 0 et 1

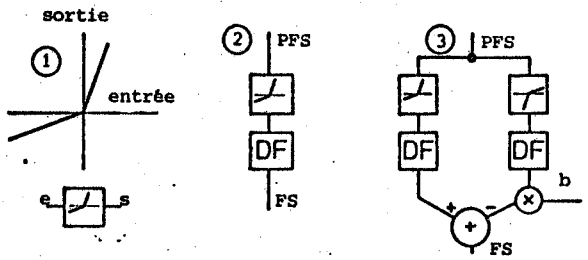
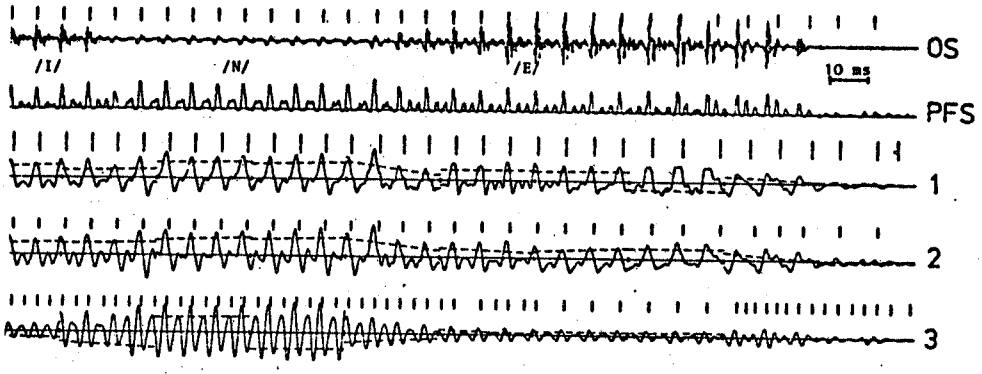


Fig.8: Fonctionnement pour signaux à bande limitée (300..3400 Hz)

OS Signal original
PFS Signal préfiltré (après l'operation non linéaire)
1-3 Signaux filtrés
1 nonlinéarité forte
2 nonlinéarité faible
3 version linéaire



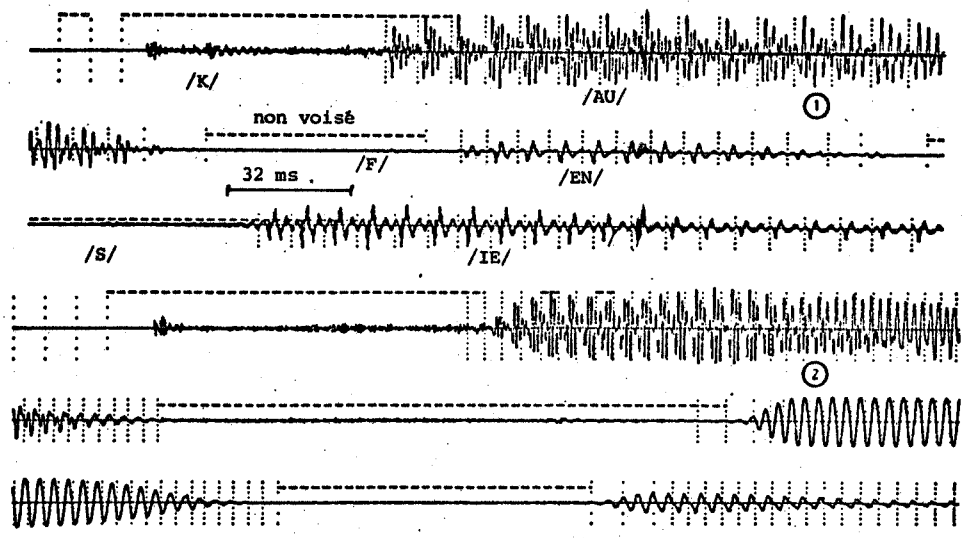


Fig.9: Examples de fonctionnement de l'algorithme

1 Locuteur: GRR
2 Locutrice: ICR

Signaux à bande limitée (300...3400 HZ)

Version non linéaire.

Parlé: Partie du début de la phrase allemande "Kaufen Sie jede Woche ..."

2.4 Contrôle et corrections. - De quelle accuratesse doit-on déterminer le formant F1? Cette question est importante en présence de signaux parasitaires et pendant des transitions rapides de F1 où la valeur nouvelle ne pourrait être mesurée qu'avec retard. Afin d'obtenir une valeur changée de l'histogramme, qui est mis à jour tout le temps, l'algorithme demande un quantité certain de passages par zéro. En conséquence, ce mesurage peut causer un retard de quelques cycles fondamentals jusqu'à ce que les parametres du filtre sont adaptés. Ce retard depend de la fréquence de F1; il sera moins long pour les hautes valeurs de F1 parce que il y a plus de passages par zéro dans le même temps. C'est-à-dire, quand la valeur mesurée de F1 est en erreur, elle a la tendence de prendre des valeurs trop hautes. Dans ce cas, F1 sera amorti insuffisamment, mais une déviation de 25% dans cette direction se peut passer sans dégradation sérieuse du fonctionnement du marquage. Dans l'autre direction - ce que se passe beaucoup moins souvent - l'algorithme est plus sensible (Fig. 4). Une déviation du zéro du FDI vers des fréquences trop basses cause un déphasage de 180^{0} de l'oscillation de F1 qui, ensemble avec l'amortissement insuffisant, détruira la structure temporelle du signal filtré et donc causera des erreurs de marquage.

Jusqu'à ici, on a néglecté l'influence de F2 quand ce formant prend une valeur basse proche de F1. Dans ce cas, aucun des filtres appliqués supprimera ce formant suffisamment (Fig. 5, à droite). A cause de cela, on réemploi l'algorithme du mesurage de F1 au signal filtré. Quand l'histogramme correspondant présente un maximum signifiant à une fréquence plus haute que F1, on applique un deuxième FDI afin de supprimer le formant F2 individuellement.

Enfin, on doit regarder un cas spécial qui est important particulièrement quand la fréquence fondamentale est haute. Dans ce cas, elle peut coïncider avec le formant F1, et on n'y doit point de tout appliquer le FDI puisque un zéro coïncidant avec la fondamentale détruirait la structure temporelle du signal filtré et causerait des erreurs sérieuses en train du marquage. A cause de cela, la distance des impulsions N du FDI n'est permise pas de surpasser un quart de la longueur de la période fondamentale expectée.

Après le marquage, l'algorithme conduit un contrôle de régularité. Dans la réalisation de simulation par ordinateur, le signal vocal est subdivisé en intervalles successifs de 25,6 millisecondes ("blocs"). Un bloc est trouvé régulaire quand les cycles fondamentaux individuels ne divergent plus de 15%. Cela contient le cycle entre la dernière marque du bloc précédant et la première marque du bloc actuel aussi que la valeur expectée (VE) de la periode fondamentale. Quand ce contrôle ne rend pas un résultat positif, il est répété sans regarder la VE; si, après cela, l'on ne trouve pas encore une régularité, il faut employer des algorithmes correctifs qui, faute de place, ne peuvent pas être discutés ici. La VE est adaptée à la periode fondamentale actuellement mesuré quand l'algorithme détecte plusieurs blocs contigus régulaires.

3. Application d'algorithme à des signaux à bande limitée.

En traitant des signaux sans distorsions, les cycles fondamentals sont déterminés avec un taux d'erreurs de moins que 2%. L'algorithme cependant demande la présence de l'harmonique fondamentale dans le signal. Cette harmonique peut être fort amortie, mais elle ne doit pas disparaître complètement; l'algorithme va défaillir quand appliqué à des signaux à bande limitée (par example signaux à bande de téléphone, 300...3400 Hz, voir fig. 8 ligne 3) sans traitement supplémentaire. A cause de cela, il faut appliquer une opération non linéaire en train du filtrage (Fig. 7). En général, une telle opération est definie par

$$y(n) = f\,(a(n)) \qquad (5)$$

y(n) dénotant le signal de sortie, a(n) le signal d'entrée, et f une fonction arbitraire de l'amplitude momentanée (par example, logarithme, valeur absolue, redressement monophasé). Cette opération reconstitue l'harmonique fondamentale dans le signal filtré; elle travaille le mieux quand il y a beaucoup d'harmoniques en signal. D'autre part, l'expériment montre qu'une opération non linéaire appliquée au signal original va dégrader la détermination de F1. A cause de cela, l'opération non linéaire est employé au PFS après la mesurage de F1. Quant au choix de l'opération non linéaire, le redressement monophasé donne des résultats acceptables (Fig. 8, voir fig. 7 pour le principe). Le redressement biphasé cependant se rapporte encore mieux aussi longtemps qu'on peut garantir que l'harmonique fondamentale manque dans le signal original; l'algorithme à redressement biphasé défaillirait quand appliqué à une voix féminine avec le fréquence fondamentale plus haute que 300 Hz. Donc, la disposition de fig. 7.3 est employé. Le PFS est exposé au redressement monophasé positif et negatif. Les deux signaux séparés résultants de cette opération sont filtrés separément par le FDI et, d'après, combinés à l'aide d'un facteur multiplicatif b qui peut varier entre 0 et 1. Quand b est égal à 0, on applique le redressement monophasé; quand b est égal à 1, le redressement biphasé est employé. La valeur b égale à 0,5 rend un bon compromis. Dans ce cas, on peut employer le filtre non linéaire à des signaux à bande limitée aussi qu'à des

signaux quels n'ont pas de distorsions.

4. Conclusions.

Un algorithme était présenté qui détermine les cycles fundamentaux du signal vocal dans le domaine temporel. L'algorithme applique une approximation simple du filtre inverse réalisé par une cascade de deux filtres digitaux à phase linéaire. Les frontières des cycles fundamentaux individuels sont directement dérivés du signal de sortie de ce filtre. En recherchant des signaux à bande limitée, on insère une opération non linéaire entre les deux étapes du filtrage. En fig. 9, on voit le resultat pour deux signaux à bande limitée (300...3400 Hz), parlés par une voix masculine et une voix feminine.

L'algorithme est programmé en Fortran IV et s'execute sur un ordinateur PDP 11/45 où il nécessite ca. 20 fois le temps reel pour traiter des signaux de parole continue. Il est possible d'évaluer cette methode en hardware spécial pour le faire travailler en temps reel; cela cependant demanderait une révision majeure des étapes de cõntrole et de correction.

Références

/1/ Hess, W.: A Pitch-Synchronous Digital Feature Extraction System for Phonemic Recognition of Speech. IEEE Transactions on Acoustics, Speech, and Signal Processing. ASSP-24 (1976), pp. 14...25

/2/ Hess, W.: Time-domain Digital Pitch Extractor for Band-Limited Speech Signals. Preprint No. 1215, Audio Engieering Society, 56th Convention, Paris, March 1977.

/3/ Des références plus détaillés peuvent être trouvées dans /1/ et /2/.

VIII èmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

DETECTEUR NUMERIQUE DE MELODIE

Ismaïl EL MALLAWANY et Jean-Frédéric ZURCHER

C.N.E.T., LANNION

• Résumé :

On décrit la simulation d'un détecteur numérique de fondamental, F_o, du signal de parole. Ce dispositif est constitué de deux chaînes de traitement, dont les fonctions sont la détection du voisement et la mesure de F_o. Le détecteur de voisement fait appel à un préfiltrage du signal vocal suivi d'un compteur du nombre de passages par zéro. Le signal est classé voisé ou non selon que ce nombre est inférieur ou non à un certain seuil. La chaîne de mesure de F_o ne livre un résultat différent de zéro que dans le cas d'un signal voisé. Cette chaîne comprend un préfiltrage, un compresseur de la dynamique, deux détecteurs des crêtes positives et négatives, et un algorithme de mesure et de correction des valeurs de F_o. Les tests entrepris ont montré la grande fiabilité de ce dispositif.

• Summary :

The simulation of a digital pitch detector is described. The system is organized into two building blocks, namely, a voicing detector and a pitch frequency estimator. The voicing decision is based on the computation of the number of zero-crossings (NZC) detected in the prefiltered speech signal. A speech signal interval is said to be voiced if the NZC remains below a given threshold. The pitch frequency estimator yields zero pitch value for unvoiced segments. The measurement of the pitch period calls for a prefiltering process, an automatic gain controller, two peak signal detectors one for each polarity of the signal, and an algorithm for the measurement and the correction of the pitch period values. The global system has been tested and proved to be highly reliable.

DETECTEUR NUMERIQUE DE MELODIE

Ismaïl EL MALLAWANY et Jean-Frédéric ZURCHER

C.N.E.T., LANNION

I - INTRODUCTION

Cet exposé présente une simulation d'un détecteur numérique du fondamental, F_0. L'objet de cette étude fut de définir les caractéristiques d'un tel appareil en s'inspirant d'un dispositif analogique [1], réalisé au Département ETA du CNET-LANNION. Le simulateur devait réunir les propriétés suivantes :

- fonctionner en temps réel
- être indépendant du locuteur
- accepter un fondamental affaibli
- être peu sensible au bruit de fond
- admettre de grandes variations de dynamique

Le simulateur est constitué de deux chaînes de traitements dont les fonctions sont la détection du voisement et la mesure de $T_0 = 1/F_0$. Le détecteur de voisement, qui est autonome, est fondé sur le calcul du nombre de passages par zéro du signal de parole, qui est préalablement mis en forme. La mesure de T_0 est basée sur le principe de la détection des crêtes significatives du signal après une opération de filtrage ayant pour objet le renforcement de F_0 et une compression de la dynamique de l'enveloppe des crêtes maximum.

L'étude du simulateur s'est articulée en quatre temps, à savoir :

- La simulation fidèle du dispositif analogique sans contraintes d'arithmétique.
- L'introduction de modifications susceptibles d'augmenter sa fiabilité et de réduire sa complexité.
- La reconversion du simulateur en un mode de fonctionnement mieux adapté au domaine numérique.
- La prise en compte des contraintes du calcul en virgule fixe avec une dynamique limitée.

II - DETECTION DU VOISEMENT

II-1 - Mode de fonctionnement

Le schéma fonctionnel du détecteur de voisement est présenté sur la figure 1. Le filtre 1, dont les caractéristiques sont présentées sur la figure 2, a pour fonction la mise en forme du signal. Cette opération de filtrage diminue considérablement l'énergie des sons non-voisés dont l'énergie est centrée vers les hautes fréquences. On sait, par ailleurs, que l'énergie des sons voisés est concentrée dans les basses fréquences et se trouve, de ce fait, moins affaiblie. Le signal filtré est amplifié par un facteur G, puis additionné au signal émis par un oscillateur. Ce dernier signal est fixé à un niveau qui détermine le seuil d'énergie en-dessous duquel tous les sons sont reconnus comme non-voisés. De ce fait, la fréquence d'oscillation doit être très élevée afin de produire

un grand nombre de passages par zéro. Dans le simulateur, la fréquence retenue est égale à la moitié de la cadence d'échantillonnage, $f_e = 1/T$; ce qui conduit à émettre une valeur fixe avec alternance de son signe.

Le détecteur des passages par zéro est composé de trois éléments :

- Un écrêteur symétrique, qui forme un ensemble avec le facteur de gain G à la sortie du préfiltrage. L'objet de cet ensemble est de renforcer le niveau des signaux de faibles énergies et de limiter le niveau des signaux à fortes énergies.

- Le filtre 2 (passe-bas ; Fc = 180 Hz) qui avec la branche supérieure et la soustraction représente un filtre passe-haut.

- Comparateur : la comparaison des deux signaux en provenance de K et du filtre 2 est réalisée par une soustraction et la transmission du bit de signe. Ce bit est comparé avec le bit de l'instant précédant. Si le bit de signe est passé de 0 à 1, le "monostable" se trouve déclenché.

Le "monostable" est constitué d'un décompteur et d'une mémoire dont le contenu est une tension négative V_a. Lors de son déclenchement, le monostable délivre N impulsions de valeur V_a à l'intégrateur, tel que $N.T \simeq 0{,}8$ms.

L'intégrateur reçoit deux valeurs s_m (V_a ou 0) qui est la sortie du "monostable" et $V_b (\ll |V_a|)$ qui est une constante positive. La sortie, s_i, de l'intégrateur sera d'autant plus faible que les passages par zéro sont plus nombreux et vice-versa. L'évolution temporelle de s_i se présente sous la forme d'une courbe en dents de scie apériodique. Un écrêteur a été placé dans la boucle d'intégration afin d'éviter que s_i n'atteigne des valeurs trop fortes ou trop faibles, ce qui aurait l'inconvénient de rendre le temps de réaction à une transition dans le type de son dépendant de la durée du type de son précédant la transition.

La décision sur le voisement ou non du signal de parole peut s'effectuer à la sortie de l'intégrateur. Cependant, cette sortie évolue irrégulièrement, ce qui signifie qu'il peut exister des oscillations autour d'une valeur de seuil. Pour supprimer les rebondissements indésirables, la sortie de l'intégrateur est lissée par un passe-bas (Fc = 35 Hz). C'est la sortie de ce filtre qui est comparée à un seuil. Si la valeur est plus grande que le seuil, le son est voisé, sinon il est non-voisé.

TESTON Bernard
10, Butte des 3 Moulins
13 - AIX-EN-PROVENCE

II-2 - Résultats

Ce détecteur a été testé sur un corpus constitué d'une longue phrase répétée par cinq locuteurs. Les résultats montrent que la décision voisé/non-voisé est très nette. Cependant, le temps de réaction du détecteur à un changement de type de son est quelque peu long. Ce retard est provoqué par deux facteurs :

- l'importance de l'écart entre la valeur initiale de la sortie de l'intégrateur au début de la transition et la valeur du seuil
- le temps de propagation dans le filtre de lissage qui est relativement long.

Il faut, cependant, noter que ce détecteur a été développé pour un vocoder à canaux et que le retard constaté dans la réponse correspond à celui introduit par les filtres des canaux, d'où la synchronisation des données spectrales avec l'information de voisement. Ce retard peut être diminué sur le plan du système global, si les traitements pour la mesure de T sont conçus bloc par bloc et non séquentiellement.

III - MESURE DE LA PERIODE FONDAMENTALE, T_0

III-1 - Versions séquentielles

On a étudié deux versions dont le mode de fonctionnement est du type séquentiel. La première est une simulation assez fidèle du dispositif analogique. La seconde (figure 3), qui est une version modifiée, a été conçue afin d'améliorer légèrement les résultats de mesure et de simplifier le mode de fonctionnement donc de réduire le volume des calculs.

Dans les deux versions le signal de parole est préalablement filtré afin de faciliter la détection des crêtes significatives. Ce préfiltrage comporte : un passe-haut (Fc=70 Hz, 24 dB/octave), un passe-bas (Fc=400Hz, 12 dB/octave) et un coupe 50 Hz.

Le signal filtré subit une compression de l'enveloppe des crêtes maximum. En effet, si la détection des crêtes significatives est facile quand cette enveloppe varie peu ou pas, il n'en va pas de même dans le cas d'évolution rapide de l'énergie du signal. Dans la version "originale", la chaîne de traitement suit deux branches distinctes mais identiques ; l'une pour la détection des crêtes positives et l'autre pour les crêtes négatives. Les deux compresseurs utilisés se caractérisent par un temps de réaction instantané à la montée et un temps de réaction relativement long à la descente. Cet effet est obtenu à l'aide d'une mémoire de crête qui commande la compression. Cette valeur mémorisée décroît à peu près linéairement après un certain retard. Dans la version modifiée, la chaîne de traitement demeure unique jusqu'à la détection effective des crêtes. Dans ce cas le compresseur unique est commandé par une fonction de "mémoire de crête" dérivée du signal redressé en double alternance.

Une deuxième opération de filtrage précède la détection des crêtes. Dans la version "originale" les deux filtres des branches identiques présentent une réponse qui centrée sur 400 Hz décroît régulièrement de part et d'autre pour atteindre -48dB à 700Hz et en dessous de -38dB pour des fréquences inférieures à 75Hz. Dans la version modifiée le filtrage est réalisé à l'aide de trois filtres: un passe-haut (Fc=70Hz,6dB/octave), un passe-bas elliptique (Fc=500Hz et affaiblissement hors-bande $\leq$-30dB) et un passe-bande (250 à 450 Hz,6dB/octave).

La détection des crêtes dans les deux cas porte séparément sur les maxima positifs et négatifs. Ainsi le détecteur de crêtes positives comporte une mémoire de crête dont la fonction perte de mémoire a un profil qui évite au mieux l'intersection avec des crêtes non significatives. Lorsque le signal d'entrée dépasse la valeur de crête en mémoire, cette dernière est instantanément raffraîchie. A chaque instant où il y a raffraîchissement le détecteur délivre une impulsion. Ainsi périodiquement le détecteur émet une suite d'impulsions dont la dernière coïncide avec le maximum de la crête du signal d'entrée. Par conséquent, les deux détecteurs de crêtes délivrent des suites de créneaux.

Des deux sorties précédantes l'une est qualifiée de plus favorable selon un certain critère. Dans la première version, ce critère est fondé sur l'intégration de la différence entre les valeurs de mémoires de crêtes positives et négatives. Si le résultat est positif c'est la chaîne des crêtes de même signe qui est dite la plus favorable. Dans la version modifiée, le critère est fondé sur l'écart entre le nombre de crêtes positives et celui des crêtes négatives. La polarité qui compte le moins de crêtes est la plus favorable pour la mesure du fondamental. Ce critère est très pertinent et facilement mesurable.

L'algorithme de correction et de mesure a pour objet d'éliminer les impulsions non significatives et de mesurer la période du fondamental. Si les valeurs mesurées ne représentent pas des valeurs possibles de T_o elles sont rejetées. Lorsque deux valeurs consécutives sont cohérentes (à 20% près) on n'admet plus de glissement de valeurs de la période supérieure à 20% par rapport à la mesure précédente.

III-2 - Version par blocs

Dans un souci d'accélérer la réponse du détecteur du fondamental et d'améliorer la fiabilité de la mesure, une version de traitement partiel par bloc a été envisagée. Cette version maintient tous les traitements séquentiels de la version modifiée jusqu'aux sorties des deux détecteurs de crêtes qui sont mémorisées dans deux registres à décalage. Ces signaux sont traités en bloc à la fin de l'intervalle d'analyse au moment où le détecteur de voisement fournit une décision définitive de voisement ou non. Dans les versions séquentielles, lors d'une transition non-voisée vers voisée le détecteur de voisement peut ne classer le segment comme voisé que vers la fin de l'intervalle et ce n'est qu'à partir de cet instant que commence la mesure du fondamental, ce qui a pour effet de retarder la réponse. Dans cette version par bloc, la mesure commence à partir de la première crête de l'intervalle d'analyse classée comme voisée, donc la mesure est initialisée plus tôt. Un autre avantage de ce type de traitement réside dans la possibilité de décider de la chaîne la plus favorable à partir de la suite des crêtes sur laquelle portera le mesure et non en fonction de l'histoire passée.

III-3 - Résultats

Les tests entrepris avec la version "originale" ont permis de vérifier la grande fiabilité du système. Cependant, lors de la transition non-voisée vers voisée le détecteur délivre parfois des valeurs erronées au début. Ces valeurs correspondent, en général, au 2ème ou 3ème harmonique du fondamental, ce qui s'explique par la caractéristique du filtre précédant le détecteur de crêtes. Après la transition on retrouve des valeurs correctes de T_o . Néanmoins, si les valeurs erronées du début sont "cohérentes", l'algorithme de "correction" qui suit tend à conserver cette erreur sur toute la longueur du segment voisé. Ce type d'erreur a été constaté dix fois sur le corpus traité.

Les erreurs soutenues ont motivé l'étude de la version modifiée. Dans ce cas encore la transition est parfois marquée par une ou deux valeurs erronées au début. Ces valeurs correspondent, en général, à des sous-multiples de la fréquence fondamentale. Néanmoins, pour le même corpus on n'a constaté qu'une erreur soutenue.

Enfin, la version par blocs a été élaborée pour accélérer la réponse du système. Ce résultat a été obtenu et la qualité des mesures relevées est celle de la version séquentielle modifiée.

IV - CONCLUSION

Nous avons décrit un système global de détection de voisement et de mesure du fondamental. Ce système a fait preuve de sa grande fiabilité. La simulation des algorithmes décrits ci-dessus existe en deux versions, l'une en virgule flottante et l'autre en virgule fixe. Dans les traitements avec précision réduite, nous avons retenu une définition de 16 bits pour les grandeurs du signal et de 12 bits pour les coefficients des filtres. Les problèmes d'ajustement de l'échelle du signal à l'entrée de chaque cellule des différents filtres ont été étudiés afin de satisfaire à la double contrainte : éviter les débordements de calculs intermédaires et maximiser le rapport S/B. Les résultats obtenus demeurent sensiblement les mêmes qu'avec la version en virgule flottante.

REFERENCE :

[1] JF. ZURCHER, M. CARTIER et LJ. BOE : Détection et mesure du fondamental 6ème Journées d'Etude sur la Parole, GALF, pp 13-21, Toulouse (Mai 1975)

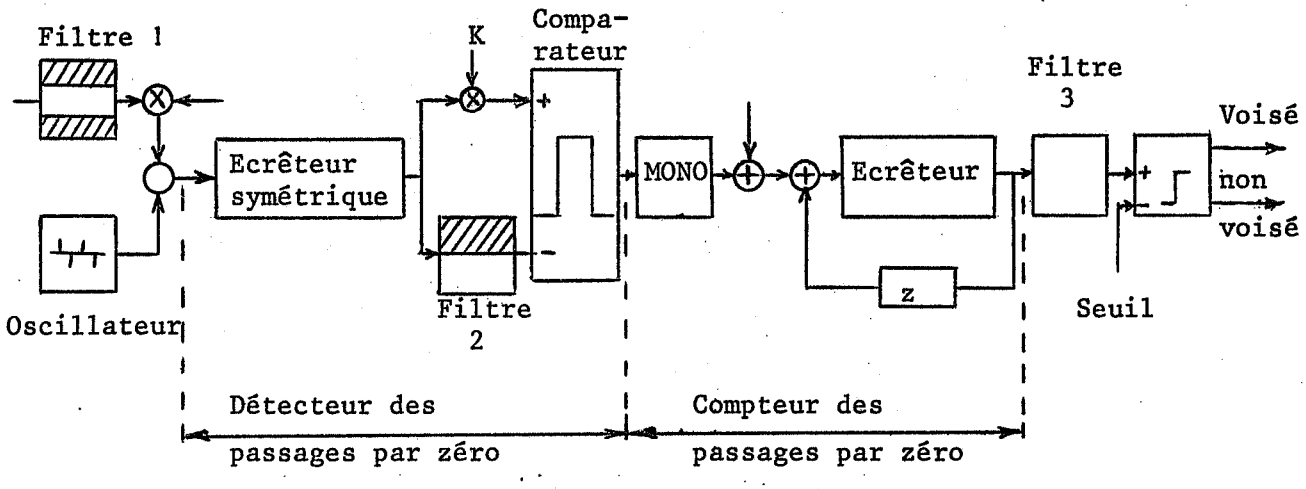


Figure 1 - Schéma fonctionnel du détecteur de voisement

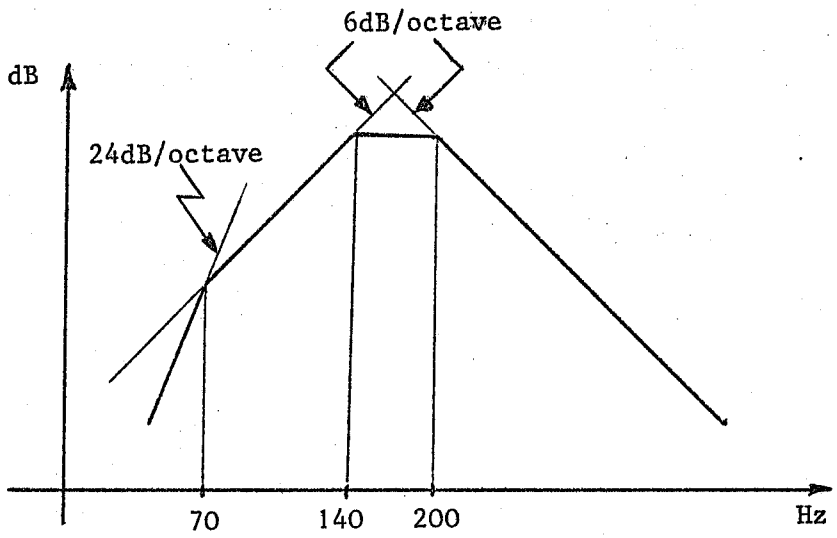


Figure 2 - Filtre de prétraitement

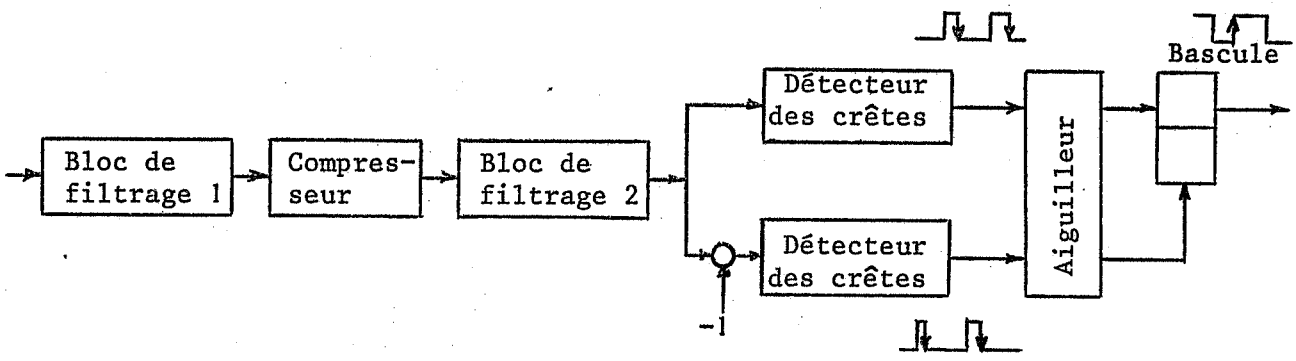


Figure 3 - Schéma fonctionnel de la chaîne de mesure du fondamental

VIIIèmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

UN ANALYSEUR-VISUALISEUR DE MELODIE A MICROPROCESSEUR

Philippe Martin
Experimental Phonetics Laboratory
University of Toronto.

• Résumé :

On décrit un analyseur de mélodie opérant le filtrage inverse du signal de parole par un banc de filtres passe-bas de configuration fixe. Ceux-ci sont sélectionnés en cours d'analyse par un microprocesseur qui assure également la mesure des périodes laryngiennes successives, la compensation des délais de propagation de groupe, la conversion en valeur fréquentielle, ainsi que diverses corrections rendues éventuellement nécessaires par la présence de bruit dans le signal.

L'analyseur décrit, qui fonctionne en temps réel dans une gamme de 70 à 1000 Hz, est couplé à un convertisseur video qui permet la présentation de deux courbes mélodiques de durée variable sur un moniteur TV ordinaire.

• Summary :

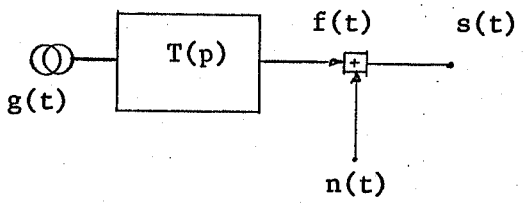
A pitch analyser is described that uses inverse filtering of the speech signal by means of a bank of low-pass filters in a fixed configuration. These filters are selected during the course of the analysis by a microprocessor which also ensures the measurement of the successive laryngeal periods, the compensation of the group delay, the conversion into frequential value, as well as the various corrections that are eventually made necessary by the presence of noise in the signal.

The analyser described, which functions in real time in a range of 70 to 1000 Hz, is connected to a video-converter which permits the presentation of two pitch curves of variable duration on an ordinary TV monitor.

Un analyseur-visualiseur de mélodie à microprocesseur

Philippe Martin

0. Comme tout appareil de mesure, un analyseur de mélodie est lié à un modèle implicite rendant compte du phénomène physique à mesurer. Dans le cas de la mélodie, le modèle est en général très simple, et se compose d'une source périodique g(t) attaquant un filtre à fonction de transfert variable T(p). Le signal obtenu f(t) s'ajoute à la fonction de bruit n(t) pour donner le signal de parole à analyser s(t).



Modèle source-filtre

Le bruit n(t) représente tout ce dont le modèle ne rend pas compte dans le signal de parole, les sources extérieures au locuteur mais aussi l'effet produit par les occlusives, les consonnes sourdes, etc.

Rapportée à ce modèle, l'analyse de la mélodie consiste alors en l'estimation des périodes successives de la fonction de source g(t).

1. Les analyseurs de mélodie peuvent se classer en deux catégories selon que, lors de chaque opération de mesure, la fonction g(t) est considérée comme stationnaire ou variable. Les dispositifs du premier groupe procèdent par prélèvement temporel du signal suivi d'une analyse fréquentielle (spectrale), ceux du second réalisent un prélèvement fréquentiel (filtrage) suivi d'une analyse temporelle. Les mesures obtenues correspondent alors respectivement à des estimations de la fréquence fondamentale (F_o) et de la fréquence laryngienne (F_L).

Ainsi, la méthode du cepstre opère par prélèvements temporels successifs, suivis d'une double analyse spectrale. Le modèle de la phonation sous-jacent présente, pour chaque prélèvement, une fonction de source stationnaire.

Le prélèvement temporel peut être précédé par diverses transformations non linéaires: élévation au carré ou au cube, simple ou double rectification, compression de dynamique (cf. Zurcher, Cartier et Boë, 1975).

Un autre méthode récente est basée sur le calcul des coefficients de prédiction linéaire pour chaque prélèvement temporel de s(t). La source du modèle implicite présente les caractéristiques d'un train d'impulsion de période variable. Malgré son élégance théorique, cette technique, très sensible au niveau de bruit présent dans le signal, ne peut fonctionner de manière simple que dans une gamme de fréquence restreinte (Markel, 1976). De plus, sa mise en oeuvre en temps réel, indispensable dans la plupart des applications de l'analyse mélodique, nécessite un matériel qui reste encore très complexe et couteux. Certaines variantes (cf. Le Roux, 1975) permettent cependant de lever quelques unes de ces restrictions.

2. L'appareil présenté ici appartient à la même catégorie d'analyseurs, et repose sur le même principe que le dispositif déjà décrit dans Léon et Martin,

(1969). Il opère par approximation variable du filtrage inverse du signal, de manière à obtenir une estimation de la source, supposée cette fois sinusoïdale. Cette technique, tout en permettant la mesure des variations fines et rapides de la fréquence laryngienne, fait appel à des filtres analogiques dont la configuration change avec la fréquence analysée, sous le contrôle d'un microprocesseur. Celui-ci assure également la mesure des périodes successives du signal filtré ainsi que diverses corrections destinées à compenser les perturbations apportées par le bruit.

Pour que le signal filtré ait une allure sensiblement sinusoïdale et ne présente que deux passages par zéro par période laryngienne, les filtres utilisés doivent, dans leur gamme de fonctionnement, apporter une atténuation telle que la condition de McKinney (1965) soit remplie. Cette condition, supposée valable pour un signal non stationnaire, énonce que l'amplitude de la fondamentale doit être supérieure à la somme des produits des amplitudes des harmoniques du signal filtré par leur rang d'harmonique, c'est-à-dire:

$$a_1 > \sum_{n=2}^{M} n \cdot a_n$$

Cette relation, qui tient compte de l'interaction du premier formant avec la fréquence laryngienne lors de la production de certains phonèmes, est satisfaite par l'emploi de filtres passe-bas à pente d'atténuation suffisamment raide (de l'ordre de 70 dB / octave). La gamme de fréquence à analyser est alors couverte par plusieurs de ces filtres aux fréquences de coupure croissantes et disposés en parallèle. La mesure du signal de sortie de fréquence la plus basse correspond alors à celle de la fréquence laryngienne.

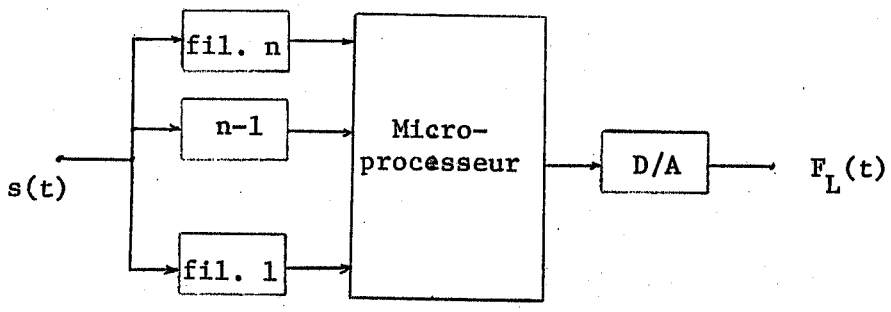


Schéma de principe de l'analyseur

3. Toute analyse fréquentielle est soumise à une relation d'incertitude liant les indéterminations dans les mesures du temps et de la fréquence: $\Delta t . \Delta f > k$ filtres, k est déterminé par la réponse impulsionnelle. Si, comme c'est le cas dans l'analyse mélodique où le signal filtré est supposé sinusoïdal, certaines caractéristiques du signal mesuré sont connues à l'avance, k peut être fortement réduit par discrétisation des mesures fréquentielles. Cette opération est réalisée par la mesure des intervalles de temps séparant deux passages successifs par zéro et par valeur croissante du signal filtré, intervalles dont l'inverse correspond à la fréquence laryngienne.
La valeur de la constante k dépend du système d'analyse considéré; pour des

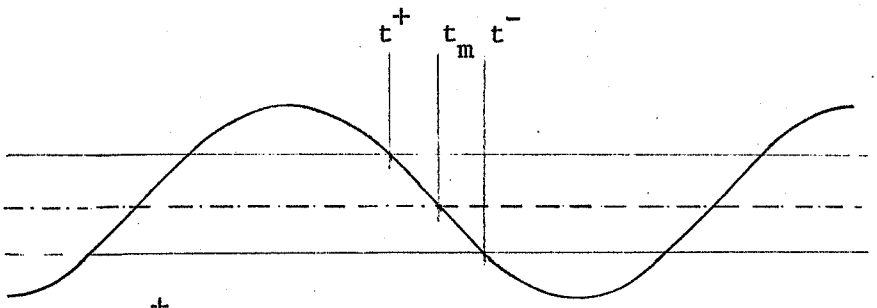
L'indétermination sur la mesure de la fréquence est alors due essentiellement à la présence de bruit à la sortie des filtres, qui 1) déplace l'instant des passages par zéro de la fonction analysée $s(t) = f(t) + n(t)$, positivement ou négativement selon la relation de phase existant entre les deux composantes de $s(t)$; et 2) rend nécessaire l'utilisation d'un seuil de détection non nul pour la mesure des périodes successives pour éviter le déclenchement des circuits lorsque $f(t) = 0$. Les valeurs mesurées des périodes deviennent alors sensibles aux variations d'amplitudes du signal filtré, d'une manière d'autant plus importante que le rapport amplitude du signal/seuil de détection est faible (ce qui est toujours le cas en début et en fin de signal). De nombreux effets de micromélodie sont dus à cet artefact qui provoque une diminution de fréquence mesurée si l'amplitude du signal est croissante, et une augmentation si elle est décroissante.

L'indétermination temporelle, i.e. l'erreur commise sur l'instant où la période laryngienne a la valeur mesurée, résulte principalement du délai de propagation de groupe qui dépend de la fonction de transfert et du rapport fréquence du signal / fréquence de coupure du filtre considéré. La valeur du déphasage entre les sorties de deux filtres de même structure mais de fréquences de coupure différentes ne sera donc pas le même, particulièrement au voisinage de la fréquence de coupure de l'un d'eux.

Aux effets du filtrage s'ajoutent alors les délais variables dus à l'exécution de l'algorithme de mesure des périodes successives et de choix du filtre approprié à un instant donné, ainsi que le temps nécessaire à la réalisation des diverses opérations destinées à corriger les autres conséquences de la non validité du modèle (traitement des consonnes occlusives, du bruit de fond, etc.).

4. Pour diminuer encore la constante d'indétermination k, on à utilisé les dispositifs supplémentaires suivants:

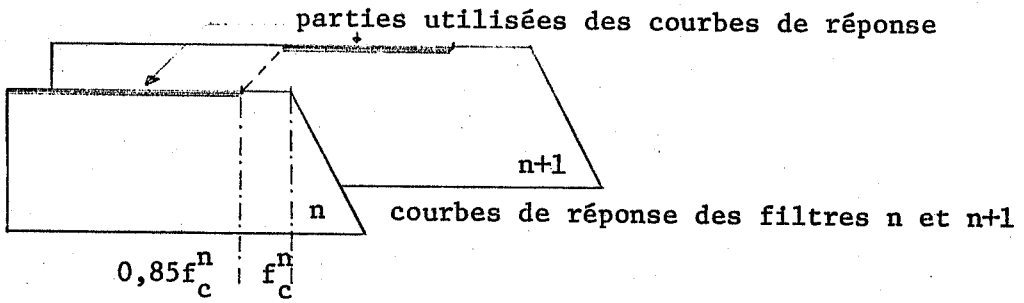
a) Les perturbations entraînées par l'emploi d'un seuil de détection non nul ont été contrecarrées par l'utilisation d'une détection symétrique, prenant nant comme repère le milieu de l'intervalle de temps entre le passage du signal filtre par un seuil négatif et positif, les seuils étant égaux en valeur absolue. Ce dispositif est préférable à l'emploi d'un compresseur de dynamique qui introduit dans le signal filtré un niveau variable de bruit dont les effets sont très marqués pour les signaux de faibles amplitude;



t^+ : déclenchement au seuil positif
t^- : déclenchement au seuil négatif
t_m : temps moyen

b) Un filtre passe-haut précédant le banc de filtres passe-bas et de fréquence de coupure égale à 70 Hz, élimine les composantes de n(t) de fréquence inférieure à cette limite;

c) Grâce à l'algorithme de sélection, les filtres passe-bas peuvent être utilisés dans la partie où la courbe de délai de groupe présente le moins de variations (pour $f < 0{,}85\ f_c$). Dès que la fréquence mesurée dépasse cette valeur, le signal utile est pris à la sortie du filtre de fréquence de coupure immédiatement supérieure dans le banc de filtres. De plus, l'emploi de filtres de Legendre du 8ème degré permet d'obtenir des courbes de délai de groupe présentant relativement peu de variations;



d) La différence entre les délais présentés lors du passage d'un filtre à l'autre est compensée par une ligne à retard gardant en mémoire les périodes successives des 2 signaux, le choix de la période finale se faisant à la sortie de ces lignes;

e) Un algorithme prédictif compense les effets des écarts par rapport au modèle, en comparant la valeur mesurée aux contraintes établies sur le signal (période valant entre 14.2 et 1 ms, écart d'une valeur par rapport à la valeur précédente inférieure à 20%, durée du signal voisé supérieur à 5 périodes fondamentales, etc.) Si la fréquence mesurée ne répond pas à un de ces critères, elle est, soit égalée à zéro, soit remplacée par les valeurs double ou moitié, ce qui entraîne une nouvelle vérification;

f) Enfin, un programme de lissage évalue la moyenne de 2 valeurs consécutives de la fréquence, si leur écart est inférieur à un certain seuil.

5. L'appareil réalisé présente 3 sections bien distinctes:

a) Un banc de 7 filtres passe-bas de type Legendre du 8ème degré, (auxquels s'ajoute 1 filtre passe-haut supplémentaire destiné à réduire la bande passante du bruit), et dont les fréquences de coupure s'échelonnent de 114 à 1000 Hz;

b) Une unité de calcul centrée sur un microprocesseur 8080A, à temps de cycle de 500 ns, travaillant avec une mémoire morte de 1024 mots contenant le programme d'instructions, et une mémoire active de 512 mots;

c) Un convertisseur video assurant le codage de 2 courbes mélodiques indépendantes obtenues par échantillonnage de 256 valeurs de 8 bits, sous la forme d'un signal TV composite (standard de 625 lignes) utilisable sur n'importe quel moniteur de télévision.

Toutes les opérations de mesure et d'amélioration du coefficient d'incertitude sont réalisées sous le contrôle du microprocesseur.

Le passage, par valeur croissante, d'un seuil positif ou négatif du signal de sortie de chacun de 7 filtres passe-bas, provoque l'inscription directe, dans la mémoire active, de la valeur d'un compteur de 16 bits dont la base de temps est de 1 µs. Le seuil de déclenchement de ce mécanisme est déterminé par la position du potentiomètre de volume placé à l'entrée du système, qui constitue le seul élément de réglage de l'appareil.

Le microprocesseur assure ensuite le calcul des périodes successives des signaux de chaque canal en évaluant la différence existant entre deux valeurs consécutives des moyennes des instants de déclenchements négatifs et positifs inscrits dans la mémoire (ou d'une approximation de cette moyenne si un des déclenchements n'a pas eu lieu).

Après passage dans une ligne à retard destinée à compenser les différences de temps de propagation de groupe existant entre les différents filtres, le microprocesseur effectue un choix de la période convenable en considérant les valeurs affichées à la sortie de la ligne à retard selon des valeurs décroissantes (correspondant à des fréquences de coupure des filtres qui vont en croissant). Une valeur ne sera retenue que si elle correspond à la portion effectivement utilisable du filtre considéré (la limite supérieure équivalent à $0.85\ f_c$).

La période obtenue est ensuite convertie en valeur fréquentielle à 10 bits et placée dans une pile contenant les 2 dernières fréquences mesurées. Deux tests de régularité comparent alors la dernière valeur aux précédentes, de manière à s'assurer que les contraintes imposées au modèle sont effectivement satisfaites. S'il en est ainsi, la valeur obtenue après lissage portant sur 2 fréquences successives est affichée comme valeur finale, et transmise aux convertisseurs analogique et video. Le délai dû à ces derniers tests est ainsi minimisé. Si le test ne réussit pas, un essai de substitution avec une valeur double ou moitié de la fréquence mesurée est tenté, mais cet artifice introduit en général dans la courbe une distorsion peu satisfaisante pour les opérations de recherche phonétique, pour lesquelles l'affichage d'une valeur nulle est préférable.

Outre les circuits de génération des impulsions de synchronisation ligne et trame, le convertisseur video contient essentiellement 2 registres à décalage de 8 bits, à 256 positions, qui renferment chacun 256 valeurs de la fréquence mesurée. Ces valeurs sont prélevées selon une fréquence réglable extérieurement de manière à obtenir des courbes mélodiques de durée variant entre 0.1 et 10 secondes. Ces registres sont commandés par une horloge opérant à 5 MHz, qui pilote également un compteur à 8 bits.

A un instant donné, la coïncidence des valeurs présentées à la sortie d'un des registres de décalage et du compteur de lignes des circuits de synchronisation provoque l'apparition d'un signal video. L'examen des 256 valeurs échantillonnées est donc effectué pour chacune des 512 lignes d'une image complète, les lignes restantes étant générées séparément.

Des circuits accessoires permettent également l'effacement automatique des

courbes stockées précédemment dès l'apparition d'un nouveau signal (i.e. après un silence de 1 s) ou après le prélèvement de 256 échantillons d'une analyse en cours.

6. L'appareil décrit travaille en temps réel, et fournit des valeurs successives des fréquences fondamentales automatiquement, dans une gamme de 70 à 1000 Hz, selon les conditions définies par le modèle implicite qui préside à sa construction. Le système de filtrage inverse à configuration variable, destiné à tenir compte des variations rapides du tractus vocal, constitue une approximation du procédé à coefficients prédictifs, où le filtre utilisé est continuellement modifié.

Par l'emploi de diverses techniques de compensation, les incertitudes dans les mesures fréquentielle et temporelle, inhérentes à l'emploi de filtres et à la présence de bruit, sont très fortement minimisées.

Enfin, la visualisation instantanée sur moniteur TV permet une exploitation extrêmement souple et rapide de l'appareil pour la recherche phonétique et l'enseignement des facteurs prosodiques.

REFERENCES

Léon, P.R. et Martin, Ph. (1969) Prolégomènes à l'étude des structures intonatives, Didier, Montréal, Paris, Bruxelles.

Le Roux, J. (1975) Une méthode synchrone d'extraction en temps réel du fondamental, Actes des 6èmes Journées d'Etude sur la Parole, GALF, Toulouse, pp. 1-11.

Markel, J.D. and Gray, A.H., Jr. (1976) Linear Prediction of Speech, Springer - Verlag, Berlin, Heidelberg, New-York.

McKinney, N.P. (1965) Laryngeal Frequency Analysis for Linguistic Research, Communication Sciences Laboratory, Report No 14, Ann Arbor, Mich.

Martin, Ph. (1973) Réalisation d'un analyseur de mélodie fonctionnant en temps réel en vue de la recherche phonétique et l'enseignement, thèse de doctorat en sciences appliquées, Université Libre de Bruxelles.

Zurcher, J.-F., Cartier M. et Boë, L. J. (1975) Détection et mesure du fondamental, Actes des 6èmes Journées d'Etude sur la Parole, GALF, Toulouse, pp. 13-21.

VIII èmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

ETUDE ET REALISATION D'UN VOCODER A CODAGE DIRECT DE LA MELODIE

S. EL SAMKARI - F. HAMELIN

(Laboratoire d'Automatique de Besançon)

• Résumé :

On décrit un vocoder à canaux dans lequel l'information de hauteur fait l'objet d'un traitement séparé, destiné à fournir un codage très fin de la courbe mélodique, et une restitution fidèle de celle-ci.

On utilise à cette fin une chaîne glottométrique et un synthétiseur de mélodie réalisés antérieurement au L.A.B.

Les applications de cette réalisation concerneront le rôle des traits prosodique dans la transmission et la synthèse de la parole en général, ainsi que le "vo-codage" de langues à tons.

• Summary :

We present a channel vocoder which performs a very sharp encording and accurate reproduction of the pitch by separate processing.

For that purpose we use a glottometrical system and a melody synthesizer premously developped at the L.A.B.

This vocoder will be used both for studying the influence of prosodic features on the transmission and synthesis of speech and for "vo-coding" of tone languages.

ETUDE ET REALISATION
D'UN VOCODER A CODAGE DIRECT DE LA MELODIE

1 - INTRODUCTION

Nous présentons ici le principe d'un Vocoder en cours d'achèvement au Laboratoire d'Automatique de Besançon. Pourquoi réaliser un tel dispositif alors qu'il en existe maintenant de nombreux exemplaires [1] , parfaitement opérationnels ?

L'analyse du signal de parole et son codage numérique ont généralement comme but l'économie de stockage ou de télétransmission ; le décodage et la synthèse doivent ultérieurement permettre de restituer un signal intelligible pour l'auditeur. On peut mettre en oeuvre à cette fin des principes très variés, qui constituent encore une voie de recherche très ouverte.

Notre but est différent, et notre ambition plus modeste : nous souhaitons seulement développer un outil de Laboratoire spécialement adapté à l'étude fine du rôle des phénomènes glottiques dans la parole, permettant à la fois une analyse directe, et une analyse par synthèse.

La caractéristique essentielle de l'appareillage est le captage direct du signal glottique, et son traitement par un analyseur logarithmique instantané de grande définition (huitième de ton). Parallèlement le spectre de fréquence du signal phonique capté par un microphone est analysé d'une manière relativement classique par un banc de filtres passe-bande.

La synthèse s'effectue en utilisant un banc de filtres convenablement excité par un générateur de bruit et/ou par un synthétiseur de mélodie au 8° de ton, programmable numériquement période par période [2] .

L'acquisition des résultats de l'analyse et l'émission des données de synthèse est effectuée par miniordinateur, selon un mode asynchrone pour la mélodie, synchrone pour le spectre.

Ce système devrait s'avèrer un instrument de choix pour des études précises sur l'effet des variations de hauteur du fondamental laryngien : intonation, micromélodie, langues à tons, etc... .

Nous décrivons sommairement ci-après la structure de l'analyseur et sa procédure d'acquisition ; puis la structure du synthétiseur et sa procédure de commande.

2 - REALISATION DE L'ANALYSEUR (fig.I)

2.1. La chaîne d'analyse spectrale

Elle comporte la succession des modules suivants :

- amplification et préaccentuation ;
- filtrage : 2 filtres passe-bande d'une octave suivis de 14 filtres passe-bande de $\frac{1}{3}$ d'octave, couvrent la plage de 62 Hz à 6.500 Hz ; ces filtres sont d'ordre 4 ; les deux premiers sont destinés à déterminer l'énergie du fondamental ;
- détection d'amplitude : 16 extracteurs de valeur de crête donnent la plus grande valeur atteinte sur le canal considéré pendant la période d'échantillonnage fixée à 20 ms ;

- multiplexage et conversion analogique-numérique : une horloge à 800 Hz commande la scrutation successive des 16 voies d'analyse, leur codage numérique et la remise à zéro des détecteurs de crête.

2.2. La chaîne glottométrique

Nous ne reviendrons pas sur le détail de cette chaîne, présentée aux Journées d'études "Communication parlée" de Nancy (1976), et qui comporte :

- un électroglottographe de Fabre, fournissant un signal basse-fréquence représentatif des variations de l'impédance haute-fréquence de la glotte, mesurée au niveau du cartilage thyroïde ;
- un fréquencemètre logarithmique instantané donnant directement, sous forme digitale, la hauteur de la vibration laryngée repérée en octaves, tons et huitièmes de ton par rapport à une fréquence de référence (ajustable) ; la mesure est effectuée période par période, et il est peut être plus significatif de dire que l'appareil fournit le logarithme du rapport de chaque cycle glottique à une période de référence. Ceci permet une grande précision temporelle dans la détection du voisement, ainsi qu'une grande sensibilité dans l'appréciation des fluctuations de la hauteur relative du fondamental laryngé.

2.3. Acquisition des résultats de l'analyse

En fin de la chaîne d'analyse de la parole nous avons donc deux séries d'informations numériques. Ces informations peuvent être soit transmises directement sur une chaîne de synthèse de la parole, soit stockées dans une mémoire pour une restitution différée.

Nous nous sommes penchés sur la seconde solution qui permet aussi de retraiter les résultats et d'en observer les effets sur la synthèse finale.

A l'aide d'un calculateur numérique nous avons donc réalisé :

1°) l'acquisition synchrone dans une pile PP de mémoires, des informations numériques provenant des seize canaux multiplexés. A la fréquence d'échantillonnage de 800 Hz cela permet un enregistrement d'un peu plus de 5 s pour une pile de 4 K mots.

2°) l'acquisition asynchrone, en mode prioritaire, dans une autre pile PM de mémoires, des informations numériques provenant de l'analyseur de mélodie (octaves-tons-huitièmes de tons).

3°) en vue d'une resynchronisation de ces ensembles d'informations lors de la synthèse, l'absence de mélodie-due aux sons fricatifs ou aux pauses- doit être repérée. Nous considérons qu'il y a absence de voisement lorsque la plus grande période To, que peut délivrer l'Analyseur de mélodie est dépassée. (C'est la fréquence de multiplexage qui nous sert de base de temps). Dans ce cas on enregistrera dans la pile PM un repère d'arrêt.
De même on enregistrera dans la pile PP un repère de marche lorsque la mélodie apparaît.

3 - LA CHAINE DE SYNTHESE (fig.II)

Elle est constituée de deux voies :

- synthétiseur de mélodie,
- synthétiseur de spectre.

3.1. Synthétiseur de mélodie

Un synthétisuer programmable peut être réalisé à partir d'un analyseur de mélodie : en comparant le signal de sortie à la valeur de la fréquence affichée ; (les différentes valeurs affichées pouvant être fournies par lecture d'une mémoire de stockage).

Le signal de coïncidence qui, après mise en forme, constitue la sortie, remplace le signal d'entrée pour déclencher l'analyseur et commander la lecture d'une nouvelle valeur.

3.2. Synthétiseur de spectre

Cet élément de la chaîne de synthèse réalise la sommation des sorties d'un banc de filtres identique à celui employé dans la chaîne d'analyse. Les filtres sont excités, à des niveaux fixés par l'ordinateur à travers une batterie de multiplieurs :

- soit par un générateur de bruit blanc en absence de voisement,
- soit par le synthétiseur de mélodie, ceci dans le but d'améliorer le naturel de la voix.

En effet les variations de fréquence du fondamental provoquent un "glissement" des fréquences des harmoniques du signal et réalisent ainsi la fonction d'intonation.

3.3. Emission différée des données

L'ordinateur est ici placé en organe de commande pour :

a) - contrôler la marche du démultiplexeur,
b) - présenter à ce dernier les informations numériques d'amplitude des composantes du spectre de la parole à travers un convertisseur digital-analogique,
c) - contrôler la marche du synthétiseur de mélodie et lui présenter les valeurs demandées,
d) - bloquer ce synthétiseur, lorsqu'une valeur de la pile PM contient le repère d'arrêt, et mettre en fonction le générateur de bruit blanc,
e) - débloquer le synthétiseur, lorsqu'une valeur de la pile PP contient le repère de marche, et couper le générateur de bruit blanc.

4 - CONCLUSION

Nous souhaitons que l'exploitation de l'appareillage ainsi réalisé, permette une étude approfondie de l'influence du signal de mélodie sur l'intelligibilité de la parole synthétique émise par un vocoder à canaux.

Fig. I - Schéma de réalisation pour l'analyse

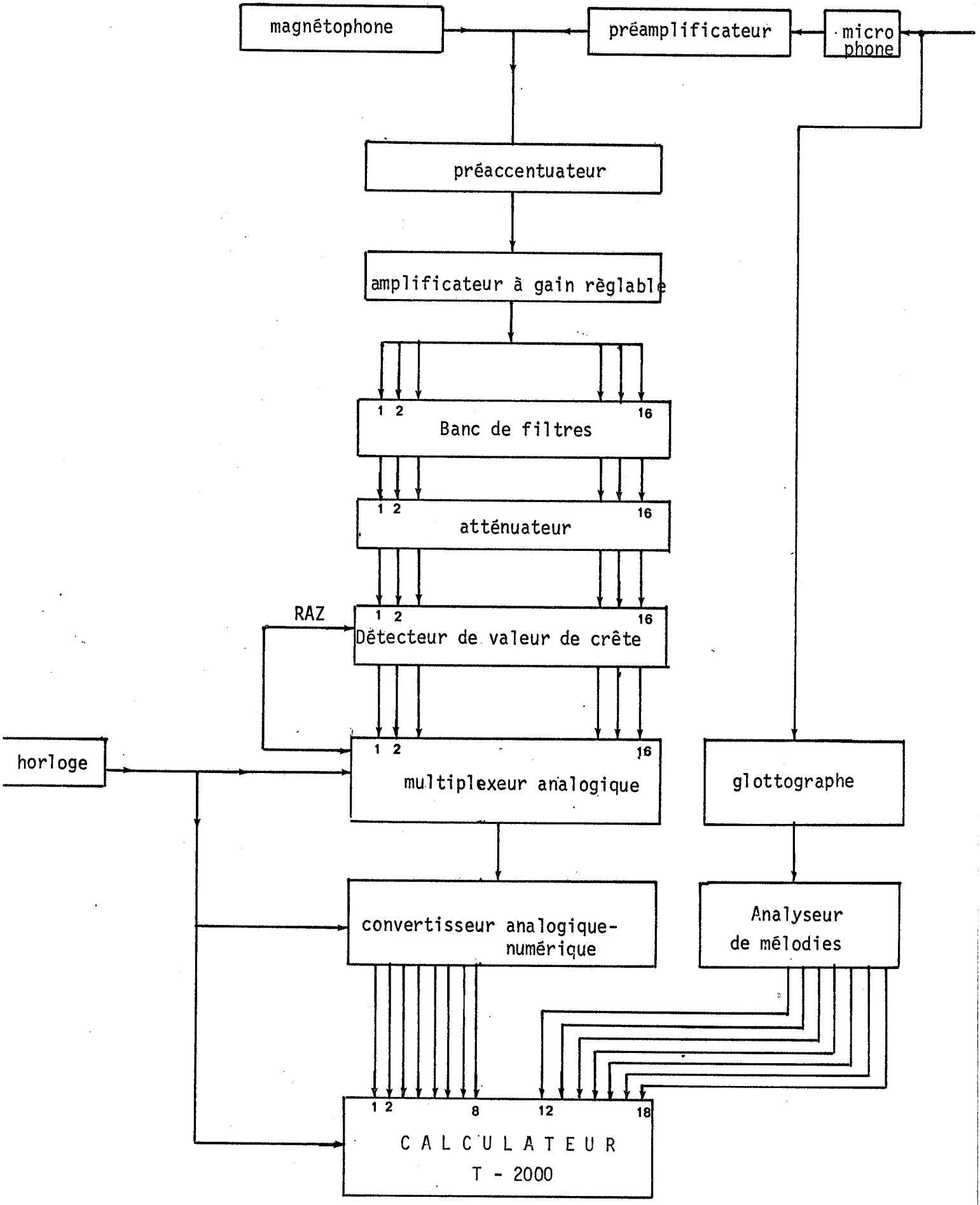
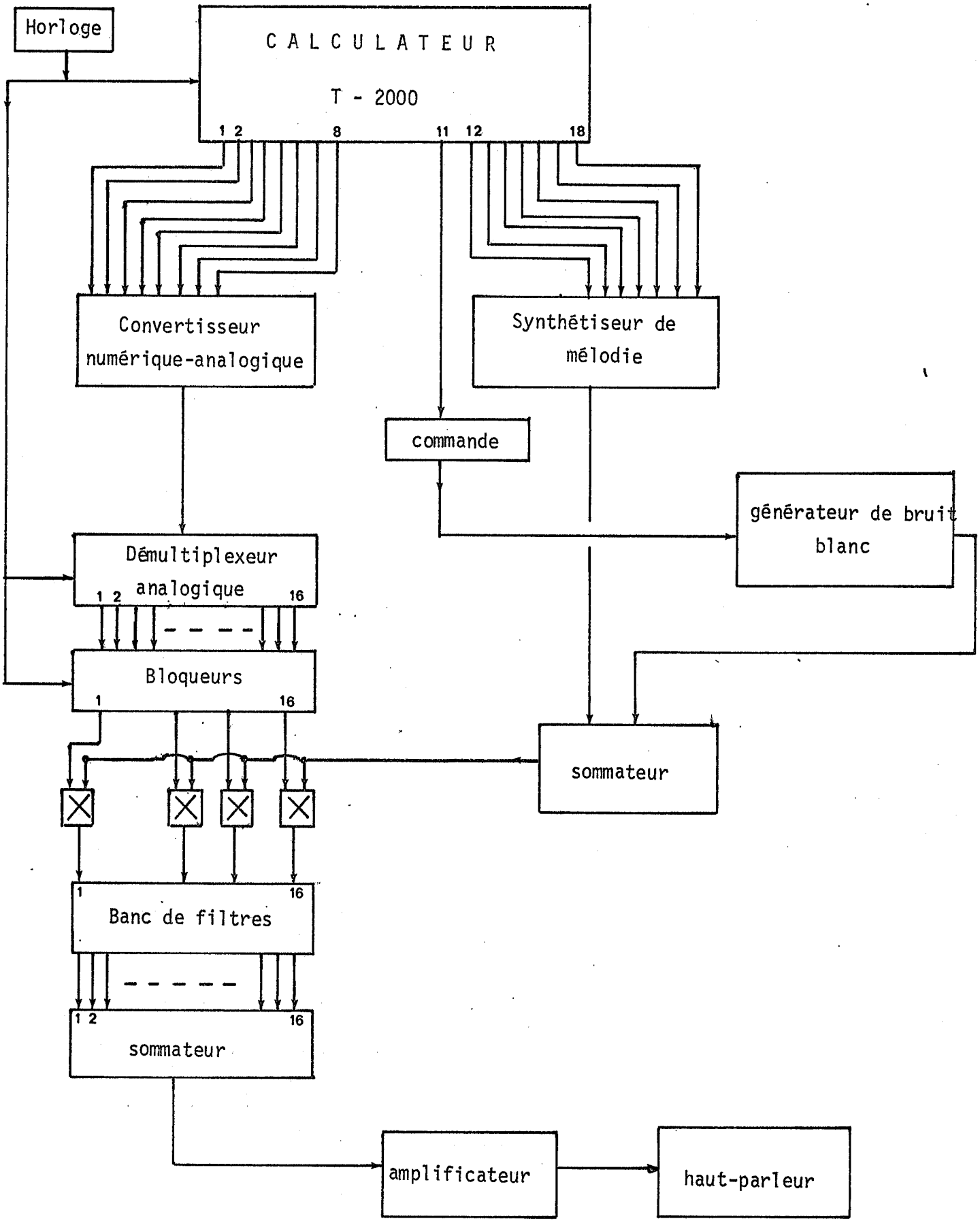


Fig. II - Schéma de réalisation pour la synthèse.



5 - BIBLIOGRAPHIE

1 - 1) H.W. DUDLEY - Fundamentels of speech synthesis
Bell Telephone Laboratories 1955

2) R. CARRE - Etude et réalisation d'un détecteur des paramètres de la parole
Thèse : Grenoble - 1966

3) J. MAX - Traitement du signal
Masson - 1972

4) J.L. FLANAGAN - Speech analysis synthesis and perception
Springer-Verlag - 1972

5) NAMBURI, JAIN - RAMAKRISHAN - 16- Channel CEERI vocoder
Bell-telephone - April 1972

6) E. VILLCHUR - Signal processing to improve speech intelligibility imperceptive deafness
Foundation for hearing aid research, woodstock, New-York - 15 November 1972

7) HASKEW - KELLY - MC KINNEY - Resultats of study of linear prediction vocoder
IEEE Conference april 1972

8) J. POTAGE - Le vocoder ciphon
7ème J.E.P. Nancy 1976

2 - 1) JOHN - N. HOLMES - The influence of glottal wave-form on the naturalness of speech from a paralle formant synthesizer
IEEE Transaction on audio and electroacoustics,
Vo, Au 21, N° 3, June 1973

2) P. ANDRE - M.FILLEAU - F. HAMELIN - Glottométrie en temps réel - application
7ème J.E.P. Nancy - 19-21 Mai 1976

3) P. ANDRE - F. LHOTE - Analyseur-synthétiseur de fréquence musicale
14th Conference on acoustics - Tatranska-Lomnica (T.S.S.R.) Octobre 1976

4) EL-MALLAWANY - Approches à la détection et à l'analyse sur l'intervalle de fermeture de la glotte
7ème J.E.P. Nancy 1976

VIII èmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

UN SYSTEME DE DETECTION AUTOMATIQUE DU FONDAMENTAL ET DE L'INTENSITE.

B. TESTON et M. ROSSI
Institut de Phonétique d'AIX-EN-PROVENCE

• Résumé :

Le système de détection automatique des paramètres prosodiques que nous décrivons comprend un détecteur de mélodie, et un détecteur d'intensité couplés sur un calculateur T 1600. Le détecteur de mélodie est dérivé des appareils élaborés pour les études phonétiques.

Le détecteur d'intensité est un détecteur efficace logarithmique. Le calculateur traite en temps réel les données fournies par les détecteurs.

Les résultats sont visualisés sur une imprimante rapide qui fournit les courbes mélodiques et d'intensité; et pour chaque période le temps cumulé en ms et les valeurs de Fo en Hz, et de I en dB.

• Summary :

The automatic prosodic-parameter detection-system which we describe makes use of a pitch-meter and an intensity-detector connected to a T 1600 computer. The pitch-meter derived from instruments elaborated in phonetic research.

The intensity-detector is a RMS circuit. The computer treats the date supplied by the detectors in real time.

The results are displayed on a high-speed printer which supplies the pitch and intensity curves and for each period the total time in ms. and the values of Fo in Hz and of I in dB.

UN SYSTEME DE DETECTION AUTOMATIQUE DU FONDAMENTAL ET DE L'INTENSITE.

B. TESTON et M. ROSSI

Le système de détection des paramètres prosodiques que nous décrivons est constitué d'un mélodimètre et d'un intensimètre reliés d'une part à un enregistreur galvanométrique (Oscillomink), d'autre part, à un calculateur T 1600.

I. LE MELODIMETRE

Il est issu d'une famille d'appareils spécialement réalisés pour l'étude phonétique. Le principe de fonctionnement du mélodimètre est bien connu depuis les premiers détecteurs, en particulier, ceux de HUNT (1935) et de GRÜTZMACHER et al. (1937) à l'étranger, et de CARRÉ et al. (1963) en France ; il est très bien décrit dans LEON et MARTIN (1970).
Notre appareil comprend 3 parties :

1) les circuits de mise en forme des signaux dont on veut mesurer la fréquence fondamentale,
2) les circuits d'extraction du fondamental,
3) les circuits de conversion fréquence/amplitude et de visualisation.

Les circuits de mise en forme des signaux sont constitués de cellules indépendantes qui effectuent les distorsions non linéaires suivantes : détections quadratiques, double et simple alternance pour renforcer l'amplitude du fondamental ; écrêtages de crête et de centre pour nettoyer le signal ; amplification logarithmique pour la compression de dynamique. L'efficacité de nombreux circuits a été contestée. Mc KINNEY (1965) après les avoir simulés sur calculateur est peu convaincu de leur utilité. Toutefois, si l'on considère le caractère souvent médiocre des enregistrements sur lesquels travaille le phonéticien, et le timbre de certaines voix qui passent mal au mélodimètre, l'utilisation d'une ou plusieurs fonctions non linéaires permet d'améliorer considérablement le résultat. Nous n'avons pas la place ici d'énumérer les cas particuliers où chacune des fonctions peut se révéler utile.

L'extraction du fondamental est réalisé à l'aide d'un filtre passe-bas dont la fréquence de coupure et la chute par octave peuvent être modifiées respectivement entre 60 et 1000 Hz et 12 et 36 dB.

Ce filtre BUTTERWORTH est caractérisé par un temps de propagation de groupe relativement faible. Le choix de la fréquence de coupure et de la chute/octave est opéré ponctuellement.

La détection du fondamental est réalisée au moyen d'un détecteur de seuil positif, négatif ou nul dont le niveau de déclenchement est réglé généralement aux limites du bruit de fond.

La conversion fréquence/amplitude qui permet d'obtenir la courbe, image de la variation mélodique, est effectuée par comptage, au moyen d'une horloge, des périodes instantanées et par conversion numérique/analogique au moyen d'un convertisseur codé sur 16 bits (pour l'enregistrement galvanométrique).

Ce dernier donne un signal de 10 volts pour une fréquence de 1 KHZ, 5 volts pour 0.5 KHZ et 1 volt pour 0.1 KHZ. Sa précision est de 2.10-3. Cette précision, qui peut paraître excessive, nous est imposée en particulier par les recherches sur la microprosodie et la perception.

Le signal des variations mélodiques est envoyé simultanément :

1) à un calculateur T 1600
et 2) à un enregistreur galvanométrique SIEMENS par l'intermédiaire du convertisseur numérique/analogique.

Avant d'analyser, l'opérateur calibre la sortie analogique à l'aide d'un oscillateur variable, par bonds de 20 Hz, de 60 à 1 000 Hz.

II. L'INTENSIMETRE

Il est essentiellement constitué d'un détecteur de valeur efficace vraie de 60 dB de dynamique. Ce détecteur est caractérisé par un facteur de crête de 8 avec une précision de $\pm$ 1 % qui s'améliore jusqu'à $\pm$ 0.2 % pour un facteur de crête de 2. Ces valeurs sont très satisfaisantes pour la détection du signal de parole. La constante de temps d'intégration peut être choisie entre 1 ms et 1 s.

Parallèlement à la détection efficace, l'opérateur peut choisir une détection linéaire ou quadratique qui donnent respectivement une image de la pression acoustique et de l'énergie.

La dynamique de sortie est logarithmique ou linéaire. Un filtre de pondération A, B ou C, norme ISO, peut être utilisé avant la détection.

L'intensimètre est équipé d'un oscillateur de calibration de 1000 Hz et d'un atténuateur variable entre 0 et 99 dB par bonds de 1 dB.

La dynamique de sortie logarithmique est codée de façon à ce que l'on obtienne 8 volts pour 80 dB, 5 volts pour 50 dB, 1 volt pour 10 dB, etc...

Le signal des variations d'intensité est envoyé simultanément sur l'enregistreur galvanométrique et le calculateur T 1600.

III. TRAITEMENT INFORMATIQUE DES PARAMETRES PROSODIQUES

La connexion est opérée sur un calculateur T 1600 par l'intermédiaire d'un coupleur universel qui reçoit les valeurs des périodes (bits 2 à 16). On fournit également au coupleur l'information sur la présence du voisement (bit 1 à 1).

L'acquisition du signal d'intensité est réalisée au moyen d'un convertisseur analogique-numérique codé sur 10 bits sur une voie de la chaîne d'acquisition.

Le cadencement de l'échantillonnage est effectué par la périodicité du signal pendant le temps de voisement et par une horloge à 100 Hz pendant les silences. La durée du temps de voisement et des silences peut être mesurée par une horloge connectée à l'entrée de comptage du calculateur ; ce calcul de temps n'est toutefois pas possible ou se révèle être compliqué, si l'on utilise la chaîne d'acquisition, puisque dans ce cas on doit masquer les interruptions de l'unité de comptage.

La saisie et le traitement des périodes et de l'intensité se font en temps réel grâce à un programme en Assembleur.

Avant de recueillir les données sur le coupleur universel et la chaîne d'acquisition, on détermine une fenêtre dont les limites sont fournies au système par l'opérateur, sur le télétype. Ces limites dépendent de la tessiture du locuteur et du type de phrase analysé (déclarative ou interrogative). Cette fenêtre permet d'affiner le filtrage électronique.

A chaque période, on recueille une valeur sur les bits 2 à 16 du coupleur universel (le bit 1 est occupé pendant le temps de voisement). Un ordre est immédiatement envoyé au coupleur pour indiquer que la valeur a été prise ; le système peut alors envoyer et stabiliser la valeur de la période suivante.

Dès que la valeur de la période est recueillie, elle est convertie en fréquence et rangée successivement dans une table et dans une mémoire tampon. On calcule la moyenne des quatre premières périodes consécutives du signal dont l'écart est inférieur à 30 %. Cette moyenne sert d'ancrage à une procédure de lissage qui ramène à la valeur qui précède toute durée excédant la précédente de plus de 50 %.

Après avoir masqué les interruptions de l'unité de comptage et initialisé la chaîne d'acquisition (la chaîne doit être réinitialisée au début de chaque phrase et chaque fois qu'une période n'est pas prise en compte), on recueille la valeur échantillonnée. Cette valeur est aussitôt convertie en dB et rangée dans une table. On retourne au coupleur universel pour saisir la valeur de la période suivante, etc...

Si le bit 1 retombe à 0, on recueille sur la chaîne d'acquisition les valeurs de l'intensité échantillonnées par l'horloge à 100 Hz. Le nombre de valeurs recueillies fournit la durée du silence.

Un sous-programme calcule le pas et la dynamique des courbes de mélodie et d'intensité qui seront disponibles, avec le temps cumulé, soit sur l'imprimante rapide, soit sur un écran cathodique.

La représentation sur l'imprimante rapide fournit les informations suivantes :

1° Valeur de la fréquence en Hz, de l'intensité en dB et du temps cumulé en ms pour chaque période du signal,

2° Valeurs de l'intensité en dB sur les consonnes non voisées et les silences,

3° Durée du temps de voisement (temps cumulé à la fin du voisement),

4° Durée des silences.

Ce traitement automatique

1/ fournit à l'utilisateur des résultats directement exploitables où les paramètres prosodiques de chaque période sont donnés sur un document imprimé ;

2/ permet un traitement différent des résultats selon le but recherché :-programme MELSON (avec sous-programme de lissage) pour une visualisation pédagogique des paramètres prosodiques ;

- programme MELDB (absence du sous-programme de lissage) pour une étude fine de la micromélodie et des paramètres prosodiques ;

3/ autorise le stockage des paramètres sur disque en vue de leur traitement statistique, ou de l'utilisation des schémas intonatifs naturels en synthèse de la parole.

B I B L I O G R A P H I E

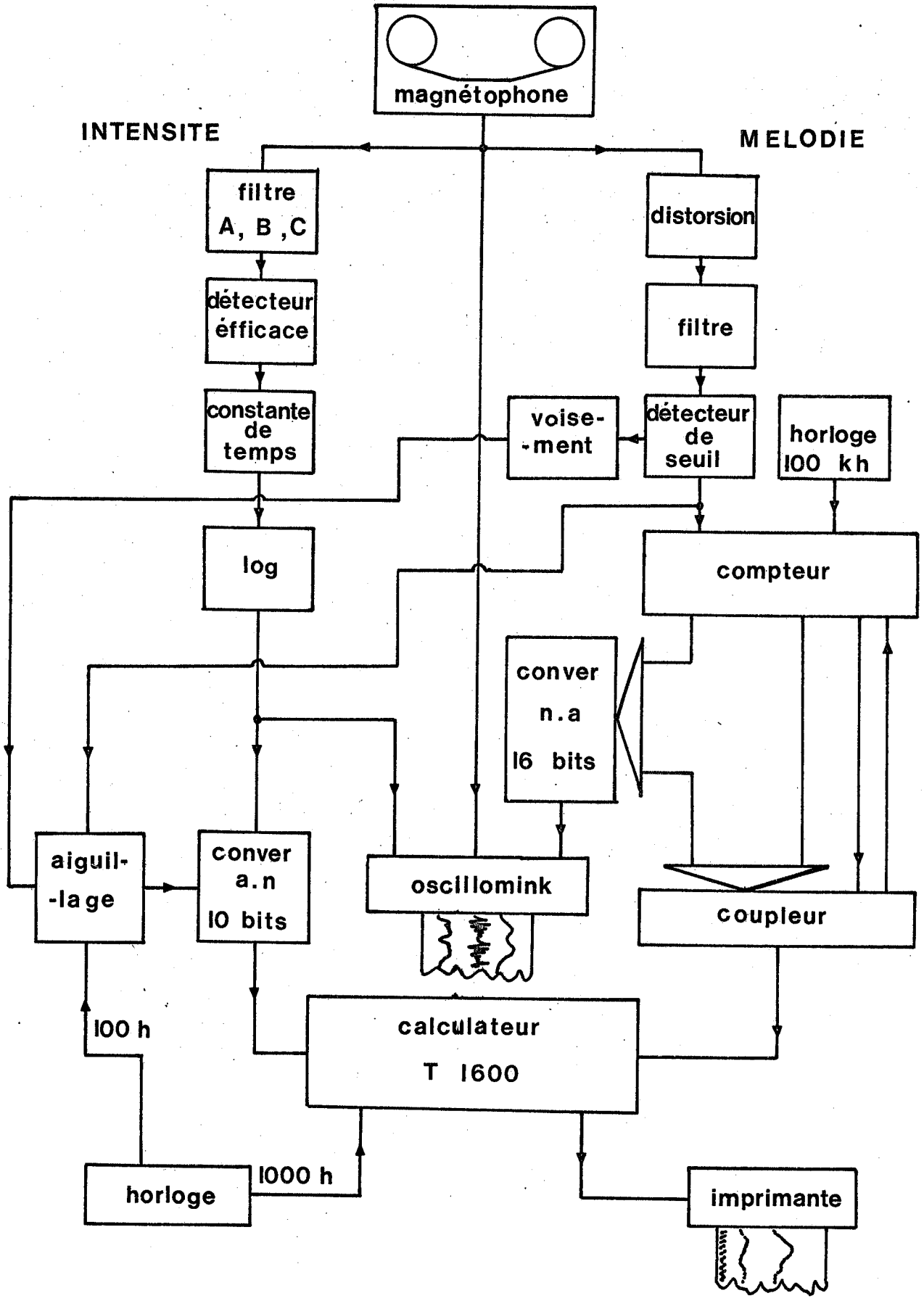
CARRÉ, R., LANCIA, R., PAILLE, J. et GSELL, R., Etude et réalisation d'un détecteur de mélodie pour l'analyse de la parole, L'onde électrique, 434, pp. 556-562, 1963.

GRUTZMACHER, M. und LOTTERMOSER, W., Uber ein Verfahren zur trägheitsfreien Dufreichnung von Melodiekurven, Akustische Zeitschrift, 5es Heft, Band 2, 2er Jahrgang, pp. 242-248, 1937.

HUNT, F.V., A direct-reading frequency meter suitable for high speed recording, The review of scientific instruments, 6, pp. 43-46, 1935.

LEON, P. et MARTIN, Ph., Prolégomènes à l'étude des structures intonatives, Studia Phonetica, 2, pp. 100-171, 1970.

Mc KINNEY, N.M., Laryngeal frequency analysis for linguistic research, University of Michigan, Communication Science Laboratory, Contr. n° 1224 (22), NR 049-122, 1965.


magnétophone
INTENSITE
MELODIE
filtre A, B ,C
détecteur éfficace
constante de temps
distorsion
filtre
détecteur de seuil
voise- -ment
horloge 100 kh
log
compteur
conver n.a 16 bits
aiguil- -lage
conver a.n 10 bits
oscillomink
coupleur
100 h
calculateur T 1600
horloge
1000 h
imprimante


Intensité
Mélodie
Temps cumulé
J' en p a r le

VIIIèmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

LA MESURE DU FONDAMENTAL PAR LA DETECTION DE CRETES

TECHNIQUES EMPLOYEES, RESULTATS

J.F. ZURCHER C.N.E.T. LANNION

• Résumé :

Cet exposé se réfère au dispositif de détection et de mesure du fondamental qui a été décrit lors des 6èmes Journées d'Etude sur la Parole à Toulouse.

On y précise la notion de détection de crêtes.

On donne les résultats de tests obtenus sur des phrases téléphonométriques prononcées par des locuteurs masculins et féminins dans différentes conditions.

• Summary :

In this paper a pitch-detector is discussed. Its principle was presented at 6e Journées d'Etudes, 1975.

Its peak detector is described in details.

Results obtained on sentences prononced by male and female speakers are given under Various conditions.

" Eh bien ! C'est à la fois très simple et très compliqué !..."

HERGE : Tintin au pays de l'or noir
page 55, 6ème image

LA MESURE DU FONDAMENTAL PAR LA DETECTION DE CRETES
TECHNIQUES EMPLOYEES, RESULTATS

J.F. ZURCHER, C.N.E.T. LANNION

I - INTRODUCTION

Un article précédent /1/ a décrit les principes généraux d'un détecteur de fondamental. Nous indiquons les résultats obtenus sur un dispositif dérivé de /1/ après avoir décrit de façon plus détaillée la détection de crêtes dans sa dernière version.

La "philosophie" de notre procédé réunit les avantages de divers procédés, parmi lesquels ceux de Gold /2/ et de Maissis /3/.

Nous utilisons une méthode temporelle qui associe au principe de détection de maxima divers perfectionnements : régulation d'amplitude et prise en compte de la dissymétrie d'évolution du signal, tant sur les valeurs des maxima locaux que sur la pente du signal à l'apparition de l'impulsion d'excitation.

II - LA DETECTION DE CRETES

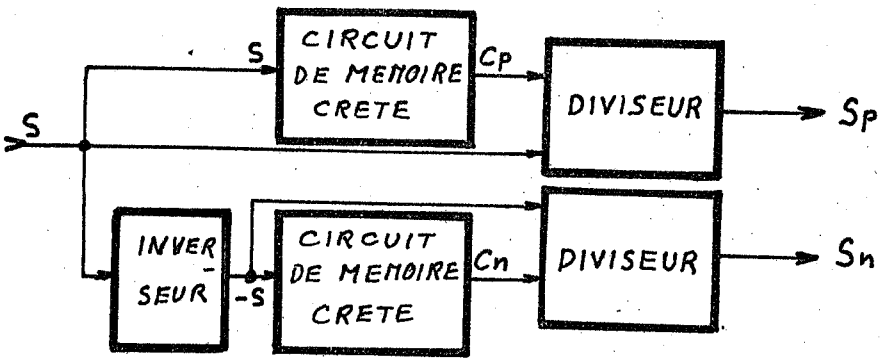
Cette technique consiste à mettre en évidence, et à détecter, les crêtes significatives de la périodicité du signal de parole. Cette opération proprement dite est précédée d'un filtrage passe-bas et d'une normalisation des amplitudes des deux enveloppes de crêtes maximum.

- Le filtrage passe-bas a sa fréquence de coupure à 350 Hz et une pente de 12 dB par octave. Ce filtrage est adapté à la mesure d'une large gamme de fréquences fondamentales. La technique de détection de crêtes utilisée ne nécessite pas un filtrage passe-bas plus sévère.

- La normalisation des amplitudes des deux enveloppes de crêtes maximum est destinée à faciliter la détection des crêtes maximum, en les alignant sur une courbe d'enveloppe de crêtes la plus rectiligne possible.

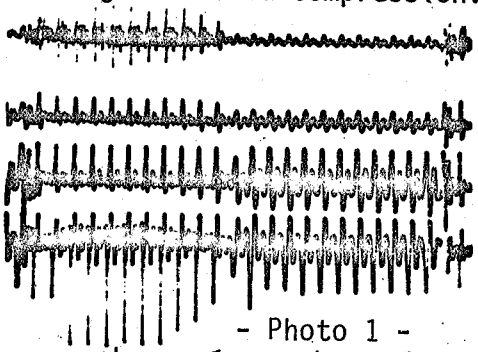
Cette opération est effectuée dans deux chaînes identiques, l'une traitant l'enveloppe des crêtes positives, l'autre l'enveloppe des crêtes négatives. Le fonctionnement peut être suivi sur le schéma de principe ci-dessous.

.../...



Le signal S (ou - S) est divisé par le signal de mémoire crête positif Cp (ou négatif Cn) dans un diviseur analogique. On obtient ainsi le signal Sp (ou Sn).

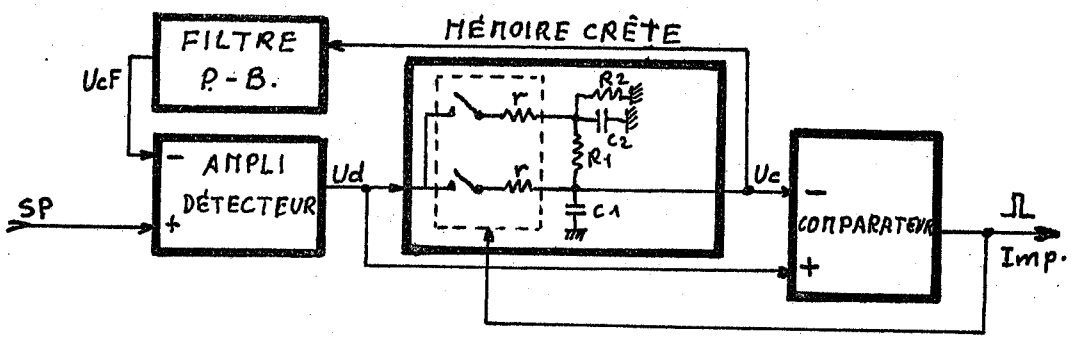
La fuite du circuit de mémoire crête se fait suivant un profil qui minimise la distorsion du signal compressé Syn : une décroissance du 2ème ordre. Si à l'issue d'un temps supérieur à 15 ms aucune crête n'a été "rencontrée" la constante de temps de décroissance est diminuée afin d'accélérer la "récupération" du signal brusquement affaibli. L'oscillogramme suivant montre les résultats du filtrage et de la compression.



- Signal de parole
- Signal filtré
- Compression de l'enveloppe positive
- Compression de l'enveloppe négative

- Photo 1 -

- Le repérage des crêtes significatives est effectué sur les deux signaux. Sp et Sn qui sont traités dans deux chaînes identiques. Le schéma de principe d'une chaîne est donné ci-dessous :



.../...

Le signal Sp (ou Sn) attaque l'entrée positive d'un amplificateur-détecteur simple alternance dont le signal de sortie Ud est connecté à un circuit de mémoire crête qui délivre un signal d'enveloppe de crête, Uc. Ud et Uc sont connectés aux entrées positives et négatives d'un comparateur. Quand la sortie Imp du comparateur est positive elle commande la fermeture des interrupteurs du circuit de mémoire crête. Tant que la tension Ud augmente, elle est légèrement supérieure à la tension Uc du condensateur C1 chargé à travers une petite résistance r, la sortie du comparateur est donc positive, et les interrupteurs sont fermés. Mais quand Ud commence à décroître, c'est la tension Uc qui est légèrement supérieure à Ud et la sortie du comparateur devient négative, provoquant l'ouverture des interrupteurs. Les condensateurs C1 et C2 se déchargent alors librement à la masse par l'intermédiaire des résistances R1 et R2. La décroissance de Uc obéit à l'équation différentielle :

$$\frac{d^2\,Uc}{dt}\,(\theta_1 \cdot \theta_2) + \frac{d\,Uc}{dt}\,(\theta_1 + \theta_2 + \theta_3) + Uc = 0$$

avec $\theta_1 = R_1 \cdot C_1$; $\theta_2 = R_2\,C_2$; $\theta_3 = R_2\,C_1$

La solution est de la forme :

$$U_c\,(t) = K_1 e^{x1\,t} + K_2\,e^{x2\,t}$$

x_1 et x_2 étant les racines réelles de l'équation

$$\theta_1\theta_2\,x^2 + (\theta_1 + \theta_2 + \theta_3)\,x + 1 = 0$$

Les conditions initiales à l'instant zéro de l'ouverture des interrupteurs sont :

$$U_c\,(0) = U_c\,\max \text{ et } U_c'\,(0) = 0$$

Elles permettent de déterminer K_1 et K_2. On trouve :

$$U_c\,(t) = \frac{x_2}{x_2 - x_1} \cdot U_c\,\max \cdot e^{x1\,t} + \frac{x_1}{x_1 - x_2} \cdot U_c\,\max \cdot e^{x_2\,t}$$

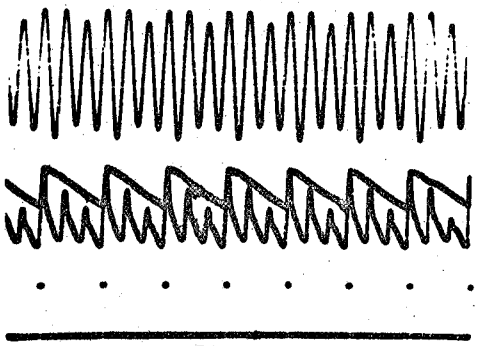
C'est la somme de deux décharges exponentielles.

La décharge résultante a comme propriété intéressante de démarrer avec une pente nulle, ce qui permet d'éviter la détection d'une crête voisine de la crête d'amplitude maximum et d'amplitude légèrement inférieure. Le signal U_c, retardé d'une milliseconde environ dans un filtre passe-bas, est appliqué en contre-réaction sur l'entrée négative de l'amplificateur détecteur, affaiblissant ainsi le signal Ud après la détection d'une crête significative. Cet affaiblissement va en diminuant au rythme de la décroissance de Uc jusqu'à la détection d'une nouvelle crête significative. Nous allons montrer que trois paramètres attachés aux crêtes sont ainsi pris en compte.

- L'amplitude : cela n'est pas original, tous les détecteurs de crête prennent ce paramètre en compte !

.../...

- La différence d'amplitude entre crêtes adjacentes : l'oscillogramme ci-dessous illustre bien cette propriété !

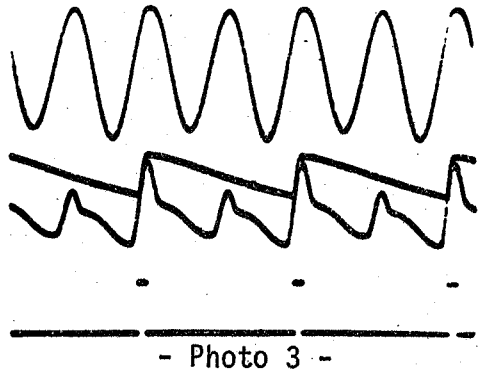


- Signal filtré et compressé
- Signal U_c
- Signal U_d
- Impulsions marquant les crêtes détectées.

- Photo 2 -

Les crêtes retenues sont celles dont l'amplitude est plus grande que celle des crêtes qui la précèdent.

- La pente du signal avant les crêtes : cette propriété est illustrée par l'oscillogramme suivant :



- Signal filtré et compressé
- Signal U_c
- Signal U_d
- Impulsions marquant les crêtes détectées.

- Photo 3 -

Les crêtes retenues sont celles dont la pente est la plus importante. Ce sont elles qui émergent le plus, avant que U_c retardée ne leur soit retranchée.

III - RESULTATS OBTENUS

Les résultats que nous présentons ici ne concernent que la mesure de la valeur du fondamental et non la fonction de voisement dont les résultats ont déjà été publiés /4/.

La dispositif a été testé à l'aide de 5 phrases utilisées en téléphonométrie :

- " Le crédit est consenti par la banque
- Le docteur a ordonné un médicament
- Le charretier fouette son cheval
- Le fantôme hante le château
- Le sculpteur taille la pierre".

Ces phrases ont été prononcées par 3 locuteurs masculins nommés A, B, C et 3 locuteurs féminins nommés D, E, F.

.../...

Lors de la prononciation des 5 phrases les fondamentaux des différents locuteurs se situaient dans les plages suivantes :

A :	85 Hz	-	160 Hz
B :	85 Hz	-	130 Hz
C :	75 Hz	-	190 Hz
D :	140 Hz	-	275 Hz
E :	150 Hz	-	275 Hz
F :	145 Hz	-	230 Hz

Les tests suivants ont été effectués :

Test I :
- Signal direct (haute fidélité).

Test II :
- Filtrage passe-haut représentant en gros la fonction de transfert d'une capsule téléphonique (deux fréquences de coupure à 150 Hz et à 250 Hz).

Test III :
- Filtrage passe-haut sévère (fréquence de coupure 300 Hz ; - 48 dB à 150 Hz et - 96 dB à 75 Hz).

Test IV :
- Superposition d'un bruit blanc multiplicatif avec un rapport signal/bruit de 10 dB.

Nous avons relevé les erreurs en pourcentage de temps par rapport au temps total de voisement pour chaque locuteur et pour chaque test.

Nous avons distingué trois types d'erreurs :

- Les erreurs se produisant à l'établissement du voisement nommées E.d.
- Les erreurs se produisant en cours d'une zone de voisement notées e.m.
- Les erreurs se produisant juste avant la fin d'une zone de voisement notées e.f.

Nous avons pris en compte les erreurs évidentes à l'oeil sur la courbe de fréquence fondamentale enregistrée se situant au delà d'une plage de 20 % autour de la courbe de fondamental lissée /5/. Une analyse précise montre d'ailleurs que la quasi totalité des petits écarts trouvés est justifiée par des variations brutales de la période fondamentale dans des zones de transition entre phonèmes différents.

.../...

		TEST I			TEST II			TEST III			TEST IV		
HOMMES	Locuteurs	ed	em	ef	ed	em	ef	ed	em	ef	ed	em	ef
	A	1,5	0	0	0	0	0	9,3	0,7	0	2,3	1,5	0
	B	0	0	0	0	0	0	10	1,8	0	0	0	0
	C	0	0	0	0	0	0	8,2	3,6	1	3,2	0,6	0
	Moyennes A,B,C,	0,5	0	0	0	0	0	9,2	2	0,3	1,8	0,7	0
	Erreurs totales Moyennes Par test (%)	0,5			0			11,5			2,5		
FEMMES	Locutrices	ed	em	ef	ed	em	ef	ed	em	ef	ed	em	ef
	D	1,1	0	0	0,9	0	0,6	3,6	1	0,7	2,3	2,3	0,9
	E	0,8	0	0	0,2	0	0	2	0,6	0	7,6	8,6	2
	F	2,1	0	0	1,7	0	0	2,4	0,3	0	7,2	5,4	0
	Moyennes D,E,F,	1,3	0	0	0,9	0	0,2	2,7	0,6	0,2	5,7	5,4	1
	Erreurs totales Moyennes Par test (%)	1,3			1,1			3,5			12,1		
MOYENNES GENERALES HOMMES+FEMMES		0,9			0,5			7,4			7,2		

Nous remarquons que les meilleurs résultats sont obtenus pour le test II. Il est intéressant de constater qu'un filtrage passe-haut léger (environ - 10 dB à 100 Hz) est très bien supporté même par les fondamentaux très bas. Il apporte même une amélioration dans la mesure où il atténue les perturbations à très basse fréquence.

Les conditions de filtrage du test III sont relativement bien tolérées par les voix féminines quoique leurs fondamentaux y soient atténués en moyenne d'environ 32 dB. Ceci amène à des conditions de fonctionnement particulièrement sévères par rapport à l'hypothèse de Léon et Martin [5] qui admettent un écart de 12 dB entre harmonique 2 et fondamental. On trouve pour les 3 locuteurs féminins un pourcentage d'erreur de 3,5 %. Par contre pour les locuteurs masculins ce pourcentage monte à 11,5 %. Cela n'est pas étonnant, le fondamental moyen étant affaibli d'au moins 60 dB dans ce test.

Le test IV montre une bonne résistance des voix masculines à un bruit blanc multiplicatif : 2,5 % d'erreurs. Par contre nous trouvons 11,5 % d'erreurs pour les locuteurs féminins.

D'une manière générale les erreurs trouvées se situent en majorité sur les attaques, en début de voisement. Deux explications peuvent être données. Il y a de temps en temps une légère anticipation du détecteur de voisement qui oblige le dispositif à donner une valeur alors que la périodicité n'est pas encore établie. D'autre part l'algorithme de correction ne fonctionne qu'après

.../...

les 2 premières périodes mesurées au début d'une zone de voisement.

IV - CONCLUSION

La méthode décrite, implantée sur trois cartes de 300 cm^2 chacune, donne des résultats qui paraissent soutenir la comparaison avec ceux de l'Etude de Rabiner et al. /6/, dans la mesure où les résultats d'expériences aussi différentes peuvent se comparer.

Il reste maintenant à :

- faire une étude analytique des erreurs résiduelles ;
- faire une étude perceptuelle des conséquences des erreurs ;
- améliorer les résultats dans le cas relativement critique de l'absence de fondamental en tenant compte du fait que l'essentiel des erreurs se situe au début des zones voisées, ce qui peut conduire à une correction en temps différé en utilisant une "postdiction" de préférence à un lissage ou une prévision.

/1/ ZURCHER, CARTIER, BOE. Détection et Mesure du Fondamental
6èmes Journées d'Etude sur la Parole - Toulouse 28-30 mai 1975.

/2/ B. GOLD and L.R. RABINER. "Parallel processing techniques for estimating pitch periods of speech in the time domain", I. Acoust. Soc. Amer. Vol 46 pp. 442-448, Aug 1969.

/3/ MAISSIS. Méthode d'Extraction du Pitch
3èmes Journées d'Etude sur la Parole 31 mai au 2 juin 1972 - LANNION.

/4/ ABRY, ZURCHER. La Détection du Voisement par les propriétés physiques résultant de l'excitation périodique du conduit vocal : Comparaison statistique de trois procédés.
6èmes Journées d'Etude sur la Parole - Toulouse 28-30 mai 1975.

/5/ LEON, MARTIN. Prolégomènes à l'Etude des Structures intonatives.
p. 145. Studia Phonetica - 2 - Didier.

/6/ L.R. RABINER, M.J. CHENG, A.E. ROSENBERG, C.A. Mc GONEGAL.
A comparative Performance Study of several Pitch Detection Algorithms.
I.E.E.E. Transactionq on Acoustics, speech, and Signal Processing,
Vol. ASSP-24, n° 5, Octobre 1976 pp 399-418.

THEME 2

Perception de la hauteur et de ses variations (microprosodie)

VIII èmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

VALIDATION D'INDICES PERCEPTIFS DE L'INTONATION PAR ANALYSE MULTIDIMENSIONNELLE

Fernand CARTON et François LONCHAMP, INSTITUT DE PHONETIQUE, NANCY

• Résumé :

Cette analyse multidimensionnelle de configurations mélodiques complète les analyses unidimensionnelles de la phrase énonciative proposées dans des travaux récents. Un filtrage a éliminé l'information syntaxique et sémantique. Nous validons 4 indices (pivot tonal, attaque, pause, position du sommet) et nous décrivons certaines interactions de ces paramètres perceptifs.

• Summary :

A multidimensional analysis of 12 intonation patterns carried by noninterrogative (assertive) french sentences was carried out to supplement data on unidimensional patterns gathered by other recent studies. Filtering was used to delete semantic and syntactic cues. 4 intonational cues were extracted : rising vs falling vowel, low vs high onset, cleft vs non-cleft sentence, position of "major continuation" within the sentence. We also discuss various interactions between these perceptual cues.

0.1 Grâce au renouveau des études intonologiques (DELATTRE, FAURE, LEON etc.), plusieurs chercheurs ont décrit du point de vue acoustico-perceptif et formalisé les structures intonatives du Français (CONTINI & BOE, DI CRISTO, MARTIN, VAISSIERE). Les diverses classes utilisées dans ces importants travaux ont été déterminées soit par transformation de données acoustiques en données perceptives, soit par analyse statistique. Une validation multi-dimensionnelle, permettant de faire varier plusieurs indices à la fois, nous a paru nécessaire pour révéler les interactions possibles. La méthode adoptée est l'Individual Scaling de CARROLL & CHANG, mise en oeuvre grâce à des programmes d'ordinateur rédigés par notre laboratoire.

0.2 Après 2 expérimentations sur corpus spontané dialectal (1974,1975) qui ont mis au point la méthode, l'expérience III (1976) a soumis au test 12 énoncés naturels filtrés non-interrogatifs (7 syllabes) de constitution syntaxique variée. L'analyse multi-dimensionnelle a permis d'extraire 4 indices, à l'exclusion de tout autre : 1 . Niveau d'attaque (cf. appel, proéminence expressive liée à des faits de phonétique textuelle peu étudiés jusqu'à présent) - 2 . Place du sommet de modulation (cf. continuation majeure, progrédience, frontière prosodique non-terminale) - 3 . Clivage par pause - 4 . Montée ou descente de la voyelle finale (cf. "pivot tonal" de FAURE, implication).

1.1 Nous avons voulu vérifier, à partir de réalisations d'une seule phrase, énonciative d'intention (implicative ou non), utilisant systématiquement ces 4 indices, que ceux-ci étaient bien extraits à nouveau par une nouvelle série de sujets "naïfs".

1.2 Ces phrases ont été lues par un phonéticien professionnel en chambre sourde. Il s'agit d'un énoncé entièrement voisé de 7 syllabes comportant une ambiguïté syntaxique :
(Jean-Louis)(voulait)((bien)manger) ; (Jean-Louis)(voulait bien(manger))
Les indices suivants ont été systématiquement variés :

Indices	A	B
1 - Sommet	tôt	tard
2 - Attaque	haute	basse
3 - Clivage	Oui	Non
4 - Voyelle Finale	Montante	Descendante

Sur les 16 (2^4) combinaisons, 4 ont été éliminées comme non naturelles par 2 phonéticiens. Il semble que l'attaque haute soit difficilement compatible avec un sommet véritable. Les 12 phrases restantes interdisent bien toute prédictibilité.

1.3 Les énoncés ont été filtrés comme pour l'expérience III (filtre passe-haut et passe-bas à 180Hz,-32 dB/oct.). Passation collective en champ libre (salle non traitée), Revox A77 et enceintes AR2a. 22 sujets, étudiants d'Université non entraînés. Scaling sur 7 positions de la différence perçue entre chaque terme de tous les couples de phrases possibles fournissant directement les distances.

2.1 L'analyse multi-dimensionnelle de ce test (jugé difficile) a permis d'extraire 4 dimensions pour lesquelles la quantité de variance expliquée est supérieure à 50 % pour 11 sujets sur 22. Aucun sujet n'a utilisé les 4 indices : tous ont utilisé 2 ou 3 dimensions. Une dimension supplémentaire ne permet pas de réduire le "bruit". Si l'on admet comme significative une dimension expliquant plus de 16% de la variance pour un sujet, on obtient le classement global suivant : la hauteur de l'attaque a été utilisée par 10 sujets, la hauteur relative de la finale par 9, la place du sommet par 6 ; seuls 3 sujets ont privilégié le clivage. Il semble que ce dernier indice corrèle assez fortement avec le niveau d'attaque.

2.2 RESULTATS (Les lettres A et B renvoient aux indices réalisés)

Stimulus	Dimension 1	Dimension 2	Dimension 3	Dimension 4
1	.09 A	-.34 B	-.96 B	.61 B
2	.99 B	-.70 B	-.29 B	1.01 B
3	1.18 B	-.23 B	-.18 B	-.94 A

4	-.02	A	2.23	A	-1.05	B	.20	B
5	.60	A	.58	A	-1.66	B	-1.39	A
6	-1.48	A	-.84	B	1.12	A	.73	B
7	-1.16	A	1.92	A	1.40	A	.63	B
8	-1.34	A	-.73	B	1.17	A	-1.23	A
9	1.00	B	-.77	B	1.10	A	1.05	B
10	1.09	B	-.24	B	.66	A	-1.34	A
11	.32	B	-.63	A	-.67	B	1.38	B
12	-1.28	A	-.26	B	-.65	B	-.72	A

La figure 1 présente une description schématique des 12 phrases tenant compte des 4 indices, qu'on comparera utilement avec les tracés du fondamental (fig. 2)

2.3 DIMENSION 1 : POSITION du SOMMET TOT/TARD. La place du sommet n'est bien définie que pour les phrases possédant une attaque basse. La phrase à attaque haute clivée (item 7) est classée comme une phrase à sommet "précoce". Les phrases à attaque haute non-clivée sont non-marquées par rapport à ce critère. Le comportement de l'item 1 est aberrant. Attaque et position du sommet ne sont donc pas des dimensions totalement indépendantes.

2.4 DIMENSION 2 : ATTAQUE HAUTE/BASSE. On note une nette séparation. L'item 11 (attaque haute) est classé parmi les attaques basses, sans doute parce qu'il est réalisé avec plateau initial. Ce n'est donc pas le seul registre de l'attaque qui importe, mais le contraste syllabe 1/syllabe 2, ce qu'un test spécifique devrait confirmer.

2.5 DIMENSION 3 : CLIVAGE/NON CLIVAGE. Nette séparation (paramètre important pour 3 sujets seulement, malgré la durée importante de certaines pauses). On note un bon parallélisme entre la durée de la pause et la position sur l'axe 3 :

item	7	8	6	9(+)	10
durée de la pause en cs	23	12	10	-	5
position	1.40	1.17	1.12	1.10	.66

(+) Pour l'item 9, il y a absence de pause de phonation ; substitution d'un allongement de la voyelle précédente, allongement de la phase d'occlusion de /b/ ; micromélodie très marquée.

2.6 DIMENSION 4 : VOYELLE FINALE MONTANTE/DESCENDANTE. Très nette séparation. On notera cependant que les montées sur les voyelles finales sont plus importantes que dans l'expérience III. Ces phrases ont pourtant été jugées très naturelles.

3.1 Pour l'intonation de la phrase énonciative envisagée sous l'angle de la communication, nous confirmons qu'un auditeur a à sa disposition 4 indices : 1° la variation de hauteur finale, très bien perçue (les auditeurs étant conditionnés par l'oxytonisme), validant le rôle particulièrement discriminant que joue le "pivot" tonal". 2° Le niveau d'attaque, tout en remarquant (item II) que le registre ne paraît pas fonctionner isolément. 3° La place du sommet pour lequel le test met en évidence une interaction avec le 4ème indice, la pause.

3.2 Toute théorie devrait inclure, selon nous, ces indices extraits par les sujets eux-mêmes. D'autres indices non extraits peuvent évidemment jouer un rôle, ne serait-ce que parce que nous avons éliminé le sémantisme. Les sujets (effet de masque ?) n'ont pas tenu compte de la largeur de bande (amplitude de la modulation) ou du débit. Mais, sous réserve d'expériences complémentaires, on peut mettre en doute la pertinence linguistique d'indices non utilisés dans des phrases de ce type. On sait cependant, au niveau de la méthode utilisée, que si un item entre fortement dans une certaine opposition, celle-ci tend à masquer le réseau des oppositions moins importantes. La multiplication d'analyses de ce type peut remédier à cet inconvénient et devrait permettre d'établir le rendement des configurations intonatives dans le "double codage" (FÓNAGY).

Cette approche, qui se veut complémentaire d'analyses acoustico-perceptives, permet une hiérarchisation des indices perceptifs et sera poursuivie, car

elle permet de ne pas simplement "tester des théories", mais de montrer comment les paramètres classiques sont intégrés les uns par rapport aux autres.

o o o o o

CARROL J.D. & CHANG J.J. (1970), "Analysis of Individual Differences in multidimensional Scaling via an N-way generalization of the Eckart-Young decomposition", Psychometrika, 35, 283-319.

CARTON F. (1974), Introduction à la phonétique du français, Paris, Bordas.

CARTON F. (1977), "Application de l'analyse multi-dimensionnelle à la caractérisation d'un français dialectal", Mélanges dédiés à Jean SEGUY, Toulouse, sous presse (rédigé 1974)

CARTON F. et LONCHAMP F. (1975), "Expériences sur la reconnaissance de traits intonatifs dialectaux", VIII° Congrès International des Sciences Phonétiques, Leeds (à paraître)

CARTON F. et LONCHAMP F. (1976), "Essai de validation psycho-linguistique de configurations intonatives par analyse multidimensionnelle", III° World Congress of Phoneticians, Tokyo, (à paraître).

CONTINI N. et BOE L.J. (1975), "Etude quantitative de l'intonation en français ; premiers résultats", VIII° Congrès International des Sciences Phonétiques, Leeds (à paraître).

DI CRISTO A. (1975), "Recherches sur la structuration prosodique de la phrase française", 6° Journées d'Etude sur la Parole, G.A.L.F., Toulouse, vol. 1, 94-116.

FAURE G. (1973), "La description phonologique des systèmes prosodiques", Studia Phonetica, VIII, Montréal, Didier, 1-16.

FÓNAGY I. et BERARD E. (1972), " Il est huit heures. Contribution à l'analyse sémantique de la vive voix", Phonetica,26, 157-192.

LONCHAMP F. (1974), "Note sur les programmes d'ordinateur mis au point au Laboratoire de Phonétique en 1973/1974", Travaux de l'Institut de Phonétique de Nancy, Vol 1, 235-256.

MARTIN Ph. (1975a), "Intonation et reconnaissance automatique de la structure syntaxique", 6° Journées d'Etude sur la parole,G.A.L.F., Toulouse, Vol. 1, 51-62.

MARTIN Ph. (1975b), "Analyse phonologique de la phrase française", Linguistics, 146, 35-68.

MARTIN Ph. (1975c), "Une grammaire de l'intonation française", Rapport d'activité, Institut de Phonétique, U.L. de Bruxelles, 9/2, 77-96.

VAISSIERE J. (1975), "Caractérisation des variations de la fréquence du fondamental dans les phrases françaises", 6° Journées d'Etude sur la parole, G.A.L.F., Toulouse, Vol. 1, 39-50.

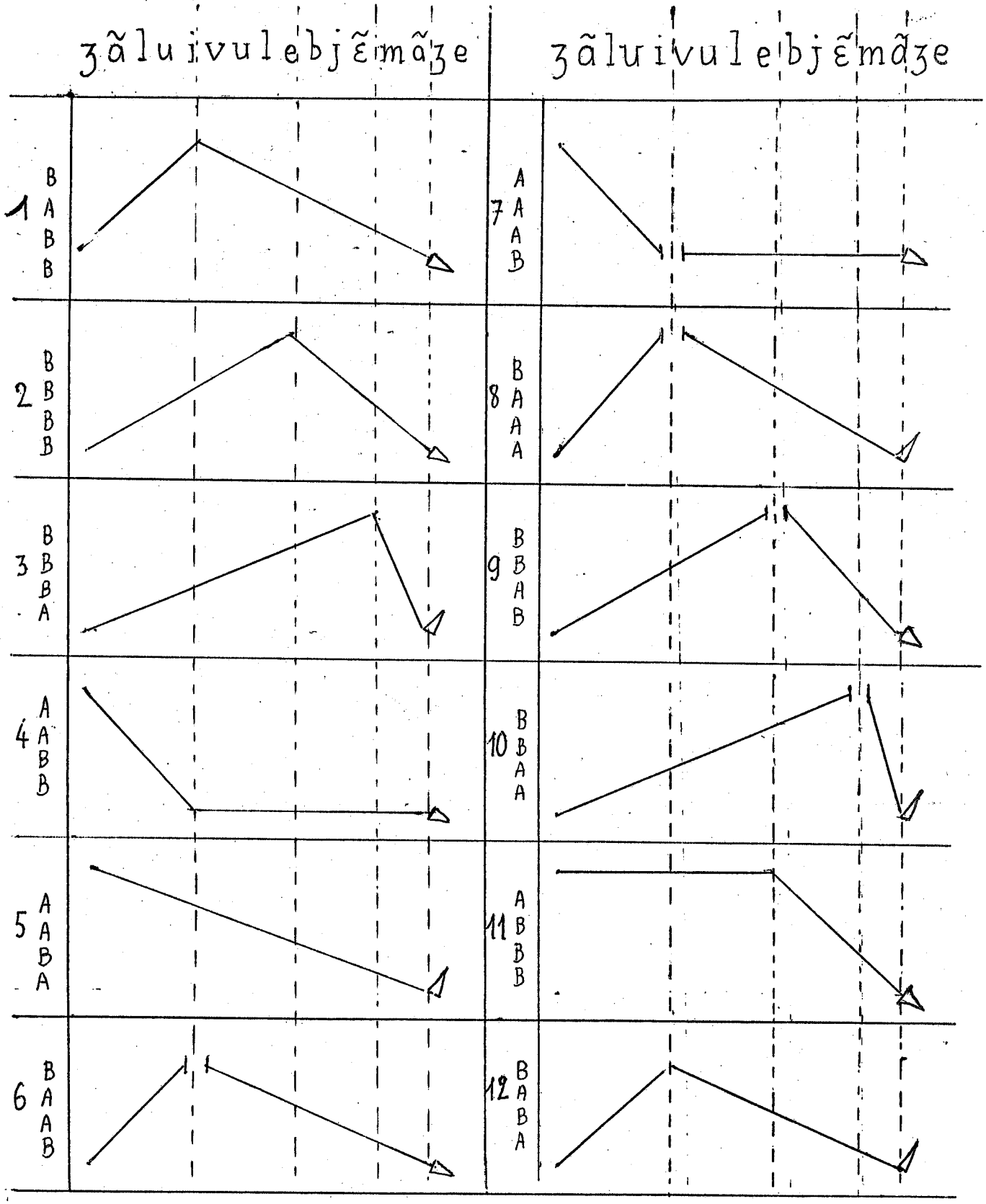


Fig. 1 Schématisation des indices perceptifs

Les lettres A et B se réfèrent au tableau 1.2 : de h. en b. sommet, attaque, clivage, voyelle finale

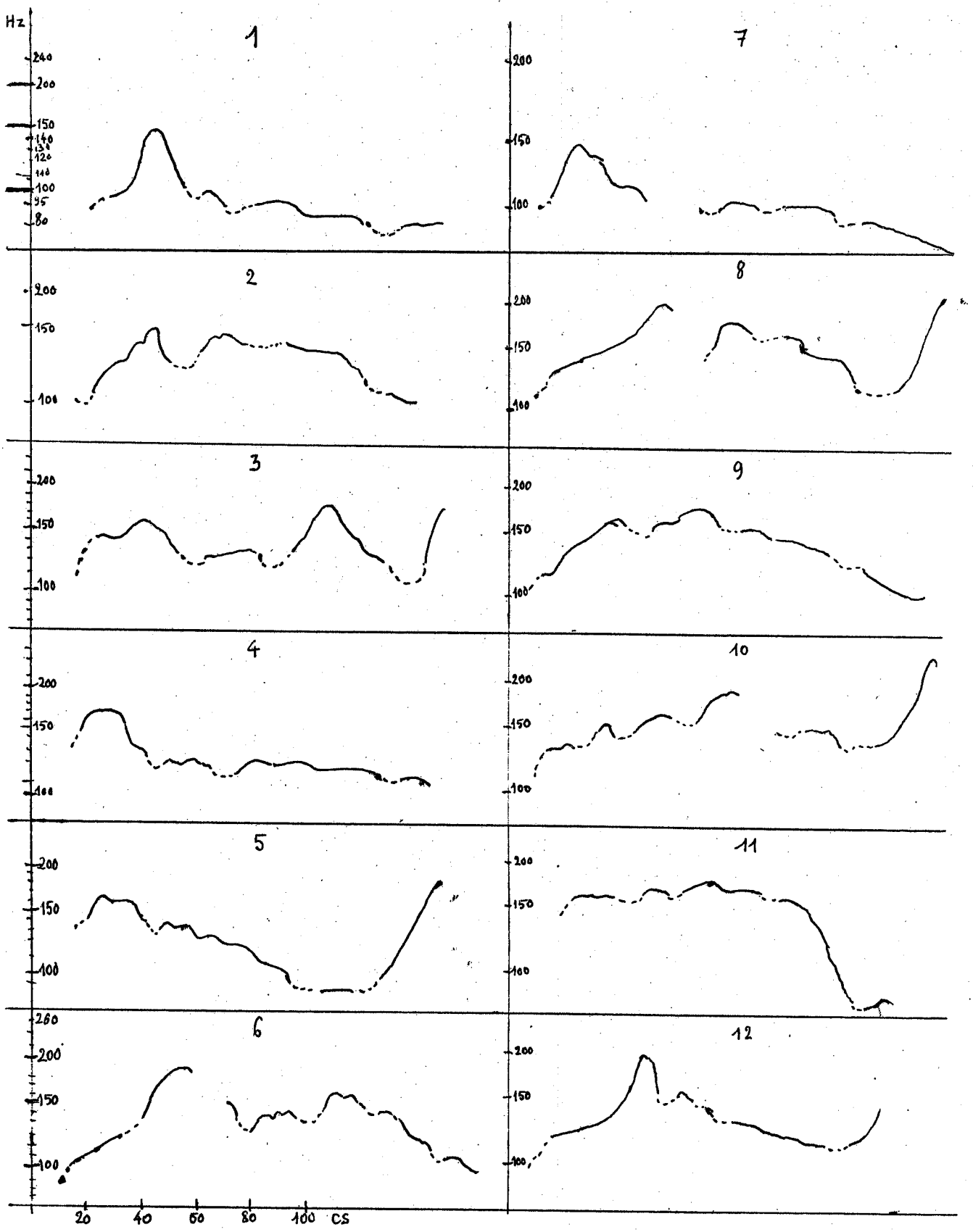


Fig. 2 Variations du fondamental

VIII èmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

LA PERCEPTION DE L'INTONATION ANGLAISE PAR DES ANGLOPHONES ET NEERLANDOPHONES

R. COLLIER

UNIVERSITÉ D'ANVERS-UFSIA, BELGIQUE ET
INSTITUUT VOOR PERCEPTIE ONDERZOEK, EINDHOVEN, PAY-BAS

• Résumé :

Dans la perception de l'intonation, c.a.d. des variations du fondamental de la langue parlée, il est possible de faire abstraction de certaines différences physiques pour aboutir à une classification globale des contours mélodiques en un nombre limité d'intonations de base. Cette faculté d'abstraction intonative a été étudiée chez des sujets anglophones et néerlandophones qui devaient "trier" vingt phrases anglaises selon leur propre critère de ressemblance mélodique.

• Summary :

In the perception of intonation the listener is capable of neglecting certain fundamental frequency variations in order to arrive at a global classification of pitch contours into a limited number of intonation patterns. This possibility of intonational abstraction was studied in English and French subjects who had to 'sort out' twenty English utterances according to their own criterion of perceptual resemblance.

LA PERCEPTION DE L'INTONATION ANGLAISE PAR DES ANGLOPHONES ET NEERLANDOPHONES

I. INTRODUCTION

Le présent rapport concerne le début d'un programme de recherches portant sur la comparaison de l'intonation du néerlandais et de l'anglais. Nous envisageons de soumettre l'intonation anglaise aux mêmes analyses perceptives que celles qui ont mené à une connaissance approfondie de l'intonation néerlandaise.

II. LE PROBLEME

Dans l'étude des aspects perceptifs de l'intonation du néerlandais nous avons entre autres abordé le problème suivant. Du point de vue acoustique nous constatons une variation infinie dans le cours du ton fondamental. Dans l'analyse perceptive certaines de ces différences physiques ne sont pas perçues, ce qui permet de styliser les variations du ton fondamental tout en respectant l'équivalence perceptive entre la stylisation et l'original. Si l'on compare des énoncés dont le fondamental a été stylisé il paraît néanmoins possible d'entendre des différences à condition d'écouter de façon analytique. Si, par contre, on écoute de façon globale, comme cela se fait dans une situation communicative, on peut atteindre un niveau d'abstraction supérieur. A ce niveau plusieurs différences entre les contours mélodiques(stylisés on non) sont jugées secondaires et considérées comme des variantes possibles dans la réalisation du même schéma intonatif. Autrement dit, la multiplicité des variantes mélodiques est réduite à un nombre limité de catégories intonatives sous-jacentes que l'on appelle les intonations de base.

Plusieurs expériences sur l'intonation néerlandaise ont eu comme objectif d'examiner la faculté d'abstraction intonative des néerlandophones (Collier et 't Hart 1972, 't Hart et Collier 1975, Collier 1975). Dans ces expériences les sujets étaient confrontés à une vingtaine d'énoncés qu'ils devaient grouper selon un critère de ressemblance mélodique.

J'ai voulu refaire la même expérience avec des sujets anglophones qui devaient à leur tour grouper des contours mélodiques de leur langue maternelle. J'ai aussi voulu étudier le comportement de sujets néerlandophones dans la même tâche, mais mis en présence d'énoncés dans une langue étrangère qu'ils connaissent mais pas de façon approfondie.

Les hypothèses suivantes ont été formulées : (1) les sujets sont capables de grouper vingt contours mélodiques différents dans un nombre de catégories nettement inférieur; (2) la catégorisation des anglophones correspond aux intonations de base décrites dans les manuels d'intonation anglaise; (3) la catégorisation des néerlandophones diffère de celle des anglophones.

En d'autres termes, la catégorisation ne se fait pas uniquement selon des critères mélodiques mais est influencée en partie par des facteurs linguistigues qui ne sont pas nécessairement communs aux deux langues.

III. L'EXPERIENCE

Comme point de départ j'ai choisi la description du système intonatif de l'anglais telle qu'elle est donnée par Halliday (1970). Dans son manuel l'auteur présente de manière assez détaillée les caractéristiques mélodiques de l'anglais de Grande-Bretagne. Il distingue sept intonations de base, appelées "Tones", et indique pour chacune un nombre de variantes. Son exposé est accompagné de bandes magnétiques contenant plusieurs exemples des Tones. Parmis ces illustrations j'ai choisi vingt phrases qui représentent au minimum deux et au maximum cinq variantes des sept Tones. Dans la Figure 1 on trouve une représentation du fondamental des vingt phrases, ainsi que leur catégorisation en 'Tones' proposée par Halliday. Les vingt phrases ont été enregistrées sur des fiches à ruban magnétique avec un enregistreur du type Language Master (Bell-Howell). De cette façon il est possible d'avoir un accès immédiat aux phrases, ce qui rend la comparaison dans n'importe quel ordre plus facile. Le premier groupe de sujets comportait treize anglophones, le second quatorze néerlandophones. Il leur était demandé de "trier" les phrases selon un critère de ressemblance mélodique, le nombre de groupes étant laissé à leur propre jugement.

IV. RESULTATS

Si l'on compte pour chacune des phrases le nombre de fois qu'elle a été groupée avec une ou plusieurs autres on obtient un score qui indique le degré de similarité entres les phrases individuelles. Les phrases jugées très similaires forment des paires ou des groupes cohérents. Dans un tel groupe il peut y avoir un membre qui a aussi un rapport (plus faible) avec une phrase dans un autre groupe. Ainsi s'établit une relation hiérarchique

entre les groupes. Il existe une méthode de calcul qui définit cette hiérarchie des groupements à partir des scores obtenus pour les phrases individuelles. (Johnson, 1967). La Figure 2 présente les resultats de la variante dite "maximale" de ce calcul. Si dans cette figure deux phrases se joignent à un niveau élevé cela indique un haut degré de similarité.

1. Sujets anglophones (voir Figure 2A)

Nous constatons que les groupements entre les niveaux 10 et 6 ne concernent pratiquement que des paires de stimuli. Les membres d'une paire font partie d'un même Tone, à l'exception de (14,19).
Ce ne sont que des représentants des Tones 1, 2 et 4 qui forment des paires. Les variantes des Tones 3, 5, 13 et 53 ne constituent jamais une paire, même pas à un niveau inférieur à 6.
Au niveau 5 toutes les variantes du Tone 1 sont groupées (1,2,3,4,5). Tone 1 est le seul dont les variantes se groupent toutes à un certain niveau. Au niveau 5 se forme aussi le groupe (12,15,16), mais celui-ci comprend des représentants des Tones 13 et 4,est donc de composition hétérogène. Il est remarquable que le groupement des variantes du Tone 1 ne se fait qu'au niveau 5, alors que déjà au niveau 6 apparaît le groupe (13,14, 19,20) qui mélange les Tones 13, 4 et 53.
Pour expliquer pourquoi des paires telles que (6,7) ou (15,16) ont été formées à un niveau élevé on ne pourrait se baser uniquement sur le degré de ressemblance physique telle qu'elle apparaît dans la Figure 1. En effet cette ressemblance est toute aussi grande entre, par exemple, les phrases 12 et 13, qui n'ont pas été prises ensemble. D'autre part il y a peu de ressemblance physique entre, par exemple, la phrase 12 et la paire de phrases (15,16) qui forment un groupe au niveau 5.

2. Sujets néerlandophones (voir Figure 2B)

La formation de paires par les sujets néerlandophones correspond bien à celle des anglophones, mais les groupements se font déjà à un niveau supérieur. Les paires se constituent de stimuli appartenant aux Tones 1,2, 4 et 5. Le groupement de plus de deux stimuli se fait également à un niveau supérieur en comparaison avec les résulats précédents. Ici aussi les sujets jugent que les variantes du Tone 1 doivent être groupées. Deux autres groupes se distinguent : (6,7,10) et (8,9,12,13). Ceux-ci ne cor-

respondent pas entièrement à la classification de Halliday, mais du point de vue physique ces groupements sont plus convaincants que les mélanges hétérogènes produits par les anglophones.

V. QUELQUES REFLEXIONS

Il est remarquable que les deux groupes de sujets s'accordent sur deux points : (I) ils placent toutes les variantes du Tone 1 dans la même catégorie; (2) ils divisent les quatre variantes du Tone 2 en deux groupes distincts, c'est à dire (6,7) et (8,9). Compte tenu des différences physiques entre ces deux types de variantes la division me paraît justifiée et refute la classification de Halliday à juste titre. Les réactions des subjects néerlandais donnent également à penser qu'il faut douter de la distinction que Halliday fait entre, d'une part, les variantes du Tone 13 (12,13) et de l'autre les contours (8,9) qui représentent en partie le Tone 2.

La classification des Tones proposée par Halliday prétend être "mélodique". Cependant on a l'impression que Halliday a aussi fait intervenir des considérations "fonctionnelles". En effet, les phrases qui illustrent l'emploi du Tone 2 sont toutes des "questions", alors que celles qui exemplifient le Tone 13 sont des "affirmations". Cela pourrait expliquer pourquoi Halliday - peut-être inconsciemment - fait une distinction entre (8,9) et (12,13), malgré leur ressemblance mélodique, alors qu'il groupe (6,7) et (8,9) qui sont nettement différents du point de vue mélodique.

Les subjects néerlandophones présentent une classification qui est plus plausible du point de vue mélodique. Mis en présence d'une langue étrangère ils se concentrent plus facilement sur l'aspect purement mélodique des stimuli, alors que les anglophones (et même Halliday dans une certaine mesure) sont divertis par l'attention qu'ils prêtent aux ressemblances interpretatives des contours mélodiques.

Bibliographie

Collier, R. (1975) Perceptual and linguistic tolerance in intonation. IRAL 13, 293-308.

Collier, R. et 't Hart, J. (1972). Perceptual experiments on Dutch intonation. Proc. 7th Int. Congr. Phon. Sci. Mouton, La Haye. 880-884.

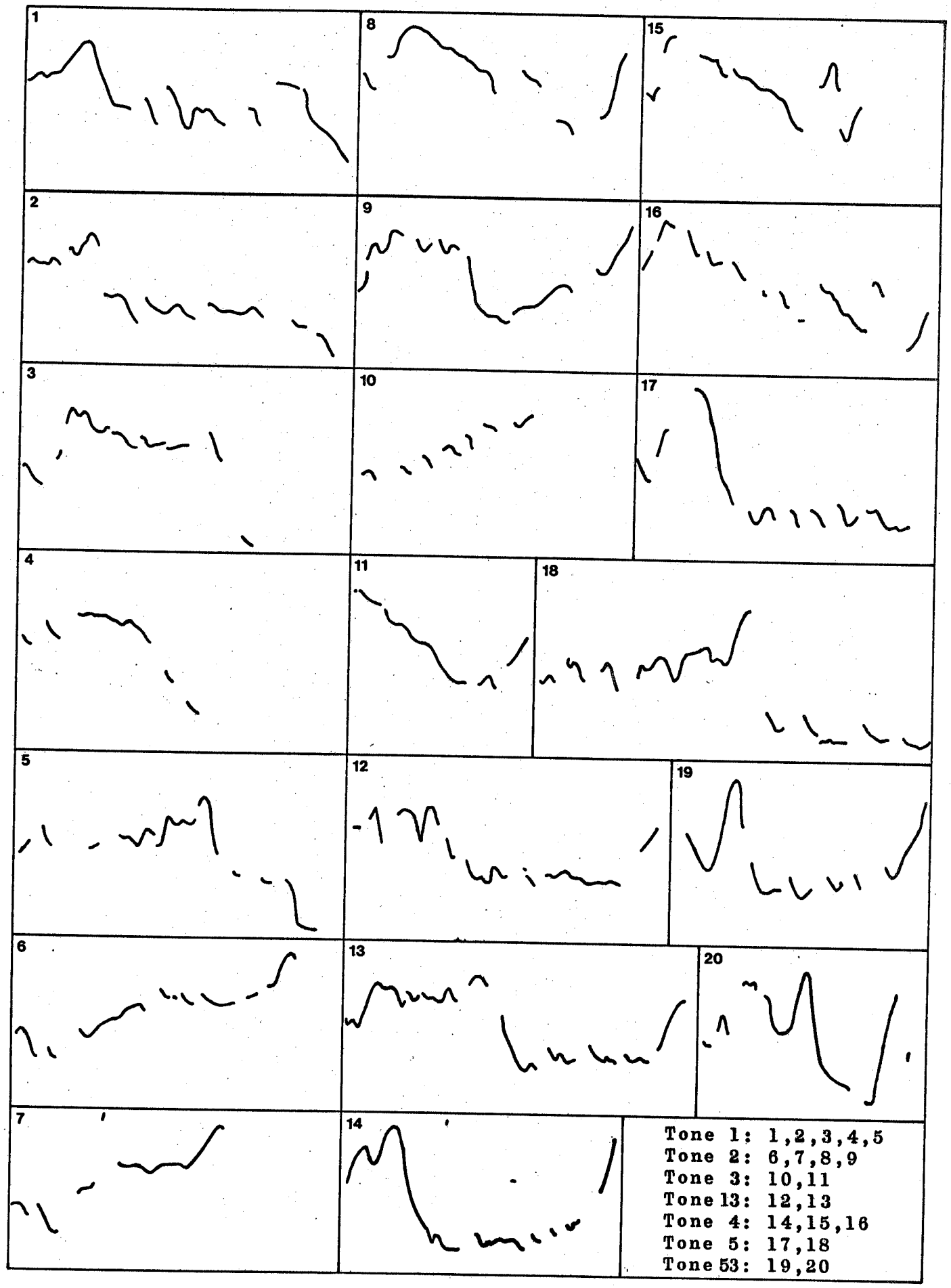
Halliday, M.A.K. (1970) A course in spoken English : intonation. Oxford University Press, London

't Hart, J. et Cohen, A. (1973). Intonation by rule : a perceptual quest. Journ. of Phon. 1, 310-327.

't Hart, J. et Collier, R. (1975) Integrating different levels of intonation analysis. Journ. of Phon. 3, 235-255.

Johnson, S.C. (1967) Hierarchical clustering schemes. Psychometrika 32. 241-254.

Fig.1 La fréquence fondamentale des vingt stimuli et leur catégorisation en 'Tones' selon Halliday (1970).



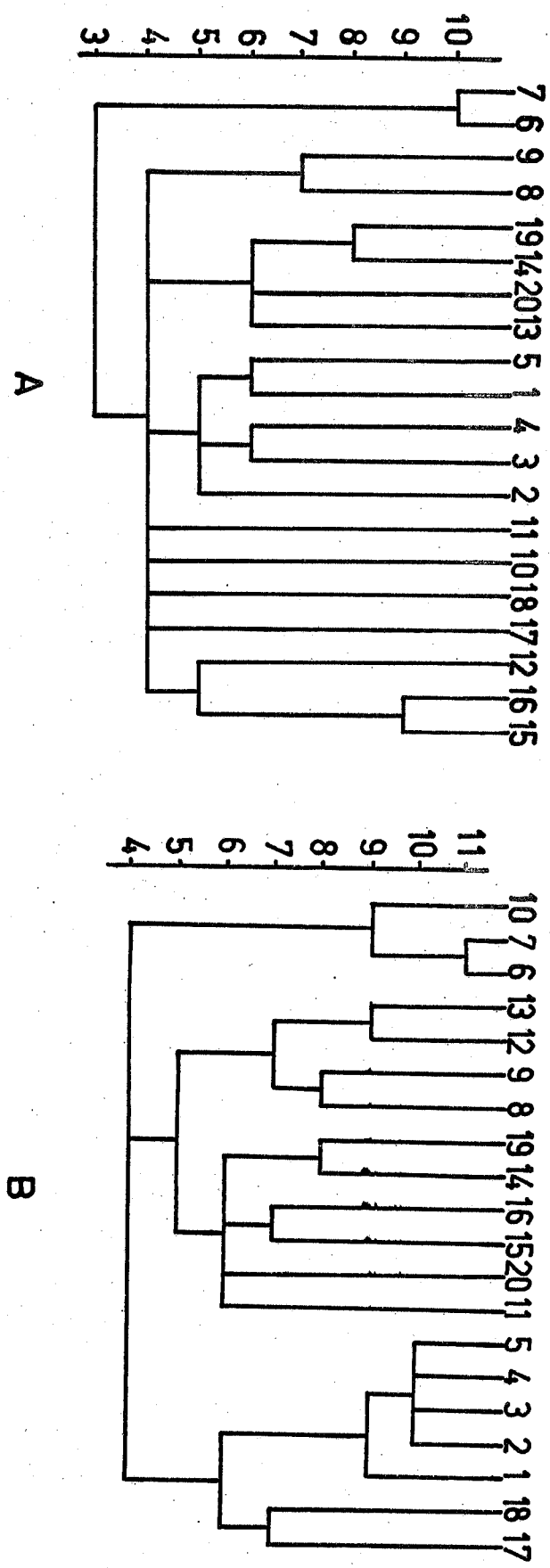


Fig. 2
Résultats de la "méthode maximale" de l'analyse hiérarchique des groupements, selon Johnson (1967) ;
(A) sujets anglophones,
(B) sujets néerlandophones.

VIII èmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

LES FAITS MICROPROSODIQUES DU FRANCAIS : VOYELLES, CONSONNES, COARTICULATION.

A. DI CRISTO et M. CHAFCOULOFF
Institut de Phonétique d'AIX-EN-PROVENCE

• Résumé :

Nous avons entrepris d'analyser les effets microprosodiques qui sont liés, d'une part à la nature des différentes unités segmentales (voyelles et consonnes) et, d'autre part à la concaténation de ces unités dans l'acte de parole.

Les premiers résultats de cette recherche sont présentés dans cette communication. Ils concernent successivement : la (Fo) intrinsèque des voyelles et des consonnes, l'intensité intrinsèque des consonnes et l'influence des consonnes sur la (Fo) des voyelles subséquentes.

• Summary :

The purpose of this study was to investigate the microprosodic effects inherent to :

- the nature of vowels and consonants,
- the concatenation of those units in the speech chain.

Our preliminary results deal with

1) the intrinsic Fo of vowels and consonants,
2) the specific intensity of consonants,
3) the influence of consonants on the Fo of adjacent vowels.

LES FAITS MICROPROSODIQUES DU FRANCAIS : VOYELLES, CONSONNES, COARTICULATION.

A. DI CRISTO et M. CHAFCOULOFF

INTRODUCTION

Nous présentons, dans cet exposé, les premiers résultats d'un programme de recherche qui porte sur les faits microprosodiques du français et dont l'objet est l'analyse des variations paramétriques intrinsèques (Fo, intensité, durée) relatives à la nature des différentes unités segmentales et à leur concaténation dans la parole.

Ces recherches ont été entreprises dans un double but. Définir, d'une part, un ensemble de règles d'effacement, qui permettrait d'éliminer du continuum acoustique les variations prosodiques conditionnées et d'isoler, ainsi, celles qui dépendent uniquement des instructions linguistiques, c'est-à-dire, de la compétence du locuteur. Elaborer, d'autre part, une grammaire des faits microprosodiques, susceptible de contribuer à l'amélioration d'un modèle de synthèse par règles ou de reconnaissance automatique de la parole continue.

Nous nous limiterons à la présentation de résultats sur : la (Fo) intrinsèque des voyelles et des consonnes, l'intensité intrinsèque des consonnes, et l'influence des consonnes sur la (Fo) des voyelles adjacentes.

1. (Fo) intrinsèque des voyelles et des consonnes.

1.1. (Fo) intrinsèque des voyelles

1.1.1. METHODE

L'analyse a porté sur les voyelles /i-y-u-ɛ-ɔ-a/, incluses dans des monosyllabes du type C_1-V-C_2. [C_1] est variable [p-t-k-b-d-g-f-s-ʃ-v-z-ʒ-l-r-m-n/], tandis que [C_2]est représentée par une consonne non voisée (occlusive ou constrictive). Les monosyllabes sont intégrés dans des phrases dont l'agencement syntaxique est le suivant :

Dét. + Nom + Complément + Copule + Adjectif

ex. La patte du chat est cassée.

Les phrases ont été enregistrées dans une chambre sourde par 4 locuteurs et traitées ensuite au moyen d'un détecteur de mélodie connecté à un ordinateur T1600, à l'aide du programme élaboré par ROSSI et TESTON (1977) (1).

La figure 1 illustre les points FoC (i.e. : Fo Cible) que nous avons sélectionnés pour mesurer les valeurs de la (Fo) sur les configurations acoustiques des différentes voyelles. On remarquera que, à la différence de nos prédécesseurs (HOUSE and FAIRBANKS, 1953 ; LEHISTE AND PETERSON, 1961 ; MOHR, 1971 ; LEA, 1972 ; SAMARAS, 1972 ; LOFQVIST, 1975), nous n'avons pas effectué de moyennes, ni retenu les valeurs de la (Fo) au sommet de la configuration. Notre méthode de mesure présente l'avantage de permettre une interprétation perceptuelle immédiate des résultats, (ROSSI, 1971a), ce qui n'est pas le cas des autres méthodes.

Les valeurs mesurées ont été l'objet d'un traitement statistique : vérification de la normalité de la distribution des divers échantillons, calcul des moyennes et écarts-types, analyse de variance.

1.1.2. RESULTATS

Nous retrouvons en français la tendance observée dans d'autres langues, à savoir que les voyelles hautes (H) /i-y-u/ sont caractérisées par une (Fo) intrinsèque plus élevée que les voyelles basses (B) /ɛ-ɔ-a/. Cette différence, qui est significative à l'analyse de variance ($F_{(1,3)} = 20.03$, $p < .05$), se situe entre 6 et 10 %.

Dans des contextes identiques, les écarts entre la (Fo) intrinsèque des voyelles H et celle des voyelles B sont égaux ou supérieurs au seuil différentiel de fréquence fondamentale, tandis que les écarts observés, d'une part, entre les différentes voyelles H et, d'autre part, entre les diverses voyelles B sont inférieurs à ce seuil (figure 2).

Ces résultats sont confirmés par l'analyse de variance, qui montre que seules les différences interclasses (H ≠ B) sont significatives.

Le second facteur significatif à l'analyse de variance concerne le voisement (ou l'absence de voisement) des consonnes adjacentes à la voyelle ($F_{(2,6)} = 5.1$, $p < .05$).

Afin d'isoler les effets dus au voisement des consonnes adjacentes, nous avons constitué un second corpus, dans lequel nous avons fait varier symétriquement le lieu (labial, dental, palatal) et le mode (occlusif/constrictif, voisé/non voisé) des consonnes adjacentes à la voyelle /a/.

---

(1) Cf. la communication de ces auteurs, dans ce volume.

L'analyse de ce second corpus confirme celle du prémier : la (Fo) de /a/ est significativement plus élevée dans un environnement non voisé que dans un entourage voisé. La différence entre ces deux contextes est de l'ordre de 4 à 6 % (tableau 1).

Aucun des autres facteurs testés par l'analyse de variance (lieu consonantique ; caractère [+ continu] ou [- continu] des consonnes adjacentes) n'est significatif. On peut en conclure que seuls les effets relatifs à la classe vocalique (H ≠ B) et au voisement des consonnes conditionnent les variations intrinsèques de la (Fo) des voyelles.

Ce résultat nous a permis d'établir une première hiérarchie des effets engendrés par diverses combinaisons des deux facteurs significatifs :

1) $\frac{C\ [-\ voisée] + V\ [+\ haute]}{C\ [+\ voisée] + V\ [+\ basse]} =$ 15 %

2) $\frac{C\ [+\ voisée] + V\ [+\ haute]}{C\ [+\ voisée] + V\ [+\ basse]} =$ 6 à 10 %

3) $\frac{C\ [+\ voisée] + V\ [+\ haute]}{C\ [-\ voisée] + V\ [+\ haute]} =$ 4 à 6 %

4) $\frac{C\ [+\ voisée] + V\ [+\ haute]}{C\ [-\ voisée] + V\ [+\ basse]} =$ 1.5 à 3 %

Ce classement montre que les effets relatifs aux deux facteurs peuvent s'ajouter ou se retrancher. Il convient donc de tenir compte, simultanément, de la nature de la voyelle et du contexte consonantique, pour définir avec précision l'importance des variations intrinsèques qui affectent la voyelle.

1.2. (Fo) intrinsèque des consonnes

1.2.1. REMARQUES

Bien que de nombreuses études aient déjà traité des caractéristiques microprosodiques des consonnes, il était assez rare, jusqu'à présent, de disposer de données quantifiées sur ce sujet. Seuls les travaux récents de BOE (1973) et de LARREUR et BOE (1973) comblent en partie cette importante lacune.

On observe généralement que les consonnes voisées se caractérisent par une configuration concave. Les occlusives médianes sont affectées de deux décrochements (implosion et détente), tandis que les constrictives ont une configuration creusée, sans décrochement abrupt.

LARREUR et BOE (1973) constatent que, dans la parole continue, ces différences tendent à s'estomper au bénéfice d'un tracé uniforme, similaire à celui des constrictives.

TESTON Bernard
10, Butte des 3 Moulins
13 · AIX-EN-PROVENCE

1.2.2. METHODE

L'analyse a été effectuée à partir de deux corpus : un corpus de parole continue et un corpus de logatomes, des séquences : /badaga/, /vazaʒa/, /lamana/, etc...

Pour mesurer les variations de la (Fo), nous avons utilisé un programme de traitement par ordinateur, de sorte que tous les résultats ont été obtenus sous la forme de listings semblables à ceux présentés dans la figure 3.

1.2.3. RESULTATS

L'étude des configurations de la (Fo) des consonnes voisées nous fournit des résultats similaires à ceux de nos prédécesseurs. Les occlusives médianes se caractérisent généralement par la présence de deux décrochements bien marqués (d1 et d2, sur la figure 3) et par une évolution nulle, positive ou négative de la (Fo) au cours de la tenue. En position initiale, ces consonnes se réalisent le plus souvent avec une (Fo) ascendante ou plate.

Les constrictives voisées ont plutôt une forme concave, en position médiane et une configuration, soit uniformément ascendante, soit concave en position initiale.

Toutefois, le résultat le plus intéressant que nous avons obtenu concerne les réalisations des consonnes intervocaliques. Notre recherche montre, en effet, que l'évolution de la (Fo) pendant la tenue est fonction du contexte prosodique dans lequel s'insère la consonne.

Si la voyelle V2 qui succède à la consonne a une (Fo) plus élevée que la voyelle V1 qui la précède, la tenue est affectée d'une pente positive. Dans le cas inverse : (Fo) V1 > (Fo) V2, la pente est négative. Elle tend à être nulle dans le cas : (Fo) V1 = (Fo) V2.

Nous avions défini antérieurement (DI CRISTO, 1977) les valeurs moyennes du décrochement de la (Fo) (paramètre Δ F) au cours de la tenue des obstruantes [- continues] (occlusives) et [+ continues](constrictives).

Ces valeurs ont été converties en quarts de ton, puis comparées à celles de BOE (1973) (cf. tableau II).

Bien que nos valeurs soient généralement inférieures à celles de BOE (1973), nous remarquons que les tendances qui se dégagent des deux études sont identiques : Δ F est plus élevé, d'une part, pour les constrictives que pour les occlusives, et d'autre part, pour les consonnes initiales que pour les consonnes médianes.

En nous fondant sur nos résultats, nous avons établi une procédure permettant de générer à la synthèse la (Fo) des consonnes voisées médianes en fonction du contexte prosodique. Cette procédure est illustrée par les figures 4 et 5.

En ce qui concerne /l/, /m/ et /n/, nous avons constaté que les consonnes se caractérisaient par la présence qua-

si constante d'un décrochement d'implosion (d1), ce dernier pouvant être suivi - ou non - d'un second décrochage (d2) au moment de la détente.

La valeur de (d1) semble être plus importante pour /m/ et /n/ que pour /l/. Les valeurs moyennes calculées pour ces deux catégories de consonnes correspondent respectivement à 8 Hz et 4,5 Hz.

2. Intensité intrinsèque des consonnes

2.1. METHODE

Les mesures de l'intensité ont été faites sur les tracés obtenus à la sortie de l'amplificateur-détecteur de LANCIA. La constante de temps relativement fine de cet appareil permet une analyse temporelle fiable de ce paramètre.

L'évaluation de l'intensité intrinsèque des consonnes n'a pas donné lieu à des mesures aussi élaborées que celles qui ont été effectuées pour la (Fo). Nous nous sommes limités, en effet, au calcul du rapport moyen entre l'intensité de la consonne voisée, mesurée au niveau le plus bas de la tenue et celle des voyelles /a/ adjacentes. Ce rapport est exprimé par le paramètre Δ i.

2.2. RESULTATS

Tous les résultats concernant cette partie de notre étude sont consignés dans la figure 6. On remarquera que :

- à l'initiale, la valeur de Δi est de - 8 dB pour /l, m, n/ et de - 15 dB pour les obstruantes (occlusives et constrictives) ;

- en position intervocalique, la valeur de Δi est de - 5 dB pour /l, m, n/ et de - 9,5 dB pour les obstruantes.

Nous pensons que ces valeurs sont directement utilisables pour la synthèse paramétrique des consonnes françaises.

3. Influence des consonnes sur la (Fo) des voyelles adjacentes.

3.1. REMARQUES

Nous avons déjà observé, dans la première partie de cette étude, que les consonnes avaient pour effet, soit d'élever (non voisées), soit d'abaisser (voisées) la (Fo) globale d'une voyelle adjacente. Nous étudierons à présent la nature et l'importance des perturbations de la (Fo) initiale des voyelles provoquées par les consonnes prévocaliques.

Plusieurs études, effectuées dans diverses langues montrent que le mode voisé ou non voisé d'une consonne influe forte-

ment sur la valeur initiale de la (Fo) d'une voyelle subséquente (LEA, 1873 ; GANDOUR, 1974 ; HOMBERT, 1976).

Au contact d'une consonne non voisée, la (Fo) initiale de la voyelle a une valeur plus élevée qu'après une consonne voisée. Cette différence est plus marquée dans les langues où les consonnes non voisées sont aspirées. A notre connaissance, aucune recherche n'a encore été entreprise pour évaluer l'importance de ce phénomène en français.

3.2. METHODE

Nous avons analysé les effets des consonnes sur la (Fo) initiale des voyelles subséquentes dans trois contextes prosodiques différents.

a) dans des syllabes inaccentuées comprises dans des séquences de logatomes prononcées en recto-tono : /pataka/, /badaga/, etc...

b) dans des syllabes accentuées non finales de groupe prosodique, mais néanmoins détachées syntagmatiquement (cf. le corpus utilisé pour l'étude de la (Fo) intrinsèque des voyelles),

c) dans des syllabes terminales de groupes progrédients.

Une technique expérimentale, permettant d'identifier avec précision la première période de la voyelle et d'éliminer ainsi les valeurs erratiques, a été mise en oeuvre (cf. DI CRISTO,1976, 266-74). Cette précaution ayant été prise, nous avons retenu les valeurs de la (Fo) aux points suivants :

- Foi-NV-1 : (Fo) initiale correspondant à la première période de la voyelle précédée d'une consonne non voisée.

- Foi-NV-2 : (Fo) initiale correspondant à la deuxième période de la voyelle précédée d'une consonne non voisée.

- Foi-V : (Fo) initiale correspondant à la première période d'une voyelle précédée d'une consonne voisée.

3.3. RESULTATS

Deux cas sont à envisager : celui des voyelles réalisées avec un ton statique (cf. plus haut a) et b)) et celui des voyelles affectées d'un montant c).

I. Si la voyelle est précédée d'une consonne non voisée, on constate qu'à la rupture de la première période (Foi-NV1) succède une chute brève ou une courte montée dont l'amorce correspond à la seconde période (Foi-NV2) et le terme à la valeur cible (FoC) de la voyelle (fig. 7 et 8).

Lorsque la voyelle est précédée d'une consonne voisée, sa (Fo) est basse à l'initiale, puis s'élève régulièrement jusqu'à la valeur cible de la voyelle.

Si nous faisons abstraction du point Foi-NV1, qui ne peut être intégré perceptuellement, nous remarquons que la présence d'une consonne non voisée n'entraîne pas une perturbation importante de la (Fo) initiale d'une voyelle subséquente. L'écart entre Foi-NV 2 et Foi-C se situe entre 2 et 6 %. Ce qui revient à dire que Foi-NV 2 et Fo C sont très proches.

En revanche, l'effet d'une consonne voisée est nettement plus marqué. La différence entre Foi-V et FoC est de l'ordre de 7 % (V.B.) à 12 % (V.H.). Mais cette différence peut atteindre des valeurs plus élevées (14 % à 28 %) quand la voyelle est affectée d'un détachement syntagmatique important (fig. 9, loc. 3 et 4). De plus, la durée moyenne de la perturbation engendrée par une consonne voisée est de 70 m/sec., alors qu'elle est seulement de 25m/sec. pour une consonne non voisée.

II. En ce qui concerne les tons montants (fig.10), la différence moyenne entre FOI-V et FOI-NV est d'environ 20 % ; mais elle peut atteindre 40 %, lorsqu'il s'agit de fortes montées.

Il se dégage avec évidence de notre travail que si les consonnes voisées affectent assez considérablement la(Fo)initiale d'une voyelle subséquente (quelle que soit la nature du ton dont elle est porteuse), il n'en va pas de même des consonnes non voisées qui n'entraînent que des modifications négligeables.

Ce dernier point nous conduit à affirmer que la voyelle précédée d'une consonne non voisée constitue une réalisation non marquée par les effets de la coarticulation prosodique. Cette réalisation peut donc être prise pour référence lors de l'interprétation fonctionnelle des tracés acoustiques.

BIBLIOGRAPHIE

BOE, L.J. (1973)
Etude de l'interaction source laryngienne-conduit vocal dans la détermination des caractéristiques intrinsèques des consonnes du français, Bulletin de l'Institut de Phonétique de Grenoble, 2 : 1-24.

DI CRISTO, A. (1977)
Contribution à l'étude de la structuration prosodique de la phrase en français moderne : les faits microprosodiques, thèse de 3ème cycle.

GANDOUR, J. (1974)
Consonant types and tones in Siamese, Working Papers in Phonetics, 27 : 92-117.

HOMBERT, J.M. (1976)
Consonant types, vowel height and tone in Yoruba, Working Papers in Phonetics, 33 : 40-54.

HOUSE, A.S. and FAIRBANKS, G. (1953)
The influence of consonant environment upon the secondary acoustical characteristics of vowels, J.A.S.A. 25 (1) : 105-135.

LARREUR, D. et BOE, L.J. (1973)
Les caractéristiques intrinsèques des consonnes voisées du français dans la parole continue, Bulletin de l'Institut de Phonétique de Grenoble, 2 : 25-29.

LEA, W.A. (1972)
Intonational cues to the constituent structure and phonemics of spoken english, Ph. D. Thesis, Purdue Univeristy, XIII : 213p.

LEA, W.A. (1973)
Segmental and suprasegmental influences on fundamental frequency contours, in : Consonant Types and Tone, HYMAN, L.M., ed., Southern Occasional Papers in Linguistics, 1, 17-70.

LEHISTE, I. and PETERSON, G.E. (1961)
Some basic considerations in the analysis of intonation, J.A.S.A., 33 (4) : 419-25.

LÖFQVIST, A. (1975)
Intrinsic and extrinsic Fo variations in swedish tonal accents, Phonetica, 31 : 228-47.

MOHR, B. (1971)
Intrinsic variations in the speech signal, Phonetica, 23 : 65-93.

ROSSI, M. (1971)
Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole, Phonetica, 23 : 1-33.

ROSSI, M. et CHAFCOULOFF, M. (1972)
Recherche sur le seuil différentiel de fréquence fondamentale dans la parole, Travaux de l'Institut de Phonétique d'AIX, 1 : 179-85.

SAMARAS, M. (1972)
Influence de l'entourage consonantique sur les variations de la fréquence laryngienne du grec moderne, Bulletin de l'Institut de Phonétique de Grenoble, 1 : 57-66.

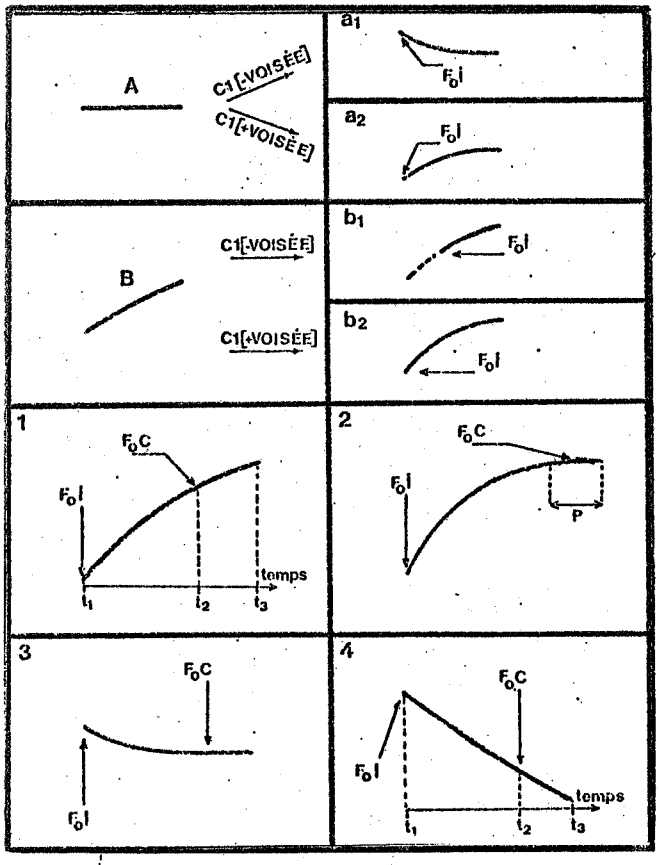


Fig.1. Influence de la consonne subséquente sur la Fo initiale (a_1, a_2, b_1, b_2).
Détermination des points (FoC)(1,2,3,4).

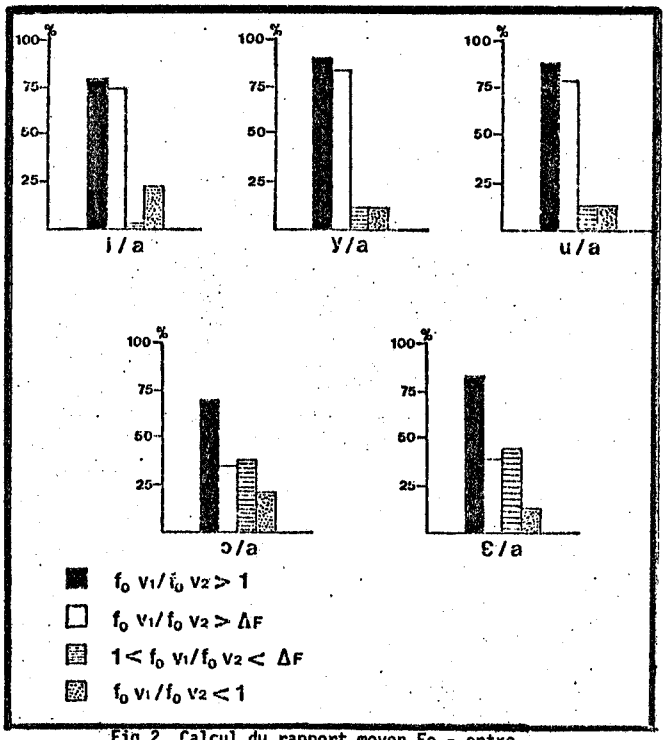


Fig.2 Calcul du rapport moyen Fo - entre
1) voyelles hautes et voyelles basses
2) voyelles basses entre elles. (Rapports exprimés par référence à Δf).

C_1-C_2	LOC 1		LOC 2		LOC 3		LOC 4	
p_ p	116	% 7	118	% 4	156	% 6	250	% 9
b_ b	108		113		146		229	
t_ t	115	3	116	3	152	4	249	6
d_ d	111		112		146		234	
k_ k	115	6	119	7	151	5	247	6
g_ g	108		111		143		232	
f_ f	114	5	119	2	160	4	253	5
v_ v	108		116		153		239	
s_ s	115	2	120	5	158	3	251	5
z_ z	112		114		152		238	
ʃ_ ʃ	114	4	121	3	156	2	253	5
ʒ_ ʒ	109		117		153		240	
	% M →	4		4		4		6

Tableau I. Variations de la Fo de /a/ en fct de C_1. C_2 (col. gauche). % d'augmentation de Fo en contexte non voisé (col. droite).

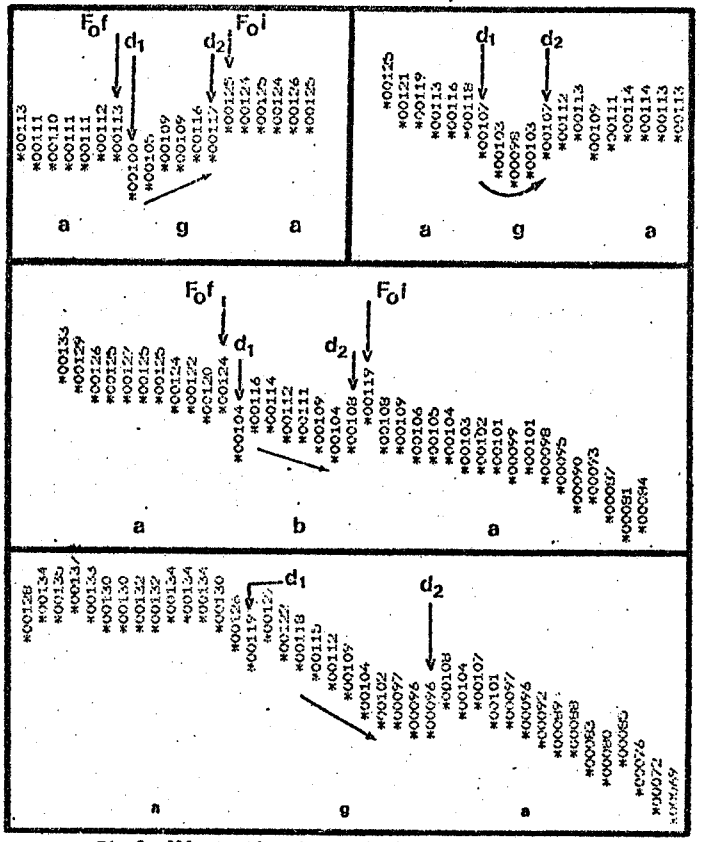


Fig.3. Illustration des variations de la Fo en contexte (consonnes voisées).
d_1 = décrochement d'implosion
d_2 = décrochement de détente.

	OBS -CONT		OBS +CONT	
	BOË	DI CRISTO	BOË	DI CRISTO
initiale	5	3.5	7.5	4.5
médiane	2.9	3.2	4.8	3.7

Tableau II. Comparaison des résultats (Valeurs de Δf exprimées en quarts de ton).

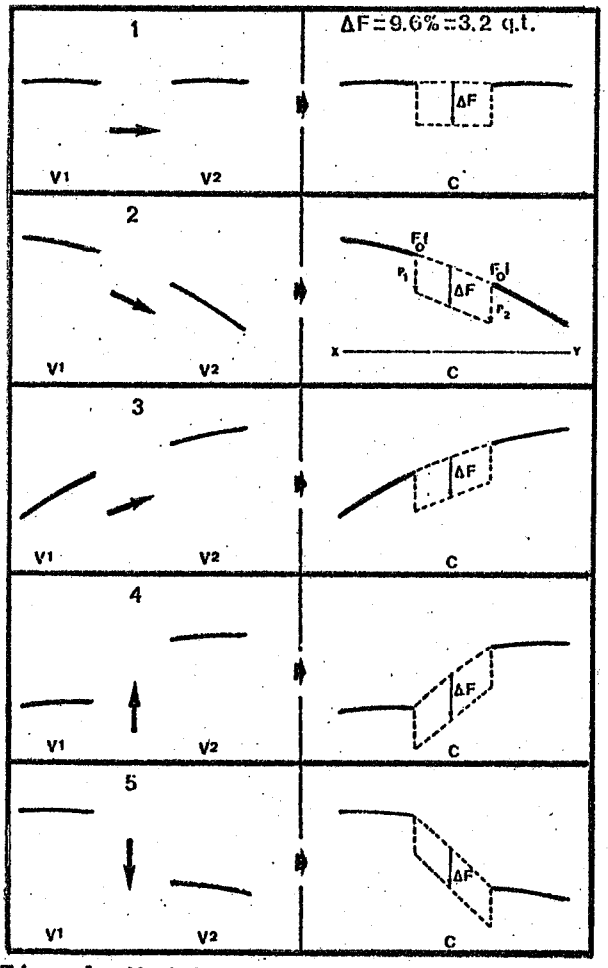


Fig. 4 Modèle de configurations de la Fo applicables à la synthèse par règles (occl. voisées intervoc.).

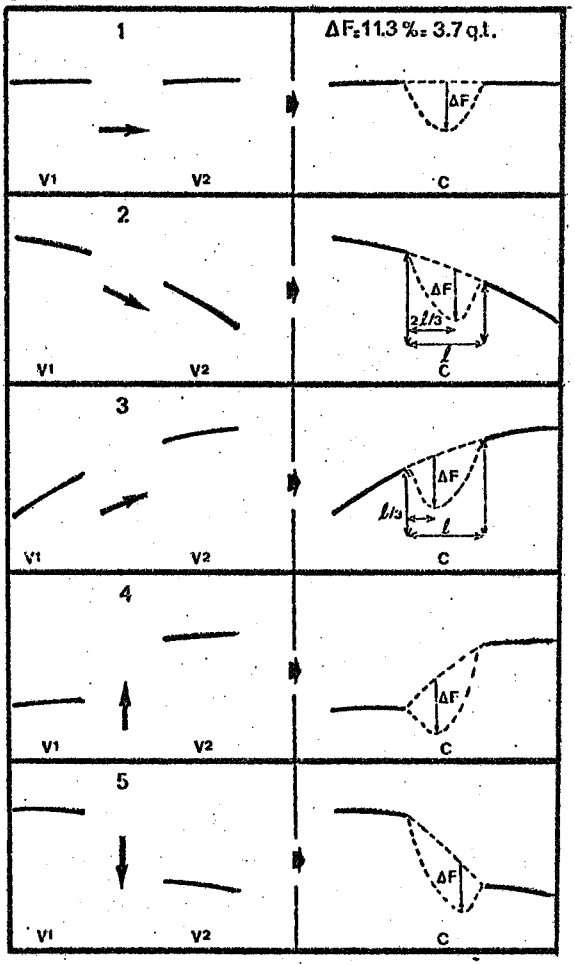


Fig. 5 Modèle de configurations de la Fo applicables à la synthèse par règles (constr. voisées intervoc.).

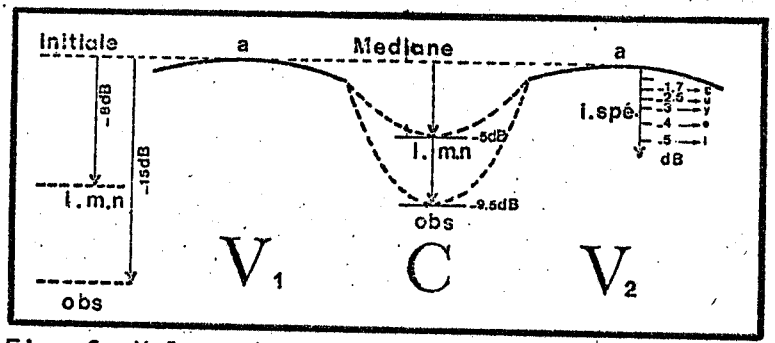


Fig. 6 Valeur des variations de l'intensité des consonnes applicables à la synthèse par règles (position initiale et médiane).

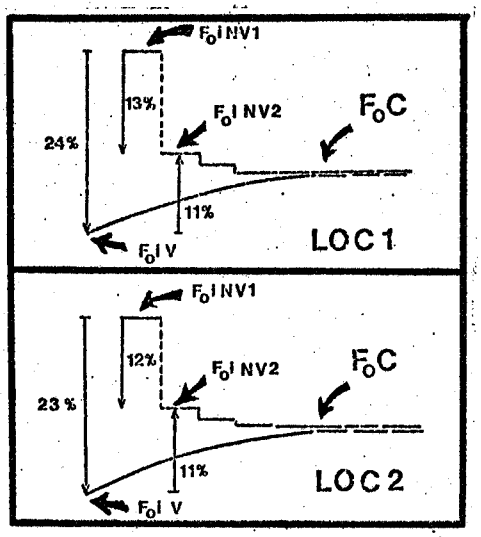


Fig. 7 Différences (en %) entre Fo I-NV_1, Fo I-NV_2, Fo I-V (voy. statiques non proéminentes).

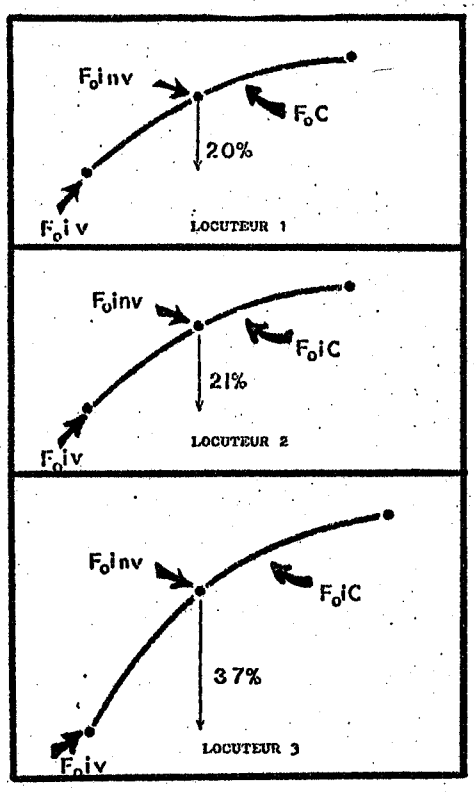


Fig. 9 % du rapport FoI-NV/FoI-V (ton montant).

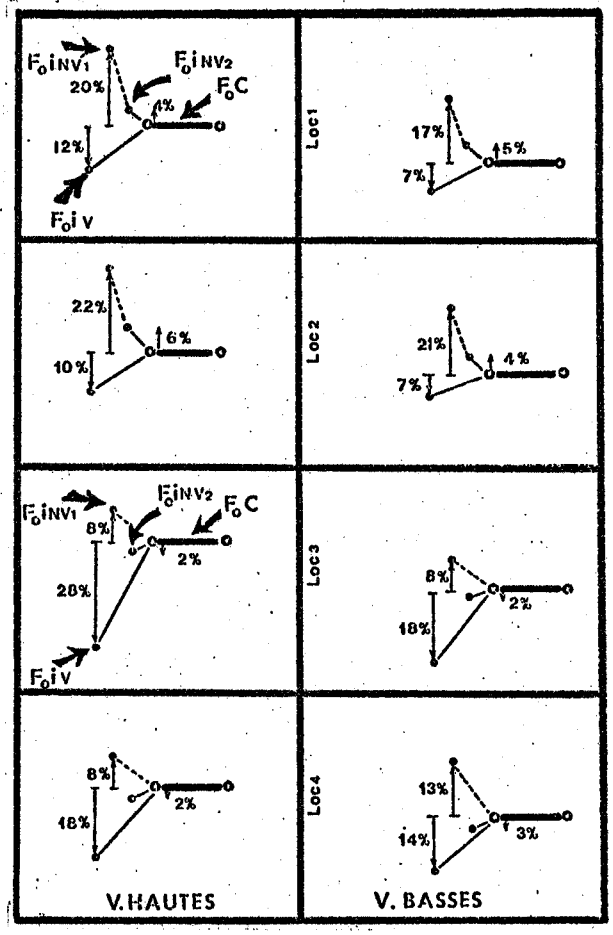


Fig. 8 Différences (en %) entre FoI-NV_1, FoI-NV_2, FoI-$_V$. (voy. statiques proéminentes).

VIII èmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

LE ROLE DE CERTAINS TRAITS PROSODIQUES DANS LA PERCEPTION DU MESSAGE +

(d'après un corpus de langage enfantin)

Gabrielle KONOPCZYNSKI
Maître-Assistant

Laboratoire de Phonétique - BESANCON

+ Texte intégral dans BULAG n° 5 - 1977 - BESANCON

• Résumé :

Une étude instrumentale et perceptuelle d'énoncés privilégiés, ne contenant que des informations suprasegmentales, donne des renseignements sur la perception de certains traits prosodiques.

1. Il y a réciprocité perceptuelle entre le contour global d'un énoncé et le contour de la syllabe finale seule, mais la détermination de la valeur sémantique altère souvent l'audition du contour final.
2. Les cas d'ambiguité perceptuelle semblent dûs à trois facteurs principaux : brièveté d'un énoncé, manque de netteté des paramètres et désaccord entre eux. On en dégage quelques principes à appliquer pour la synthèse ou la transmission des éléments prosodiques.

• Summary :

An instrumental and perceptual study of utterances giving only suprasegmental information provided us with findings concerning the perception of some prosodic features.

1. There is a perceptual reciprocity between the general contour of an utterance and its final contour, but the semantic decision often alters the final auditory shape.
2. Perceptual ambiguity seems to be due to three main factors : shortness of the utterance, lack of clearness of the specific parameters and discrepancies between them. Some principles for the synthesis or the transmission of prosodic features are discussed.

Le rôle de certains traits prosodiques dans la perception du message (d'après un corpus de langage enfantin)

G. KONOPCZYNSKI (Laboratoire de Phonétique - BESANCON)

Une analyse d'énoncés d'enfants en bas-âge (7-22 mois) peut donner des renseignements sur la perception de certains traits prosodiques. En effet, les énoncés étudiés présentent la caractéristique intéressante de ne contenir que des informations suprasegmentales : les enfants enregistrés, au nombre d'une dizaine environ, réalisent de manière très floue le niveau segmental ; la couche verbale, où l'on repère uniquement, sur des timbres vocaliques imprécis, des éléments mélodiques, des pauses et des unités accentuelles, est incompréhensible pour tout auditeur ne connaissant pas bien l'idiolecte de l'enfant et la situation dans laquelle est émis le discours. Contrairement aux discours généralement étudiés, notre corpus ne nécessite donc aucun traitement pour le dépouiller artificiellement de son contenu sémantique ; P. LEON (p. 120) signale à juste titre le danger des diverses techniques manipulant le continuum sonore car "on ne sait jamais en réalité si on ne supprime pas une partie importante de l'information ...".

1. LE CORPUS

1.1. Analyse instrumentale

Notre corpus est constitué d'une heure d'enregistrement environ, d'où ont été extraits 150 énoncés, un énoncé étant défini d'après Z.S. HARRIS comme une émission de voix entre deux pauses ; sa longueur peut varier dans d'importantes proportions : depuis l'énoncé minimal d'une syllabe, en passant par les énoncés de deux syllabes (les plus fréquents dans notre corpus, formant des "signes-phrases"), jusqu'à des énoncés de 17 syllabes d'une durée de 6 secondes. Ces énoncés, soumis à analyse instrumentale (oscillogramme, courbe d'intensité en dB, courbe mélodique en Hz) ont permis de dégager divers invariants (KONOPCZYNSKI, 1975a, 1976). Dans le présent travail seuls les paramètres de fréquence et de durée seront pris en considération, et non l'intensité. Par ailleurs, ne sont retenus que deux types d'énoncés :

- ceux qui, appréhendés par des adultes comme une modalité interrogative (abrégé en I), présentent les caractéristiques suivantes :
  - contour ascendant tant en ce qui concerne l'énoncé entier que la voyelle finale seule ;
  - registre mélodique plus élevé que celui de la modalité énonciative (gamme tempérée 3-4) ;
  - écarts mélodiques importants : - sur la voyelle finale ($\simeq$ 12, 12 quarts de tons)
    - sur l'énoncé entier ($\simeq$ 22 quarts de tons à une octave, et plus)
  - durée de la voyelle finale ($\simeq$ 19,65 cs)
  - pente P : abrupte ; nous donnerons dans cet article les chiffres, non pas de la pente P $\left\{\frac{\text{écart mélodique en Hz}}{\text{durée}}\right\}$, mais ceux de la rupture mélodique R, c'est-à-dire la même variation que P, mais exprimée en tons musicaux qui correspond mieux à l'effet auditif

$$\text{donc } R = \frac{\text{écart mélodique en 1/4 de tons}}{\text{durée}}$$

(exprimée en %, soit n 1/4 tons pour 100 cs)
R = (+) 61,66 % (moyenne)

- ceux qui, appréhendés par des adultes comme une modalité énonciative ou déclarative (abrégé en D), présentent les caractéristiques suivantes :
  - contour descendant, présentant différentes variantes : soit chutes basses (¯\), soit basses statiques (¯¯), soit chutes hautes (¯\), soit contour plat (——) ;
  - registre mélodique moyen : gamme tempérée 2-3 ;
  - écarts mélodiques réduits : sur la voyelle finale : 8,3 quarts de ton ; sur l'énoncé entier : 14 quarts de ton ;
  - durée de la voyelle finale : plus longue que dans les énoncés de type I (≃ 22,42 cs) ;
  - rupture mélodique R = (-) 35 %.

Dans ces énoncés - que l'on considère l'énoncé entier ou la voyelle finale seule - une seule direction de contour relativement stable est attestée.

1.2. Analyse auditive.

1.2.1. Dans un premier temps, tout le corpus a été soumis, pour dégrossissage, à un groupe d'adultes naïfs en linguistique, auxquels on ne posait aucune question précise. Seul un commentaire de ce qu'ils entendaient leur était demandé ; ils ont, dans l'ensemble, signalé que les énoncés semblaient être, soit de modalité interrogative, soit de modalité énonciative, soit d'une modalité à définir, mais marquée affectivement.

1.2.2. Dans un second temps, 65 des 150 énoncés analysés instrumentalement ont servi à l'élaboration de deux tests perceptuels destinés à vérifier les impressions du groupe d'adultes sus-cité, à mettre en corrélation les données instrumentales avec les données perceptives et à évaluer si l'auditeur s'appuie sur des paramètres préférentiels. Le corpus des deux tests est le même : il contient, outre des segments choisis de façon aléatoire, plusieurs énoncés identiques au niveau segmental, mais formant des paires minimales du point de vue intonatif, qui sont présentées, soit successivement, soit séparément.

1.2.2.1. Test linguistique. (T.L.)

On demande aux auditeurs de porter un jugement sur la valeur sémantique des énoncés, qu'ils doivent classer en interrogatives (I) ou non-interrogatives (D) ; il s'agit donc d'une tâche purement linguistique.

Les auditeurs : 36 étudiants bien entraînés en phonétique, ayant tous une excellente connaissance du français, quoique d'origine linguistique diverse.

1.2.2.2. Test psycho-physique. (T.P.P.)

On demande aux auditeurs d'indiquer la forme du contour de la dernière syllabe des énoncés. Nous fondant sur nos données expérimentales ainsi que sur des expériences perceptuelles de même type réalisées par HADDING (1973), trois formes de contour sont proposées : syllabe finale ascendante (↗), descendante (↘) ou plate (—). Le jugement demandé, fondé sur les caractéristiques physiques d'un énoncé, et plus précisément sur sa partie finale, peut être influencé, soit par le contour général perçu, soit par une interprétation linguistique inconsciente, et inversement, le jugement linguistique demandé en T.L. peut être contrôlé, autant par le contour global de l'énoncé que par celui de la seule syllabe finale, que par d'autres paramètres à définir.

Les auditeurs : 20 étudiants, différents des 36 du T.L., mais répondant aux mêmes caractéristiques.

1.2.3. Résultats comparés des deux tests (T.L. et T.P.P.)

Les résultats présentent de remarquables similitudes : que la tâche soit d'ordre linguistique ou psycho-physique :

1.2.3.1. les jugements se correspondent : T.L. = T.P.P. dans la plupart des cas.
1.2.3.2. à une caractéristique physique donnée correspond généralement un jugement linguistique donné :
- courbes physiquement ↗, entendues ↗, appréhendées comme I.
- courbes physiquement ↘, entendues ↘ ou —, appréhendées comme D.
- courbes physiquement —, entendues — ou ↘, appréhendées comme D.

1.2.3.3. que ce soit dans T.L. ou T.P.P., les auditeurs semblent influencés de façon quasi similaire par le contour global de l'énoncé.
1.2.3.4. cependant, dans T.L. les énoncés ambigus semblent perçus sur la base de divers critères autres que le contour, alors que dans T.P.P. les cas ambigus de T.L. perdent leur ambiguïté et la forme du contour final est perçue nettement, ce qui tend à montrer que le jugement psycho-physique est moins conditionné par la globalité d'un énoncé que le jugement linguistique.
1.2.3.4.5. Il en découle que les cas ambigus dans T.P.P. sont moins fréquents que dans T.L.

1.2.4. Résultats d'ensemble des deux tests.

Etant donné la similarité des résultats des deux tests, nous les regroupons. Les auditeurs émettent des jugements qui concordent (de 75 à 100 %) dans 87 % des cas environ. La plupart des énoncés sont donc appréhendés de façon univoque, malgré le manque d'informations segmentales, ce qui tend à confirmer la valeur linguistique des structures prosodiques en général et de celles du jeune enfant en particulier.

Cependant, 13 % des énoncés sont ambigus, soit parce que les auditeurs les classent dans une catégorie donnée avec un pourcentage inférieur à 75 %, soit qu'ils portent des jugements en totale contradiction. Cette discordance de jugement en T.L. est due, dans certains cas, à l'interférence des systèmes linguistiques, de nombreux auditeurs n'ayant pas le français comme langue maternelle (KONOPCZYNSKI, 1975b). Mais elle a également d'autres causes, que nous avons essayé de dégager sur la base de l'analyse instrumentale et de T.P.P.

2. CAUSES DE MAUVAISE PERCEPTION D'UN ENONCE

Une étude précise des analyses acoustiques permet de distinguer trois causes essentielles à cette discordance de jugement.

2.1. Brièveté de l'énoncé

Les énoncés courts d'une ou de deux syllabes sont classés dans une modalité donnée par environ 70 % des auditeurs, ce qui n'est pas le cas des énoncés plus longs où l'accord est plus unanime. Dans T.P.P. les étudiants signalent leur difficulté à repérer le contour de ces énoncés. Il semblerait donc que le manque de temps de reconnaissance et de points de référence empêche d'évaluer la forme de la courbe mélodique, et donc d'appréhender le sens général d'un énoncé.

Nous avons constaté un phénomène analogue dans des épreuves de type inverse, où il est demandé à des étudiants très entraînés, non plus de reconnaître divers types de courbes, mais de les produire (1). Les performances, très

(1) Il s'agissait, pour des étudiants anglophones, très entraînés à ce type d'exercices, de produire, après écoutes multiples du modèle, les courbes que propose HALLIDAY dans Intonation and Grammar in British English (1967).

satisfaisantes sur des phrases d'une certaine longueur, baissent de qualité sur des énoncés monosyllabiques. Que ce soit pour reconnaître/produire la forme d'une courbe mélodique, la brièveté d'un énoncé entraînant une absence de points de comparaison entrave donc la perception.

2.2. Manque de netteté des divers paramètres

2.2.1. Divers énoncés présentent des paramètres se situant en-dessous de la moyenne générale de la modalité correspondante.
Ainsi,

	Enoncé 23	Enoncé 31	Moyenne
- Ecart V.F. (2)	- 4	- 2	- 8,3
- Rupture mélodique R	- 17 %	- 20 %	- 35 %
ou			
- Ecart E.E. (énoncé 48) = - 10			- 14

La modalité de ces énoncés ne peut être définie clairement par les auditeurs. Dans T.P.P. également, les courbes finales présentant une montée ou une descente peu nettes sont confondues entre elles. Ainsi,

	T.L. modalité perçue	T.P.P. courbe perçue	Forme physique	Moyenne
énoncé 54	D 82 %	45% ↘ (35% ↘, 10% →)	Ecart V.F. - 6	- 8,3
		50% ↗	Ecart E.E. - 8	- 14,2
			R - 19,6	- 35
énoncé 27	I 95 %	45% ↗		
		50% ↘ (20% ↘, 30% →)	Ecart V.F. 10	12,12
			R 36,3	61,66

(2) Abréviations : cf. en fin d'article.

Dans ces deux cas, le contour est mal perçu en raison d'une pente trop faible. E. ULDALL (1961 : 780) fait la même remarque, quoique travaillant sur un matériel totalement différent (parole synthétique. Elle pense que l'ambiguïté "is presumably due to doubt about the relation of the contour to the total range".

2.2.2. Le cas inverse, c'est-à-dire l'exagération de certains paramètres paraît être une entrave tout aussi grande à la compréhension du message. Même si la ligne générale des paramètres correspond à celle d'une modalité donnée, mais avec des chiffres supérieurs, l'énoncé est ambigu. L'éloignement des paramètres par rapport à la moyenne gêne donc le jugement linguistique. D. ROSTOLLAND, dans ses études sur la voix criée, aboutit aux mêmes constatations : loin de rendre le message plus intelligible, les exagérations de fréquence, d'intensité, de durée introduisent des distorsions préjudiciables à la bonne transmission du message.

2.3. Manque d'accord entre les paramètres

L'énoncé 49, interprété D à 88 %, présente un contour d'E.E. descendant et un contour V.F. réalisé ascendant, mais perçu comme plat ou descendant par 95 % des auditeurs. Exemple net où le jugement linguistique altère la perception psycho-physique. Etant donné que les auditeurs n'ont pu interpréter l'énoncé comme déclaratif parce qu'ils ont entendu une descente finale, on est

en droit de conclure qu'ils ont cru entendre une descente finale en raison de leur interprétation linguistique faite sur la base d'autres critères.

D'autres énoncés (31, 32) avec contours et écarts de modalité D., mais durée de V.F. très brève , caractéristique de la modalité I, ne sont classés qu'entre 68 % et 71 % dans la catégorie D. Certains énoncés, enfin, présentent un désaccord entre l'ensemble des traits de contour, de modalité D, et le registre élevé qui correspond à I. C'est le cas des interrogatives ambiguës à contour V.F. descendant, mais débutant dans la gamme 4. La perception correcte du contour final n'aide en rien le jugement linguistique; elle est, au contraire, source de doute. De même, dans 9a, le jugement porté par les Français et les étrangers varie dans des proportions inverses (FR. = 80 % I, ET. = 44 %). Si le comportement des étrangers est logique par rapport à leurs autres réactions (appui sur les paramètres de contour), les Français ne tiennent compte que du départ élevé de l'énoncé pour le classer en I. Notre hypothèse est confirmée par divers auteurs, entre autres O. METTAS et E. ULDALL qui va jusqu'à proposer 3 modalités : énonciative, interrogative et ambiguë, les auditeurs faisant, dans cette dernière, le choix en fonction de critères préférentiels : les uns s'appuient sur le registre, les autres sur la globalité de l'énoncé.

Cependant, lorsque la plupart des paramètres d'une modalité sont nets, le jugement n'est plus autant influencé par un élément discordant. Les énoncés 11, 26, 40b, 44b, 45a, qui présentent, outre les caractéristiques de leur modalité, une courbe V.F. de type ⌒ sont interprétés I à 100 %. Pourtant T.P.P. prouve que la descente finale est perçue par 5 à 15 % des auditeurs, qui, dans T.L., négligent cet élément en raison d'une décision linguistique prise à d'autres niveaux. (3) Ainsi, l'interprétation linguistique peut prédéterminer la forme perçue, comme le démontrent aussi STUDDERT-KENNEDY et HADDING (1971) : "decisions from higher levels feed back to correct ... tentative decisions at lower levels ans to construct final percept". Nous aboutissons aux mêmes types de conclusions que les auteurs sus-cités, à partir d'un matériau moins sophistiqué, moins contrôlable aussi, mais plus difficile à interpréter en raison du manque d'informations verbales.

D'autres causes d'ambiguïté, plus subtiles, sont probablement à envisager, mais seule une étude à l'aide de la parole synthétique permettrait éventuellement de les dégager.

Il reste clair que, dans des situations où, ni le contexte, ni les éléments paralinguistiques, ni la couche verbale n'apportent d'information, la prosodie peut jouer, à elle seule, un rôle important dans la transmission de la parole. Elle est suffisante pour véhiculer le message, à condition que les divers paramètres soient le plus près possible des chiffres moyens caractérisant une modalité donnée. Pour obtenir des énoncés naturels dont l'appréhension soit aisée, tant dans la synthèse de la parole que dans sa transmission, il importe donc d'éviter les écarts par rapport à cette moyenne, d'éliminer les éventuels désaccords entre paramètres, de fournir des énoncés où les points de comparaison soient en nombre suffisant, et probablement de concentrer ses efforts sur la forme globale des contours, qui semble prédéterminer la perception des paramètres physiques.

ABREVIATIONS UTILISEES

Ecarts V.F. = écarts de fréquence sur la voyelle finale, exprimés en 1/4 de tons musicaux.
Ecarts E.E. = écarts de fréquence sur l'énoncé entier, en 1/4 de tons musicaux.
Contour V.P./V.F. = Contour Voyelle Pénultième/Voyelle Finale.
Durée V.F. : Durée de la voyelle finale en cs.

(3) Pour exemples et plus amples détails, cf. KONOPCZYNSKI - à paraître.

B I B L I O G R A P H I E

HADDING K. STUDDERT-KENNEDY M. 1973 — Are you asking me, telling me or talking to yourself ? Working Papers 7, 111-125, Lund.

KONOPCZYNSKI G. 1975a — Contribution à l'étude des structures prosodiques chez les enfants de 1 à 2 ans. Communication faite au VIIIe Congrès International des Sciences Phonétiques, Leeds, 1975. A paraître.

KONOPCZYNSKI G. 1975b — Perception (par des Français et des étrangers) des structures intonatives d'enfants français. Analyse d'un test perceptuel. BULAG n° 3, 101-126. Université de Besançon.

KONOPCZYNSKI G. 1976 — Etude expérimentale de quelques structures prosodiques employées par les enfants français entre 7 et 22 mois ; Travaux de l'Institut de Phonétique de Strasbourg, Vol. 7, 1976, 171-205.

KONOPCZYNSKI G. — A paraître. Interprétation linguistique, jugement psycho-physique dans la perception de contours intonatifs : interrelations.

LEON P. 1971 — "Où en sont les études sur l'intonation ?", Proceedings of the Seventh International Congress of Phonetic Sciences, Montréal, Mouton, 1972, 113-135.

METTAS O. 1971 — Les techniques de la phonétique instrumentale et l'intonation. Presses Universitaires de Bruxelles.

ROSTOLLAND D. PARANT C. 1973 — Distorsion and intelligibility of shouted voice. Symposium Intelligibilité de la parole. Liège.

ROSTOLLAND D. PARANT C. 1974 — Physical analysis of shouted voice. Eight Inter. Congr. on Acoustics. Londres.

STUDDERT-KENNEDY M. HADDING K. 1971a — Auditory and linguistic processes in the perception of intonation contours. Working Papers n° 5, 1-26. Lund.

STUDDERT-KENNEDY M. HADDING K. 1971b — Further experimental studies of f_o contours. Proc. Seventh Inter. Cong. Phonetic Sci, Montréal, Mouton, 1972, 1024-1031.

ULDALL E. 1961 — Ambiguity : question or statement ? or "Are you asking me or telling me ?" Proc. Fourth Int. Cong. Phonetic Sci, Helsinki, Mouton, 1962, 779-783.

VIIIèmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

VERS UNE BASE PSYCHOPHONETIQUE DE LA STYLISATION INTONATIVE

J. 't HART

INSTITUUT VOOR PERCEPTIE ONDERZOEK
EINDHOVEN PAYS BAS

• Résumé :

Nous avons fait des expériences nombreuses dans lesquelles il apparaissait possible de styliser des courbes de la fréquence fondamentale étonnamment fort, sans perdre "l'équivalence perceptuelle" avec les intonations originales.

Les limitations au pouvoir séparateur pour des fréquences changeantes rapidemment pourraient-elles expliquer cet effet?
Des considérations sur le seuil différentiel de l'étendue d'un mouvement de F_o, celui de sa position dans la syllabe, et celui de sa pente, en connexion avec le seuil de glissando montrent que les différences entre les courbes de F_o naturelles et leurs stylisations comme communément appliquées restent subliminales en grande partie.

• Summary :

How far are limitations in the perceivability of dynamic pitch phenomena responsible for the possibility of stylizing F_o-curves as has been fruitfully applied in the study of Dutch intonation?
Considerations on the differential threshold of the size of a pitch movement, that of its position in the syllable, and that of its slope, in relation to the glissando threshold, show that the majority of the differences between natural F_o-curves and their usual stylisations remain subliminal.

1. Introduction

Les régularités de l'intonation du néerlandais, que nous avons décrites précédemment ('t Hart et Cohen 1973, 't Hart et Collier 1975), sont à vrai dire des régularités plutôt dans des contours stylisés (à l'aide de l'Intonateur, l'appareil qui nous permet de remplacer l'intonation naturelle par un contour artificiel) que dans des traces de la fréquence fondamentale. Mais on a pu mettre en évidence une certaine équivalence perceptuelle entre les intonations naturelles et les contours stylisés, même si les derniers sont composés de mouvements auquels sont décernées des valeurs normalisées pour la durée et la pente. Nous avons trouvé qu'il suffit de distinguer douze mouvements intonatifs différents pour être capable de donner une transcription de tous les phénomènes qui sont pertinentes à la perception dans un corpus assez grand.

Cet inventaire de douze mouvements contient cinq types de montée, cinq types de chute, entre deux lignes de déclinaison, dont une est la ligne de base descendante graduellement pendant tout l'énoncé, et l'autre est parallèle à la première, à environ 5 semi-tons plus haut.

Le problème qui se pose est: comment se fait-il que les fluctuations diversifiées de l'intonation naturelle puissent être simplifiées à un tel degré, et que pourtant les contours stylisés soient jugés comme perceptuellements équivalents aux intonations naturelles? Dans cette communication, nous essaierons d'examiner à quel point les limitations dans la perception des fréquences changeantes peuvent expliquer ce phénomène. Nous traiterons successivement la différence juste discriminable de l'étendue des mouvements intonatifs en rapport avec la distance entre les lignes de déclinaison; la discriminabilité de la position d'un mouvement intonatif dans la syllabe; le seuil de glissando ou la pente minimale d'un mouvement discriminable d'un ton statique; le seuil différentiel de la pente.

2. Les variations de l'étendue des mouvements intonatifs

Dans notre interprétation de la mélodie de la parole nous partons de la ligne de base à laquelle se superposent les mouvements intonatifs. Les étendues de ces mouvements montrent des variations considérables. Mais il apparaît que l'oreille n'est pas très sensible à ces variations. Sur la base d'expériences sur le seuil différentiel de l'étendue de deux mouvements consécutifs ('t Hart 1974), nous pouvons estimer que les mouvements dont les étendues diffèrent de moins de 3 semi-tons (semit) peuvent être remplacés par des mouvements dont l'étendue est normalisée. Ainsi, les parties élevées de la courbe mélodique s'arrangent à une deuxième ligne de déclinaison parallèle à la ligne de base.

Nous pouvons classer les mouvements intonatifs en catégories en fonction de leur étendue: petite, autour de 2 semit; modérée, autour de 5 semit; grande, autour de 8 semit; géante, plus de 10 semit. La distance entre les lignes de déclinaison variera en échelons conformément à ces catégories.

3. La position des mouvements dans la syllabe

La sensibilité de l'oreille est si grande qu'elle permet de distinguer différentes positions d'un mouvement intonatif à l'intérieur d'une syllabe. Il y a trois expériences dont nous pouvons déduire les limitations de cette faculté. Collier (1970) demandait à ses sujets d'ajuster la position de mouvements montants pour donner proéminence optimale. En fonction du schéma intonatif auquel la montée était associée, les positions du sommet préférées étaient 38 ou 90 ms à partir du début de la voyelle. Dans une expérience dans laquelle des cartes d'un Language Master devaient être assorties et triées (Collier 1975) on a pu trouver trois catégories de position dans les syllabes où la position d'une montée avait été variée de façon continue. Hill et Reid (1976) ont

trouvé comme seuil différentiel un peu plus de 70 ms pour des positions arbitraires d'une montée dans un énoncé sans signification, mais moins de 50 ms aux endroits de transition du premier formant. Ils ont interprété ce résultat comme indiquant une perception catégorique.

Quoiqu'il n'y existe pas d'expériences comparables sur des chutes, nous croyons que, en général, il n'est pas possible de discriminer plus de trois positions d'un mouvement intonatif dans une syllabe. Notre inventaire des mouvements pertinents à la perception contient trois montées et trois chutes qui sont distinguées selon leur position: le mouvement est presque achevé au début de la voyelle, il commence un peu plus tard que ce début, ou il commence 50 ms avant la fin de la partie voisée de la syllabe.

4. La pente d'un mouvement intonatif

4.1 Le seuil de glissando

Un changement de fréquence n'est pas toujours perçu comme un ton glissant. Particulièrement, un changement pourrait se dérouler dans une durée tellement courte que même un changement assez rapide ne puisse pas être différencié d'un ton statique. D'autre part, pendant une durée plus longue, la fréquence pourrait changer si peu que le changement ne soit pas perçu.

Plusieurs expériences ont été faites sur la perceptibilité des changements de fréquence, en fonction de la durée. Les résultats de Sergeant et Harris (1962), Pollack (1968), Rossi (1971) et Klatt (1973) ne semblent pas comparables, parce qu'on a utilisé des sinusoïdes de 1500 ou 125 Hz, ou la voix naturelle ou synthétique à 125 Hz. Mais si l'on recalcule leurs données en semit par seconde au lieu de Hz par seconde, les résultats se prêtent à une présentation élégante sur une figure, quel que soit le domaine de fréquences dans lequel ils sont obtenus. Sur la Fig. 1, on a ajouté un point à 10 semit/s dans 100 ms, sur la base des données de Shower et Biddulph (1931), et quelques points obtenus par nous mêmes au cours d'une expérience préliminaire. Les points ne montrent pas une dispersion trop grande; les écarts des données de Pollack et de Rossi peuvent être dus au choix des sujets: ceux de Pollack par exemple, étaient étudiants en musique très entraînés, et Rossi (communication privée) a obtenu un seuil plus bas dans des expériences plus récentes.

La pente de la ligne de déclinaison varie entre 1 et 5 semit/s. Par conséquent, la déclinaison est inaudible pendant une syllabe, mais audible dès que la durée de l'énoncé dépasse environ une demi-seconde.

4.2 Le seuil différentiel de la pente

Pollack a essayé de mesurer ce seuil différentiel, malgré la difficulté essentielle qu'on ne le peut pas mesurer purement (Pollack 1968, p.58, Experiment IV). Il a utilisé des sinusoïdes avec une fréquence centrale de 707 Hz. Nous avons recalculé ses données en semit/s pour les pentes entre 1 et 64 semit/s et les durées entre 50 et 400 ms. Pour le rapport des pentes qui sont juste discriminables, on obtient des valeurs comprises entre 1,5 pour les pentes rapides de durée longue et 6 pour les pentes faibles et courtes.

Nous utilisons ces valeurs pour construire un diagramme dans lequel est indiquée, en partant du seuil de glissando (á la base des moyennes des valeurs de Fig. 1), chaque pente suivante qui est tout juste discriminable de la précédente. Ensuite, nous portons des régions autour de 2, autour de 5 et autour de 8 semit, les catégories de l'étendue des mouvements. Ainsi, nous pouvons voir combien de pentes sont discriminable dans chaque catégorie et

dans une durée donnée (Fig. 2). Il apparaît que dans la région des étendues modérées on ne peut pas discriminer plus d'une seule pente, sauf si la durée est assez longue. Seulement avec des étendues petites on pourrait discriminer plus de deux pentes (voir à 250 ms).

La seule expérience de ce type avec une voyelle synthétique, celle de Klatt (1973), donne un seuil différentiel un peu plus bas que la Fig. 2 le prédirait. Cette écart peut être dû au type de stimulus que Klatt a utilisé, dans lequel une différence de pente était alliée à une différence entre les fréquences initiales et finales.

Ce résultat nous fait comprendre pourquoi les petites chutes dans des syllabes non accentuées que l'on voit dans des enregistrements de la fréquence fondamentale peuvent être remplacées par une ligne de déclinaison dont la pente correspond seulement avec leur différence mutuelle en fréquence moyenne, mais qui est plus faible que celles des chutes individuelles.

En outre, il explique pourquoi notre inventaire de mouvements intonatifs, qui sont normalisés à des étendues modérées, n'a que besoin de deux pentes différentes.

5. Discussion et conclusion

Nous avons exécuté de nombreuses stylisations des courbes mélodiques de toutes sortes d'énoncés arbitraires en néerlandais. En essayant d'exprimer autant de variations caractéristiques que possible avec le moins de moyens que possible, nous constations à notre étonnement que seulement une douzaine de mouvements différents sont suffisants pour donner une approximation fidèle des traits les plus importants de l'intonation néerlandaise.

En y regardant de plus près, on constate que ce nombre relativement petit provient des limitations à la discriminabilité des phénomènes dynamiques de la fréquence. La différence juste discriminable de la pente est telle que les lignes de déclinaison donnent la même impression que les variations réelles. Il y a trois (ou exceptionellement quatre) catégories de distances entre la ligne de base et la ligne haute. Nous avons essayé de rendre plausible que les différents mouvements de l'inventaire qui couvrent entièrement cette distance correspondent précisément aux différentes façons discriminables auxquelles la fréquence peut aller de l'une à l'autre ligne de déclinaison. Seulement pour les étendues petites il y aurait la possibilité de discriminer plus de pentes.

Il ne reste que deux mouvements de l'inventaire qui ne couvrent pas la distance entière, mouvements qui sont appelés mi-grands, une demi-montée et une demi-chute. La première monte au-dessus de la ligne supérieure de déclinaison, la chute tombe d'en haut et s'arrête à mi-chemin. Donc, ces deux mouvements ne sont définis qu'à l'égard de la ligne haute, pas à l'égard de la ligne de base. Dans le dernier cas ils seraient interpretés comme mouvements entiers entre des lignes de déclinaison à demi-distance. C'est parce que les mouvements voisins déterminent la distance locale entre les lignes de déclinaison, que l'on interprète ces mouvements comme mi-grands. Il est clair que le seuil différentiel de l'étendue des mouvements intonatifs prévient une division d'une distance modérée plus fine qu'un partage en deux. Cela ne serait que possible pour les distances plus grandes.

Ces considérations ne donnent pas une justification complète du processus de la stylisation intonative, mais elles montrent que l'équivalence perceptuelle entre les courbes mélodiques naturelles et les contours stylisés est fondée essentiellement sur les limitations du pouvoir séparateur de l'oreille pour les fréquences changeantes. Cela impliquerait que la possibilité de décrire

les traits perceptuellement pertinents de l'intonation à l'aide de douze mouvements seulement n'est pas limité au néerlandais.

Bibliographie

Collier, R., 1970, The optimum position of prominence lending pitch rises, IPO Annual Progress Report 5, p.82-85.

Collier, R., 1975, Perceptual and linguistic tolerance in intonation, Int. Review of Applied Linguistics 1975/4, p.293-308.

't Hart, J. et Cohen, A., 1973, Intonation by rule: a perceptual quest, J. of Phonetics 1, p.309-327.

't Hart, J., 1974, Discriminability of the size of pitch movements in speech, IPO Annual Progress Report 9, p.56-63.

't Hart, J. et Collier, R., 1975, Integrating different levels of intonation analysis, J. of Phonetics 3, p.235-255.

Hill, D.R. et Reid, N.A., 1976, An experiment on the perception of intonational features, Research Report nr. 76/6/6, Dept. of Computer Science, The University of Calgary.

Klatt, D.H., 1973, Discrimination of fundamental frequency contours in synthetic speech: implications for models of pitch perception, J. Acous. Soc. Amer., 53, p 8-16.

Pollack, I., 1968, Detection of rate of change of auditory frequency, J. Exptl. Psychology 77/4, p.535-541.

Rossi, M., 1971, Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole, Phonetica 23, p.1-33.

Sergeant, R.L. et Harris, J.D., 1962, Sensitivity to unidirectional frequency modulation, J. Acoust.Soc.Amer. 34, p.1625-1628.

Shower, E.G. et Biddulph, R., 1931, Differential pitch sensitivity of the ear, J.Acoust.Soc.Amer. 3, p.275-287.

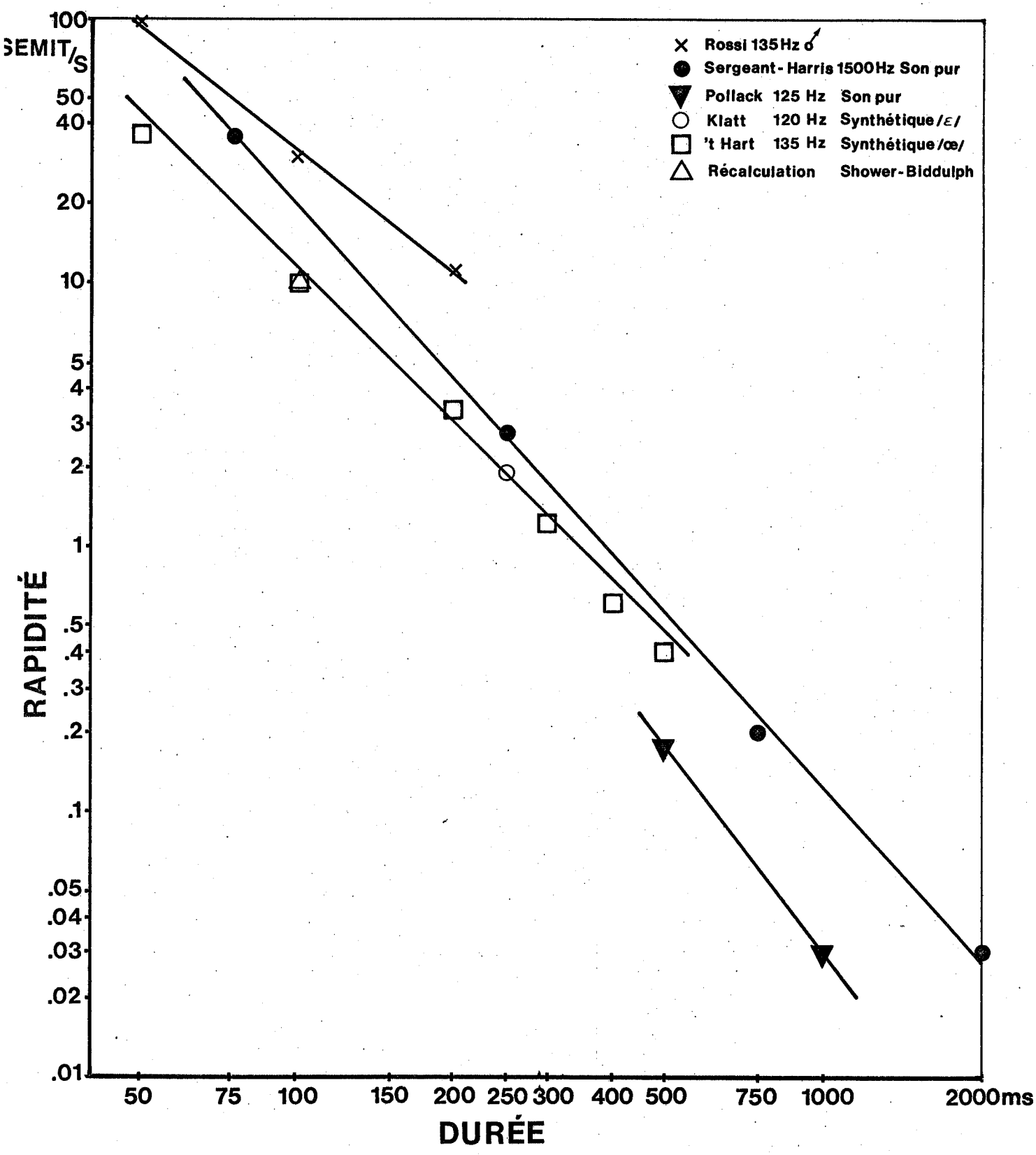


Fig. 1. Seuils de glissando trouvés par auteurs divers.
La rapidité nécessaire d'un mouvement d'une durée donnée
pour qu'on puisse distinguer ce mouvement d'un ton statique.

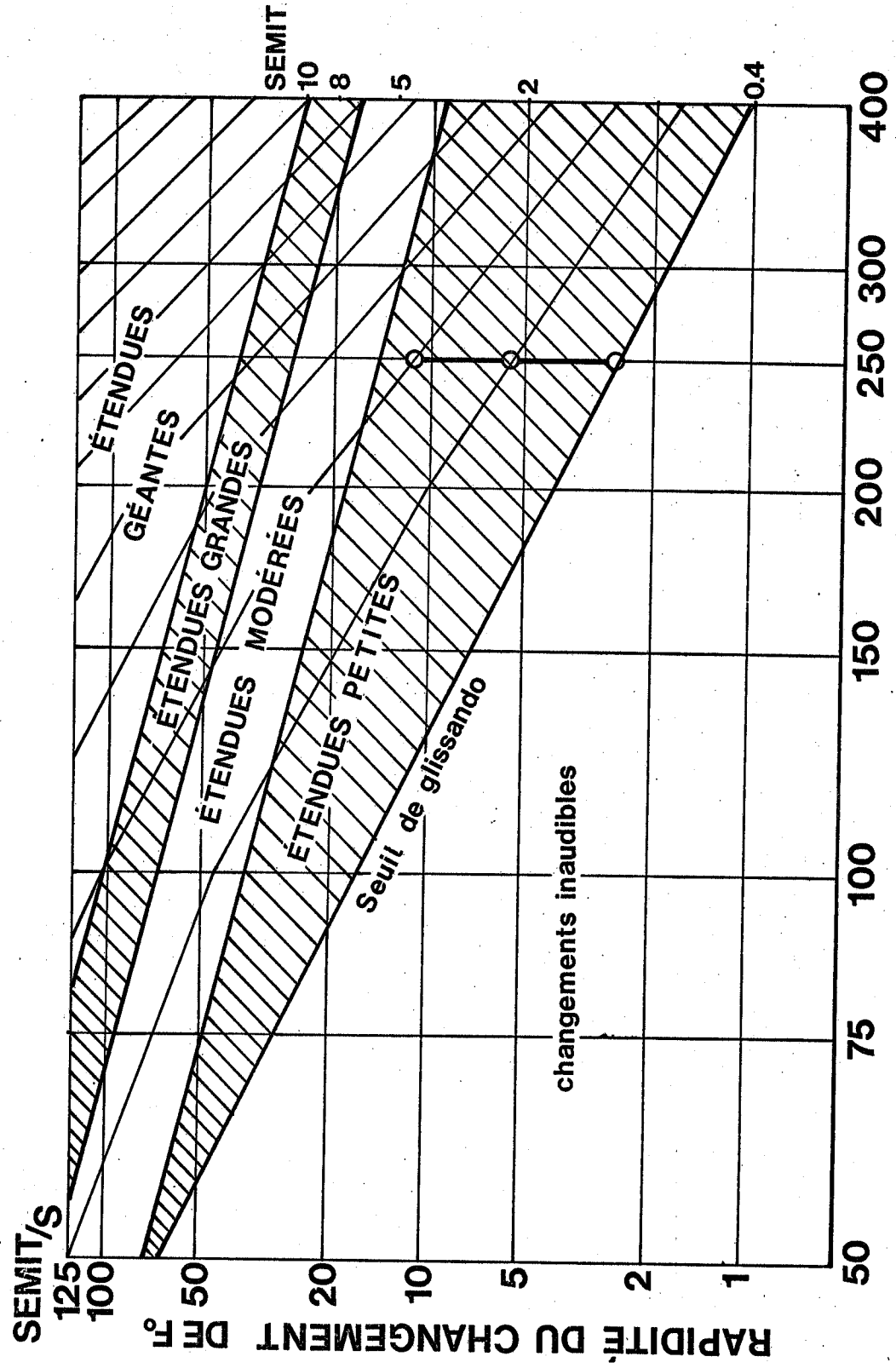


Fig. 2. Régions des pentes discriminables en fonction de la durée, dans des catégories différentes de l'étendue. L'exemple à 250 ms indique que pour les étendues petites il y a trois pentes discriminables: 3, 6 et 12 semit/s environ.

VIII èmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

ON THE INTERPRETATION OF RAW FUNDAMENTAL FREQUENCY TRACINGS

NINA THORSEN

INSTITUTE OF PHONETICS
UNIVERSITY OF COPENHAGEN

• Résumé :

Cet article décrit une procédure qui permet de décomposer les tracés compliqués du fondamental (Fo) en ses composants constituants, à savoir 1) le contour intonatif de la phrase, 2) les configurations de Fo liées aux groupes d'accent, 3) les mouvements du Fo appartenant aux segments individuels (voyelles), 4) les différences de Fo inhérentes aux segments individuels (voyelles) et 5) les variations coarticulatoires du Fo. Les problèmes inhérents a une telle procédure sont discutés aussi bien que les suppositions concernant les modèles physiologiques de la production du signal physique et ses corrélats perceptuels.

• Summary :

This paper describes a procedure with which complex raw fundamental frequency (Fo) tracings can be dissolved into their constituent components, namely 1) the intonation contour of the sentence, 2) the Fo patterns associated with stress groups, 3) Fo movements pertaining to individual segments (vowels), 4) intrinsic Fo differences between segments (vowels), and 5) coarticulatory Fo variations. The problems inherent in such a procedure are discussed as are assumptions of underlying physiological production models of and perceptual correlates to the physical signal.

On the interpretation of raw fundamental frequency tracings

A phonetic investigation of fundamental frequency (Fo) may concern phenomena at various levels, such as the over-all intonation contour(s) of sentences, the Fo patterns associated with stressed and unstressed syllables (in languages where stress and Fo are interrelated), intrinsic Fo differences between segments, Fo movements associated with individual segments, and Fo variations arising from the interaction between neighbouring segments (coarticulatory variations). The contributions from each one of these phenomena may vary between languages, between dialects, and between speakers but, to some extent, they all interfere with each other to make for the rather complex Fo courses we observe in the outputs of our analyzers, and thus, no matter what the focal point of interest may be, the effect of the others upon it should be accounted for, in order to arrive at a stringent and relevant description and also to create a certain simplicity in the frequency dimension.

We may attain simplicity in the time dimension by disregarding parts of the Fo course, which presupposes a reliable delimitation. Segmentation is also necessary to create line-up points for the averaging of several repetitions of a given utterance and is in itself a problem which, however, is not considered any further here.

The outcome of the processing of Fo tracings is often (tacitly) assumed to reflect either an underlying physiological production model or a perceptual correlate to the physical signal - but, clearly, either of these goals can only be approached in a round-about fashion which involves guess-work and inferences made from observations of similarities and dissimilarities among and between sounds in various environments.

I shall try to exemplify some of the problems inherent in separating, in raw Fo tracings, the contributions from each of the factors mentioned in the first paragraph, stressing the fact that they are relevant to "manual" as well as computerized treatment of data. The problems are the experimenter's and the answers to them can only be supplied by him.

In a recent investigation (Thorsen 1976) of Danish intonation (in non-emphatic, non-emotional speech) - which was also to illuminate the relationship between linguistic stress and Fo - it was found necessary, first of all, for the averaging of repetitions of the same utterance, to line up the traces to each new beginning of a stretch of voicing in the sentence. The time-axis is thus, of course, distorted.

Only the Fo course of vowels and (voiced) sonorants were included in the subsequent treatment, assuming that the Fo course in voiced obstruents is irrelevant for the perception of pitch patterns and -contours (associated with stress groups and sentences, respectively), the ultimate goal of the investigation being a description as close to the perceived pitch course as possible.

The assumption of the irrelevance of the course of Fo in voiced obstruents will have to be tested in perceptual experiments, but it seems justified in the light of a case like the one below (which is but one out of many): The word stavelserne ('the syllables' (the stress is on the first syllable)) may appear in any of the three shapes depicted in Fig. 1 (Note that Danish does not have an opposition between unvoiced and voiced alveolar fricative, but /s/ may be voiced between voiced sounds, especially in rapid speech.)

(a) a rise in the first two syllables, continuing through the [s̬], followed by a fall in the last two syllables,

(b) a rise in the first two syllables, a fall and a rise in the [s], and a fall in the last two syllables, which corresponds to the fall in (ă),

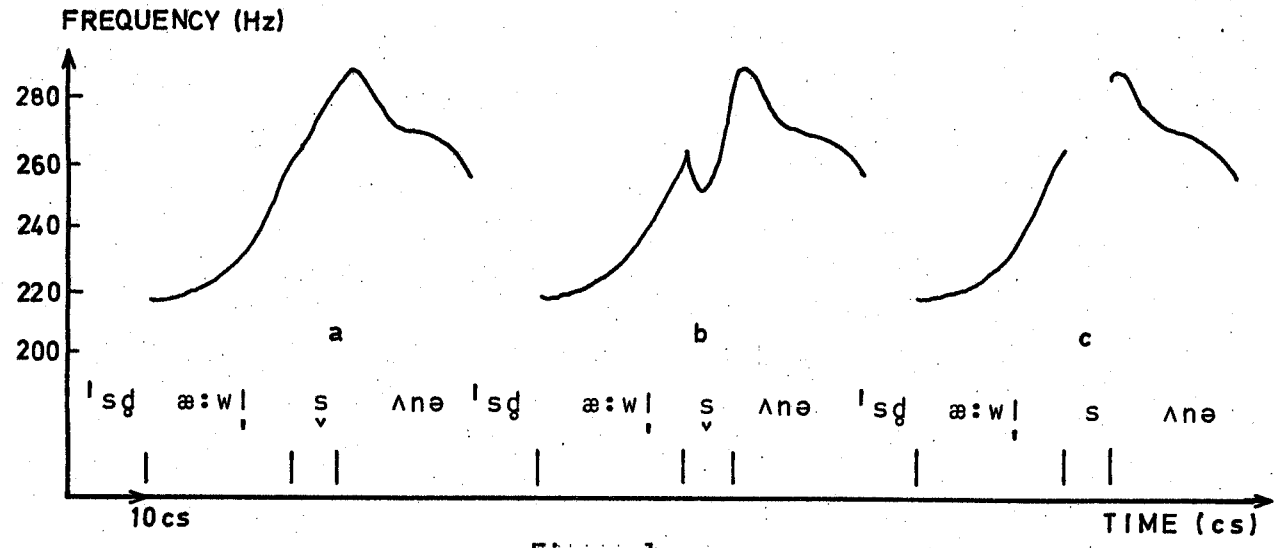


Figure 1

Fo tracings of three recordings of the word stavelserne (female subject).

(c) a rise in the first two syllables, 'silence' during the [s], a jump upwards to the fall in the last two syllables, which corresponds to (a) and (b).

The basic production model might look somewhat like (a) and is realised as such (in rapid speech) when the glottis remains "closed" and the vocal cords vibrate during the [s] and, probably, when at the same time the constriction at the alveolar ridge is fairly loose. It is modified as in (b) when the vocal cords vibrate around a somewhat more open position during the [s], a modification which is not voluntary but due to mechanico-acoustico-aerodynamical factors. Finally, the model may be realised as in (c) if the glottis is wide open during the [s], impeding vocal cord vibrations. Unless one's attention is specifically drawn towards the detection of voicing in an /s/, the three editions are likely to be perceived in the same fashion, i.e. there is as much information for the listener in (c) as in (a), and the Fo course in the (voiced) obstruent is thus irrelevant in the sense that it passes unnoticed - and (c) may thus be an approximation to the perceptual model.

In vowels one often finds, after an aspirated stop or an [s] or [f], i.e. unvoiced consonants produced with an open glottis and with forceful airflow, one or two vibrations which are between 10 and 40 Hz above the succeeding ones. Such vibrations were left out, cf. the line of argument above for /s/.

The pruning of the Fo course is, regrettably, not uncontroversial. How much of a movement in a vowel surrounded by (voiced and unvoiced) consonants can be ascribed to influence from these consonants? In other words, what does the intended Fo course - the production model - look like before it is superimposed by involuntary modifications? And, once again, what are the perceptual correlates (which are not necessarily identical to the production model)? To my knowledge, there is no sure way to compensate accurately for such coarticulatory Fo variations since, as yet, too little is known about the physiological and aero-dynamical mechanisms involved to allow us to quantify these variations. - But there are ways around this problem. One (which I chose to follow) is to disregard those portions of the Fo course where coarticulatory variations occur, i.e. mainly at the boundaries between sounds of the obstruent and sonorant classes, and prune away the obstruents plus bits of the adjoining sonorants (but how large bits is difficult to decide). This is probably permissible if the outcome of the processing is to reflect a perceptual correlate to the signal. Another

possibility is to interpolate an Fo course through all voiced obstruents so as to smoothly connect the courses in the adjoining sonorants and thus approach what may be the underlying production model.

The Fo patterns associated with stressed and unstressed syllables were attacked next - on the basis of a material of reference words of two and three syllables ([ɓʰiɓʰi(ɓʰi)]) where only the position of the main stress varied (They are nonsense words but they are all possible words in Danish.). These words were embedded, in various positions, in declarative sentences. By employing the same consonants and vowels in the reference words, observed differences in Fo courses must be due to different stress distributions. (It turned out that a stressed syllable and all succeeding unstressed syllables in the same non-compound sentence, irrespective of intervening morphological and syntactical boundaries, constitute the unit for an Fo pattern, which can then be described as a low stressed syllable followed by a tail of higher and falling unstressed syllables, cf. the dotted line in Fig. 2,d.) But the problem remains to demonstrate that words of different and varying segmental structures exhibit the same Fo patterns as do the reference words, since intrinsic and coarticulatory Fo variations interfere.

A tabulation of intrinsic Fo differences between segments (vowels) must be performed for each individual in an experiment since the magnitude of these differences varies. It may be as large as 35 Hz for females, and as small as 10 Hz for males, between [i:] and [ɑ:], cf. Reinholt Petersen (1976). But the magnitude of these differences seemingly also varies with degree of stress (being much larger in stressed syllables), with the short/long distinction (being larger in long vowels), and with position on the intonation contour (being larger at the top than at the bottom of the contour).

Once these intrinsic differences in various conditions have been established for each subject, it becomes possible, in sentences consisting solely of short vowels and unvoiced consonants, to simply move the Fo course in a given vowel up or down the frequency axis, as the case may be, - and only then does the similarity to the Fo patterns of the reference words become apparent. - But sentences consisting also of words of long vowels and of vowels surrounded by sonorants, which exhibit longer stretches of voicing (abbreviated "voiced words" in the following) have to be dealt with as well. It is possible to reduce the long continuous stretches of voicing in "voiced words" to a succession of shorter ones if we assume that identical tonal patterns shall recur in "voiced words" and reference words. (This assumption must of course be verified or falsified in perception experiments.) One can compare (under identical conditions) the Fo courses in "voiced words" and reference words with identical stress patterns after corrections for intrinsic Fo differences have been performed. Parts of the Fo courses in the "voiced words" are then concurrent with the Fo courses in the reference words, and if the remainder of the Fo courses in the "voiced words" is disregarded, "voiced words" are seen to follow the same Fo patterns as exhibited by the reference words (and other words with short vowels and unvoiced consonants). (This is in itself an indirect support of the assumption of the recurrence of identical tonal patterns in reference words and "voiced words".)

Fo movement associated with individual segments (vowels) must be accounted for, but in this case results will probably be valid for the whole ensemble of speakers of the same dialect. (In Advanced Standard Copenhagen Danish all short vowels exhibit falling movements and all long vowels exhibit falling-rising movements, but the fall is of greater extent than the rise. The only exception seems to be the first stressed syllable in a sentence, which may be rising. The direction of the movement is unaffected by surrounding consonants, but the extent of the fall is slightly greater after aspirated stops and unvoiced fricatives than

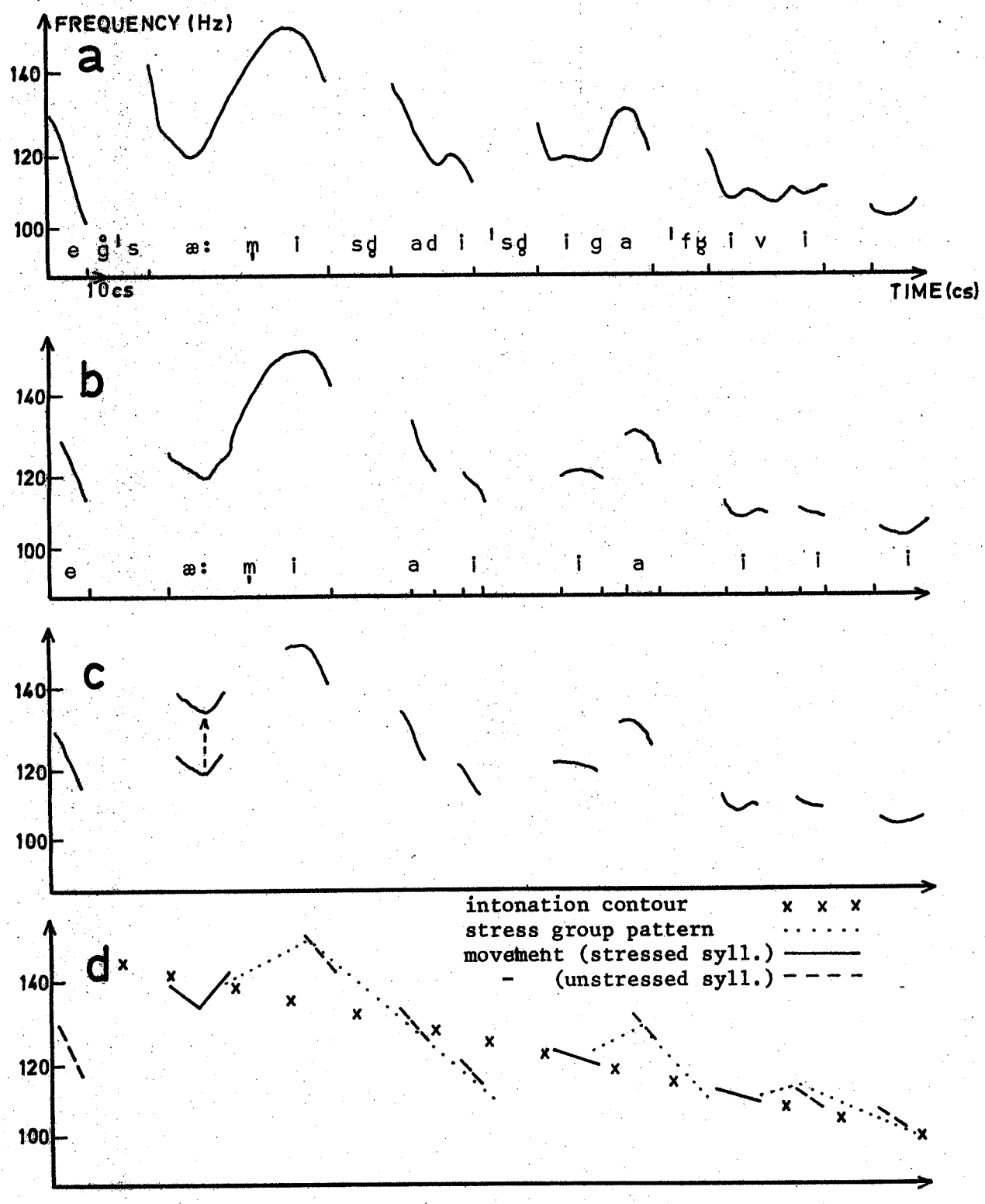


Figure 2

Fo tracing of a sentence 'Eksamen i statistik er frivillig.' (The examination in statistics is non-compulsory.) a: raw signal, b: pruned signal, c: reduced and corrected signal, d: stylized signal. (Male subject.)

after unaspirated stops and sonorants, cf. Reinholt Petersen (1976). If a movement is not heard as such, but rather as a level pitch, it can be represented as such, but should the level then correspond to the beginning, the middle, or the end of the movement? This is, of course, a problem which has engaged phoneticians for years, cf. e.g. Rossi (1971). - If a movement is perceived as a movement, it may be preferable to preserve it in the description, especially if it is to serve comparative purposes. Different vowel Fo movements may well be one of the important prosodic distinctions between different Danish dialects.

When Fo patterns associated with stress groups are recurrent entities (allowing, however, for context dependent modifications, as long as they are predictable) the intonation contour can be defined narrowly as the course described by the stressed syllables alone, cf. the "crossed" line in Fig. 2,d.

Given the results of experiments and considerations as outlined above, a raw Fo tracing can be broken down into constituent components, as illustrated in Fig. 2. First "irrelevant" passages (cf. above) are pruned away (b). Corrections for intrinsic Fo differences between (stressed) vowels are performed and, simultaneously, longer voiced stretches are reduced (via comparisons with reference words in identical context) to a succession of shorter ones (c). In this case movements associated with individual vowels are preserved but it may be justified to reduce them to level courses. Finally, the tracing may be stylized (d).

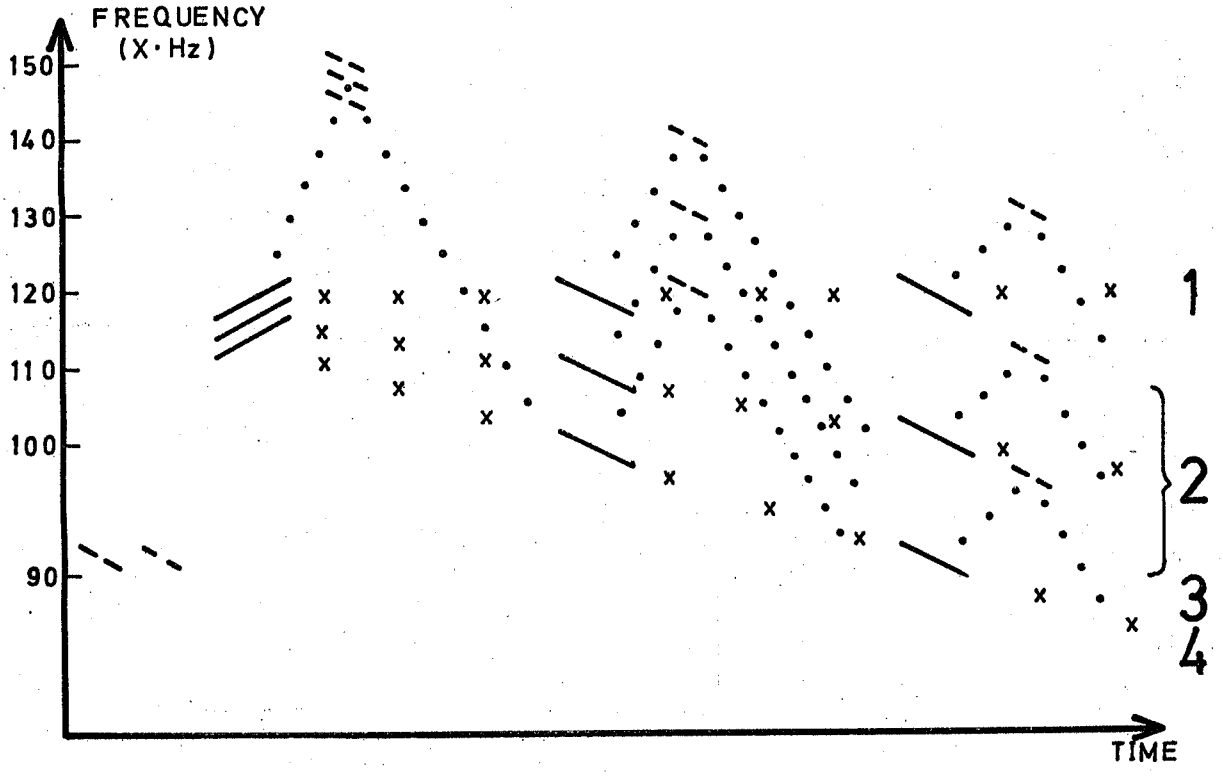


Figure 3

Model for Fo in short sentences in Advanced Standard Copenhagen Danish. 1: Intonation questions, 2: Questions with word order inversion and Non-terminal periods (variable), 3: Questions with interrogative particle, 4: Declarative sentences. See further the legend to Fig. 2,d.

On the basis of a relatively comprehensive material it has been possible, following the procedure just described, to postulate a preliminary model (Fig. 3) for Fo in various types of short sentences in Advanced Standard Copenhagen Danish.

A scale has been tentatively indicated. For X equal to one it would fit a rather low male voice. For other speakers, X will have to be determined, which is probably not a simple procedure. A very coarse approximation is $X = \frac{a}{100}$ where 'a' is the frequency in Hz of the second (of three) stressed syllable(s) in a declarative sentence. This value of X, however, predicts too high values for the highest unstressed syllables for female voices.

The model is supposedly predictive as well as descriptive. Given a sentence with a certain intonation contour (determined by the syntactical and semantical nature of the sentence) the stress group patterns are superimposed on the contour and the appropriate vowel movements are added (if they are not prescribed by the model). The result is modified by intrinsic Fo differences pertaining to individual segments (vowels) and, finally, is turned into a continuous course by smoothly connecting the vowels, with interruptions, of course, for unvoiced consonants (and with dips for voiced obstruents, but this last step needs further clarification). For a more detailed account, see Thorsen (1976).

References

Reinholt Petersen, N. (1976) "Intrinsic fundamental frequency of Danish vowels" Annual Report of the Institute of Phonetics, University of Copenhagen, vol. 10, 1-28.

Rossi, M. (1971) "Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole" Phonetica, vol. 23, 1-33.

Thorsen, N. (1976) "An acoustical investigation of Danish intonation: preliminary results" Annual Report of the Institute of Phonetics, University of Copenhagen, vol. 10, 85-148.

VIII^èmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

QUELQUES EXPERIENCES D'ANALYSE PERCEPTUELLE EN FRANÇAIS

Jacqueline Vaissière

C.N.E.T.

Lannion

• Résumé :

Cette communication décrit quelques expériences d'analyse perceptuelle en Français. La courbe originale du fondamental d'un groupe de mots (voir Fig.1) a été progressivement simplifiée, puis les contours de mots ont été remplacés par d'autres contours. Le résultat de ces modifications a été livré au jugement de quinze auditeurs. Les conclusions concernent la perception de la micromélodie (Fig.4), l'effet auditif produit par la substitution d'un contour caractéristique de mot de quatre syllabes par un autre (Fig.2), et par la diminution ou la suppression de la baisse du fondamental durant les mots grammaticaux, entre deux mots lexicaux de quatre syllabes (Fig.3).

• Summary :

This paper describes experiments in perceptual analysis in French. The fundamental frequency contour of a given sentence (see Fig.1) has been simplified, step by step, and then different Fo pattern have been superimposed on the four-syllables words. 15 listeners have participed in this experiment. The results concern the perception of the micro-variations during the voiced plosives and fricatives (see the stimuli on Fig.4), the perceptual effect obtained by the substitution of different Fo patterns on the four-syllables lexical words (see the stimuli on Fig.2), and the perception of the diminution or suppression of the Fo valley during the lexical words located between two four-syllables lexical words (see the stimuli on Fig.3).

QUELQUES EXPERIENCES D'ANALYSE PERCEPTUELLE EN FRANCAIS.

Une analyse acoustique des faits prosodiques, aussi fine soit-elle, ne permet pas de tirer des conclusions sur le fonctionnement des faits intonatifs d'une langue (1). Une étude perceptuelle est nécessaire. Les courbes de fréquence du fondamental (F_o) présentent un certain nombre de détails qui ne sont pas percus par l'auditeur. L'équipe de Eindhoven utilise pour son étude de l'intonation hollandaise un appareil appelé 'intonator' (2), qui permet de remplacer la courbe de fondamental originale d'une phrase par une autre courbe choisie par l'expérimentateur (3). Le résultat des transformations est livré au jugement d'auditeurs. Seule une telle méthode permet sans doute de déterminer les mouvements pertinents du fondamental pour un auditeur, mouvements susceptibles de jouer un rôle dans la perception de la phrase.

Bien que le fondamental ne soit qu'un des trois principaux facteurs prosodiques (les deux autres étant la durée et l'intensité), la plupart des études perceptuelles (comme par ailleurs les études acoustiques sur la phrase) s'y consacrent, sans doute pour une question de moyens d'expérimentation: il est plus aisé de faire varier le profil du fondamental, que la durée ou l'intensité des phonèmes qui composent une phrase.

Certains chercheurs ont effectué un certain nombre d'expériences d'analyse perceptuelle sur le francais. Rossi (4) a étudié le seuil de variations tonales pour les sons de la parole (les résultats restent valables pour d'autres langues que le français). Rossi et Chafcouloff (5) ont fait des recherches sur le nombre de niveaux pouvant être identifiés par un auditeur pour une langue telle que le français. Larreur et Boë (6) , Ausseterre et Di Cristo (7) ont étudié du point de vue perceptif les types intonatifs tels que la question, l'énonciation et l'interrogation. Lhote, Filleau et Grange ont testé la reconnaissance de phrases à partir des seuls patrons intonatifs (8).

Les expériences que nous reportons dans cette communication concernent la différence auditive obtenue en opérant une simplification plus ou moins poussée de la courbe originale de la fréquence du fondamental d'une phrase donnée; suppression des variations de micromélodie durant les plosives et fricatives voisées, stylisation du contour mélodique des mots (dans cette expérience , les mots concernés ont quatre syllables), substitution de différents contours mélodiques sur un même mot, et diminution ou suppression de la baisse du fondamental sur les mots grammaticaux, entre deux mots lexicaux de quatre syllabes.

MATERIEL:

Pour confirmer les conclusions de son analyse acoustique sur l'anglais par une analyse perceptuelle, Maeda (9) a écrit un programme permettant de réaliser des expériences semblables à celles de Eindhoven. La voix est analysée par une technique similaire à la prédiction linéaire (ce qui permet d'obtenir une voix synthétique de bonne qualité), et le contour du fondamental peut être modifié à volonté à l'aide d'un light-pen sur une tablette. C'est ce programme de M.I.T. que nous avons utilisé pour créer nos stimuli.

AUDITEURS:

Quinze français, appartenant au département SST du CNET à Lannion ont accepté de servir comme sujets dans ces expériences. Aucun de ces auditeurs n'avait participé auparavant à des expériences de perception.

CORPUS:

La figure 1 représente la courbe du fondamental originale de la phrase de référence: "Il a contribué à la majorité des progrès technologiques...".(Cette phrase se termine par une intonation montante car elle n'est pas terminée). Les

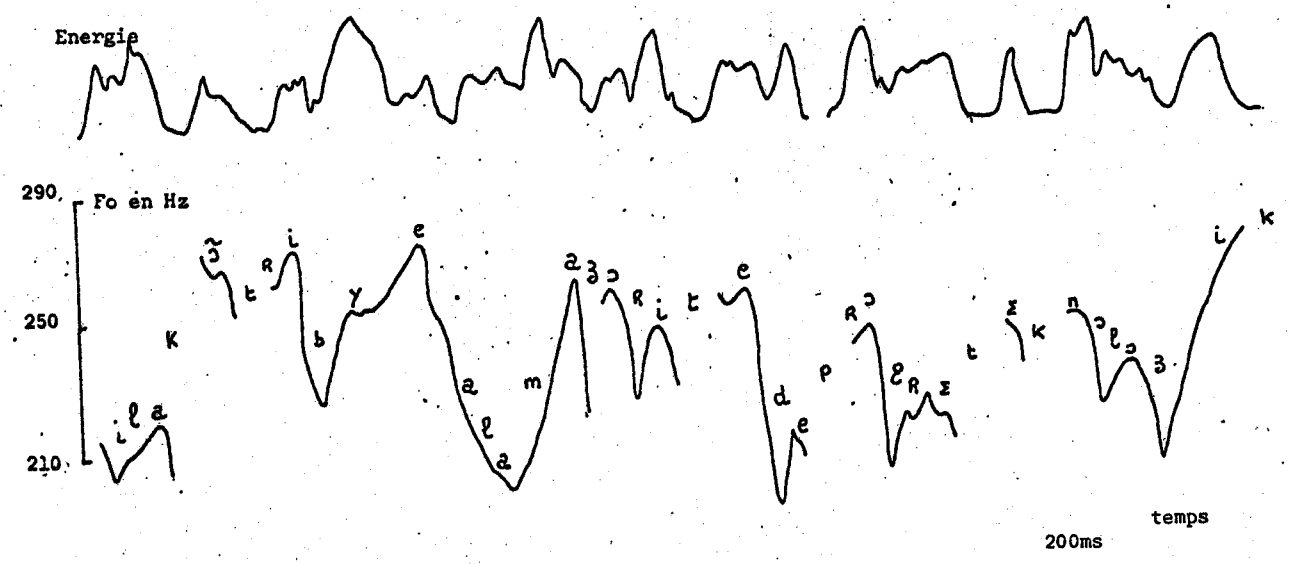


Figure 1: Fréquence du fondamental original de la phrase. On observera les variations importantes du fondamental durant les plosives et fricatives voisées, le creux caractéristique concommitent aux mots grammaticaux entre les mots lexicaux, la montée initiale de Fo entre la voyelle qui précède les mots lexicaux et la première voyelle de ces mots, le schéma sensiblement équivalent des mots 'contribué', 'majorité' et technologiques'. Cette phrase a été choisie pour la régularité des contours de mots, et leur longueur.

stimuli ont été présentés par paire au groupe d'auditeurs (3 groupes de 5 auditeurs). Pour chaque paire de stimuli (présentée deux fois), les sujets devaient répondre aux quatre questions suivantes:
(1) les deux stimuli sont-ils pareils?
(2) Y en a-t-il un de meilleur? Lequel?
(3) Y en a-t-il un d'anormal? Le ou lesquels?
(4) Si vous entendez une différence entre les deux stimuli, essayez de la définir.

RESULTATS:

1) Effet de la suppression de la micromélodie.(Fig.4)

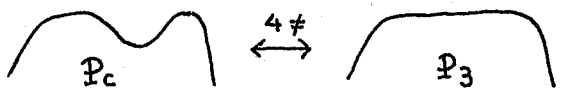
Le stimuli avec la courbe originale du fondamental (Fig.1) et le stimuli illustré à la Fig.4, ont été présentés par paire. Sur le second stimulus, la courbe du fondamental a été obtenue en joignant les valeurs du fondamental d'un sommet vocalique à une autre.

12 locuteurs sur les 15 n'ont pas perçu de différences entre les deux stimuli. Sur les 3 locuteurs qui ont perçu une différence, 2 ont effectivement jugé le premier comme le meilleur , et un locuteur a noté que le second stimulus semblait un peu "rauque".

En conséquence, si les variations de micromélodie semblent avoir une influence sur la qualité des consonnes (10), leur suppression n'est cependant pas perçue par une large majorité des auditeurs au niveau de la phrase.

2) Modifications du contour global de Fo sur les mots lexicaux de quatre syllabes.(Fig.3)

a) Effet de la suppression du creux sur la pénultième des mots longs, caractéristique de ce locuteur. (Fig.2)

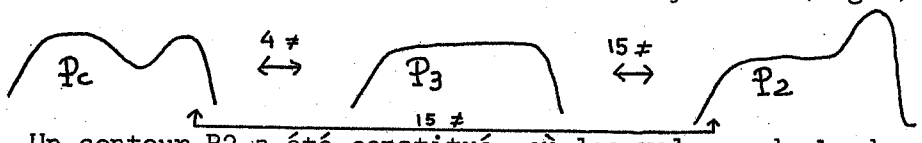


Nous avons pu observer, lors de l'analyse acoustique de deux textes lus par ce locuteur, une tendance à une incursion du fondamental vers le bas sur la pénultième des mots longs: la dernière syllabe est prononcée avec des valeurs du fondamental sensiblement égales à celles des premières syllabes du mots (Voir par exemple l'évolution du fondamental sur les mots 'contribué', 'majorité' et 'technologiques' sur la Fig.1).

Nous avons supprimé ce creux et schématisé le contour des deux premiers mots lexicaux de la phrase par un trapèze à trois côtés (Contour P3 (11)). Le contour résultant est l'un des contours représentés à la Fig.2.

11 auditeurs n'ont pas perçu la suppression du creux, et sur les 4 auditeurs qui ont perçu une différence entre les deux contours, 2 ont jugé le premier comme meilleur et 2 le second. L'ensemble des locuteurs n'a donc pas été sensible à la simplification poussée des contours intiaux. La différence de 35Hz entre la pénultième et la dernière syllabe n'a en général pas été entendue.

b) Comparaison des contours avec creux sur la pénultième (Pc), plat (P3) et avec un pic de 35hz sur la dernière syllabe. (Fig.2)

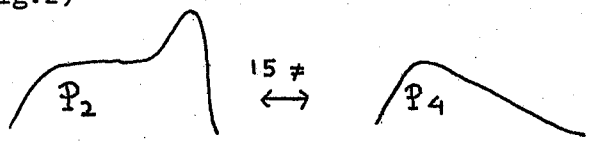


Un contour P2 a été constitué, où les valeurs de la dernière syllabe atteignent 35Hz de plus que les valeurs sur la pénultième (la valeur de 35Hz a été choisie, par référence au creux initial de 35Hz sur la pénultième). La différence entre le contour P2 et Pc, est que dans le premier cas, les valeurs de Fo sur la pénultième sont les mêmes que durant les syllabes précédentes.

Si les deux premiers contours (Pc et P3) ont été jugés en général comme semblables (voir a)), il n'en est pas de même pour les deux derniers qui ont été trouvés différents par tous les locuteurs. 13 locuteurs ont jugé Pc et P2 comme différents.

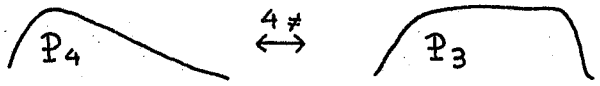
Ces résultats viennent renforcer les premières observations, à savoir que le contour des mots longs est perçu de façon globale, et que la hauteur de la dernière syllable n'est pas perçue seulement en fonction de la hauteur sur la pénultième, mais en fonction des hauteurs perçues sur toutes les autres syllabes du mot.

c) Contour P2 et contour avec des valeurs de Fo décroissantes sur tout le mot (P4). (Fig.2)



Tous les auditeurs sont unanimes: les deux contours sont perçus comme différents (sans doute, intonation montante, versus intonation descendante).

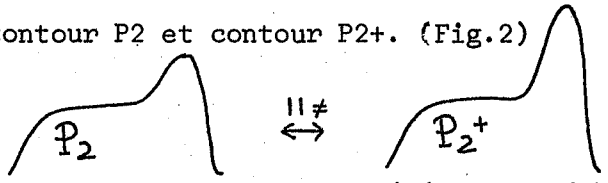
d) Contour P et contour P3. (Fig.2)



Le résultat est interessant. La majorité des locuteurs (11) n'ont pas percu de différence entre les deux contours.

Ces deux schémas, trouvés lors d'une analyse acoustique, sont confondus en un seul patron intonatif. P2 s'opposerait à la fois à P3 et P4 (intonation montante versus intonation non montante).

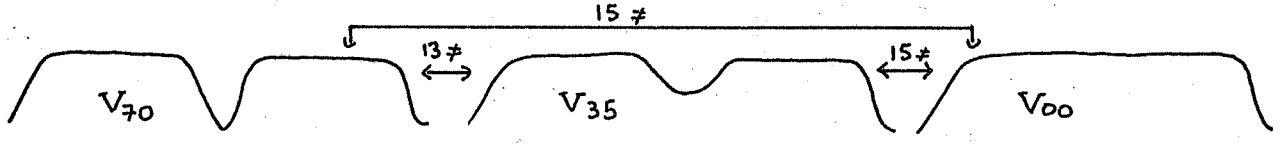
e) Contour P2 et contour P2+. (Fig.2)



Le saut du fondamental de 35Hz a été augmenté à 70Hz. 11 auditeurs ont percu la différence.

Alors qu'une baisse du fondamental de plus de 35Hz sur la dernière syllabe par rapport à une courbe plate n'a en général pas été percue(P4 versus P3), une élévation de 35Hz par rapport au contour plat (P2 versus P3) a été perçue. Et bien que les contoursP2 et P2+ soit tous les deux similaires dans leur forme générale, ils sont ressentis comme différents. Il est donc évident que des différentes importantes en valeur absolue peuvent ne pas être perçues (elles ne sont pas importantes du point de vue auditif), alors qu'à certains endroits de la chaine des différences moins importantes en valeur absolue peuvent avoir une plus grande importance auditive.

3) Modifications du contour de Fo entre deux mots lexicaux longs, sur les mots grammaticaux. (Fig.3)



Le creux initial de 70 hz environ entre les deux mots lexicaux, sur les mots grammaticaux (V70), a été réduit à 35Hz (V35), puis supprimé (V00).

La diminution a été nettement entendue par la majorité des auditeurs (13) (V70 versus V35), et la suppression par la totalité des auditeurs (V70 versus V00). 9 locuteurs ont ressenti une différence entre les deux derniers contours, qui ont été jugés 'anormaux'.

L'oreille semble particulièrement sensible à une diminution du creux de Fo à cet endroit, alors qu'elle ne l'était pas à une suppression du creux de 35Hz sur le pénultième. Cela confirme les conclusions précédentes.

Mais pourquoi tous les locuteurs sont-ils presque unanimes à reconnaitre V35 comme anormal (13 auditeurs sur 15) et V70 (15 auditeurs), alors qu'ils s'étaient montrés très tolérants lors de l'écoute des stimuli sur les mots. La plupart des locuteurs ont noté une trop grande 'platitude' de la phrase. Est-ce parce que 10 syllabes ont été prononcées sans variations mélodiques, est-ce parce les mots grammaticaux sont prononcés avec une hauteur trop élevée?... Cette expérience limitée ne nous permet pas de répondre à ces questions.

CONCLUSION:

De même que les conclusions d'une analyse acoustique dépendent fortement du corpus de phrases étudiées, du rythme d'élocution, de la longueur des mots, et des locuteurs,..., une analyse perceptuelle ne permet pas toujours de déboucher

sur des résultats sûrs: la perception des stimuli verbaux (comme des autres stimuli du reste) dépend de nombreux facteurs. La modalité de structure des stimuli, état du sujet, facteurs émotionnels, technique expérimentale, conditions environnantes, complexité de la tâche (12) influent sur les conclusions. Nous ne prétendons donc pas donner des résultats valables pour le français en général, mais dans les limites de nos expériences, nous pouvons conclure:
(1) que les locuteurs ne perçoivent en général pas une suppression de la micromélodie au niveau de la phrase, quand leur attention n'est pas dirigée vers la compréhension des consonnes;
(2) que le contour des mots de quatre syllabes semble être perçu globalement, et que la hauteur de la dernière syllabe de ces mots est perçue non seulement en fonction de la hauteur de la pénultième, mais également en fonction de la hauteur des syllabes précédant la pénultième;
(3) qu'une intonation plate sur le mot ou descendante n'est en général pas perçue comme différente, mais qu'une intonation plate ou descendante semble s'opposer nettement à un contour où les valeurs du fondamental sur la dernière syllabe sont supérieures à celles des syllables précédentes (pas seulement la pénultième);
(4) et qu'à amplitude égale, la perception d'un changement dépend du sens de ce changement (une augmentation du fondamental étant plus facilement percue qu'une diminution par rapport à une intonation plate sur les syllabes précédant la syllabe où sont opérées les modifications) et elle dépend aussi de la localisation de ce changement dans la phrase.

NOTES ET BIBLIOGRAPHIE:

(1) Notons cependant qu'une analyse acoustique est suffisante pour des applications telles que l'utilisation des paramètres prosodiques en reconnaissance, du moins jusqu'à un certain degré. De même, pour des applications telles que la synthèse par règles, une analyse perceptuelle est souhaitable, afin de ne pas générer des détails non audibles, mais elle n'est pas strictement nécessaire.

(2)Willems, L.F., The Intonator, I.P.O., Ann. Progr. Rep. 1, 1966, pp 123-5.
(3) Cohen, A., and t'Hart, On the Anatomy of Intonation, Lingua 19, 1967, pp 177-92.
(4) Rossi, M., Le Seuil de Glissando ou Seuil de perception des variations tonales pour les sons de la parole, Phonetica, 23, 1971, pg. 1-33.
(5) Rossi, M., et Chafcouloff, M., Les niveaux intonatifs, Trav. Inst. Phon. d'Aix, Vol. 1, 1972, pg. 167-76.
(6) Larreur, D., et Boë, L.J., Synthèse paramètrique de la phrase énonciative en français, 5ième J.E.P., Orsay, Vol. 2, 1974.
(7) Ausseterre, D., et Di Cristo, A., Recherches psycho-sémantiques sur l'intonation de la phrase française, Trav. Inst. Phon. d'Aix, Vol. 1, 1972, pg. 61-8.
(8)Lhote, E., Filleau, M., et Grange, M.F., Reconnaissance de patrons intonatifs, 6ièmes J.E.P., Toulouse, 1975, pg. 28-37.
(9) Maeda, S., A characterization of American English Intonation, Ph.D. Thesis, M.I.T., 1976.
(10) Larreur, D., et Boë, L.J., Etude de l'influence des variations de la fréquence laryngienne sur l'intelligibilité et la qualité des consonnes sonores générées par vocodeur, Bull. Inst. Phon. Grenoble, 2, 1973.
(11) Vaissière, J., Further Note on French Prosody, Progress report 115, Research Laboratory of Electronics, M.I.T., 1975, pp. 251-61.
(12) Slama-Cazacu, T., Quelques principes psychologiques et psycholinguistiques dans l'étude de la perception de la parole, 6th Cong. Phonetics Sciences, Prague, 1967, pp. 837-40.

FIGURES: (voir page suivante).

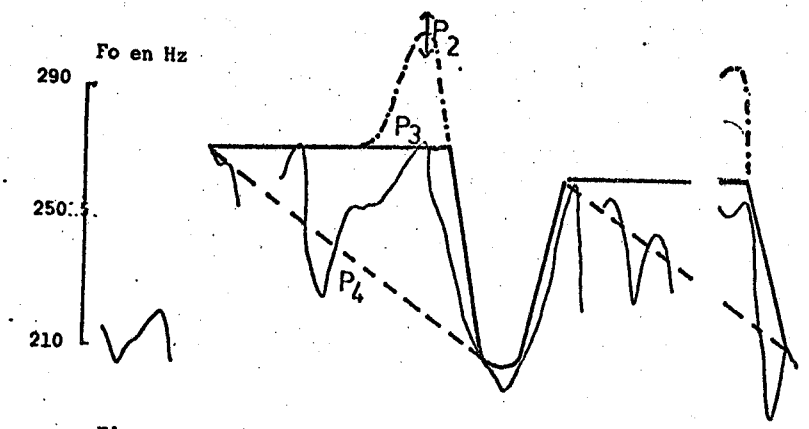


Figure 2: Fréquence du fondamental des stimuli pour la comparaison de l'effet acoustique obtenu par la substitution de contours de mots différents: montée de Fo sur la dernière syllabe (P2), descente de Fo sur tout le mot (P4), et à la fin de la dernière syllabe.

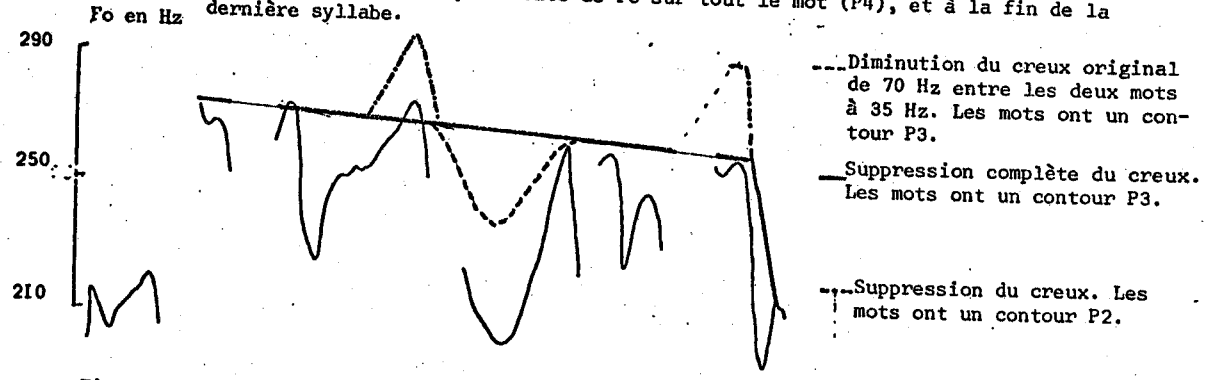


Figure 3: Fréquence du fondamental des stimuli pour l'évaluation de l'importante du V de fondamental caractéristique entre deux mots longs: diminution de l'échancrure, suppression complète, et compensation de la suppression par l'addition d'un pic sur la dernière syllabe des mots lexicaux.

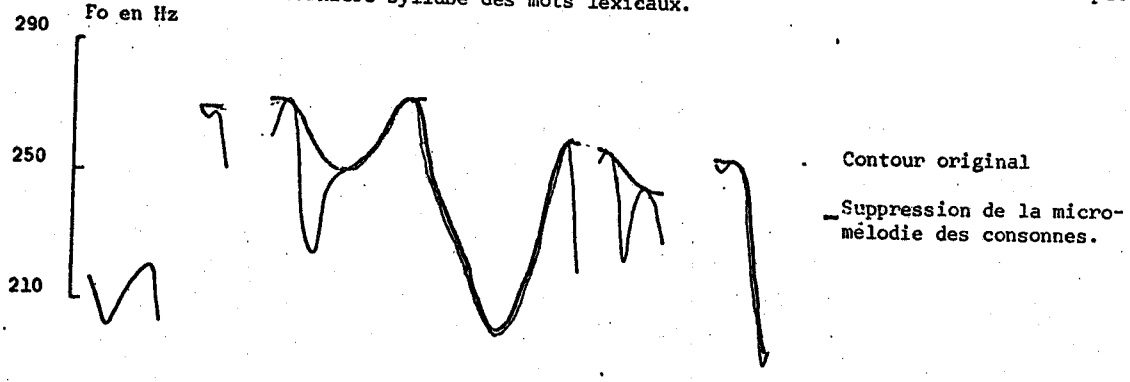


Figure 4: Fréquence du fondamental de l'un des stimuli pour l'étude de l'effet perceptuel obtenu en supprimant les variations du fondamental durant les plosives et les fricatives voisées.

Corrélations entre l'intensité, la durée et la fréquence fondamentale

VIIIèmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

RELATION ENTRE L'EVOLUTION DE LA FREQUENCE LARYNGIENNE ET DE L'INTENSITE POUR LA PHRASE ENONCIATIVE DU FRANCAIS. PREMIERS RESULTATS

P. GRAILLOT
Département E.T.A
C.N.E.T. Lannion

&

L.J. BOË
Institut de Phonétique
de Grenoble

Résumé :

Le mécanisme du fonctionnement phonatoire du larynx n'implique pas obligatoirement une relation entre les variations de la fréquence laryngienne et de l'intensité du signal vocal. Cependant plusieurs auteurs font état de certaines similitudes entre les évolutions de ces deux variables. Le but de ce travail préliminaire est de préciser si une telle relation existe et, si tel est le cas, de chiffrer statistiquement les limites de sa validité. Cette étude a été menée dans le cadre de la phrase énonciative du français.

Summary :

The phonatory behaviour of the larynx does not implicate a systematic relation between the variations of laryngeal frequency and sound pressure level. But authors have noted similarity between the evolutions of these parameters. The aim of this preliminary study is to state if such a relation exists and, if yes, to give data concerning its validity. This work has been conducted in the frame of enunciative sentence of French.

RELATION ENTRE L'EVOLUTION DE LA FREQUENCE LARYNGIENNE ET DE L'INTENSITE POUR LA PHRASE ENONCIATIVE DU FRANCAIS. PREMIERS RESULTATS

P. GRAILLOT & L.J. BOË

L'intonation est réalisée et perçue à l'aide de plusieurs paramètres parmi lesquels l'évolution de la fréquence laryngienne (F_0) et de l'intensité (I) jouent un rôle essentiel. Généralement les auteurs s'accordent à reconnaître que c'est F_0 qui est le facteur le plus important. De fait, la très grande majorité des études instrumentales consacrées à l'intonation portent sur des tracés de l'évolution de la fréquence fondamentale du signal de parole.

Les modifications du régime de fonctionnement des cordes vocales dépendent principalement de la pression subglottique (P_s) et de la tension des cordes vocales (T). L'intensité de la parole est liée, pour sa part, beaucoup plus au premier paramètre physiologique qu'au second[1]. On voit donc que le mécanisme de fonctionnement du larynx n'implique pas obligatoirement une relation entre les variations de F_0 et celle de I : les chanteurs peuvent très bien émettre une note indépendamment de son niveau sonore. C'est le résultat d'un apprentissage de dissociation que n'exige pas le code de la communication parlée, contrairement à celui de la musique. Il a été noté, et cela depuis bien longtemps[2], que les locuteurs ont tendance à "hausser le ton" lorsqu'ils parlent plus fort. Plusieurs auteurs ont signalé des ressemblances entre les évolutions de F_0 et de I (voir par exemple DENES & MILTON-WILLIAMS, 1962), précisé leurs domaines d'évolution (FANT, 1959; FAIRBANKS, 1960; BOË & RAKOTOFIRINGA, 1975), et même chiffré des coefficients de corrélation (RIGHINI, 1968).

Une application a été tentée dans le cadre de la synthèse; la fréquence d'excitation d'un vocodeur étant déduite de l'intensité pendant 90% du temps (JOOS, 1962). Le but de ce travail est de préciser, dans le cadre de la phrase énonciative du français, s'il existe bien une relation entre F_0 et I et dans l'affirmative de préciser les limites de sa validité par une évaluation statistique.

---

1. On pourra se reporter pour ce problème à la communication de GUERIN & BOË dans ces mêmes Journées d'Etudes.

2. Par exemple par ARCHYTAS de Tarente (IVe siècle A.C.).

PROCEDURE ET RESULTATS

Nous avons utilisé 5 phrases[1] qui ont été enregistrées avec une élocution naturelle par 8 locuteurs (1 femme et 7 hommes). Les valeurs de F_ℓ calculées à partir de la période codée sur 256 niveaux[2] et celles des variations du niveau d'intensité (256 niveaux couvrant une dynamique de 48 dB) ont été échantillonnées toutes les 5 ms.

Les résultats numérisés, stockés sur bandes magnétiques, ont été imprimés avec, par ligne :

- le numéro de l'échantillon,
- la variation de l'énergie dans 6 canaux[3] , ceci pour permettre la délimitation et la transcription des sons,
- les variations du logarithme de l'intensité (niveau d'intensité),
- les valeurs de F_ℓ,
- deux astérisques permettant de suivre graphiquement les évolutions des deux dernières variables.

Pour déterminer globalement la relation entre les évolutions de la fréquence laryngienne, nous avons envisagé de calculer le coefficient de corrélation entre ces deux variables. Mais la durée d'intégration nécessaire pour l'obtenir est fonction de la largeur spectrale des signaux (MAX, 1972). Si l'on admet que l'on a affaire à des signaux gaussiens identiques l'erreur d'estimation est :

$$\varepsilon = \frac{1,6}{\sqrt{B.T}}$$

B : largeur du spectre du signal
T : temps d'intégration

En ce qui concerne le signal d'intensité, B est de l'ordre de 20 Hz ce qui donne avec $\varepsilon = 5\%$ un temps voisin de 50 s. bien supérieur à celui dont nous disposions par locuteur, la durée voisée pour chacun d'entre eux étant de l'ordre de 5 s. Aussi avons-nous adopté une procédure plus adaptée aux données. Pour chaque échantillon nous avons cherché la valeur de K telle que :

$$\log I - K \log F_\ell = 0$$

l'écart type $\sigma(K)$ reflétant la latitude avec laquelle log I et log F_ℓ se superposent au cours du temps. Nous avons choisi de travailler avec les log.

---

1. Le crédit est consenti par la banque - Le docteur a ordonné un médicament - Le cocher fouette son cheval - Le fantôme hante le château - Le sculpteur taille la pierre.

2. ZURCHER (1972/73)

3. 100-430 Hz, 430-800 Hz, 800-1400 Hz, 1400-2000 Hz, 2000-2700 Hz, 2700-3400 Hz, Les filtres font partie d'un appareillage élaboré pour la reconnaissance automatique de la parole (CARTIER & al., 1974 et 1976).

des deux variables pour nous débarasser des problèmes de valeurs absolues puisqu'il y avait plusieurs locuteurs à la fréquence fondamentale différente et qui n'avaient pas forcément lu le corpus au même niveau d'intensité.

Les résultats sont présentés au tableau I par locuteur (le n° 8 est féminin) avec n le nombre d'échantillons avec lesquels ont été effectués les calculs. On peut constater que les valeurs concernant la moyenne et l'écart-type de K sont très homogènes d'un locuteur à un autre. Pour l'ensemble des locuteurs $\overline{K} = 75,5$: compte-tenu du codage adopté une variation d'une octave de F_ℓ correspond en moyenne à une évolution d'intensité de 4,3 dB.

D'autre part, $\frac{\sigma(K)}{\overline{K}}$ est assez élevé puisqu'il évolue entre 39 % et 45 % ce qui laisse à supposer que la relation qui lie les 2 variables étudiées n'est pas systématique comme d'ailleurs on peut l'observer sur les tracés.

Locuteur	n	$\overline{F_\ell}$	$\overline{K}$	$\sigma(K)$
1	1509	133	76	31
2	1421	108	76	28
3	1258	129	76	34
4	2646	125	77	30
5	640	120	78	28
6	1927	120	80	31
7	1037	126	59	-
8	1190	189	82	-

TABLEAU I

Autre façon de mettre en évidence la relation entre F_ℓ et N.I, nous avons établi par locuteur, des histogrammes pour lesquels, par tranche de 2 Hz de F_ℓ était relevé le nombre d'occurrences et calculé le niveau d'intensité moyen et l'écart type de ce niveau. Un tracé correspondant est présenté pour le locuteur 8 (figure 1).

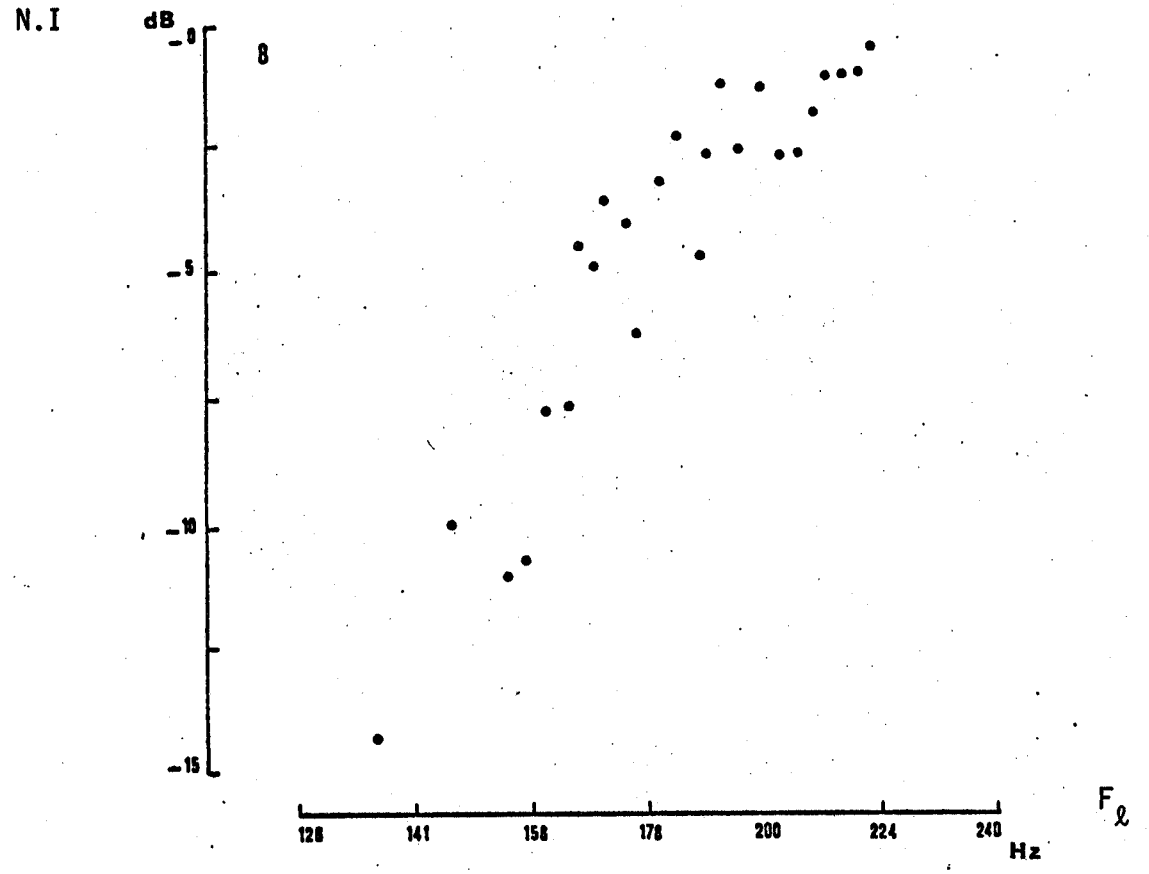


Figure 1

Histogramme des niveaux d'intensité (dB relatifs) par tranche de F_ℓ de 2 Hz

On constate sur tous les tracés une très nette similarité entre les histogrammes, les fréquences élevées correspondent toujours bien aux niveaux les plus intenses avec une certaine latitude , les $\frac{\sigma(N.I)}{\overline{N.I}}$ varient entre 10 et 40 %.

REMARQUES

Ces résultats préliminaires (les statistiques globales ne sont pas d'une grande utilité pour un signal comme la parole) permettent de mettre en évidence que :

- en moyenne, il existe bien une relation entre les évolutions de la fréquence laryngienne et celles du niveau d'intensité.

- cette relation est très voisine d'un locuteur à un autre : une variation d'une octave correspondant à environ 4 dB, ce qui est relativement faible compte-tenu de certains résultats obtenus précédemment.

En fait il faut bien voir que :

- l'intensité est une fonction continue du temps qui évolue entre un niveau maximum et un niveau de bruit (dépendant des conditions d'enregistrement) qui est le résultat d'une intégration sur 20 ms environ; alors que F_ℓ , à partir du moment où elle est détectée (seuil d'intensité),s'établit à une valeur donnée avec une constante de temps plus petite. La corrélation entre N.I et F_ℓ ne peut donc être qu'inexistante pour les faibles niveaux d'intensité. C'est d'ailleurs ce qui a été mis en évidence par un autre type d'analyse (BOË & RAKOTOFIRINGA, 1975).

- si les variations de F_ℓ sont dues à des modifications de pression subglottique, le niveau d'intensité va évoluer aussi dans le même sens; ce n'est pas le cas si le régime de vibration des cordes est accéléré par une croissance de la tension des cordes vocales car celle-ci n'a que peu de répercussion sur le niveau d'intensité (GUERIN & BOË, 1977).

Ce travail doit être suivi d'une étude permettant de délimiter les séquences de parole pour lesquelles N.I et F_ℓ ont des évolutions comparables et de les spécifier dans le cadre de l'analyse des faits supra-segmentaux.

REFERENCES

BOË, L.J. & RAKOTOFIRINGA, H. (1975). A Statistical Analysis of Laryngeal Frequency : its Relationship to Intensity Level and Duration.
Language & Speech 18, pp. 1-13.

CARTIER, M., COURBON, J.L., DAGORNE, R., DOUCEN, R., LUCAS, J.J. (1974).
Capteur pour la reconnaissance de six sons.
2° Rapport intermédiaire de contrat C.R.I. n° 73034.

CARTIER, M., COURBON, J.L., LUCAS, J.J., DAGORNE, R., DOUCEN, R. (1976).
C.H.A.R.L.E.S., Appareil de reconnaissance d'un vocabulaire artificiel.
Recherches Acoustiques, CNET - Lannion.3, pp. 131-147.

DENES, P. & MILTON-WILLIAMS, J. (1962). Further Studies in Intonation.
Language & Speech 5, pp. 1-14.

FAIRBANKS, G. (1960). Voice and Articulation Drillbook.
2 ed. Harper & Row Pub. New-York.

FANT, G. (1959). Acoustic Analysis and Synthesis of Speech with Applications to Swedish.
Ericsson Technics 1.

GUERIN, B. & BOË, L.J. (1977). La régulation de la vibration des cordes vocales : simulation à l'aide d'un modèle à deux masses.
8° J. E. d'Etudes du Groupe Communication Parlée du GALF (dans ce volume)

JOOS, M. (1962).
Proc. 2nd Texas Conference in Problem of Linguistic Analysis in English.

MAX, J. (1972). Méthodes et techniques de traitement du signal et applications aux mesures physiques. Tome I.
Masson & Cie. Paris.

ZURCHER, J. (1972/73). Dispositif de détection et de mesure du fondamental de la parole humaine.
Analyse et Synthèse de la parole, CNET - Lannion.1, pp. 7-15.

Rôle des traits prosodiques dans la transmission et la synthèse de la parole

VIII èmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

SUR LA COMMANDE PAR LA GLOTTE D'UNE PROTHESE DES MEMBRES SUPERIEURS

PAR

P. ANDRE et F. LHOTE

Laboratoire d' Automatique de Besançon

• Résumé :

On étudie ici la possibilité d'une commande vocale "stricto sensu"d'une prothèse destinée à des amputés bilatéraux des membres supérieurs.

On considère la vibration des cordes vocales du patient comme une source de messages volontaires susceptibles d'être utilisés comme codes d'ordres pour l'exécution de mouvements.

Il s'agit donc d'élaborer au niveau du larynx un langage artificiel parfaitement identifiable,distinct de la parole,et pouvant être produit bouche fermée (captage par un glottographe suivi d'un analyseur de mélodie en temps réel).On envisage la création d'un vocabulaire fini constitué de motifs mélodiques,dont on précise les traits distincts (nécessaires à la reconnaissance),et les règles d'association (nécessaires à l'exécution d'ordres complexes).

• Summary :

We study here the possibility to perform a strictly vocal control of a prothesis intended for both arms amputated persons.

We consider the patient's vocal strings vibration as a source of deliberate messages capable of being used as orders codes for controlling movements.

It is necessary to elaborate at larynx's level an artificial linguage allowing a reliable recognition,differing from speech and producible with closed mouth (picking up being performed by a glottograph associated to a real time pitch-meter).We intend to create a finite vocabulary constituted by pitch motifs and for which we point out the distinctive features (necessaries for recognition) and linking rules (required for expressing complex orders).

SUR LA COMMANDE PAR LA GLOTTE D'UNE PROTHESE DES MEMBRES SUPERIEURS

P.ANDRE - F.LHOTE

1 - ROBOTS D'ASSISTANCE AUX HANDICAPES DES MEMBRES SUPERIEURS

La nécessité d'une assistance permanente met les tétraplégiques et les amputés bilatéraux des membres supérieurs en état de totale dépendance vis à vis d'une tierce personne et peut entrainer des conséquences psychologiques désastreuses.

Ceci justifie largement l'intérêt porté à la mise au point de téléthèses pour tétraplégiques (Projet SPARTACUS de l'I.R.I.A.) et de prothèses pour les amputés bilatéraux (Projet E.N.S.C.M.B. suscité par le Docteur MAZOYER).

Ces dispositifs,que nous appellerons plus généralement robots d'assistance médicale,doivent pouvoir être commandés par le handicapé,ils comprendront de ce fait:

-Un système mécanique susceptible d'assurer les principales fonctions des membres supérieurs:le "Manipulateur"

-Un système de traitement d'information,le "Coordinateur" qui interprête les ordres donnés par l'utilisateur et commande les actionneurs du manipulateur en assurant leur coordination en fonction du mouvement demandé.

Le manipulateur anthropomorphique(genre MA 23 du C.E.A.)comporte 7 degrés d'action:2 au niveau de l' "épaule",2 à celui du "coude",2 au "poignet" et 1 pour la "main".Pour cette dernière on n'a à demander que l'ouverture et la fermeture:on utilise soit une simple pince à limitation automatique de l'effort de serrage,soit un dispositif du genre"Main de BELGRADE" définissant lui-même l'effort suffisant pour empêcher le glissement.On dispose donc de 6 degrés de liberté pour déplacer et orienter la "main" dans l'espace.

Nous avons choisi de pouvoir faire demander par l'utilisateur deux types de gestes.

a- Des gestes enregistrés:

Ce sont soit des gestes complexes,répétitifs ou semi-répétitifs, soit des gestes tels que le handicapé se trouve dans l'impossibilité de donner un ordre au cours de leur exécution,soit encore des actions particulières(allumage de lampe,déclenchement d'une sonnerie,déplacement d'un fauteuil roulant....)

Ces gestes seront engendrés par des programmes pré-enregistrés; la sélection se fera en demandant le défilement d'un système d'affichage séquentiel des programmes disponibles et en l'arrêtant lors de l'apparition du programme désiré.

b- Des gestes élémentaires:

Ils permettent le déplacement et l'orientation de la main dans l'espace ainsi que la préhension d'objets.

On peut,par exemple,considérer les huit gestes suivants,exécutables chacun dans deux sens:

- 3 translations de la main: X Y Z,selon 3 axes orthogonaux(1vertical et 2 horizontaux)
- 3 rotations de la main: A B C,
- 1 translation le long de l'axe principal de la main, celle-ci jouant le rôle de pointeur
- 1 ouverture/fermeture de la main

Pour ces gestes,dont la succession pourra donner un geste complexe quelconque,on modifiera un seul paramètre à la fois en gardant tous les autres fixes.

Nous avons choisi cette stratégie dans le but d'éviter au handicapé d'avoir à analyser mentalement un geste complexe pour ensuite en donner une définition précise,presque mathématique,difficile à exprimer.

Dans tous les cas il s'agira de mouvements relatifs dont l'utilisateur contrôlera visuellement l'exécution;il pourra en demander à tout instant l'arrêt puis la remise en route dans le même sens ou en sens inverse sans avoir à le redéfinir.

On disposera par ailleurs de deux vitesses d'exécution;tout mouvement commencera à vitesse lente et pourra être ensuite accéléré puis,éventuellement,décéléré.

Pour que le handicapé puisse exprimer ses ordres et que le coordinateur puisse les interprêter sans ambigüité,il faut:

- Définir un langage de commande aussi simple que possible mais permettant une grande souplesse d'utilisation;
- Choisir une source de signaux physiologiques volontaires en tant que support des messages de commande.

Dans ce qui suit nous donnons un exemple de langage de commande adapté au support de message que peut constituer le signal laryngien.

2 - SIGNAUX PHYSIOLOGIQUES DE COMMANDE

Plusieurs supports sont envisageables et font actuellement l'objet d'études,essentiellement dans le cadre du projet SPARTACUS.

2 - 1 La parole

Ce serait le meilleur moyen de communication.Sa reconnaissance et sa possibilité d'utilisation dans ce cadre sont explorées,notamment au L.E.A. de NANCY.Ses avantages sont incontestables:captage facile,ordres donnés en clair,pas d'apprentissage.Cependant à l'heure actuelle sa reconnaissance ne semble pas encore opérationnelle;de plus le système de captage est sensible au parasitage par l'environnement (bruits,conversations) et,si le patient veut participer à une conversation,il doit désactiver le coordinateur.

2 - 2 Les mouvements d'yeux,de paupières,de mâchoires etc.

Ils présentent trois inconvénients importants:le dispositif de captage engendre une gêne,certains mouvements sont incontrôlés et l'émission d'un message entraîne des mimiques qui peuvent être désagréables pour le handicapé et pour son entourage.

2 - 3 Les mouvements de pieds et/ou d'orteils

Ils ne sont utilisables que pour des amputés bilatéraux et seulement lorsqu'ils sont assis.

2 - 4 Les mouvements de la tête et des épaules

Ils sont exploités au Centre d'Automatique de LILLE et à l'IRIA et donnent des résultats très intéressants(consignes analogiques pour des asservissements en position du manipulateur) mais leur captage impose pratiquement au handicapé d'être assis dans un fauteuil, ce qui n'est pas gênant dans le cas des tétraplégiques;on peut également craindre une certaine fatigue de l'utilisateur.De plus ils ne suffisent pas à coder tous les types d'ordres possibles.

2 - 5 Les signaux myo-électriques

Ils ajoutent aux deux premiers inconvénients mentionnés ci-dessus en 2-2 un contrôle incertain par le patient ainsi qu'un traitement délicat.Ils peuvent s'ajouter aux signaux précédents (2-4) mais l'ensemble engendre encore un langage insuffisant.

2 - 6 Le signal laryngien

2-6-1 Avantages et inconvénients

La modulation de la vibration des cordes vocales nous semble intéressante à plus d'un titre,àcommencer par le confort que procure l'utilisation de cette source habituelle de signaux de communication.

L'émission d'un message peut être discrète puisqu'exécutable bouche fermée.

Le parasitage par l'environnement est quasi inexistant.

Le captage est assez simple au moyen d'un glottomètre ou d'un microphone de contact.

De plus l'apprentissage peut être bref puisque,comme nous le verrons par la suite,le nombre d'ordres peut être très réduit ,tout en conférant au robot une grande latitude de mouvements,grâce à l'emploi de règles simples et peu nombreuses pour la construction d'un message.

L'inconvénient majeur de ce signal réside dans le fait qu'il apparait aussi lorsque le patient parle.On sera donc amené soit à choisir un codage éliminant les interférences avec la parole,soit, plus simplement,à disposer d'un signal,de préférence non vocal,permettant d'activer ou de désactiver le coordinateur.

2-6-2 Caractéristiques utilisables

On peut à priori envisager de moduler plusieurs paramètres du signal laryngien pour coder une information:

- Le spectre est à écarter parce que difficilement maîtrisable et,de surcroît altérable (affections de la gorge).

- La durée seule:sa modulation en utilisant un codage du genre Morse ne semble pas convenir:ce type de code est difficile à retenir et engendre des messages longs à émettre.

- La hauteur absolue du fondamental:sa modulation n'est possible que par des sujets éduqués capables de "chanter" juste bouche fermée,ce qui n'est pas facile;de plus la répétitivité dans le temps est douteuse.

- La hauteur relative du fondamental est intéressante.On peut en effet associer des variations simples de hauteur,que nous appellerons traits mélodiques,pour former des motifs mélodiques correspondant à des mots;il est alors possible,avec un nombre réduit de traits,de former un grand nombre de mots.

2-6-3 Modulation choisie

Pour simplifier l'émission des ordres et la mémorisation du langage,nous avons choisi de combiner des variations ternaires de hauteur relative avec une modulation binaire en durée.

Les variations de hauteur permettent de définir trois traits simples:

-trait constant: note tenue quelconque,en admettant éventuellement une variation d'un demi ton de part et d'autre de la valeur moyenne.

-trait ascendant: sera considéré comme tel tout voisement continu dont la variation de hauteur sera croissante sur un intervalle d'au moins deux tons.Représentation: /

-trait descendant: symétrique du précédent(variation décroissante.
Représentation: \

Les variations de durée permettent de distinguer deux traits constants différents:

-trait constant bref: d'une durée inférieure à une demie seconde.
Représentation: +—+

-trait constant long: d'une durée supérieure à une seconde.
Représentation: ——

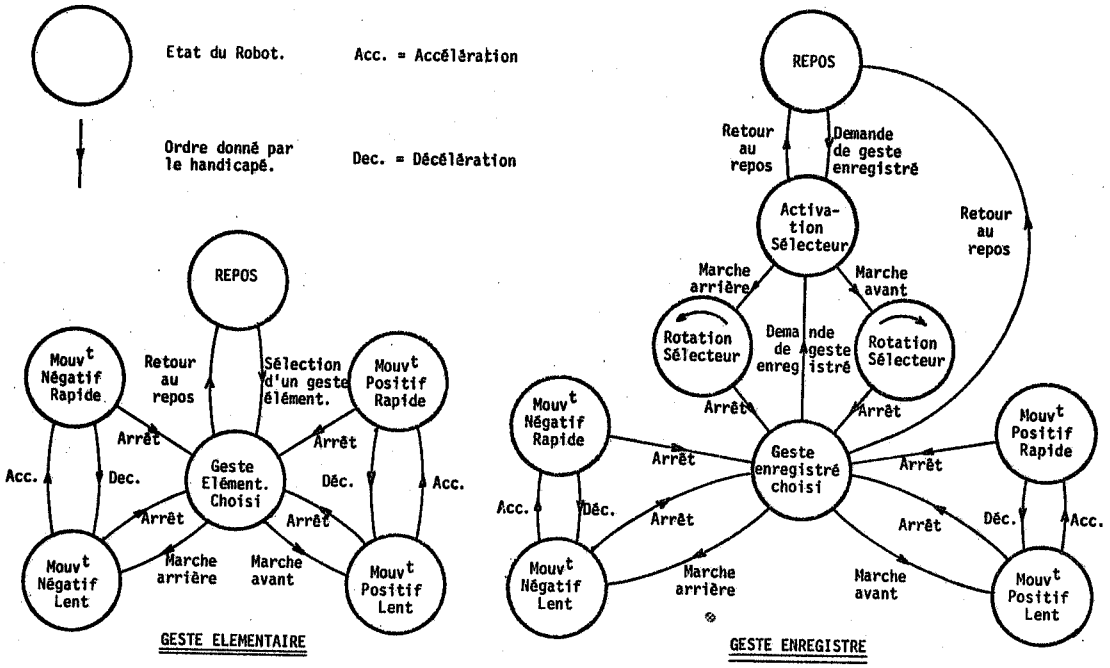
3 - STRUCTURE DU LANGAGE

3 - 1 SYNTAXE

On considère comme message ou phrase la suite des ordres nécessaires à définir et à conduire un geste (élémentaire ou enregistré).

Entre deux phrases,le manipulateur est arrêté dans la dernière position atteinte,tandis que le coordinateur se trouve dans l'attente d'un nouveau message.

Le support choisi imposant une présentation série de l'information et la commande devant être assurée au fur et à mesure de l'exécution du mouvement,nous avons été conduits à exprimer le déroulement temporel des gestes par les graphes de la figure ci-dessous,qui représentent les modifications de l'état du robot entrainées par les ordres donnés par le handicapé.



GRAPHE DES TRANSITIONS ENTRE ETATS DU ROBOT

On peut déduire de ce graphe les règles syntaxiques suivantes:

- -Toute phrase commencera par la demande d'un geste (élémentaire ou enregistré).
- -Suivra l'indication de sens (qui déclenchera le mouvement à petite vitesse).
- -Une demande d'inversion de sens ne pourra intervenir qu'après un ordre d'arrêt.
- -Une modification de la vitesse ne sera autorisée qu'en cours de mouvement.
- -Une décélération ne pourra être demandée qu'après une accélération.
- -Un signal d'arrêt sera pris en compte à tout instant.
- -Toute phrase se terminera par un ordre de mise au repos.

Une telle structure de phrase autorise tous les enchaînements souhaitables,voire des séquences de longueur infinie,en ne requérant qu'un vocabulaire limité.

3 - 2 VOCABULAIRE

Les phrases seront constituées d'une suite de mots (représentant

les ordres)séparés par des silences.

Un mot sera obtenu par enchaînement continu (sans arrêt de voisement) de traits mélodiques choisis parmi les quatre définis en 2-6-3

3-2-1 Lexique

On doit tout d'abord disposer d'autant de mots que de gestes possibles:

Déplacement X------Mot Mx Rotation A------Mot Ma
Déplacement Y------Mot My Rotation B------Mot Mb
Déplacement Z------Mot Mz Rotation C------Mot Mc
Déplacement dans l'axe de la main D----------Mot Md
Ouverture/Fermeture de la main P-------------Mot Mp
Geste enregistré E---------------------------Mot Me

Compte tenu des règles syntaxiques énoncées plus haut,on peut utiliser le même mot de Stop pour l'arrêt et la remise au repos: Mot Ms.

De plus une modification de la vitesse ne peut être demandée qu'en cours de mouvement donc après spécification du sens;ces deux groupes d'ordres,sens et modification de la vitesse,sont par ailleurs de même nature:sens positif ou négatif,aller "plus" ou "moins" vite. On peut donc utiliser pour les deux groupes la même paire de mots,leur ordre chronologique levant l'ambigüité sémantique:

M+ -- Sens positif / Accélération M- -- Sens négatif / Décélération

La commande peut ainsi être assurée au moyen d'un vocabulaire limité à 12 mots.

3-2-2 Morphologie (Codage mélodique)

Les codes de mots doivent être aussi simples et brefs que possible afin de faciliter leur mémorisation et leur réalisation par le handicapé.

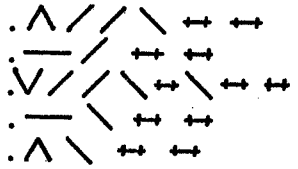
Pour optimiser le temps d'émission d'une phrase et le confort du patient,les mots les plus brefs seront Ms,M+ et M- (effet immédiat), les plus longs ou les plus complexes correspondant aux demandes dont la fréquence d'occurrence est la plus basse.

Une évaluation de la décroissance de celle-ci permet de définir par exemple les codages donnés par le tableau ci-dessous:

Ms	M+	M-	Mp	Mx	My	Mz	Md	Ma	Mb	Mc	Me
⟷	↗	↘	—	∧	∨	≈	≈	∧	∨	≈	≈

A titre d'illustration nous donnons une suite de cinq phrases permettant d'avancer la main ouverte vers un objet,le saisir,le déplacer à droite,le lâcher et reculer la main:

Avance: Mx M+ M+ M- Ms Ms
Serrage: Mp M+ Ms Ms
Déplacement à droite: My M+ M+ M- Ms M- Ms Ms
Lâcher: Mp M- Ms Ms
Recul: Mx M- Ms Ms



Nous n'aborderons pas ici le problème de la reconnaissance automatique des motifs et des phrases par le coordinateur,mais il est clair que les règles génératives sont aussi à la base des algorithmes d'interprétation des messages.

4 - CONCLUSION

La commande d'un robot d'assistance médicale par le signal glottique semble donc assez facile à mettre en oeuvre.

Le point sur lequel on pouvait avoir des inquiétudes était le temps nécessaire à l'émission (séquentielle) des messages;or la demande la plus longue (mouvement enregistré) ne demande pas plus de deux secondes pour un sujet particulièrement lent,ce qui est parfaitement acceptable.

La structure du langage n'est pas compliquée:

- Règles élémentaires et peu nombreuses
- Vocabulaire réduit n'exigeant du handicapé ni aptitudes particulières ni long apprentissage.

Ce dernier peut d'ailleurs être grandement facilité par visualisation des motifs mélodiques au moyen d'un analyseur de mélodie (1)

++++++++++++

BIBLIOGRAPHIE

(1) P.ANDRE - M.FILLEAU - F.HAMELIN
"Glottométrie en temps réel - Applications"
7° Journées d'étude sur la Parole - NANCY Mai I976

(2) P.ANDRE - F.LHOTE
"Analyseur-Synthétiseur de fréquences musicales"
14th Conference on Acoustics - BRATISLAVA Octobre I976

(3) J.M.BADOZ - G.MONIN-BAROILLE - J.PEPIOT - J.P.SASSARD
"Prothèse de bras"
Projet de fin d'études E.N.S.C.M.B. 1976

(4) J.CLOT - Z.STOJILJKOVIC
"Fonctionnement d'une prothèse de main équipée d'un revêtement sensible"
Colloque BIOMECA II - TOULOUSE Novembre 1976

(5) I.R.I.A.
"Le Projet pilote SPARTACUS"
Bulletin de liaison de la recherche en Informatique et Automatique - 5 Novembre 1976

(6) H.H.KWEE
"Quelques repères de commande pour tétraplégiques"
Projet pilote SPARTACUS - Rapport Novembre 1976

VIII èmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

PROSODIE AUTOMATIQUE POUR LA SYNTHESE PAR DIPHONEMES

C. CHOPPY, J.S. LIENARD

Laboratoire d'Informatique pour la Mécanique
et les Sciences de l'Ingénieur (C.N.R.S.)

• Résumé : La prosodie relève simultanément de plusieurs niveaux d'analyse, verbaux (phonétique, lexical, syntaxique et sémantique) ou non verbaux (esthétique, diagnostique). On tente de montrer ici, par analyse et synthèse, que les principales indications prosodiques implicites dans un texte écrit peuvent être déduites de considérations simples sur les mots ou groupes de mots et sur les ponctuations.

On montre tout d'abord qu'il existe une prosodie moyenne - ou normative - adoptée par plusieurs locuteurs à la lecture d'un même texte. Ce résultat permet de dégager des règles pour un algorithme de prosodie automatique intégré au processus de synthèse par diphonèmes mis au point au L.I.M.S.I.

Le texte est découpé en groupes de mots à partir de critères portant sur la longueur des mots(en syllabes) et sur une quarantaine de mots grammaticaux. Les paramètres prosodiques (F_o et durée) sont obtenus par combinaison de schémas élémentaires, les uns relatifs aux groupes de mots, les autres relatifs à l'ensemble de la phrase et aux ponctuations. L'algorithme ne comporte pas d'analyse syntaxique ; il peut être utilisé quelle que soit la structure de la phrase, même si celle-ci est agrammaticale, et ne requiert aucune intervention manuelle autre que la frappe du texte écrit.

• Summary : Several levels are jointly relevant in prosody : verbal levels (phonetic, lexical, syntactic and semantic) and non verbal levels (aesthetic, diagnostic). The attempt made here is to show, by means of analysis and synthesis, that the main prosodic indications, that are implicit in a written text, may be derived from simple considerations on words, word groups and punctuations.

First, it is shown that a standard-prosody is taken by several speakers reading the same text. This result allows to bring out a set of rules for an automatic prosody algorithm which is combined with the diphone synthesis process elabored at L.I.M.S.I.

The text is partitioned in word groups using criteria based on word length (in syllables)and on a list of frequent grammatical words. Prosodic parameters (F_o and duration) are obtained by combining elementary patterns, these patterns being associated with word groups, or with the whole sentence and punctuations. There is no syntactic analysis in the algorithm : it can be used whatever the sentence structure, even if agrammatical. The only manual intervention required is the typing of the text.

I - Introduction

Le synthétiseur utilisé au L.I.M.S.I. fonctionne sur le principe de synthèse par diphonèmes (bib. 1, 2). La voix synthétisée est parfaitement intelligible, mais de qualité assez artificielle, inconvénient dû d'une part au type de synthétiseur (Icophone, type relecteur de sonagramme), et d'autre part à l'absence de variations prosodiques.

L'étude présentée ici est l'élaboration d'un programme qui commande les variations prosodiques pour le synthétiseur. Ce programme est intégré au processus de synthèse et permet d'obtenir automatiquement une synthèse vocale comprenant des variations prosodiques.

Cette prosodie est déduite du texte écrit ; le traitement doit être rapide et peu encombrant de façon à pouvoir être intégré à une unité à réponse vocale qui fonctionne en temps réel, tel l'Icophone V (bib. 3).

La prosodie qu'on cherche à obtenir est une prosodie moyenne composée des variations nécessaires pour que la voix synthétique "sonne français", et que l'intelligibilité et le naturel du message soient améliorés.

La prosodie relève de plusieurs niveaux d'analyse : phonétique, lexical, syntaxique, sémantique, esthétique, diagnostique.

Les variations de tous ces paramètres engendrent une infinité de prosodies différentes ; on peut toutefois observer certaines constantes dans la prosodie de plusieurs locuteurs, notamment lors de la lecture d'un même texte.

Une première partie de cette étude consiste à dégager par analyse cette prosodie moyenne réalisée en moyenne. Les grandeurs étudiées sont la durée et la fréquence. Le relevé et le traitement des données mettent en oeuvre une triple normalisation. Le traitement met en évidence une prosodie moyenne représentative des variations prosodiques communes à tous les locuteurs.

Dans la deuxième partie nous décrivons un algorithme qui permet d'engendrer une telle prosodie. Le programme est intégré au traitement de synthèse à partir du texte, et est entièrement automatique, c'est-à-dire que la seule donnée est le texte écrit en français, à l'exclusion de toutes autres indications.

II - Mise en évidence d'une prosodie moyenne (bib. 4)

Le corpus utilisé est constitué par les enregistrements de 6 locuteurs, 3 hommes et 3 femmes, qui ont lu un même texte à voix haute. Ce texte est à contenu purement informatif et comprend des phrases de longueurs et de structures diverses. Les locuteurs n'ont pas d'accent étranger ou régional, et n'ont pas bafouillé ou sauté de mots au cours de cette lecture.

Les enregistrements sont dépouillés au sonagraphe. Les données en temps et en fréquence sont mesurées par rapport à chaque unité phonétique prononcée (fig. 1) : pour chaque phonème, on mesure en abscisses l'instant T_1 de début du phonème, l'instant T_3 de fin du phonème, et l'instant T_2 correspondant au maximum de fréquence (si un maximum n'apparaît pas, T_2 est pris au milieu du segment $[T_1 , T_3]$). On mesure également les fréquences F_1 , F_2 , F_3 , correspondant aux instants T_1 , T_2 , T_3 . Cette méthode de mesure des paramètres par rapport aux unités phonétiques correspondantes est originale, elle diffère considérablement des relevés classiques en fréquence par rapport au temps qui ne permettent pas une comparaison interlocuteurs. C'est là notre première normalisation.

Sur les courbes ainsi obtenues la comparaison n'est toujours pas très aisée. La figure 2 montre un relevé de fréquence pour deux locuteurs et pour une même phrase. L'ensemble des valeurs prises par la courbe (2) est inférieur à celui observé pour (1), et les variations pour (2) sont de faibles amplitudes alors qu'elles sont plus importantes pour (1).

La normalisation en fréquence consiste pour chaque courbe à fixer une valeur arbitraire pour le maximum, soit 110, et une valeur pour le minimum, soit

10, et à calculer par interpolation linéaire les valeurs intermédiaires. Les valeurs maximum et minimum fixées sont les mêmes pour toutes les courbes. Sur la figure 3 sont tracées les courbes de fréquences normalisées correspondant aux courbes de fréquence de la figure 2. Les zones de variations communes aux deux locuteurs apparaissent par ce traitement de normalisation. Les courbes normalisées sont facilement comparables et peuvent faire l'objet d'un calcul de moyenne.

La courbe de fréquence moyenne (fig. 4) calculée sur les fréquences normalisées des 6 locuteurs est représentative des variations communes adoptées par tous les locuteurs. C'est un indicateur qui permet de décrire la prosodie en réduisant au minimum l'aspect individuel, et en mettant en évidence les variations essentielles, communes à plusieurs locuteurs.

Un traitement identique est appliqué en ce qui concerne la durée, à savoir normalisation et calcul de courbe moyenne sur les courbes normalisées. Là encore la courbe moyenne de durée est une bonne image des variations communes à tous les locuteurs (fig. 5).

Dans certains cas on peut observer deux stratégies prosodiques différentes. Par exemple, pour les phrases à syntaxe interrogative il y a une remontée fréquentielle finale chez la moitié des locuteurs étudiés (fig. 6). Dans ce cas, en effet, cette marque prosodique de l'interrogation est redondante et facultative. La courbe moyenne n'est alors pas représentative d'une prosodie moyenne adoptée par tous.

Ce type de courbes doit être utilisé avec précaution, et il faut vérifier sa représentativité sur les courbes individuelles. Une fois cette vérification faite, on dispose d'une bonne image de prosodie moyenne dans la majorité des cas.

La mise en évidence de cette prosodie moyenne permet d'isoler, parmi les différentes prosodies possibles, les variations essentielles dont il importe de rendre compte pour la prosodie d'un synthétiseur.

III - Algorithme de prosodie automatique

Cet algorithme déduit du texte la prosodie commandée au synthétiseur, il est conduit à partir de règles qui font la relation entre une prosodie et le texte support de cette prosodie. Il est intégré au processus général de synthèse par diphonèmes du L.I.M.S.I. (bib. 1, 2), où on peut distinguer (i) l'étape de conversion du texte écrit en suite phonétique, (ii) la conversion de la suite phonétique en parole.

1. Variations de fréquence

La courbe de variation de fréquence pour une phrase énonciative (fig. 7) est caractérisée par une allure générale comprenant 3 niveaux fréquentiels (bib. 5, 6), et des groupes de variation autour de cette allure générale.

Pour une phrase interrogative, la courbe de fréquence diffère par une remontée finale de fréquence (bib. 7).

Les règles établies rendent compte de l'allure générale, des groupes de variation, et des particularités associées aux ponctuations (bib. 4, 8).

Les groupes de variation ne portent généralement pas sur chaque mot du texte, mais sur des groupes de mots. La première étape du traitement est le découpage du texte en groupes de mots qui sont supports de ces groupes de variations.

1. a) Constitution de groupes de mots

Pour une même phrase plusieurs découpages en groupes de mots sont possibles. Tous devront comprendre au moins un "mot-idée", un mot qui contient une information, et les mots qui l'accompagnent, généralement des mots grammaticaux.

Pour que le traitement prosodique ne soit pas trop grossier, il faut éviter

de prendre des groupes de mots trop importants.

Les critères de regroupement des mots sont leur longueur (c'est-à-dire le nombre de syllabes), et leur appartenance à une liste de 40 mots grammaticaux fréquents.

Ainsi, par exemple, pour la phrase : "Elle a acheté une robe très courte.", "Elle" et "a" appartiennent à la liste et sont regroupés à "acheté", "une" appartient à la liste et est regroupé à "robe", "très" est monosyllabique et est regroupé à "courte", ce qui donne : " Elle a acheté une robe très courte. ".

La phrase est ensuite convertie en suite phonétique.

1. b) Détermination des extremums de groupes

Les maximums de fréquence pour les groupes de variation sont situés sur la première voyelle pour les groupes en position finale (suivis de ponctuation) et sur la dernière voyelle prononcée pour les groupes en position non finale.

Dans cette étape, le programme effectue le repérage en abscisses du début, du maximum et de la fin du groupe de variation, ainsi pour la phrase précédente les indices correspondants sont : / ɛl a aʃə te yn rob trɛ kuurt. /

1 8 9 12 13 14 16 22

8

1. c) Elaboration de la courbe finale

Différentes étapes conduisent à la courbe de variation de fréquence : calcul des schémas intonatifs des groupes de variation, calcul du schéma d'allure générale, micromélodie, traitement des ponctuations.

Le calcul des schémas des groupes de variation est effectué linéairement à partir des indices calculés dans l'étape 1.b) et des trois niveaux fréquentiels de l'allure générale (fig. 7).

Le calcul du schéma d'allure générale est effectué de même en prenant le point maximum sur le maximum du premier groupt (fig. 8).

Les schémas de groupes et le schéma d'allure générale sont combinés par une moyenne géométrique.

Un traitement micromélodique affecte aux consonnes une valeur inférieure à celle des voyelles.

Un traitement de ponctuations introduit des modifications selon la ponctuation rencontrée.

Pour un point final, la descente de fréquence est accentuée à partir de l'avant dernière voyelle prononcée.

Pour un point d'interrogation, une remontée finale de fréquence a lieu sur la dernière voyelle de la phrase.

Pour un membre de phrase suivi de virgule, une remontée de fréquence a lieu sur la dernière voyelle du membre de phrase.

Un lissage est enfin appliqué à la courbe obtenue (fig. 8).

2. Rythme : durée et pauses

Pour des raisons de construction et de codage le maniement des variations de durée est beaucoup moins souple que celui des variations de fréquence.

Les variations de rythme sont pour l'instant limitées à (i) un redoublement de la voyelle finale des mots précédant une ponctuation (traitement prévu initialement et effectué par le programme de transcription phonétique), (ii) un allongement des voyelles en fin de groupes de mots.

Des pauses sont introduites par le programme qui effectue la conversion de la suite phonétique en parole, pauses qui interviennent au niveau des ponctuations, leur durée varie en fonction du type de ponctuation rencontré.

3. Résultats - Discussion

3. a) Résultats

Deux critères ont été utilisés pour évaluer les résultats obtenus : un critère graphique de comparaison de courbes et des tests auditifs.

En ce qui concerne le critère graphique, il s'agit de comparer la courbe de fréquence calculée par l'algorithme, et la courbe moyenne des courbes normalisées (§ II) sur 6 locuteurs. On constate (fig. 9) que les variations majeures sont bien reproduites par l'algorithme.

Pour les tests auditifs, quelques problèmes étaient posés par la qualité trop artificielle de synthèse, gênante pour évaluer les résultats relatifs à la prosodie. Les auditeurs ont entendu des phrases synthétisées sans prosodie puis avec prosodie et ont porté dans l'ensemble une apréciation positive sur les phrases synthétisées avec prosodie.

En résumé, la prosodie réalisée est assez voisine de la prosodie moyenne dégagée par analyse et elle semble améliorer la qualité de synthèse pour une large majorité d'auditeurs.

3. b) Discussion

L'utilisation d'une moyenne sur plusieurs locuteurs, qui peut paraître artificielle, nous a permis de dégager des éléments prosodiques représentatifs pour la langue au delà des particularités individuelles des locuteurs. Il serait souhaitable toutefois, de pouvoir effectuer le même type d'expérimentation pour un nombre plus grand de locuteurs.

Le programme de traitement prosodique ne comprend pas d'analyse syntaxique. Il ne s'agit pas ici de nier toute importance de la syntaxe, mais de dire que le rôle de la syntaxe n'est pas premier et unique. Les questions que nous posons sont (i) d'ordre théorique : aucune démonstration n'a été faite qu'un enfant de 3 ans dirigerait sa prosodie à partir d'une connaissance innée (ou acquise dans les premières années de la vie) de la syntaxe, (ii)d'ordre pratique : la prise en compte d'une analyse syntaxique dans un tel système de synthèse pose des problèmes importants relatifs à la rapidité du traitement et à l'encombrement de taille mémoire, et ne permettrait pas de traiter n'importe quelle phrase (phrases agrammaticales). Il semble que pour des phrases ambigues, la prise en compte de la syntaxe serait un élément déterminant pour la prosodie. Dans la plupart des cas, cette prise en compte de la syntaxe doit être assortie d'une analyse sémantique. Mais le cas des phrases ambigues qui sont peu nombreuses, est un cas particulier et ne saurait être généralisé.

IV - Conclusion

L'algorithme décrit ci-dessus déduit du texte écrit la prosodie par une série d'opérations simples. Cette simplicité du traitement permet d'envisager son intégration dans une unité à réponse vocale qui fonctionne en temps réel.

Cet algorithme est intégré au traitement général de synthèse pour obtenir automatiquement une synthèse vocale comprenant des variations prosodiques. La prosodie réalisée est une prosodie moyenne qui améliore le naturel et l'intelligibilité du message. Nous n'avons pas cherché à reproduire une voix particulière, mais des variations prosodiques communes à plusieurs locuteurs, qui puissent servir de référence pour un traitement automatique du français.

Bibliographie

(1) E. LEIPP, M. CASTELLENGO, J.S. LIENARD - La synthèse de la parole à partir de digrammes phonétiques - C.R. du 6e Congrès International d'Acoustique - Tokyo, Août 1968

(2) J.S. LIENARD, D. TEIL - Les éléments phonétiques et la traduction automatique du message écrit en message parlé - Revue "Automatisme", n° 10, Octobre 1970

(3) D. TEIL - Conception et réalisation d'un terminal à réponse vocale - Thèse de Docteur-Ingénieur - Université Paris VI, 1975
(4) C. CHOPPY - Introduction de la prosodie dans la synthèse vocale automatique Thèse de Docteur-Ingénieur - Université Paris VI, 1977, Mars
(5) P. DELATTRE - Studies in French and Comparative Phonetics - Mouton et C° - London, The Hague, Paris, 1966
(6) L. BOE - Synthèse paramétrique de la phrase énonciative en français - 5e Journées d'Etude sur la Parole, Orsay, 1974
(7) A. GUNDSTROM, P. LEON - Interrogation et Intonation - Studia Phonetica 8, Didier, 1973
(8) C. CHOPPY, J.S. LIENARD, D. TEIL - Un algorithme de prosodie automatique sans analyse syntaxique - 6e Journées d'Etude sur la Parole - Toulouse, 1975

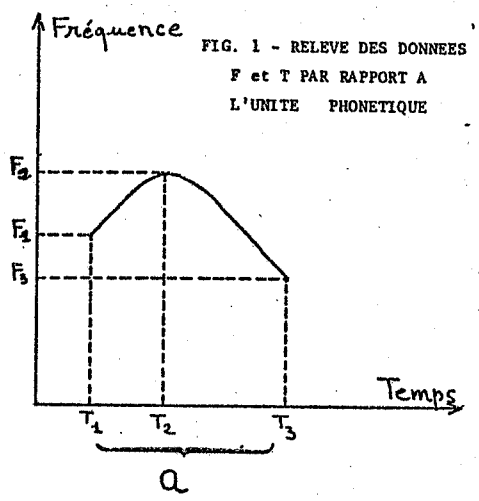


FIG. 1 - RELEVE DES DONNEES F et T PAR RAPPORT A L'UNITE PHONETIQUE

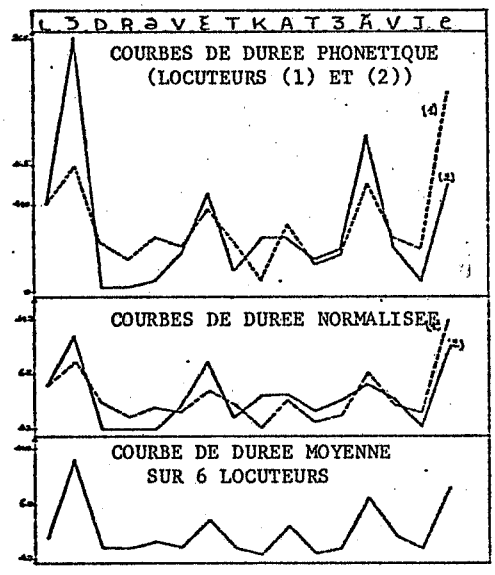


FIG. 5 TRAITEMENT DES DONNEES EN DUREES

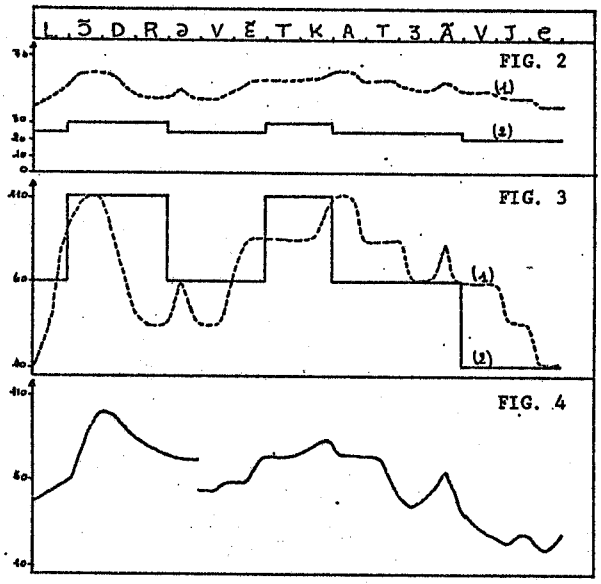


FIG. 2 - COURBES DE FREQUENCE PHONETIQUE POUR LES LOCUTEURS (1) ET (2)
FIG. 3 - COURBES DE FREQUENCE NORMALISEES POUR (1) ET (2)
FIG. 4 - COURBE DE FREQUENCE MOYENNE CALCULEE SUR LES COURBES NORMALISEES DE 6 LOCUTEURS

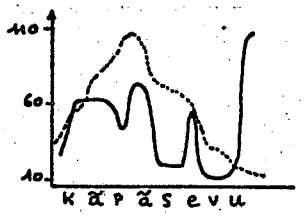


FIG. 6 - COURBES DE FREQUENCE NORMALISEES POUR LA PHRASE : "QU'EN PENSEZ-VOUS?"

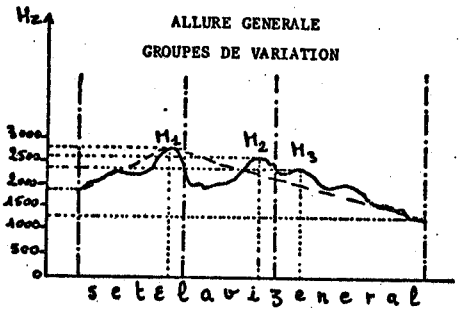


FIG. 7 - COURBE DE VARIATION FREQUENTIELLE (HARM. N° 8) POUR UNE PHRASE ENONCIATIVE M_1, M_2, M_3, MAX. DE FREQUENCE DES GROUPES

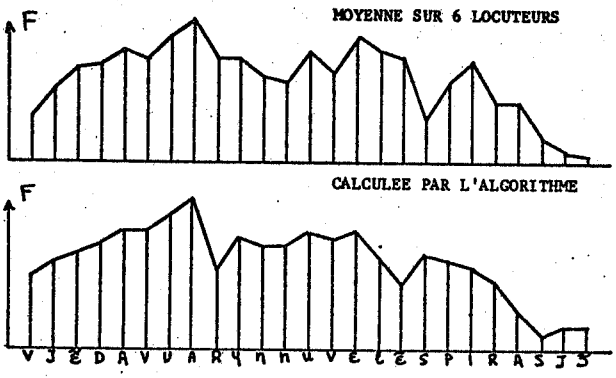


FIG. 9 - COMPARAISON ENTRE LA COURBE DE FREQUENCE CALCULEE PAR L'ALGORITHME ET LA COURBE DE FREQUENCE MOYENNE SUR 6 LOCUTEURS POUR "...VIENT D'AVOIR UNE NOUVELLE INSPIRATION".

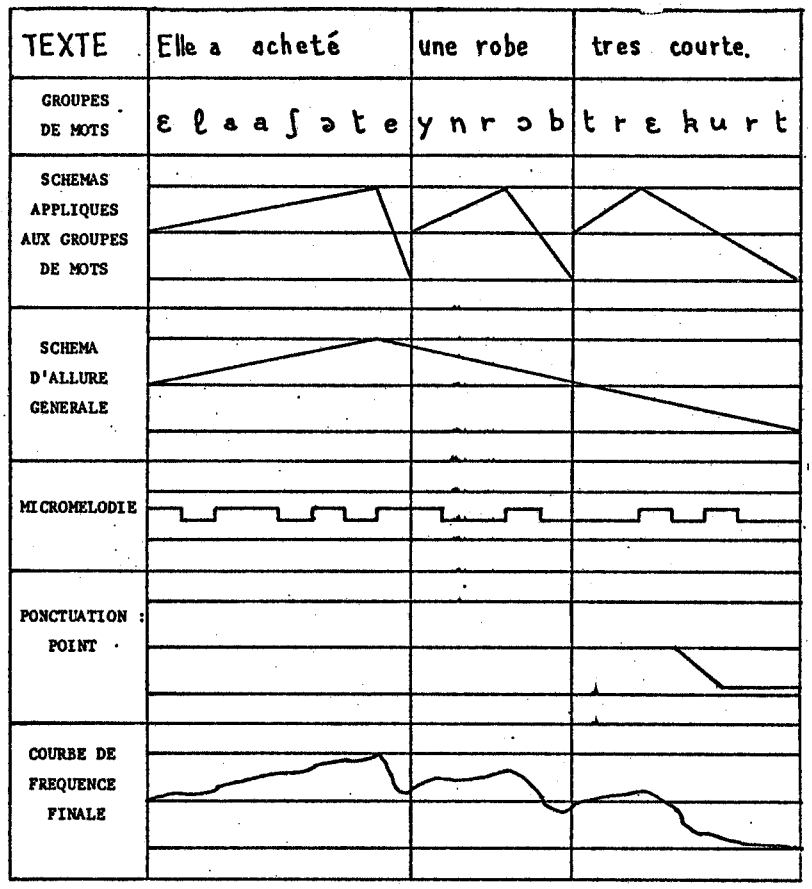


FIG. 8 - PRINCIPE D'ELABORATION DE LA COURBE DE FREQUENCE

VIII èmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

ANALYSES DES REDONDANCES DE SYSTEMES SYMBOLIQUES ET DEGRES DE VOISEMENT

Jean A. Dreyfus-Graf

Genève (Suisse)

RÉSUMÉ

L'onde acoustique de la parole est considérée comme résultant de l'interaction entre une onde modulante, dont les degrés de voisement (pitch et souffle) sont variables, et une onde modulée dont les formants (et anti-formants) sont modifiés par ces degrés de voisement. La démodulation permet de séparer l'onde modulante de l'onde modulée. L'onde de la parole véhicule 50.000 bits/sec téléphoniques vers l'oreille, mais le cerveau n'accepte que 10 bits/sec linguistiques. Comment simuler cette transformation 5000:1, et parvenir à une reconnaissance optimum de la parole ? L'analyse des redondances de systèmes symboliques est utilisée comme fil directeur. Nous distinguons les débits d'information spectraux et symboliques (bit/sec), les néguentropies (bit/symb), les redondances et les anti-redondances (bit/bit) qui caractérisent les contraintes (règles ou sources de connaissance). L'analyse des redondances mesure l'importance relative de leurs composantes, telles que spectre, dynamique, durée, mélodie, voisement, signes phonétiques, compléments linguistiques. Elle peut être appliquée à toutes les formes de la parole : continue ou isolée, naturelle ou codée, avec des vocabulaires illimités ou restreints.

SUMMARY

The acoustic speech wave results from the interaction between a modulating wave, with various voicing degrees (pitch, breath) and a modulated wave, whose formants (and anti-formants) are influenced by these voicing degrees. Demodulation allows to separate the modulating wave from the modulated one.
The speech wave carries 50.000 telephonic bits/sec towards the ear, but the brain accepts only 10 linguistic bits/sec. How may a machine simulate such a 5000:1 transformation, in order to approach an optimum speech recognition ?
The redundancy analysis of symbolic systems is used as a guide-line. We distinguish the spectral and the symbolic information capacity (bit/sec), the negentropy (bit/symbol), the acceptance rhythm (symb/sec), the redundancies and anti-redundancies (bit/bit), which characterize the relevant constraints (rules or knowledge sources). Redundancy analysis measures the relative sizes of their components, such as spectrum, dynamics, duration, pitch, voicing degree, phonetic signs, linguistic complements. It may be applied to any connected or isolated speech form, with unlimited or restricted vocabularies.

Analyses des redondances de systèmes symboliques et degrés de voisement

Jean A. Dreyfus-Graf, Genève (Suisse)

1. Degrés de voisement, mélodie et débits d'information minimaux

Avant de développer l'analyse des redondances de systèmes symboliques, il faut faire 2 remarques préliminaires :

a) Degrés de voisement et mélodie

La plupart des systèmes d'analyse spectrale, proposés récemment pour parvenir à la reconnaissance automatique de la parole, sont basés sur une distinction dichotomique entre sons "voisés" et "non-voisés", c'est-à-dire entre l'excitation "périodique" (par l'onde glottale) et l'excitation "non-périodique" (ou "stochastique", par le souffle) du "conduit vocal", dont les résonances sont nommées "formants". Parmi ces systèmes, on peut citer la "déconvolution homomorphique"[1] et la "prédiction linéaire"[2]. Pourtant une observation élémentaire dénote que la distinction entre sons "voisés" et "non-voisés" n'est pas primordiale : en effet, la parole peut être plus ou moins chuchotée, impliquant des degrés de voisement très divers, sans que l'identification des sons phonémiques s'en trouve radicalement affectée. Par exemple, les voyelles sont reconnaissables aussi quand elles sont chuchotées[3].

Nous proposons de considérer l'onde acoustique de la parole comme résultant de l'interaction entre une onde modulante, dont les degrés de voisement (pitch et souffle) sont variables, et une onde modulée, dont les formants (et anti-formants) sont modifiés progressivement par ces degrés de voisement. Une transmission téléphonique éventuelle limite le spectre de l'onde modulée entre 200 et 3400 Hz, tout en préservant ses régions formantiques principales. Elle ne limite pas, vers le bas, le spectre de l'onde modulante, dont les composantes significatives (partiellement sub-formantiques[4]) peuvent être les suivantes :

a) 400 à 80 Hz pour la fréquence fondamentale de sa partie voisée (mélodie)
b) 80 à 40 Hz pour les composantes de sa partie bruitée (souffle)
c) 40 à 30 Hz pour les composantes de sa partie roulée (r)

Nous proposons de récupérer les composantes sub-formantiques de l'onde modulante par démodulation des composantes formantiques à l'onde modulée. Ceci en comparant leurs actions physiques (= énergie X temps) relatives[5]. Dès lors, les spectres de formants et anti-formants peuvent être corrigés en fonction des degrés de voisement par les valeurs significatives de la mélodie et du souffle.

b) Débits d'information spectraux et phonétiques minimaux

Quels sont les débits minimaux d'information spectrale qui sont encore admissibles pour la reconnaissance automatique de la parole ? On doit faire appel à l'oreille humaine[6], donc à la reconnaissance subjective (l'intelligibilité), en fonction de la réduction progressive du débit d'information fourni par un "vocoder". Son débit d'information C(bit/sec) = B(bit) X F(Hz) dépend de la quantité d'information $B = \log_2 N$ (où N = canaux X niveaux) et de la fréquence moyenne d'exploration F(Hz), qui est la valeur inverse de la fenêtre de temps (ms).

Quels sont les nombres minimaux de signes phonétiques qu'on peut admettre pour une première transcription des variations spectrales ? Pour le français continu, nous admettons $32 = 2^5$ signes alphabétiques (phonémique) et $32 = 2^5$ signes auxiliaires (prosodie), avec une cadence de 10 symboles par seconde. Il en résulte un débit d'information phonétique D_p = 10(bit/symb) X 10(symb/sec) = 100 (bit/sec), représentant une réduction supplémentaire 1:24 par rapport au débit minimal d'information spectrale D_v = 2400(bit/sec).

Le vocoder CIPHON[7] admet un débit d'information C = 1200(bit/sec) = 32(bit) X 37,5(Hz), "sans dégradation notable de l'intelligibilité". Selon le TABLEAU I, nous adoptons un débit minimum [6,8] de vocoder D_v = 2400(bit/sec) = 30(bit) X 80(Hz), où nous séparons les 5 composantes suivantes : fréquences (formants plosifs et non-plosifs), énergie (dynamique), durée (rythme), mélodie (pitch), voisement (degré).

TABLEAU I	DEBITS D'INFORMATION SPECTRAUX D_v ET PHONETIQUES D_p MINIMAUX (bit/sec) base : vocoder D =2400(bit/sec), fréquence moyenne exploration F_e=80(Hz)							
VARIATION DE:	FREQUENCE F(Hz) spectre,formants E(J)= f(F)		ENERGIE dynamique E(J)=f(T)	DUREE T(s)	MELODIE pitch F_0(mel)	VOISEMENT (souffle) E_0(J) %		TOTAL
VOCODER	plosifs	non-plos.	globale	rythm.	tons	degré		
canaux x niveaux = N_v	8 8 (2) 64(16)	32 8 x 256 x	1 16 16 x	1 4 4 x	128 2 256 x	1 4 4	= 10^9	
$\log_2 N_v = B_n$	6 (4)	8	4	2	8	2	bit	30
fréqu. F_e	80(120)	80	80	80	80	80	Hz	80
débit D_v	480	640	320	160	640	160	bit/sec	2400
PHONETIQUE	phonèmes		prosodie					
32 phonèm.=2^5	5(bit/symb)		--					
32 symb.aux.=2^5	--		5(bit/symb)					
cadence moy.	10(symb/sec)		10(symb/sec)					
débit D_p	50		50				bit/sec	100

Toutefois les analyses spectrales et phonétiques minimalisées ci-dessus ne concernent qu'une partie des éléments nécessaires à la reconnaissance de la parole.
Peut-on trouver une méthode d'optimisation plus générale ?

2. L'oreille, le cerveau et les systèmes symboliques

Les machines reconnaissant la parole doivent simuler non seulement certaines fonctions d'analyse spectrale de l'oreille, mais aussi diverses fonctions de mémoire, de calcul et de logique du cerveau. Comment sélectionner les meilleures combinaisons de ces fonctions ? Certains auteurs attachent plus d'importance à l'oreille[9], d'autres au cerveau[10]. Concernant la reconnaissance de la parole continue, avec de grands vocabulaires, ce domaine ressemble encore à un labyrinthe, présentant plusieurs entrées possibles, mais aucune sortie satisfaisante. En effet, les taux d'erreur[6,11] restent plus grands que 10%, même en temps différé, alors qu'ils devraient devenir plus petits que 0,1% en temps réel (soit 100 fois meilleurs). Sinon, la commande vocale ne peut pas concurrencer la commande manuelle des machines.
Nous allons examiner l'entrée nommée "débit d'information", mesurable en bits/sec.

En direction de l'oreille, l'onde de la parole peut véhiculer 144.000 bits/sec acoustiques, ou 50.000 bits/sec téléphoniques. Mais le cerveau n'accepte que 100 bits/sec phonétiques, voire 10 bits/sec linguistiques. Comment optimiser finalement cette transformation 5000 : 1 ? Nous considérons la parole comme un système symbolique, et la machine de reconnaissance comme un transformateur de débit d'information.
Nous proposons "l'analyse des redondances" comme fil conducteur général.

3. L'analyse des redondances : théorie et formules

Le TABLEAU II résume la théorie et les principales formules (1) à (7) concernant l'analyse des redondances[6,8]. Selon les formules (1) et (2) nous distinguons 2 débits d'information : le débit spectral D_d et le débit symbolique D_s (bit/sec). Le débit d'information spectral D_d transmet d degrés d'amplitude à la fréquence d'échantillonnage élevée $F_d = 2.L(Hz)$, où L est la largeur de bande du canal transmetteur. Par contre, le débit d'information symbolique D_s transmet s symboles à la fréquence basse (acceptée par le cerveau) F_s (symb/sec). La jonction entre ces 2 débits est opérée par l'entropie négative N_s (bit/symb), si tous les symboles sont équiprobables. S'ils ne sont pas équiprobables, la formule (3) définit l'entropie négative moyenne, ou néguentropie N (bit/symb). Selon la formule (4) la redondance R_{xy} mesure la transformation de néguentropie quand un débit d'information D_x se transforme en un autre, (plus faible) D_y.

Toute composante de redondance caractérise une contrainte (règle ou source de connaissance) correspondante, telle que "phonémique", "phonétique", "prosodique", "lexicale", "syntaxique" ou "sémantique". La valeur de chaque composante de redondance est comprise entre un maximum "un" et un minimum "zéro". Comment évaluer la résultante de plusieurs redondances, connectées par exemple en parallèle (comme "phonémiques" et prosodiques"), ou en série (comme "lexicales" et "syntaxiques") ? Dans ce but, et selon la formule (5), nous utilisons les anti-redondances $A_{xy} = 1 - R_{xy}$, qui sont les compléments à "un" des redondances. La résultante de redondances en série ou en parallèle est exprimée par le produit ou par la somme des anti-redondances correspondantes, selon les formules (6) et (7).

TABLEAU II | FORMULES POUR L'ANALYSE DES REDONDANCES DE SYSTEMES SYMBOLIQUES

a) Débit d'information spectral D_d d'un canal (largeur de bande L(Hz)), transmettant d degrés d'ammplitude à la fréquence $F_d = 2.L(Hz)$:

(1) $\boxed{D_d(\text{bit/sec}) = B_d(\text{bit}).\ F_d(Hz)}$ où $B_d = \log_2 d$ (bit), ou dynamique (dB)

b) Débit d'information symbolique D_s d'un canal transmettant s symboles équiprobables (p_i=0,5) à la fréquence F_s (symb/sec), avec $N_s = \log_2 s$ (bit/symb) = entropie négative

(2) $\boxed{D_s(\text{bit/sec}) = N_s(\text{bit/symb}).F_s(\text{symb/sec})}$

Si les s symboles ne sont pas équiprobables, avec $p_i \neq 0,5$:

(3) $\boxed{N = -\sum p_i.\log_2 p_i\ (\text{bit/symb})}$ = entropie négative moyenne = néguentropie

c) Redondance R_{xy}, quand le débit d'information $D_x = N_x.F_x$ (bit/sec) passe à un autre débit d'information (plus petit) $D_y = N_y.F_y$(bit/sec), avec $F_x = F_y$ et $N_x \geqslant N_y$

(4) $\boxed{R_{xy} = 1 - N_y/N_x\ (\text{bit/bit})}$, où N_y/N_x= rapport des redondances
= anti-redondance = $A_{xy} = 1 - R_{xy}$

d) Anti-redondance (= 1 - redondance)

(5) $\boxed{A_{xy} = 1 - R_{xy} = N_y/N_x\ (\text{bit/bit})}$

e) Redondances $R_1, R_2, \ldots R_n$ en série = S_n
La résultante S_n de n redondances $(R_1, R_2, \ldots R_n)$ en série correspond au produit des n anti-redondances $(A_1, A_2, \ldots A_n)$

(6) $\boxed{S_n = 1 - (A_1.A_2 \ldots A_n)}$

f) Redondances $R_1, R_2, \ldots R_n$ en parallèle = P_n
La résultante P_n de n redondances $(R_1, R_2, \ldots R_n)$ en parallèle correspond à la somme des n anti-redondances $(A_1, A_2, \ldots A_n)$

(7) $\boxed{P_n = 1 - (A_1 + A_2 + \ldots + A_n)}$

4. Analyse des redondances du français continu (25000 mots)

A titre de premier exemple, selon le TABLEAU III, nous appliquons les formules (1) à (7) à l'analyse des redondances devant orienter la reconnaissance du français continu, comprenant un dictionnaire de 25.000 mots. Selon la formule (1), le débit d'information acoustique initial D_a = 144000 (bit/sec) tombe succesivement à D_t = 50000, D_v = 2400, D_p = 100, D_c = 10 (bit/sec), après avoir été soumis aux contraintes d'une transmission téléphonique (t), d'un vocoder (v), des signes phonétiques (p) et de compléments linguistiques (c). (Vocoder signifie "analyseur spectral minimal". Les "contraintes" peuvent être nommées aussi "règles" ou "sources de connaissances"). Cependant, du début à la fin, le débit d'information symbolique D_s = 10 (symb/sec) reste inchangé. En application des formules (2) et (3), la néguentropie initiale est N_a = 144000(bit/sec) : 10 (symb/sec) = 14400 (bit/symb). Elle tombe successivement à N_t = 5000, N_v = 240, N_p = 10, N_c = 1 (bit/symb). En appliquant les formules (4) et (5), les anti-redondances A et les redondances R deviennent : $A_{at} = N_t/N_a = 5000/14400 = 0,35$, $A_{tv} = N_v/N_t = 0,048$,

$A_{vp} = N_p/N_v = 0,042$, $A_{pc} = A_c/A_p = 0,1$, $R_{at} = 1-A_{at} = 1-0,35 = 0,65$, $R_{tv} = 0,952$,

$R_{vp} = 0,958$, $R_{pc} = 0,9$.

Comme les 4 composantes de redondances sont connectées en série, nous appliquons la formule (6) pour obtenir la redondance résultante R_{ac}. Celle-ci est exprimée par le complément à "un" du produit des anti-redondances (bit/bit):

$R_{ac} = 1 - (A_{at}.A_{tv}.A_{vp}.A_{pc}) = 1 - A_{ac} = 1 - (0,35.0,048.0,042.0,1) = 0,99993$

Les pourcentages de contraintes exprimés par les composantes de redondances sont les suivants : téléphone 18,8%, vocoder 27,5%, signes phonétiques 27,7%, compléments linguistiques 26%. Prenant la sortie de la transmission téléphonique comme point de départ, les ordres d'importance des 3 autres contraintes sont les suivants : vocoder (analyseur spectral minimal) 34%, signes phonétiques 34%, compléments linguistiques 32%.

5. Analyse complémentaire des redondances du français continu

L'analyse des redondances du français continu peut être détaillée comme suit :

a) La néguentropie N_v = 240(bit/sym) du vocoder résulte des 2 composantes "en parallèle", $N_{v1} + N_{v2} = N_{v7}$ = 112(bit/symb) pour "phonémique", et des 4 composantes $N_{v3} + N_{v4} + N_{v5} + N_{v6} = N_{v8}$ = 128 pour "prosodique" (incluant "dynamique", "durée", "mélodie" et "voisement").
En appliquant la formule (7), la redondance R_{tv} est exprimée par le complément à "un" de la somme des anti-redondances :
$R_{tv} = (A_{v1} + A_{v2} + ...A_{v6}) = 1 - (A_{v7} + A_{v8}) = 1 - (112+128)/5000 =$

$1 - 0,048 = 0,952.$

b) La néguentropie N_p = 10 (bit/sym) des "signes phonétiques" résulte des 2 composantes "en parallèle", N_{p1} = 5 (bit/symb) pour "signes alphabétiques" (correspondant à "phonémique"), et N_{p2} = 5 (bit/sym) pour "signes auxiliaires" (correspondant à "prosodie"). La redondance résultante R_{vp} est exprimée par le complément à "un" de la somme des anti-redondances :
$R_{vp} = 1 - (5+5)/240 = 1 - 0,042 = 0,958.$
Ces contraintes (ou sources de connaissances) permettent la "segmentation" et la réduction à la néguentropie alphabétique N_a = 5 (bit/symb), avec l'aide de l'anti-redondance intermédiaire $A_{pa} = N_a/N_p = 5/10 = 0,5.$

TABLEAU III — ANALYSE DES REDONDANCES DE LA PAROLE FRANÇAISE (25.000 mots)

débit d'information: acoustique D_a=spectral D_d= 144000 (bit/sec)
linguistique D_c=symbolique D_s= 10 (bit/sec)

néguentropie: acoustique N_a= 14400(bit/symb), linguistique N_c= 1 (bit/symb)

anti-redondance $A_{ac} = N_c/N_a = A_{at} \cdot A_{tv} \cdot A_{vp} \cdot A_{pc} = \frac{5000}{14400} \cdot \frac{240}{5000} \cdot \frac{10}{240} \cdot \frac{1}{10} = \frac{1}{14400} = 0{,}694.10^{-4}$

redondance $R_{ac} = 1 - A_{ac} = 1 - (0{,}35.0{,}048.0{,}042.0{,}1) = 1 - 0{,}694.10^{-4} = 0{,}99993$ (bit/bit)

18,8 % — 27,5 % — 27,7 % — 26 %

ACOUSTIQ./TELEPHON.
$A_{at} = N_t/N_a$ = 5000/14400= 0,35
R_{at}= 1-0,35= 0,65

TELEPHON./VOCODER
$A_{tv} = N_v/N_t$ = 240/5000=0,048
R_{tv}=1-0,048

VOCODER/PHONETIQ.
$A_{vp} = N_p/N_v$ = 10/240 =0,042
R_{vp}=1-0,042=0,958

PHONETIQ./COMPLEM.
$A_{pc} = N_c/N_p$ = 1/10 = 0,1
R_{pc}= 1-0,1 = 0,9

ACOUSTIQ. D_a=144000(bit/sec)
TELEPHON. D_t= 50000 "

VOCODER D_v=2400(bit/sec), N_v=240(bit/sec)
=112+128

PHONEM.PLOSIFS
N_{v1}=48(bit/symb)
A_{v1}= 48/5000 = 9,6.10^{-3}

PHONEM.NON-PLOS.
N_{v2}=64(bit/symb)
A_{v2}= 64/5000 = 12,8.10^{-3}

PHONEMIQUE
N_{v7}=48+64= 112
A_{v7}= 112/5000 = 22,4.10^{-3}

DYNAMIQUE (dB)
N_{v3}=32(bit/symb)
A_{v3}=32/5000 = 6,4.10^{-3}

DUREE (sec)
N_{v4}=16(bit/symb)
A_{v4}= 16/5000 = 3,2.10^{-3}

MELODIE (Hz)
N_{v5}=64(bit/sym)
A_{v5}= 64/5000 = 12,8.10^{-3}

VOISEMENT(dB/dB)
N_{v6}=16(bit/symb)
A_{v6}= 16/5000 = 3,2.10^{-3}

PROSODIE
N_{v8}=32+16+64+16 = 128
A_{v8}= 128/5000 = 25,6.10^{-3}

SIGNES PHONETIQUES
D_p= 100(bit/sec)
N_p=5+5=10(bit/sym)

SIGNES ALPHABET.
N_{p1}= 5(bit/symb)
A_{p1}= 5/240 = 20,8.10^{-3}

SIGNES AUXILIAIR.
N_{p2}= 5(bit/symb)
A_{p2}= 5/240 = 20,8.10^{-3}

SEGMENTATION PHONETIQ./ALPHABET.
$A_{pa} = N_a/N_p$ = 5/10 = 0,5

COMPLEMENTS LINGUIST.
D_c= 10 (bit/sec)
N_c= 1 (bit/symb)
$A_c = A_{c1} \cdot A_{c2} \cdots A_{c5}$ = 0,1

CONTEXTUEL
N_{c5}=1 (bit/symb)
A_{c5}= 1/1,2 = 0,83

SEMANTIQUE
N_{c4}=1,2(bit/sym)
A_{c4}= 1,2/1,5 = 0,8

GRAMMATICAL
N_{c3}=1,5(bit/sym)
A = 1,5/3 = 0,5
(>100 règles)

LEXICAL
N_{c2}= 3(bit/symb)
A_{c2}= 3/4,2 = 0,71

OCCURENTIEL
N_{c1}=4,2(bit/sym)
A_{c1}= 4,2/5 = 0,84

c) La néguentropie finale N_c = 1(bit/symb) résulte de l'adjonction "en série" des 5 composantes caractérisées par : "occurentiel", "lexical", "grammatical", "semantique" et "contextuel". L'anti-redondances résultantes A_c est le produit $A_{c1}.A_{c2}.A_{c3}.A_{c4}.A_{c5}$ de leurs composantes.

On retrouve ainsi la redondance $R_{pc} = 1 - N_c/N_p = 1-0,1 = 0,9$, complétant finalement la redondance globale (entre le départ "acoustique" et l'arrivée "linguistique"), qui était $R_{ac} = 0,99993$ (bit/bit).

6. Analyse des redondances du français restreint (100 mots isolés)

Si la parole se compose seulement de 100 mots français isolés, par exemple, l'anti-redondance "lexicale" $A_{c2} = 0,72$ (=3/4,2) diminue à $A'_{c2} = 0,45$ (=6,6/14,6), tandis que les anti-redondances des autres compléments linguistiques augmentent à "1" ($A'_{c1}=A'_{c3}=A'_{c4}=A'_{c5}=1$), c'est-à-dire qu'ils disparaissent. Partant du niveau téléphonique, les pourcentages des autres redondances deviennent R'_{tv}=39%, R'_{vp}=39%, R'_{pc}=22%.

7. Analyse des redondances de la parole codée (phonocode SOTINA)

Concernant l'analyse des redondances de la parole codée (par exemple, selon le phonocode SOTINA), on peut développer un tableau[8] analogue au TABLEAU III précédent, mais présentant les différences suivantes :

a) Le débit D_v du vocoder (analyseur spectral minimal) tombe de 2400 à 800 (bit/sec)

b) La néguentropie "phonétique" N_p tombe de 10 à 3 (bit/symb), et la "segmentation" spéciale disparait.

c) La néguentropie "linguistique complémentaire" N_c augmente de 1 à 2,5 (bit/symb).

En partant du niveau téléphonique, les pourcentages des 3 autres redondances deviennent : R''_{tv}=47%, R''_{vp}=45%, R''_{pc}=8%.

8. Temps différé et temps réel, traitement analogique ou digital

En temps différé, les mémoires d'ordinateurs peuvent tourner une partie des difficultés rencontrées dans la reconnaissance de la parole. Par contre, en temps réel, les difficultés fondamentales, inhérentes aux lois physiques, restent les mêmes, quelle que soit la technologie, digitale ou analogique, utilisée. Cependant, les miniaturisations actuelles (circuits intégrés, micro-processeurs) permettent des réalisations compactes et économiques. Elles sont proposées pour diverses machines reconnaissant la parole naturelle (phonétographes) ou codée (phono-décodeurs OTEA, SOKINA, UTOSANERI).

9. Conclusions

Les "redondances" fournissent des images quantitatives des "contraintes" (nommées aussi "règles" ou sources de connaissance"). Leurs importances relatives orientent le chemin à suivre pour parvenir à une reconnaissance optimale de la parole. Les pourcentages relatifs de "l'analyse spectrale" (ou "vocoder"), des "signes phonétiques" et des "compléments linguistiques", concernant les 3 exemples discutés, sont résumés dans le TABLEAU IV :

TABLEAU IV	analyse spectrale	signes phonétiq.	complém. linguist.
a) parole française continue (25000 mots)	34 %	34 %	32 %
b) français restreint (100 mots isolés)	39 %	39 %	22 %
c) parole codée (SOTINA, 25000 mots)	47 %	45 %	8 %

L'analyse plus détaillée a montré que les 2 composantes "phonémique" (N_{V7}=112) et "prosodique" (N_{V8}=128) sont également importantes, environ, pour parvenir à l'étape "signes phonétiques". En comparant les dimensions des "mémoires", exigées par les "compléments linguistiques" (tels que "lexicaux", "grammaticaux", etc) avec les redondances et taux d'erreur correspondants, on obtient des critères d'efficacité additionnels. Pour être acceptables, les taux d'erreur ^{6,8} doivent rester inférieurs à 0,1%, et ils exigent une utilisation équilibrée des "contraintes", donc des "redondances". "Utilisation" signifie essentiellement "séparation" : par exemple, la "dynamique" et le "degré de voisement" doivent être séparés initialement des autres espèces de "redondances", car elles ne peuvent plus être récupérées efficacement dans la suite des opérations. Les "redondances" constituent une chaîne fonctionnelle, dont il faut respecter la suite naturelle. La parole actuelle est en effet le résultat d'une succession de contraintes, expérimentées par l'homme, au cours des millénaires de son évolution.

REFERENCES

1. A.V. Oppenheim, R.W. Schafer, "Digital Signal Processing", Prentice-Hall, Englewood Cliffs, New Jersey, 1975, p.511-519
2. J.D. Markel, A.H. Gray, Jr "Linear Prediction of Speech", Springer-Verlag, New York, 1976, p.129-189
3. J.A. Dreyfus-Graf "Spectres phonétographiques de la parole chuchotée", Revue d'Acoustique, Paris, No 1, 1968
4. " " " "Phonétographe et sub-formants", Bull. Techn.PTT, Berne, No 5, 1961, p.41-59
5. " " " "Reconnaissance de parole et segmentation" 7èmes Journées d'Etude sur la Parole, GALF, Nancy 1976
6. " " " "Machines commandées par la Parole humaine", Volume ISR 32, Birkhäuser-Verlag, Basel 1976, p.351-360
7. J.S. Bourgenot, C. Dechaux, J. Potage "Le vocodeur CIPHON", Revue Techn. Thomson-CSF, Vol.7, No 4, Paris, déc.1975, p. 755-804
8. J.A. Dreyfus-Graf "Redundancies of Symbolic Systems", IEEE, ICASSP, Hartford (Conn.), 1977
9. W.S. Hess "... Feature Extraction System for Phonetic Recognition of Speech", IEEE, ASSP, Febr. 1976
10. F. Jelinek "Continous Speech Recognition by Statistical Methods" Proc.IEEE, Man-Machine Comm. by Voice, Apr.1976
11. D.R. Reddy "Speech Recognition by Machines : A Review" Proc. IEEE, Man-Machine Communic. by Voice, Apr.1976

Separata sur demande. Adresse de l'auteur : 5 Av. Grenade, CH-1207-Genève (Suisse)

VIII èmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

ANALYSE DE LA STRUCTURE INTONATIVE DE LA PHRASE EN FRANCAIS

DANIELLE LARREUR ET FRANCOISE EMERARD CNET LANNION

• Résumé :

La simple concaténation de diphones permet d'obtenir une parole de synthèse intelligible mais totalement dépourvue de naturel.

Pour cette raison, nous avons réalisé un traitement prosodique qui attribue à chaque mot d'un message une configuration intonative et temporelle spécifique sur chacune de ses syllabes en fonction de la longueur du mot et de la position de ce dernier dans l'énoncé.

Nous présentons ici une formalisation des manifestations prosodiques observées à l'analyse d'un vaste corpus lu par un locuteur féminin ainsi que la quantification des évolutions de F_o sur certains points considérés comme significatifs dans l'énoncé.

• Summary :

Synthesis by dyad concatenation produces an intelligible speech, but the lack of prosodic features gives the speech an unnatural and unpleasant sound.

For this reason, we realized a prosodic processing which gives to each word of the message a melodic and temporal pattern depending upon the length of the word and upon its position inside the sentence.

Here are presented a formalization of prosodic features measured by analyzing a large corpus read by a female speaker and the quantization of F_o evolutions at certain syllables considered as pertinent in the sentence.

ANALYSE DE LA STRUCTURE INTONATIVE DE LA PHRASE EN FRANCAIS

Danielle LARREUR et Françoise EMERARD CNET LANNION

Dans le cadre général de l'automatisation des Centres de Renseignements, nous avons tenté de réaliser un système de synthèse opérationnel, fonctionnant en temps réel et produisant une parole non seulement intelligible mais aussi relativement naturelle.

Pour parvenir à ce résultat, nous avons effectué un certain nombre de choix qui nous ont amené à retenir comme type de synthétiseur le vocodeur à canaux, comme élément minimal de parole le diphone, et comme moyen d'amélioration de la qualité la mise au point d'un traitement systématique de la prosodie (caractéristiques intrinsèques, accent, intonation).

Dans cette optique, deux étapes importantes ont été franchies :

* la constitution d'un dictionnaire de 1200 diphones extraits à partir de mots naturels enregistrés par un locuteur féminin, analysés à 4800 eb/s. et stockés en mémoire d'un calculateur,
* une modélisation permettant une reconstitution ab nihilo (par rapport aux données d'entrée du système) des faits prosodiques suffisamment souple pour pouvoir s'adapter à des constructions syntaxiques de complexités différentes ; nous avons utilisé comme points de repère un double système de marqueurs :
  - inscrits en mémoire sur chaque diphone pour repérer les frontières consonantiques et vocaliques,
  - insérés en nombre limité dans la chaîne phonétique pour signaler la fin des mots considérés comme significatifs du point de vue prosodique.

On comprend que l'élaboration du traitement prosodique ait nécessité une analyse préalable pour dégager les composantes acoustiques qui, dans la parole continue, ont pour fonction d'assurer la réalisation des faits prosodiques et plus précisément des faits intonatifs, ainsi que le découpage du message en groupes syntaxiques pour aider au décodage sémantique de l'énoncé.

Cette analyse a porté sur un corpus de quelques 400 phrases de type énonciatif, impératif et interrogatif, lues par un locuteur féminin à un débit considéré comme moyen (Boë et al., 1975 : 160 mots par minute) et construites de façon à posséder d'une part des mots au nombre de syllabes recouvrant toutes les possibilités de la langue française, d'autre part une composition syntaxique très variée.

Nous allons résumer dans cet exposé les observations induites de l'analyse pour ce qui concerne les manifestations prosodiques de certaines réalisations vocaliques situées en des points de l'énoncé énonciatif français considérés comme significatifs - on trouvera par ailleurs les résultats détaillés de cette analyse : Emerard, 1977 -

Nous appelerons règles des variations systématiques de F_0 que nous avons retenues pour élaborer le traitement automatique de la prosodie en synthèse.

Les points de l'énoncé que nous étudions concernent la voyelle finale du message, d'une proposition non située en fin de message, du syntagme situé avant le verbe (groupe nominal $_1$), du syntagme situé après le verbe (groupe nominal $_2$), du groupe verbal.

Le GN_1 et le GN_2 pourront se décomposer en ce que nous nommerons groupe de sens c'est-à-dire en un groupe de mots unis très directement et très profondément par le sens, les limites entre deux groupes de sens pouvant résulter de la présence d'une préposition ou d'une locution prépositive, de la présence d'une virgule, d'un pronom relatif, d'une conjonction de subordination ou de coordination.

A/ Phrases de type GN_1 + Verbe :

. Règles obligatoires :

Si GN_1 est composé d'un pronom, on constate une descente intonative en fin de phrase mais un schéma intonatif "neutre" pour les pronoms précédant le verbe ; il n'existe jamais de pause entre le pronom sujet et le verbe ; au contraire, si GN_1 est un mot lexical, sa voyelle finale présente toujours une montée intonative.

. Règles facultatives (en fonction du débit du locuteur) :

Allongement du dernier segment du mot composant GN_1 et pause entre GN_1 et le verbe (la pause ne peut exister que si un allongement de durée est réalisé sur le segment final d'un mot).

Cette différence systématique dans le schéma intonatif de dernière syllabe de GN_1 selon qu'il s'agit d'un mot grammatical ou d'un mot lexical, permet d'éviter les ambiguïtés homophoniques, par exemple l'évolution de F_0 permet de distinguer entre /j'en mange/ et /Jean mange/ .

B/ Phrases GN_1 + GV + GN_2 (ex : "les médicaments sont vendus en pharmacie") :

. Règles obligatoires :

Dans tous les exemples de phrases constituées d'un syntagme sujet simple (sans épithètes ni groupes de sens), d'un groupe verbal, d'un syntagme complément simple (sans épithète ni groupe de sens,) on observe une montée intonative sur la syllabe finale du sujet, un schéma de F_0 descendant en fin de phrase, et une montée intonative sur la syllabe finale du groupe verbal. (Exception faite des verbes copules et des verbes transitifs à sens très général : "faire", "aller" "pouvoir...).

. Règles facultatives :

Le locuteur peut, selon son débit, réaliser ou non une démarcation du GN_1 et du groupe verbal par la production combinée d'un allongement de durée sur la dernière syllabe du GN_1 et d'une pause ; il peut également - par la production des mêmes paramètres - réaliser ou non une démarcation entre le groupe verbal et le groupe nominal 2 ; mais il est bien évident qu'aucune démarcation n'est réalisable si le groupe verbal - de par sa nature grammaticale - connaît un schéma intonatif descendant.

Les phrases de même structure dans lesquelles le GN_1 et le GN_2 comportent des épithètes présentent des schémas prosodiques qui pour les mêmes points clés ne sont pas modifiés par la présence d'épithètes successifs ("le gentil petit chien a mangé des escargots") : tous les épithètes, qu'ils soient situés dans le GN_1 ou dans le GN_2, qu'ils soient un ou plusieurs, ont un schéma descendant neutre et possèdent une durée de dernier segment égale à la durée d'un segment inclus dans un mot ; on constate d'autre part que ce n'est pas la nature du mot (épithète, substan-

tif...) qui conditionne son évolution prosodique mais sa position dans l'énoncé (la même conclusion se dégage des travaux de OLIVE, 1975) : dans la phrase "le vieux facteur fait sa tournée", le substantif "facteur", parce qu'il est situé avant le groupe verbal, prend le schéma intonatif propre aux mots situés dans ce contexte ; au contraire, dans la phrase "le vieux facteur prétentieux fait sa tournée matinale", c'est l'épithète "prétentieux" qui occupe la position propre à lui octroyer un schéma intonatif montant ; le substantif facteur incorporé dans le syntagme nominal sujet n'a plus de situation intonative privilégiée, et la pente mélodique de sa dernière syllabe a même signe que celle des autres mots inclus dans le syntagme.

. Règles obligatoires :

C'est la position dans l'énoncé, et non la nature grammaticale d'un élément linguistique qui détermine son schéma intonatif.

Tous les mots intégrés dans un même groupe nominal (qui ne possède qu'un groupe de sens) ont un schéma de F_0 de pente négative quel que soit le groupe nominal auquel ils appartiennent (GN_1 ou GN_2).

. Il n'y a pas de règles facultatives : il n'est pas possible d'insérer une pause entre les éléments du groupe, sauf manifestation d'une insistance particulière sur un des termes de l'énoncé : dans ce cas, ce phénomène se réalise par l'action conjuguée des trois paramètres de durée, d'intensité et de fréquence fondamentale.

C/ Phrases GN_1 (composée de plusieurs groupes de sens) + GV + GN_2.

ex : "les voyageurs pour Calcutta sont attendus à la passerelle".

. Règles obligatoires :

- Tous les mots appartenant au GN_1, qu'ils soient situés en fin de groupe de sens (non situé à la fin du GN_1) ou à l'intérieur d'un groupe de sens, possèdent un schéma intonatif descendant ; si le signe de la pente est identique, nous verrons que les niveaux de F_0 et leur amplitude sont différents dans l'un et l'autre cas.
- Il n'y a jamais de pause, sauf insistance particulière, entre deux mots quand le second est subordonné au premier pour en limiter le sens (exemple : un cor de chasse, un verre de vin).

. Règles facultatives :

- Des pauses peuvent apparaître dans le GN_1 entre deux groupes de sens quand le second est introduit par une préposition ou une locution prépositive (dans, depuis, chez, avec...) ou quand il constitue une proposition subordonnée. Dans ce cas, les pauses succèdent naturellement à un schéma intonatif de dernière syllabe de type descendant et à un allongement de durée sur le dernier segment du mot.

D/ Phrases GN_1 + GV + GN_2 (composé de plusieurs groupes de sens introduits par une préposition ou par une locution prépositive ou composé d'une subordonnée).

ex : "le vilain corbeau fatigué tenait dans son bec un camembert usagé."

Dans toutes les phrases du corpus qui présentent cette structure, on

observe une montée intonative sur la dernière syllabe du mot qui marque le terme du premier groupe de sens. Nous insistons sur le fait que cette observation ne concerne que les phrases dont le GN_2 possède deux groupes de sens.

D'autre part, alors que les phrases composées d'un seul groupe de sens dans GN_2 présentent un schéma intonatif montant sur la dernière syllabe du groupe verbal, les énoncés de cette catégorie (deux groupes de sens) font constater une modification du signe de la pente intonative sur la syllabe finale du groupe verbal : dans tous les cas, cette pente est négative.

Pour ce type de phrases, on peut énoncer les règles suivantes :

. Règles obligatoires :

- la montée intonative en fin de GN_1 demeure, ainsi que la descente intonative de fin de phrase énonciative.
- la syllabe finale du premier groupe de sens présente toujours une montée de F_0 (une restriction cependant quand GN_2 se compose de deux groupes de sens dont le dernier constitue un complément déterminatif, en général introduit par /de/).
- le signe de la pente intonative pour la syllabe finale du groupe verbal est transformée : la pente est toujours négative.
- il n'y a pas de pause entre le groupe verbal et le GN_2.

. Règles facultatives :

- on observe souvent dans un corpus lu une pause qui assure le découpage de la structure syntaxique entre les deux groupes de sens. La pause n'existe que si un allongement de durée caractérise le segment final du premier groupe de sens.

Quand GN_2 est composé de plus de deux groupes de sens ("le génie des alpages a envoyé les moutons chez des amis au bord de la mer"), on revient au schéma intonatif d'origine où le groupe verbal n'est suivi que d'un seul groupe de sens : la syllabe finale du groupe verbal présente un schéma intonatif de type montant.

En fait, on peut résumer de façon très simple les contours finals des groupes de sens dans le GN_2, puisque l'on observe, quel que soit le nombre des groupes de sens dans ce syntagme, une alternance parfaite entre schémas montants et schémas descendants depuis le début du syntagme verbal jusqu'à la fin de la phrase.

Signalons enfin que ces groupes de sens peuvent très bien être suivis de propositions subordonnées ; chacune des propositions réalisant :

. la décomposition en groupes nominaux et groupes verbaux,

. et l'alternance dans le signe de la pente de F_0,

pour les éléments qui la composent.

Les frontières essentielles de l'énoncé sont représentées par les pauses :

- pause qui met un terme à une proposition finale : dans ce cas la pause fait suite à un schéma intonatif descendant et à un allongement de durée sur le dernier segment de mot.
- pause qui conclut une proposition non finale ; elle succède à un schéma intonatif montant et à un allongement de durée sur le dernier segment de mot.

Dans le cas d'une proposition non située en fin d'énoncé, on trouve toujours sur la syllabe finale du dernier groupe de mots, un schéma intonatif de type montant ; et pour les groupes qui précèdent - mais seulement à partir du groupe verbal car le GN_1 n'est pas modifié - on observe des schémas inverses de ceux constatés dans une proposition de fin d'énoncé.

E / Les schémas intonatifs des propositions de même nature qui sont reliées par une conjonction de coordination ("la tempête s'éloigne et les vents sont calmés").

Le principe est toujours le même : on dégage de cet énoncé deux grandes unités. L'une non finale, se termine donc par un schéma montant qui ordonnance l'alternance pour les unités du niveau inférieur qui la composent. L'autre finale, se termine par un schéma de F_0 descendant et règle pour les unités du niveau inférieur qui sont en elles l'alternance dans le signe de la pente de F_0 ; il n'est pas besoin de faire référence à la nature syntaxique des mots.

Pour conclure, nous dirons qu'à notre avis, ce qui relève de l'arbitraire, du conventionnel, dans l'évolution de F_0, c'est le signe de sa pente qui permet de définir une organisation de la phrase en deux constituants majeurs répartis de part et d'autre d'une frontière située immédiatement avant le groupe verbal. On retrouve ici la notion bien connue de thème et de propos. Ce qui relève des variantes individuelles, conscientes ou non, est lié au débit du locuteur et à son intention de bien marquer (ou non) la structure syntaxique de l'énoncé (pauses subséquentes à un allongement de durée du dernier son).

Il ne nous est pas possible dans le cadre de cette étude de rendre compte de l'analyse d'autres structures énonciatives : celle-ci aurait confirmé ce que nous avons tenté de montrer : une alternance remarquable dans le signe de la pente de F_0 sur trois niveaux successifs de décomposition d'un énoncé :

1°/ niveau de la proposition : finale / non finale.

2°/ niveau des groupes majeurs : GN_1 / GV / GN_2.

3°/ niveau des groupes de sens des groupes majeurs : si groupe majeur = schéma de F_0 montant ⟶ groupe de sens = schéma de F_0 descendant.

Il existe un quatrième niveau qui se situe sur n'importe quel point de l'énoncé et qui présente une évolution identique de F_0 : un schéma de type descendant.

Ces résultats sont à rapprocher de ceux proposés par Martin (1973, 1974, 1975, 1976) mais en diffèrent par deux points essentiels :

1°/ Pour Martin : "un contour mélodique d'une unité ne peut prendre de valeur que par contraste avec le contour mélodique qui marque l'unité de niveau immédiatement supérieur et avec celui-là seulement "... ..." en français cette opposition de contours d'unités composantes et d'unité composée est réalisée essentiellement par un contraste de pente mélodique (et secondairement par un

contraste d'amplitude de variation mélodique".

Ainsi, par exemple, la phrase "le gros chien mange le fromage" donne le parenthésage suivant :

(((le gros) (chien)) (blanc)) ((mange) (le fromage).

Toutes les phrases de même construction syntaxique que nous avons analysées ne présentent jamais le contour détecté ici sur l'adjectif /gros/ et caractérisé par les traits montant, court, restreint (faible variation de F_0).

Pour notre part, nous n'avons trouvé de schéma montant de F_0 dans le syntagme précédant le groupe verbal que sur la première voyelle de certains mots pour réaliser ce que l'on appelle "l'accent didactique" ; un mot considéré comme sémantiquement important présente sur sa première voyelle, une montée intonative et sur sa seconde voyelle, une descente de F_0 qui a pour niveau de départ une valeur fréquentielle plus haute (d'environ ^{0} 1/2 ton) que la valeur d'arrivée de la première voyelle. Nous pensons que la succession "le gros chien blanc" appartient à la même unité de signification et que tous les mots inclus dans le même groupe de sens du syntagme nominal situé avant le verbe présentent à leur terme une pente de F_0 négative.

2°/ Le second point de désaccord porte sur l'affirmation suivante : "les contours des unités d'un même niveau sont souvent réalisés avec une amplitude correspondant à l'importance de l'unité dans la composition de l'énoncé : ces unités étant en général rangées par ordre décroissant d'importance....les contours sont réalisés avec des amplitudes décroissantes au fur et à mesure qu'elles se rapprochent du contour de référence final".

Nous pensons, (de par les résultats de l'analyse de corpus) que si le locuteur réalise une pause après le GN_1 et le groupe verbal, on constate sur la syllabe finale de l'un et l'autre groupe, un allongement dans la durée du dernier segment, et l'on observe également un schéma intonatif identique sur les syllabes situées en fin de chaque groupe, ou bien un schéma dont les différences de F_0 ne sont dûes qu'aux caractéristiques intrinsèques du dernier segment vocalique ; mais à voyelle identique, les schémas intonatifs ne semblent pas présenter de différence significative qui tienne à leur position par rapport au "contour de référence final".

Les résultats de l'analyse quantitative des évolutions de F_0 pour le locuteur choisi et pour les trois types de phrases est présentée au ^{0} Tableau 1 ; les points étudiés sont les suivants :

-Pour les phrases énonciatives :

(1) valeurs de F_0 dans les monosyllabiques situés en initiale de phrase.

(2) valeurs de F_0 en fin de mots plurisyllabiques à schéma intonatif descendant, c'est-à-dire les mots appelés "non marqués".

(3) valeurs de F_0 en fin de mot à schéma intonatif descendant, situé en fin de groupe de sens du syntagme pré-verbal.

(4) valeurs de F_0 à la fin de l'avant dernière syllabe d'un mot $\geq$ 3 syllabes et valeur de F_0 à l'attaque de la dernière syllabe, quand ce mot est situé avant une pause et possède un schéma intonatif montant.

(5) valeurs de F_0 en fin de mot plurisyllabique avec un schéma intonatif montant et situé avant une pause.

(6) valeurs de F_0 en fin de mot à schéma intonatif montant mais non suivi d'une pause.

(7) valeurs de F_0 dans les monosyllabiques (non points clés) dans la phrase.

(8) valeurs de F_0 à la fin de l'avant dernière syllabe d'un mot $\geqslant$ 3 syllabes et valeur de F_0 à l'attaque de la dernière syllabe quand ce mot est situé avant une pause et possède un schéma intonatif descendant.

(9) valeurs de F_0 en fin de mot plurisyllabique avec un schéma intonatif descendant et situé avant une pause (fin de phrase).

- Pour les phrases intérrogatives :

(10) valeurs de F_0 à la fin du mot qui porte la marque de l'interrogation dans les phrases introduites par un mot ou un groupe de mots interrogatifs.

(11) valeurs de F_0 à la fin du mot qui porte la marque de l'interrogation dans les phrases construites selon le même modèle syntaxique que les phrases énonciatives ou avec inversion du sujet et du verbe.

(12) valeurs de F_0 en fin de phrase interrogative quand la marque de l'interrogation est déjà présente dans la phrase.

(13) valeurs de F_0 pour les monosyllabiques.

- Pour les phrases impératives :

(14) valeurs de F_0 pour la voyelle finale d'un mot à intonation montante situé avant une pause.

(15) valeurs de F_0 pour la voyelle finale de phrase.

L'ensemble des observations faites à l'analyse ont été utilisées pour réaliser le traitement prosodique à la synthèse et donne des résultats que l'on peut qualifier globalement de satisfaisant.

Bien que les règles d'évolutions prosodiques aient été vérifiées pour deux autres locuteurs, il ne faut pas oublier que l'analyse du corpus n'a porté que sur un seul locuteur ; il est par conséquent évident que leur généralisation sur un plan phonétique exigerait une vérification sur un grand nombre de sujets.

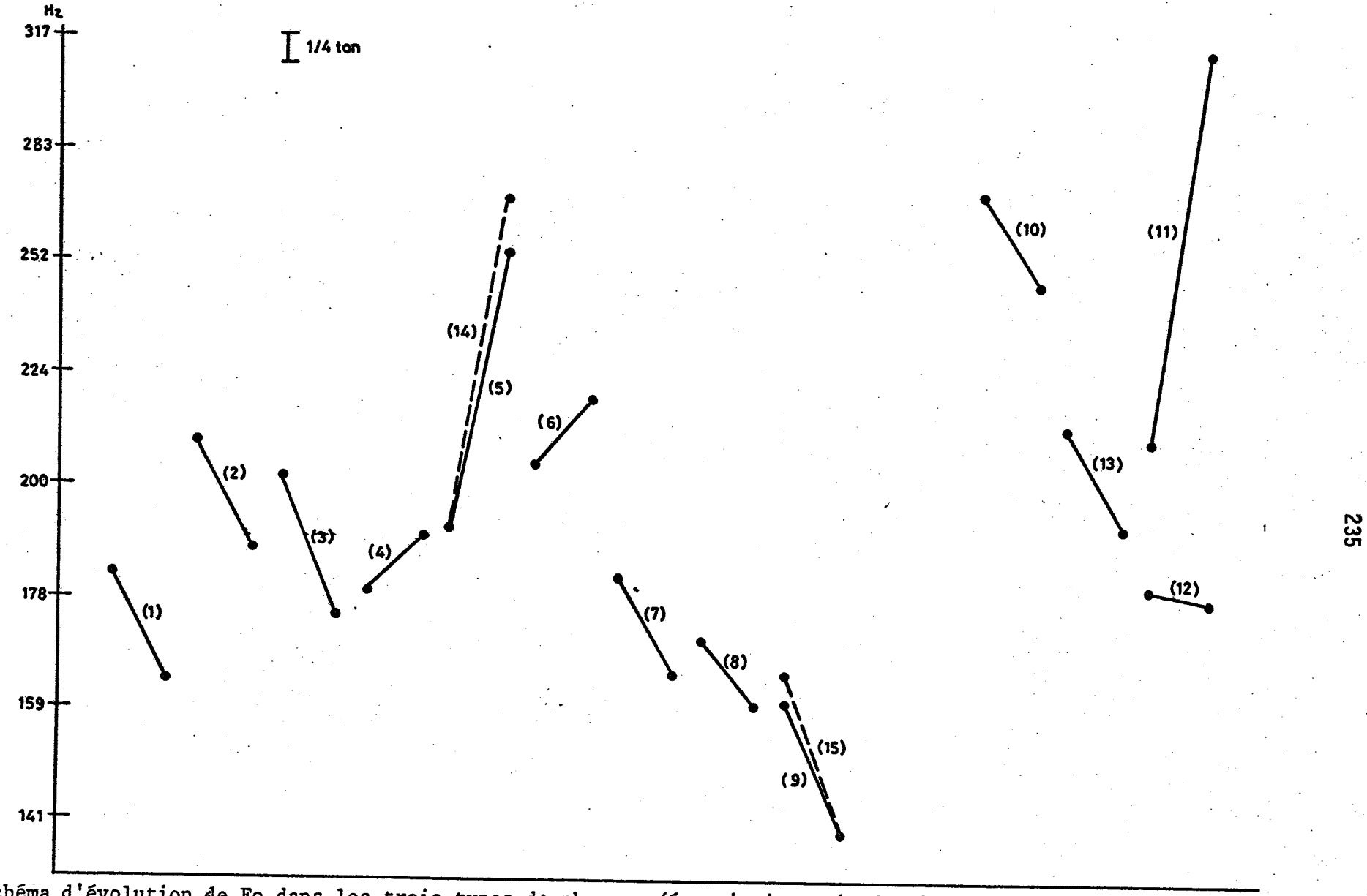


Schéma d'évolution de Fo dans les trois types de phrases (énonciatives, impératives et interrogations) et pour les voyelles des catégories étudiées (les catégories sont présentées autant que possible, dans l'ordre de leur apparition).

- TABLEAU I -

Bibliographie :

Boë,L.J., Contini, M., et Rakotofiringa, H., Etude statistique de la fréquence laryngienne, Phonetica, 32, 1-23, 1975.

Emerard, F., Synthèse par diphones et traitement de la prosodie, Thèse de 3è cycle, Grenoble, 1977.

Martin, Ph., Les problèmes de l'intonation, Recherches et Méthodes, Langue française, 19, 4-32, 1973.

Martin, Ph., Intonation et reconnaissance automatique de la structure syntaxique, 6è Journées d'Etude sur la parole, GALF, Toulouse, 51-62, 1975.

Martin, Ph., Eléments pour une théorie de l'intonation. Rapport d'activités de l'Institut de Phonétique (Bruxelles) n° 9/1, 97 - 126, 1974 - 75.

Martin, Ph., Une grammaire de l'intonation de la phrase. Rapport de l'Institut de Phonétique (Bruxelles) n° 9/2, 1975.

Martin, Ph., Analyse phonologique de la phrase française, Linguistics, 146, 35-69, 1975.

Olive, J.P., Fundamental frequency rules for the synthesis of simple declarative English sentences, JASA, 53, 2, 476-482, 1975.

-=-=-=-=-=-=-=-

VIII èmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

TRAITS PROSODIQUES DISTINCTIFS DE CERTAINES ATTITUDES INTELLECTUELLES ET EMOTIVES

Ivan FONAGY CNRS

Jacques SAP Laboratoires de Marcoussis

Résumé :

On a tâché de déterminer à l'aide de variantes synthétisées de deux phrases françaises les traits prosodiques signalant certaines attitudes intellectuelles et émotives.

Summary :

We tempted to define by means of synthesized variants of two French sentences the distinctive prosodic features of some intellectual and emotive attitudes.

TRAITS PROSODIQUES DISTINCTIFS DE CERTAINES ATTITUDES INTELLECTUELLES ET EMOTIVES

1. L'APPAREILLAGE PERMETTANT LA SYNTHESE DE PAROLE

L'ensemble permettant la synthèse de parole est construit autour d'un mini-ordinateur sur lequel sont connectés le synthétiseur de parole et un certain nombre d'autres périphériques d'usage général.

1.1. Le synthétiseur

L'organe de production de parole est un synthétiseur à formants du type de ceux étudiés en Suède par FANT, en Ecosse par LAWRENCE et en France par CARRE.

Il a été réalisé à l'aide de circuits numériques, ce qui permet une bonne linéarité de l'ensemble et en particulier une très bonne représentation de la fréquence du fondamental qui est représentée par sept éléments binaires.

La parole est produite en temps réel, l'ordinateur fournissant au synthétiseur un jeu de onze paramètres toutes les huit millisecondes.

1.2. Analyse de la parole

Un dispositif permet d'acquérir de la parole et de la stocker sur disque. Un certain nombre de programmes d'ordinateur permet d'analyser automatiquement la parole afin d'en extraire les paramètres nécessaires à la synthèse. Par la suite on aura toute liberté pour modifier n'importe quel paramètre. La synthèse totale "ab nihilo" est possible.

1.3. Mode opératoire

Les paramètres de pilotage du synthétiseur sont stockés dans la mémoire centrale de l'ordinateur ; toutes les directives nécessaires à leur manipulation sont issues d'un télétype.

On peut afficher les courbes représentatives de n'importe quel paramètre sur une unité de visualisation à tube cathodique.

Toute modification de ces paramètres est possible à l'aide de commandes issues du clavier. On peut ainsi donner rapidement toute forme voulue au contour mélodique, aux formants, à l'intensité vocale... etc...

2. SYNTHESES ET TESTS

2.1. Triangle assertif et interrogatif "Il est venu"

Le français parisien connaît un certain nombre de clichés mélodiques, c'est à dire de schèmes d'intonation qui sont réalisés avec une marge de fluctuation très étroite. (FONAGY et BERARD 1973, FONAGY, GUZMAN et BERARD 1976).

Un de ces schèmes mélodiques, particulièrement fréquent dans la conversation familière, a une forme triangulaire et se greffe sur les trois dernières syllabes des phrases ou propositions, ou sur des phrases trissyllabiques. Le caractère "chantonné" de ces clichés facilite leur présentation musicale (figure 1).

Une montée d'une quarte augmentée est suivie d'une descente d'une tierce majeure. Or, ce schéma recouvre, à vrai dire, deux formes d'intonation différentes : un triangle assertif tel qu'il peut apparaître dans : Il est venu ! rappelant le schème mélodique de la raillerie enfantine (Na-na-na !) ; mais aussi dans des questions incrédules comme : Il est venu ?, avec l'implication : "mais qu'est-ce que tu me chantes là ?".

Selon des tests de perception faits à partir de phrases repiquées d'enregistrements de pièces de théâtre et de conversations, les auditeurs arrivent, dans la plupart des cas, à distinguer hors contexte le triangle assertif du triangle interrogatif. C'est à partir de ce matériel présenté au cours de tests préalables que nous avons pu formuler certaines hypothèses concernant les traits distinctifs des deux schèmes semblables pour la forme et opposés quant à leur contenu.

Les triangles généralement perçus comme assertifs sont caractérisés par trois niveaux mélodiques relativement stables ; les triangles interrogatifs par une ligne mélodique montante, dans la deuxième et/ou dans la troisième syllabe. Cette divergence dynamique est masquée par le fait que les "noyaux mélodiques" - les parties de la courbe de fréquence fondamentale qui coïncident avec les maxima d'intensité - sont pratiquement les mêmes dans les deux schèmes.

Au cours d'un premier test nous avons présenté vingt variantes qui se distinguaient par leurs lignes mélodiques intra-syllabiques, d'une part, et par les intervalles mélodiques entre les trois syllabes, d'autre part. Les sujets ont été invités à déterminer le degré de naturel des phrases synthétisées en concentrant leur attention à l'aspect prosodique. Ces tests (que nous nous proposons d'analyser ailleurs, en rapport avec d'autres clichés mélodiques synthétisés) ont montré que les sujets étaient extrêmement sensibles, dans leurs jugements, aux écarts mélodiques. Pour tester le caractère modal des variantes, nous n'avons retenus que les variantes jugées suffisamment naturelles par la plupart des sujets.

Au cours d'un deuxième test, nous avons présenté trois types de configuration mélodique :

(a) Un schème à trois niveaux relativement stables : une montée d'un quart de ton dans la première syllabe (de 182 à 188 Hz), une montée d'un demi-ton dans la deuxième (de 290 à 298 Hz), et une descente d'un demi-ton (de 218 à 206 Hz) dans la troisième syllabe.

(b) Le deuxième schème était plus dynamique : montée de huit demi-tons dans la deuxième syllabe (de 188 à 308 Hz), la deuxième syllabe se composait de deux niveaux stables, celui de /y/ dépassant de deux demi-tons celui de /n/.

2. INTONATION ET SITUATION "Il est huit heures"

Des tests d'interprétation faits avec 14 variantes de la phrase "Il est huit heures" prononcées par Marie Claude MESTRAL et avec 18 variantes de la phrase "Il a une voiture" (CALBRIS 1973), suggèrent que la façon de prononcer des phrases banales polyvalentes permet dans certains cas d'identifier le sens précis de l'énoncé et la situation censée l'avoir déclenchée. Ainsi la phrase Il est huit heures dite d'une certaine manière peut évoquer la personne qui rappelle à une personne l'heure tardive, avec gentillesse, ou, celle qui s'affole en s'apercevant qu'elle manquera probablement le train, etc. (FONAGY, BERARD 1972).

Pour mettre en évidence l'apport de la mélodie phrastique, nous avons reproduit les courbes de fréquence fondamentale des variantes qui ont été plus ou moins correctement identifiées par les sujets des tests. Nous avons présenté huit variantes (dont l'une est à deux reprises) à deux groupes de sujets, un groupe de 12 et un autre groupe de 26 étudiants en linguistique, des deux sexes.

Comme il ressort du tableau 2 résumant les résultats des tests de perception, la distribution des réponses n'est aléatoire que dans le cas de la variante 8° reproduisant la courbe mélodique du monologue intérieur ("Voyons... il est maintenant 8 heures..."). Ceci indique que le débit (l'allongement des syllabes - surtout celle de la dernière) qui n'a pas été reproduit, est un élément essentiel dans la signalisation du monologue interne.

Le niveau élevé de la dernière syllabe et le ton progressif qui monte de 144 à 186 Hz (ce qui correspond à une courbe que Pierre DELATTRE - 1966 - considère comme celle de la "continuation majeure") n'a pas suffit à prêter un caractère bien déterminé à la phrase.

Une distribution serrée ne signifie pas nécessairement que le message suggéré a été effectivement reçu. Nous pouvons noter deux confusions d'attitude ou plutôt deux cas d'homonymie mélodique (cf. Milan Romportl 1974, 137-166) : le rappel énergique est interprété dans 16 cas sur 38 comme l'expression de l'affolement (phrase 2°). La courbe de fréquence fondamentale, reproduisant celle de la variante affolée, a été perçue dans 19 cas comme l'expression d'un rappel énergique. Dans les deux cas, nous avons taché de suivre en même temps la courbe d'intensité des variantes originales de Marie Claude MESTRAL. Les durées des deux variantes naturelles étant tout à fait comparables, il faut supposer que le timbre joue un rôle important dans le signalisation de ces deux attitudes. L'affolement se caractérise surtout par un voisement imparfait, une voix aérée. L'absence de ce critère a pu diriger les votes vers "rappel énergique" dans le cas de la phrase 2°. Les deux courbes de fréquence sont, en effet, assez semblables. Celle reproduisant l'appel énergique se distingue par une forte montée dans la deuxième syllabe (est) simulant un accent d'insistance (figure 5). La montée très importante (224---350---156 Hz) est précédée, par contre, dans la phrase 7° (imitant l'affolement) par une "tenue" mélodique, un niveau égal.

Les variantes 3°, 5°, 6° révèlent une autre homonymie mélodique : le rappel tendre est systématiquement confondu avec l'expression d'une attitude justificative ("Allons donc ! Il n'est que huit heures !". Les deux attitudes se reflètent également dans une mélodie progressivement descendante en quart

.../...

(c) Un troisième schème qui se distinguait du deuxième par une montée finale plus caractérisée de quatre demi-tons, en ligne droite, de la troisième syllabe (figure 2-4).

Les écarts mélodiques entre les "noyaux" des syllabes étaient les mêmes pour les trois schèmes mélodiques. On a associé la même courbe d'intensité aux trois variantes, et l'on n'a changé ni les durées, ni les structures formantiques. Nous avons profité du nombre restreint des variantes pour tester chaque item à trois reprises, dans un ordre aléatoire. Les signes A1, A2, A3 ; B1 B2 B3 ; C1 C2 C3 du tableau 1 renvoient à trois (et non pas à neuf) phrases synthétisées.

Comme il apparaît au tableau 1, la variante A caractérisée par des niveaux relativement stables a été perçue comme une assertion par 21-25 sujets sur 26. La variante C par contre a été interprétée comme une question par 19/21 sujets sur 26. Le mouvement intrasyllabique - la montée mélodique dans la deuxième et troisième syllabe - suffit, donc à distinguer le "triangle interrogatif" du "triangle assertif" malgré l'identité du cadre mélodique, déterminé par les rapports mélodiques des "noyaux" syllabiques. C'est probablement l'absence de la montée mélodique dans la partie vocale de la dernière syllabe qui est responsable des résultats moins nets obtenus dans le cas du schème B, jugé comme interrogatif par 14-18 sujets sur 26.

Les sujets avaient à choisir en même temps parmi quatre attitudes (émotives ou intellectuelles). Cette fois, il y a eu sensiblement plus d'hésitation et d'incertitude dans le choix. Il apparaît, toutefois, que les variantes "dynamiques" C et B évoquent plutôt une attitude incrédule ; la variante A, par contre, fait plutôt penser à une attitude sereine.

Perception de la modalité et de l'attitude suggérée : Tableau 1

Schèmes mélodiques présentés	Modalité		Attitudes		
	Assertive	Interrogative	Neutre	Joyeux	Incrédule
A 1	21	5	9	11	6
B 1	12	14	6	4	16
C 1	6	20	3	1	22
A 2	25	1	11	10	5
B 2	12	14	7	2	22
C 2	7	20	3	-	23
A 3	24	2	12	8	6
B 3	8	18	4	1	21
C 3	5	21	3	2	21

de ton (figure 6). Le rappel tendre se distingue par une faible remontée dans le mot heure. Cet écart ne paraît pas suffisant pour distinguer effectivement l'expression mélodique des deux attitudes. (Ajoutons, que nous avons rencontré de pareilles confusions au cours des tests faits avec les variantes naturelles).

L'intonation caractérise par contre très efficacement la voix du garçon d'hôtel réveillant le client (phrase 4° - figure 7) et l'absence d'attitude particulière : le ton neutre (phrese 1°).

	A neutre	B rappel énergique	C rappel tendre	D garçon d'hôtel	E On se justifie "Allons donc il n'est que 8 h !"	F affolement	G monologue (on calcule)
1° (A)	29	3	1	1	-	-	4
2° (B)	-	18	-	1	1	16	2
3° (C)	10	-	14	1	8	-	4
4° (D)	-	13	-	21	-	3	-
5° (E)	9	-	12	-	9	-	5
6° (E')	1	-	18	-	15	-	4
7° (F)	-	19	-	2	1	15	1
8° (G)	4	6	2	2	8	4	12
9° (D)	-	11	-	23	1	1	1

TABLEAU 2

REFERENCES

(1) CALBRIS G. 1973. Recherche pour une formation auditive et inductive. Thèse Sorbonne nouvelle Paris III.

(2) DELATTRE P. 1966. Les dix intonations de base du g français moderne. French Review 40, 1-14.

(3) FONAGY I et BERARD E. 1977. "Il est huit heures". Contribution à la sémantique de la vive voix. Phonetica 26, 157-192.

(4) FONAGY I. et BERARD E. 1973. Questions totales simples et implicatives. Studia Phonetica 8, 53-97.

(5) FONAGY I., GUZMAN M. et BERARD E. 1976. Comment mesurer l'accent d'intonation. Travaux de l'Institut d'Etudes Linguistiques et Photonétiques. Paris, 41-60.

(6) FONAGY I. et MAGDICS K. 1963. Emotional patterns in intonation and music. Z. Phonetik 16, 293-326.

(7) LEON P. 1971. Essais de phonostylistiques - Paris : Didier.

(8) ROMPORTL M. 1973. Studies in phonetics. Prague : Academia.

(9) MASSOT J.J. Etude et réalisation d'une unité de réponse vocale - Automatisme - tome 16 n° 12 décembre 1971

(10) ROGER G., KARSENTI E., SAP J. Application des procédés de visualisation à l'étude d'un synthétiseur paramétrique - Automatisme - tome 18 n° 3 mars 1973

(11) ROGER G., SAP J. Système d'analyse des signaux de parole. Extraction des paramètres utilisés pour la synthèse "à formants". Automatisme - décembre 74.
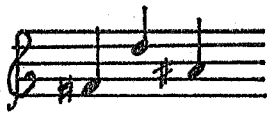


Fig. 1 : Schème musical d'un "cliché mélodique" triangulaire du français parisien.

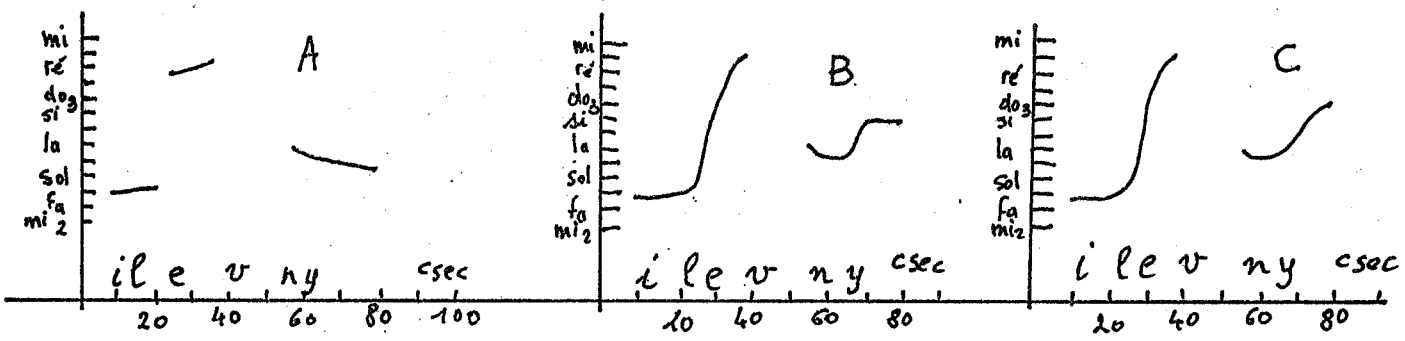


Fig. 2 : Courbe de fréquence fondamentale de la phrase synthétisée "Il est venu", modèle A.

Fig. 3 : "Il est venu" modèle B

Fig. 4 : "Il est venu" modèle C

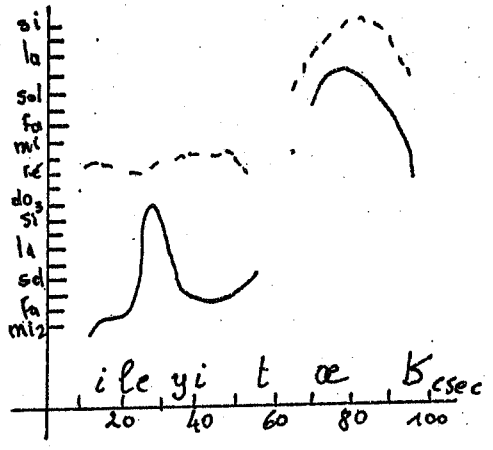


Fig. 5 : Ligne continue : courbe de fréquence de la phrase synthétisée "Il est huit heures", variante "rappel énergique".
--- ligne hachée : variante "affolée".

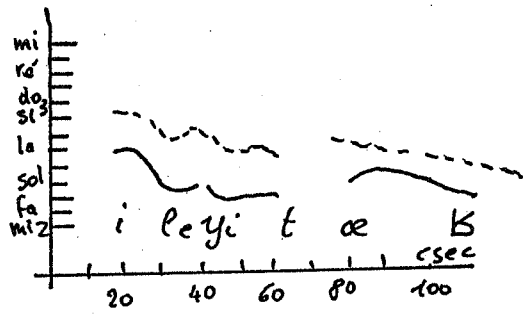


Fig. 6 : Ligne continue : "Il est huit heures", variante "rappel tendre"
---- ligne hachée : variante "justificative" ("Mais voyons il n'est que huit heures").

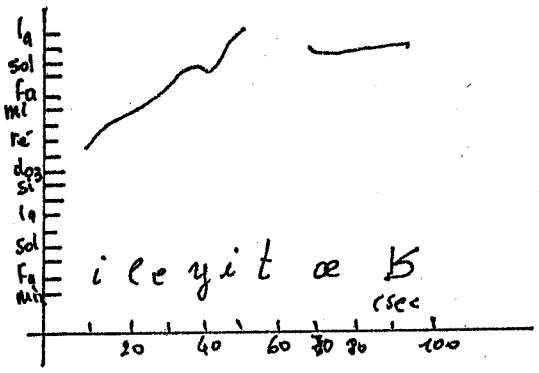


Fig. 7 : "Il est huit heures", variante "le garçon d'hôtel réveille le client

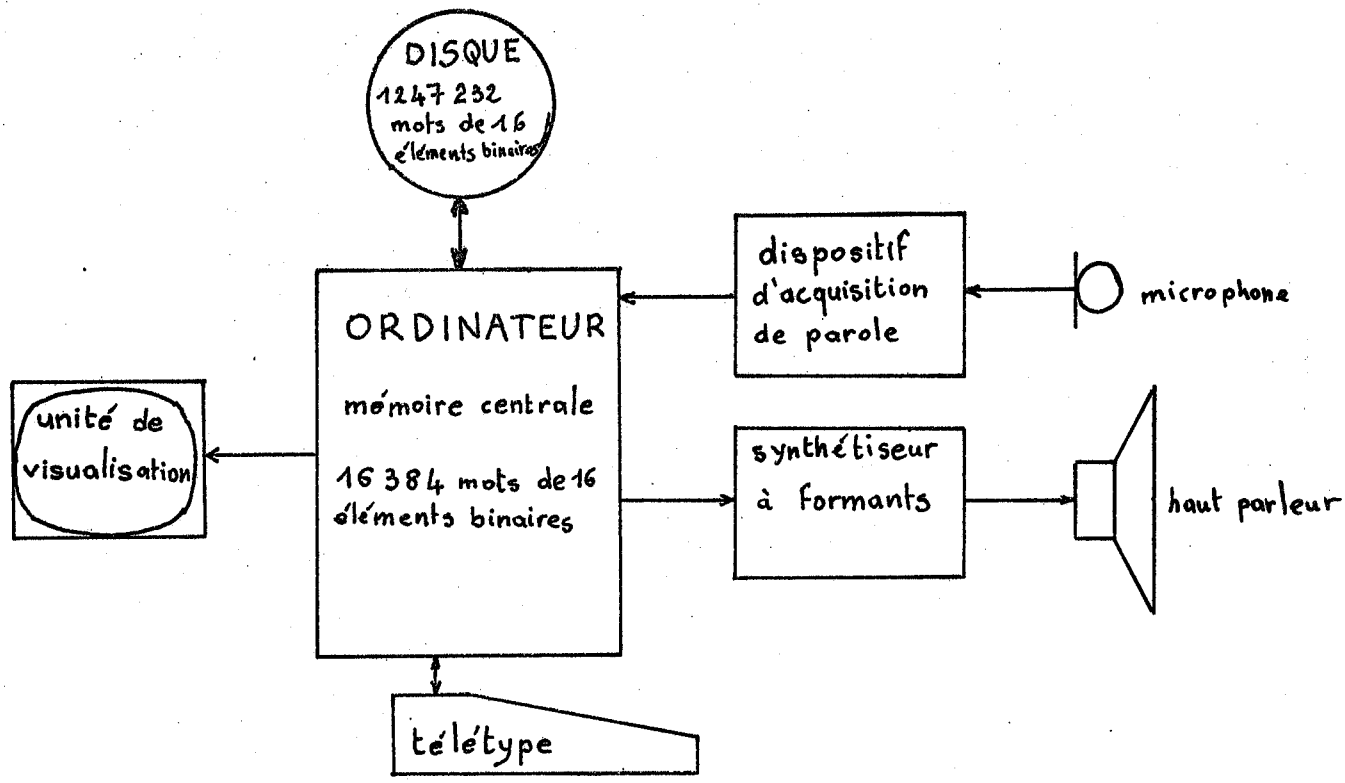


Fig. 8 : Schéma général de l'appareillage de synthèse de parole

VIII èmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

UTILISATION D'UN MODELE PROSODIQUE POUR LA SYNTHESE PAR CONCATENATION DE MOTS EN FRANCAIS ET EN ANGLAIS.

Philippe MARTIN
Experimental Phonetics Laboratory
UNIVERSITY OF TORONTO

• Résumé :

On présente des règles pour la synthèse des facteurs prosodiques de la phrase en français et en anglais. Celles-ci opèrent à partir d'un parenthèsage définissant la structure prosodique indépendamment ou non de la structure syntaxique.

• Summary :

Rules for the synthesis of prosodic features in French and English are presented here. These rules are applied to a bracketing which defines the prosodic structure of the sentence, independently or not of the syntactic structure of the sentence.

Utilisation d'un modèle prosodique pour la synthèse par concaténation de mots en français et en anglais

Philippe Martin

Intonation et syntaxe

L'énoncé peut être considéré comme formé de quatre lignes parallèles, de nature sémantique, syntaxique, syllabique et prosodique. Chacune de ces lignes est constituée par une séquence d'éléments minimaux, classés hiérarchiquement selon des structures indépendantes les unes les autres, mais présentant entre elles certains points d'ancrage définis par des règles d'association déterminées.

Les rapports entre les structures syntaxique et prosodique sont décrits par des règles d'accentuabilité des unités syntaxiques et de neutralisation des contours prosodiques (Martin, 1976a, 1976b, 1977a). Les mécanismes d'association que ces règles impliquent correspondent à la fois aux hypothèses fortes ou faibles que l'on peut émettre à propos du rapport intonation syntaxe.

Le processus de synthèse des paramètres prosodiques de la phrase visant à générer des séquences qui soient acceptables et non une variante déterminée, l'hypothèse d'un rapport étroit entre les structures syntaxiques et prosodiques est toujours valable, dans les limites assignées par les règles d'accentuabilité. Si toutes les unités accentuables sont effectivement accentuées, donc pourvues de contours mélodiques à synthétiser, cette hypothèse correspond de plus à une mise en oeuvre fortement redondante du code linguistique semblable à celle présente dans la lecture ou le discours formel. La compréhension du signal synthétisé risque donc de s'en trouver améliorée.

Manifestation des contours

Phonétiquement, les contours manifestant l'accent d'une unité syntaxique sont réalisés par les traits acoustiques de pente, d'amplitude de durée et de niveau de départ de la variation mélodique relative à la voyelle accentuée. Ces traits sont utilisés selon une hiérarchie qui peut varier d'un locuteur à l'autre, le caractère montant ou descendant restant, en français et en anglais, le trait le plus important.

Dans une application de synthèse de la mélodie, cette hiérarchie peut être fixée une fois pour toutes. Elle correspondra, dans l'ordre, aux traits de sens de la pente, amplitude, durée et niveau de départ de la variation de F_o pour le français, et de sens de la pente, durée, niveau de départ et amplitude de variation de F_o pour l'anglais, de manière à s'accorder avec les manifestations les plus courantes que l'on peut observer.

Grammaire de l'intonation

Plusieurs procédés de synthèse font appel à des règles générant des séquences de contours prosodiques à partir de la structure syntaxique (super-

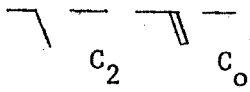
ficielle) de la phrase (par exemple, pour le français, Emerard et Larreur, 1976; pour l'anglais, Levine, 1976). Ces règles formalisent les observations empiriques réalisées à partir de l'analyse de corpus plus ou moins importants, mais limités en général à des structures simples.

Une formalisation plus complète peut être obtenue en conférant aux contours le rôle de marques indiquant les relations de présupposition qui existent entre les unités de la structure prosodique, et qui la définissent implicitement (Martin, 1975a; 1976c).

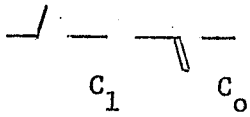
En français, il n'y a qu'un seul type de relation corrélatif de ces marques: un contour montant (descendant) indique que l'unité à laquelle il appartient présuppose à droite l'unité portant un contour descendant (montant), et présentant une caractéristique de durée et/ou d'amplitude supérieure ou égale. Si cette caractéristique est plus petite, ou si les deux contours sont de même sens, les unités prosodiques qui y sont rattachées sont dans un rapport de non-présupposition réciproque.

En anglais par contre coexistent deux types de relations de sélection (i.e. de présupposition unilatérale) entre unités prosodiques. La relation du premier type est indiquée par une opposition de niveau de départ et/ou de durée entre deux contours qui sont tous deux descendants. Celle du deuxième type est signalée par une opposition contour montant, contour descendant, le contour montant prédédent toujours le contour descendant, et s'opposant à lui par le niveau de départ et/ou la durée, plus grande pour le contour relatif à l'unité présupposée. L'utilisation de l'un ou l'autre de ces deux mécanismes est liée à la présence éventuelle d'un rapport de présupposition sémantique ou syntaxique existant entre les deux unités syntaxiques correspondant à ces contours.

Ainsi, (English teacher) "professeur d'anglais", de structure Nom +



Nom, présente deux contours descendants C_2 et C_o, et s'oppose à (English teacher) "professeur anglais", de structure Adj + Nom, pourvu d'un



contour montant sur la première unité.

TESTON Bernard
10, Butte des 3 Moulins
13 - AIX-EN-PROVENCE

Règles pour le français et l'anglais

Ces principes se formalisent par des règles simples opérant sur les contours spécifiques des deux systèmes. Ceux-ci sont décrits par des matrices phonologiques rendant compte de la hiérarchisation des traits phonétiques la plus courante. Les règles dérivent la séquence de contour de la phrase à partir d'un parenthésage décrivant la structure prosodique. Elles sont de type $C_j \longrightarrow C_i\ C_j$, C_j et C_i étant respectivement les contours relatifs à l'unité

divisée et à la nouvelle unité ou aux nouvelles unités apparaissant dans la structure: $(C_j) \longrightarrow (C_i)\ (C_j)$.

Pour le français, la matrice phonologique est:

	C_o	C_1	C_2	C_3	C_4	C_5	C_6
Long	+	-	-	-	-	-	-
Montant		+	-	+	-	+	-
Ample		+	+	-	-	-	-
Restreint				-	-	+	+

C_o est le contour final de modalité (descendant si l'énoncé est déclaratif ou impératif, montant s'il est interrogatif). Les traits acoustiques binaires utilisés sont: la longueur du contour, le sens de la variation mélodique et son amplitude (décrite par les deux traits Ample et Restreint).

La règle de division d'une unité prosodique est alors:

$$(C_j) \longrightarrow (C_{j + 2k - 1})\ \ (C_j)$$

avec j = 0, 1, ..., 6

et k = différence de niveau entre les unités de contours j et (j + 2k - 1)

L'application de cette règle à la structure prosodique

(( )( ))(( )( ))

donne successivement

(C_o)

niveau 1 ((C_1)(C_o))

niveau 2 ((C_2)(C_1))((C_3)(C_o))

Pour l'anglais, la matrice est

	C_o	C_1	C_2	C_3	C_4	C_5	C_6
Long	+	-	-	-	-	-	-
Montant		+	-	+	-	+	-
Bas		+	+	-	-	-	-
Haut				-	-	+	+

Les traits Bas et Haut se rapportent au niveau de départ des contours.

Les règles sont alors:

1. s'il n'y a pas de présupposition syntaxique ou sémantique:

$$(C_j) \longrightarrow (C_{j + 2k})\ \ (C_j)$$

2. s'il y a présupposition syntaxique ou sémantique:

$$[C_j] \longrightarrow [C_{j + 2k - 1}]\,[C_j] \qquad \text{si j est pair}$$

$$[C_j] \longrightarrow [C_j]\,[C_{j + 2k - 1}] \qquad \text{si j est impair}$$

Appliquées à la structure (( )([ ][ ])( ))

elles donnent:

(C_o)

niveau 1 ((C_2)(C_2)(C_o))

niveau 2 ((C_2)([C_3][C_2])(C_o))

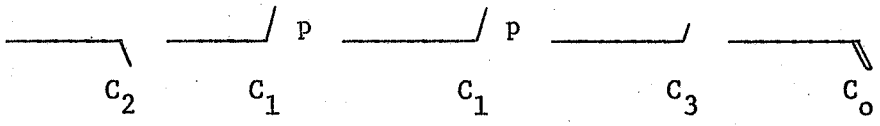
A ces dérivations s'ajoutent les marques d'insistance, portant soit sur une unité particulière (notation *), et réalisée par exemple par un contour montant descendant de faible amplitude placé sur la première voyelle (cf. Emerard et Larreur, 1976), soit sur les contours de premier niveau (notation ","), et réalisé par une pause succédant à l'unité concernée, et par une remontée du contour précédant si celui-ci est final et descendant, ou par un contour montant sur la voyelle finale si celle-ci est inaccentuée.

Règles d'accentuabilité

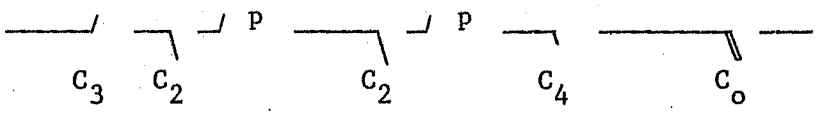
Les règles d'accentuabilité, qui définissent les conditions d'association des structures syntaxiques et prosodiques, font appel aux relations de présupposition existant entre les unités syntaxiques. On se contentera de préciser ici que toute unité syntaxique est accentuable sauf si elle est solidaire d'une autre unité ou si elle contracte un double rapport de présupposition avec d'autres unités. Dans ces deux derniers cas, c'est en général la dernière unité impliquée dans ces relations qui reçoit l'accent: ex. il ⟷ part'ait (unités solidaires), Mar'ie et P'ierre (conjonction présupposant les éléments qui l'entourent).

Les deux exemples qui suivent illustrent l'application de ces règles dans l'hypothèse d'un rapport étroit entre intonation et syntaxe.

(p = pause) ((La soeur)(de Paul)),(Annabelle),((est allée)(au cinéma))



([Paul's][sister]),(Annabella),((went)(to the movies))



Synthèse de la courbe mélodique

La mise en oeuvre de ces règles dans un dispositif de synthèse par concaténation de mots (i.e. d'unités syntaxiques minimales) nécessite une analyse préalable, par la méthode des coefficients prédictifs, des unités du dictionnaire préalablement segmentées à l'aide d'une visualisation de la forme temporelle du signal sur oscilloscope. Les frontières des voyelles et consonnes sont marquées par des repères indiquant la nature de l'élément suivant: voyelle, consonne sourde, consonne sonore type "v", consonne sonore type "—". Ces marques servent d'indication pour la génération de la micromélodie: les voyelles et consonnes sonores type — (l, m, n, r final) ne sont pas modifiées par rapport à F_o moyen, (= F_o de début de contour), les consonnes sourdes ont F_o = o, et les consonnes de type v (b, d, g, z, ʒ, ŋ, r intervocalique) reçoivent une variation en creux de F_o. Aux valeurs de ces coefficients s'ajoute celle du gain calculé par la méthode de Markel (1976).

Lors de la synthèse, les unités sont représentées par un ensemble de m x p coefficients obtenus par analyse préalable, auxquels s'ajoutent les informations sur le découpage syllabique ainsi que les indications sur les variations de l'intensité de la source. Les paramètres prosodiques à définir, essentiellement les variations du fondamental et de la durée vocalique, sont alors assignés de la manière suivante:

- modification des durées vocaliques, afin de réduire la durée des unités

selon les niveaux où elles se trouvent dans la structure prosodique affichée à l'entrée par parenthésage des unités lexicales (cf. Levine, 1976). Il n'y a pas de changement s'il s'agit de l'unité finale de la phrase;

- insertion des pauses, proportionnelles au niveau des unités qu'elles précèdent, et aussi au nombre d'unités de la phrase, donc au nombre de parenthèses formantes situées à droite des unités considérées;

- calcul des contours théoriques et inscription des variations mélodiques correspondantes sur la dernière voyelle de chaque unité, selon des caractéristiques proportionnelles au niveau moyen de F_o désiré et à la durée de la dernière voyelle dans le dictionnaire;

- placement d'un contour plat F_o sur toute la longueur restante de l'unité, égal au niveau de départ du contour placé sur la dernière voyelle, et modifié selon la catégorie des consonnes présentes: vallée avec point d'inflexion situé à - 15% de F_o moyen pour consonnes de type v, pas de modification pour consonnes de type —, et fréquence fondamentale nulle pour consonnes sourdes;

- inscription des virgules éventuelles, réalisées par une pause supplémentaire et par une variation montante du contour final descendant éventuel, réalisée au 2/3 de la durée de la voyelle, ou par un contour montant (du type C_3) sur la voyelle finale si celle-ci n'est pas accentuée;

- inscription des marques d'insistance éventuelle, par un contour montant - descendant placé sur la première voyelle de l'unité minimale syntaxique voulue;

- maintien des variations d'intensité inscrites dans le dictionnaire (pour les consonnes voisées, les rapports d'amplitude des deux sources est fixé à 2/3 - 1/3), sauf pour le dernier élément qui se voit assigner une chute de 4 dB sur la dernière voyelle.

Les structures syntaxique et prosodique étant définies indépendamment, le système proposé est très souple. Dans l'hypothèse d'un lien étroit intonation - syntaxe, le parenthésage prosodique peut être obtenu à partir d'une analyse syntaxique automatique basée sur les relations de présupposition entre unités lexicales (cf. Martin, 1976b).

REFERENCES

Emerard, F. et Larreur, D. (1976) Synthèse par diphones, Recherches / Acoustique, CNET, Lannion, III, pp. 293-314.

Levine, A. (1976) Report on Prosodic Research, Technical Report no 272, Inst. for Math. Studies in the Social Sciences, Stanford U., California.

Markel, J.D. and Gray, A.H., Jr (1976) Linear Prediction of Speech, Springer, New-York.

Martin, Ph. (1975a) Analyse phonologique de la phrase française, Linguistics (146) pp. 35-68.

Martin, Ph. (1975b) Eléments pour une théorie de l'intonation, R.A. de l'Institut de Phonétique ULB, Bruxelles, 9/1, pp. 97-126.

(1976a) L'accentuation en français. Théorie présuppositionnelle, R.A. de l'Institut de Phonétique ULB, Bruxelles, 10/1, pp. 75-82.

(1976b) Synthèse par règle de l'intonation de la phrase, Actes des 7èmes J.E.P., GALF, Nancy, pp. 205-214.

(1976c) About a Theory of Intonation, Phonologica 1976, Akten der 3 Internationalen Phonologie-Tagung, Wien, 1-4 Sept. 1976.

(1977a) Une théorie syntaxique de l'accentuation en français, in Studia Phonetica 13, Didier, Montréal (à paraître).

(1977b) Phonosyntaxe et phonosémantique vont en bateau (Questions de phonosyntaxe et de phonosémantique en français), Linguisticae Investigationes, II (à paraître).

VIII èmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

LA PLACE DES ELEMENTS VOISES DANS L'ACQUISITION DU LANGAGE

N. TIRANDAZ ET C. BERGER-VACHON

Service du Professeur MESNARD

UNIVERSITE DE LYON I

• Résumé :

Dans cet article, on se propose d'étudier l'acquisition des phonèmes par l'enfant ainsi que l'évolution du voisement avec l'âge.

Plusieurs stades sont considérés :

- les comptines enfantines (3-5 ans)
- le premier livret du français (une correspondance avec le persan est indiquée)
- le second livret du français
- le "français fondamental" de l'adulte

Cette étude montre que l'importance du fondamental diminue avec l'âge, surtout au niveau du premier phonème.

• Summary :

This paper is an approach to phonem acquisition by the child and voicing evolution.

Several ages are seen :

- the child ditties (3-5 years old)
- the first book for reading in french, a parallel is made with the persian (6 years)
- the second book for reading in french (7 years)
- the "fondamental french" for adult

It seems, in this study, that the importance of voicing diminishes as years pass, particulary for the first phonem.

I - INTRODUCTION

La reconnaissance de la parole et d'un langage en particulier, d'une part, et la modelisation d'un comportement verbal de l'autre, sont les deux axes inducteurs de ce travail. L'idée directrice, en ce qui concerne la reconnaissance est la constatation physiologique de la maturation du cerveau enfantin médicalement terminée vers l'âge de 5 ans. L'acquisition des sons par ce cerveau enfantin peut nous permettre l'élaboration de schémes de reconnaissance [1] par la machine qui est , de toute évidence, moins évoluée que le cerveau, fut-il enfantin. La seconde idée, celle du comportement verbal et la recherche d'un modèle le définissant complète la première.

La genèse de l'acquisition des sons peut être résumée, d'après Dewy (cité par G.A. Miller 1956), comme suit :

- Les sons émis par le nourrisson et le jeune enfant sont des éléments à partir desquels se développe le discours.
- Les vocalisations sont utilisées avant le mot.
- On comprend le mot avant de pouvoir l'utiliser.
- L'enfant possède un répertoire limité jusqu'à 3 ans. A cet âge le processus de l'acquisition s'accélère.
- Les substantifs apparaissent avant le verbe.

Ces traits descriptifs méritent d'être confrontés à la mesure. Pour atteindre cet objectif, nous avons choisi d'étudier l'acquisition des sons, tout au long du processus de l'apprentissage verbal de l'enfant et nous avons retenu trois époques qui nous paraissent importantes :

-3-5 ans : c'est la fin de la prime enfance
- 6 ans : l'âge de la scolarité obligatoire
- 7 ans : la fin du cours préparatoire

Pour chacun de ces âges nous avons pris comme échantillons d'étude, le matériel servant à l'apprentissage : les comptines de la maternelle, le premier livret de lecture et le deuxième livret de lecture du cours préparatoire, respectivement. Deux autres échantillons ont complété ceux-là : le premier livre de lecture des écoles publiques iraniennes et le "vocabulaire fondamental" de Henmon et Pirenne qui représente 50% de tout texte écrit de quelque importance.

Nous avons classé les phonèmes en deux catégories :

- ceux communs au français et au persan qui représente l'ensemble consonantique français et cinq voyelles et une semi-voyelle (/j/)
- ceux qui sont particuliers à chaque langue, que nous avons appelés les particularismes.

D'autre part, nous avons distingué deux catégories de phonèmes : les voisés et les muets [2] . Par rapport aux travaux déjà effectués [3] dans l'étude statistique de la parole, nous devons signaler le fait que notre étude ne s'intéressant qu'à l'acquisition, nous n'avons pas tenu compte de la fréquence des mots, chaque mot n'ayant été répertorié qu'une fois.

Pour chaque échantillon nous avons dressé le tableau matriciel dont les lignes représentent la position d'un phonème dans le mot et les colonnes se rapportant au phonème, ainsi l'élément Cj signifie que le phonème j a été rencontré C fois dans la position i. Pour illustrer cela, nous indiquerons ci-dessous, le tableau dressé pour l'échantillon des 7 ans.

(voir tableau page suivante)

positions \ phonème	occlusives						frictatives						voyelles									particularismes															
	muettes			voisées			muettes			voisées																											
	p	t	k	b	d	g	f	s	ʃ	v	z	ʒ	l	m	n	r	i	e	a	o	u	j	ɛ	y	ɸ	œ	ə	ɛ̃	ɑ̃	ɔ̃	œ̃	w	ɥ	ʁ	ɲ
1	112	61	108	60	25	33	66	89	35	37	1	28	39	67	18	44	10	21	43	25	1	2	5	2	-	5	-	8	8	3	1	2	2	-	961
2	11	14	17	7	5	8	4	8	1	10	1	1	29	7	6	115	79	34	133	64	99	26	58	22	19	6	29	21	44	24	1	48	6	2	959
3	24	59	37	24	24	9	12	66	12	28	27	16	69	27	27	106	38	19	58	31	8	29	31	8	11	2	5	20	34	10	1	10	2	13	897
4	9	37	27	5	15	4	12	34	18	10	12	8	23	22	20	55	45	23	51	53	10	42	38	21	7	8	5	20	26	27	1	15	-	4	707
5	2	36	5	5	10	9	4	22	3	1	11	5	18	12	13	59	17	22	35	21	4	34	32	11	11	8	6	7	18	22	-	4	-	4	470
6	2	12	3	1	6	1	5	14	2	2	7	2	12	12	10	48	14	15	18	3	1	20	20	5	8	7	3	4	6	12	-	1	-	1	277
7	-	6	3	-	1	-	1	8	-	-	5	1	7	3	6	18	4	6	13	7	-	15	5	-	4	3	-	1	6	2	-	1	-	-	126
8	-	3	-	-	-	-	6	4	-	-	1	-	5	4	3	6	3	2	1	-	-	9	5	-	-	1	-	3	4	8	-	-	-	-	68
9	-	1	-	-	-	-	-	-	-	-	-	-	-	-	2	4	1	-	-	1	-	4	-	-	-	1	-	1	-	6	-	-	-	-	21
10	-	-	1	-	-	-	-	-	-	-	-	-	-	-	1	1	-	-	-	-	-	-	-	-	-	-	-	1	-	3	-	-	-	-	7
Σ	160	229	201	102	86	64	110	245	71	88	65	61	202	154	106	456	211	142	352	204	123	181	194	69	60	41	48	86	146	117	4	81	10	24	4493

Tableau de distribution des phonèmes - Français 7 ans

II - LES RESULTATS OBTENUS

Pour chaque échantillon, nous allons résumer les résultats obtenus, en ce qui concerne le voisement, sous la forme d'un tableau de 7 colonnes. La première colonne, P, indique la position dans le mot considéré. La deuxième colonne donne par ordre décroissant les phonèmes les plus fréquents dans cette position dont la fréquence cumulée atteint au moins 50% des effectifs. Les troisième et quatrième colonnes donnent le nombre respectif des voyelles et des consonnes de cette liste et la cinquième colonne indique la fréquence cumulée tandis que la sixième colonne donne la fréquence relative du premier phonème (f_1) de la liste. Quant à la dernière colonne, elle donne le taux des phonèmes voisés dans la position considérée (f_v). La ligne G donne ces mêmes indications pour l'ensemble des phonèmes de l'échantillon, toutes positions confondues.

II.1 - Les comptines

P	phonèmes	V	C	f_i	f_1	f_v
G	a , r , o , i , e , s , t , ã	5	3	58, %	11,2%	82%
1	s , d , l , r , p	1	5	50,2%	12,3%	66%
2	a , a , i , ə	4	0	50,6%	14,1%	96%
3	r , a , t , z	1	3	50,9%	17,6%	84%
4	a , r , o , i , e	4	0	52 %	11,2%	82%

On remarquera la prédominence des sons voisés en 2ème position et la forme CVCV très nette du langage.

II.2 - Français : premier livret

P	phonèmes	V	C	f_i	f_1	f_v	observations
G	a r i l o m n	3	4	57,4%	11,5%	79,3%	t = 6,6
1	p m s t f l	0	7	54,9%	11,8%	58,1%	
2	a i o r	3	1	55,8%	22,2%	95,9%	
3	r a l n i m	2	4	51%	14,6%	76,2%	t = 9
4	i e o r a	4	1	51%	15,9%	85%	t = 8,4

Le voisement de la 2ème position reste prépondérant, la forme CVCV demeure, les éléments non voisés du langage apparaissent principalement en 1ère position.

La consonne t qui apparait avec une fréquence relative au moins égale au dernier de la liste en 1ère, 3ème et 4ème position est indiqué en marge du tableau.

II.3 - Persan 6 ans

P	phonèmes	V	C	f_i	f_1	f_v	observations
G	ɑ a r e n m i	4	3	52,4%	11,2%	83,5%	pour la ligne G
1	b m d p s t k	0	7	51%	8,9%	65,6%	b et d équi-
2	ɑ o e i u	5	0	58%	19,2%	96,7%	valents aux 2
3	r h n ʃ b	0	5	51%	21,8%	82%	derniers phonèmes
4	ɑ a e i	4	0	50,6%	18,8%	89,8%	

Le voisement de la 2ème position reste au même niveau que les autres échantillons. Si nous regardons les fréquences des phonèmes par rapport à l'ensemble des phonèmes communs aux deux langues, les particularismes de chacune étant exclus, nous constatons que la fréquence des sons voisés est de

76,9 % pour le français. Ceci nous amène à remarquer que les particularismes de chaque langue interviennent peu dans l'acquisition verbale à ce niveau là et selon notre échantillon.

II.4 - Français : deuxième livret

P	phonèmes								V	C	f_i	f_1	f_v	observations
G	r	a	s	t	i	o	k	l	4	5	51,2%	10,2%	77,4%	
1	p	k	s	f	t	b			0	6	51,7%	11,7%	51%	m à 7%
2	a	r	u	i	o				4	1	51,2%	13,9%	94,3%	
3	r	l	s	t	a	i	k	a	3	5	52,2%	11,8%	76,6%	
4	r	o	a	i	j		t	s	5	3	50,5%	7,8%	80,6%	

La concentration des non voisés en première position s'accompagne d'une dispersion plus grande tandis que la forme CVCV reste vraie malgré une atténuation qui s'accentue avec l'âge et avec la longueur des mots.

II.5 - Adulte

P	phonèmes								V	C	f_i	f_1	f_v
G	d	r	t	s	v	l	ə	e	3	5	53,1%	10,6%	81,2%
1	s	t	p	d	v	l	m		0	7	60%	14,1%	62,4%

Nous n'avons pas donné les autres positions car ces mots outils sont généralement courts (bi ou triphonémiques).

A signaler cependant que le taux de voisement est de 94,7 pour la 2ème position.

III - CONCLUSION

En comparant les lignes de tous les échantillons, on constate une certaine constance :

- le nombre d'éléments phonémiques constituant 50% des fréquences cumulées (7 pour les comptines, 8 pour l'enfant de 7 ans)
- le taux de voisement, l'introduction des éléments non voisés avec l'âge. A ces constatations intuitivement compréhensibles viennent s'ajouter :

a) le maintien de la forme CVCV du langage qui donne sur notre façon d'ordonnancement des phonèmes une sorte de " balancement" sur le diagramme des fréquences (alternance consonne - voyelle), par exemple pour le persan :

En 1ère position, sur 7 phonèmes, 6 sont dans le premier tiers du diagramme, tandis qu'en 2ème position les 5 phonèmes les plus fréquents sont dans le dernier tiers.

Ce "balancement" s'atténue, comme on pouvait s'y attendre, avec la longueur des mots.

b) l'inversion des sons voisés sur la 1ère position entre les comptines et les autres échantillons ; pour les comptines on a 2 éléments sur 5 muets, tandis que pour le vocabulaire de 7 ans on en a 6 sur 6.

c) la diminution progressive du voisement avec l'âge

d) la constance du taux de voisement au-delà de 90% pour la 2ème position.

Ceci nous permet de penser que pour la reconnaissance de la parole par un cerveau enfantin (jusqu'à l'âge de 5 ans, correspondant à la fin de la maturation) un code CVCV avec des consonnes voisées semble favorable.

Plus généralement, selon le schéma Piagetien [4] on peut dire qu'on assiste à deux phases d'assimilation et d'accomodation dans le comportement verbal. Assimilation des sons voisés et accomodation des sons non voisés à partir du commencement de l'acte verbal. Ce qui laisse penser qu'un modèle adaptatif tenant compte de ce double phénomène est un premier pas vers cette modélisation.

REFERENCES

1 BERGER-VACHON C. : "Conception d'une entrée vocale automatique" Thèse de Doctorat es Sciences - Lyon - 1975

2 ROSSI M. : "La matrice phonologique de JACKOSON, FANT et HALLE" Journées d'études sur la parole - avril 1971

3 HATON J.P. : "Contribution à l'analyse, la paramétrisation et la reconnaissance automatique de la parole" Thèse d'état - 5 janvier 1974

4 PIAGET J. : "Six études de psychologie" DENOEL/GONTIER - 1964

THEME 3

Intonation et syntaxe

VIIIèmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

TRAITS, INDICES ET PROPRIETES DU VOISEMENT :

VERS UNE PHONOLOGIE DES UNITES INFRA-PHONEMIQUES

Implications pour la détection automatique et perceptive

C. ABRY & L.J. BOË

Institut de Phonétique de Grenoble

• Résumé :

L'état d'avancement des recherches en phonétique sur le trait de voisement et ses indices perceptifs ainsi que les performances auxquelles parviennent les nouveaux procédés qui traitent la plus vieille détection en reconnaissance de la parole, ont rendu nécessaire que soit spécifié le type d'unité linguistique dont chacun traite un ou plusieurs corrélats. traits, indices et propriétés ne doivent plus être confondus :
Nous proposons, en tant que cadre, une théorie particulièrement bien adaptée à définir les unités linguistiques : la grammaire générative transformationnelle.
Nous évoquons aussi quelques implications concernant les niveaux auxquels se situent les détections perceptive et automatique.

• Summary :

Past ten years findings in phonetics on the features of voicing and its multiple cues as well as the sophistication of new apparatuses which perform nowadays the oldest detection in automatic speech recognition have made necessary to specify which type of linguistic unit correlate each one is processing : i.e. features, cues and properties are no more to be confused. We propose, as a frame, stratificational generative grammar, a device particularly well fitted for such linguistic specification.
Some implication concerning the levels at which perceptual and automatic recognition are performed will be also discussed.

TRAITS, INDICES ET PROPRIETES DU VOISEMENT : VERS UNE PHONOLOGIE DES UNITES INFRA-PHONEMIQUES

Implications pour la détection automatique et perceptive

C. ABRY & L.J. BOË *

L'état d'avancement des recherches sur la multiplicité des indices de la corrélation p/b, ..., f/v, etc. vient poser de nouveaux problèmes à tous ceux qui se préoccupent de relier la structure linguistique à sa principale manifestation, la communication parlée.

DU TRAIT A L'INDICE

En phonologie, à la belle époque de la biunivocité entre le trait de voisement et son corrélat acoustique, le voice-bar, l'existence d'un concurrent tendu/lâche, avait déjà alimenté un long passé de controverses. N'opérant pas avec l'indice, les raisonnements ne pouvaient concevoir que l'un des deux traits soit simplement un indice de l'autre, ou mieux, que les deux soient des indices du trait : la seule alternative possible pour l'unité minimale de description était d'être pertinent ou redondant. Ces façons de voir ont encore cours de nos jours, et si quelques phonéticiens se sont préoccupés de donner des règles de réalisation des traits, aucune théorie phonologique ne s'est jusqu'à présent souciée d'intégrer la multiplicité des indices. La phonologie de descend pas encore au-dessous du trait.

En phonétique, ce sont les études acoustiques et perceptives qui, depuis HASKINS[1], ont véritablement fait de l'indice un acquis. Mais là aussi, les tentatives pour relier biunivoquement structure phonologique et manifestation phonétique, qui se révèlent dans la quête d'un indice prépondérant, montrent bien que l'hypothèque phonologique n'est pas facile à lever. La publicité donnée à la seule dimension V.O.T., et la recherche de l'en-deçà physiologique de son invariant articulatoire , qui subsumerait les différents indices acoustiques[2], est certainement l'exemple le plus frappant de la persistance de cette démarche. Ainsi, il aura fallu attendre plus de 10 ans après HASKINS, pour que des chercheurs européens reprennent les stratégies expérimentales sur les indices dans l'optique d'une véritable simultanéité[3-4].

---

* Cet article a bénéficié des apports implicites de D. CREISSELS (linguistique) M. CARTIER & F. ZURCHER (acoustique), W. SERNICLAES & M. WAJSKOP (psycho-acoustique).

La discussion de la relation entre trait et indice se retrouve au coeur des recherches sur les détecteurs perceptifs, lorsqu'il s'agit de déterminer quel est le niveau de spécialisation de ceux-ci : leur unité de traitement est-elle alors le trait[5], l'indice[6] ou encore, non contradictoirement, les deux[7].

Si la détection automatique, elle, peut sembler pour l'instant, à l'abri de ces discussions théoriques, c'est que les problèmes de la relation son - phonème y sont rarement explicitement posés. Après avoir cru détecter directement les unités du linguiste, ou du moins les symboles de sa transcription, les ingénieurs se sont créés, faute souvent de propositions linguistiques en ce domaine, différentes unités "bien à eux", le "diphonème" par exemple, quitte à redécouvrir l'allophone rebaptisé récemment pseudo-phonème.

Pour prendre l'exemple de la détection qui est à la fois la plus ancienne et la plus opérationnelle - celle du voisement - il est maintenant clair, à la lumière de ce qui vient d'être dit plus haut, qu'il ne s'agit pas de la détection d'un trait mais tout au plus de l'indice "vibration laryngée" de ce trait. Cette confusion est loin d'être simplement terminologique, puisqu'un programme de reconnaissance qui comprend une détection classique du voisement, se révèlera incapable de distinguer [p] de [b̥], [b] et [p̬], et par suite, d'identifier sous leur "forme de base" un nombre élevé de mots, du simple fait de l'assimilation de sonorité, courante en français.

DE L'INDICE A LA PROPRIETE

Si les recherches en psychophonétique sur les indices nous ont appris à ne pas les confondre avec le trait, un examen quelque peu approfondi des nombreux procédés qui détectent la fonction de voisement, montre qu'il ne s'agit pas en fait de la détection d'un indice, mais bien de différentes propriétés de cet indice. Celles-ci sont nombreuses, presqu'autant que les indices, mais tous les procédés de détection, en étant parfois fort différents, n'exploitent pas pour autant des propositions différentes. A notre connaissance la seule tentative de classification des procédés, sur la base de ces propriétés, se trouve dans PIROGOV[8]. Ce dernier en énumère 5 pour les sons à vibrations laryngées : (I) quasi périodicité, (II) non équi-répartition temporelle de la puissance à court terme, (III) distribution a-normale des valeurs instantanées, (IV) prépondérance de la puissance au bas du spectre, (V) structuration formantique définie. On peut à notre avis en rajouter deux autres qui se différencient à la détection. Les détecteurs de présence d'énergie au bas du spectre (VI) qui n'exploitent pas la même propriété que les comparateurs (IV) et le comptage du nombre de passage à zéro (VII) qui donne davantage d'informations sur la répartition spectrale (il est plus efficace que (IV) pour [z] et [ʒ]). Les disparités de détection des procédés[9-10] font apparaître que les différentes propriétés ne sont pas des corrélats invariants de l'indice "vibration laryngée". Ce n'est que la bonne corrélation d'une seule propriété, la périodicité, qui peut créer une telle illusion.

LA PLACE DES TRAITS, DES INDICES ET DES PROPRIETES DANS UNE THEORIE LINGUISTIQUE

Avant de se demander quelle peut être la définition relative de ces unités dans la structure linguistique, il conviendrait de poser la question de l'opportunité de les y introduire. Les arguments que l'on peut utiliser à cette

fin vont dépendre de la conception que l'on peut avoir de la structure linguistique. Du strict point de vue de la phonologie classique, doivent être retenus du signal tous les éléments susceptibles de faire changer le contenu linguistique du message : la pertinence du son est d'ordre sémantique. La phonologie générative a modifié cette conception : est retenu dans la représentation phonétique du signal tout ce qui peut permettre de remonter aux représentations morphophonologiques de base : la pertinence n'est plus directement sémantique, mais elle est commandée par les exigences des représentations morphémiques et des règles de dérivation.

Dans la première conception, il suffira d'avoir démontré, par des procédures linguistiques ou des tests psycho-phonétiques, que l'indice et la propriété sont pertinents, pour que ces unités aient un statut phonologique. S'il n'y a aucun problème pour décerner un tel brevet aux indices, il n'en va pas de même pour les propriétés qui, elles, n'ont pas bénéficié jusqu'à présent de la même attention.

Dans la seconde optique (générativiste) il faudra montrer que la négligence des indices et des propriétés ne permet plus (ou moins bien) d'extraire du signal les éléments indispensables pour remonter au trait nécessaire à l'application des règles. Par exemple le changement de *d* dans *grand ami* (assimilation de trait de voisement) peut être traité avec la même règle que dans *grande folle* (assimilation partielle c'est-à-dire d'indice). Mais ne pas faire la différence [d̥] et [t] parce que l'on confond trait et indice montre que la phonologie générative, si elle veut continuer à travailler sur des propriétés du signal, prévisibles par les règles de dérivation, devra elle aussi descendre au-dessous du trait.

Dans les limites de cet exposé nous avons choisi de ne pas aborder en priorité les problèmes que poserait l'introduction des indices et des propriétés dans le cadre de la grammaire générative transformationnelle. Ceci, non pas parce que le modèle se révèlerait incapable de les résoudre sans modifier profondément sa composante phonologique, mais parce que nous estimons que le cadre qu'il propose n'est pas très propice à dégager le statut linguistique des unités sur lesquelles opèrent les règles étant donné que l'existence linguistique de ces unités n'est fondée qu'en fonction des exigences pratiques de ces mêmes règles.

Par contre, le modèle stratificationnel[11-14] (M.S) est particulièrement bien adapté aux définitions des relations entre unités linguistiques, nous avons choisi la version qui a le mieux élaboré sa phonologie. Ce modèle est aussi génératif mais avant d'envisager son fonctionnement, il nous faut rappeler les deux notions fondamentales de tout M.S. : les différences de rang et les différences de niveau entre unités.

DIFFERENCE DE RANG

Après ce que nous avons dit des trois unités (trait, indice et propriétés) on admettra sans peine qu'il y ait entre elles une différence dite de rang : comme le phonème est composé de traits, le trait est composé d'indices et l'indice de propriétés. A ce stade, les raisons pour ne pas composer directement le phonème en propriétés sont de même nature que celles qui font reculer tout un chacun devant une description de phrases directement en suite de phonèmes, sans passer par le morphème et/ou le mot. Comme le trait peut servir à composer plusieurs phonèmes, un indice peut-il entrer dans plusieurs traits, une propriété dans plusieurs indices ?

Nous n'avons pu recenser que peu d'exemples. Peut être ce manque n'est-il que provisoire et tient-il au fait qu'aucun trait n'a été aussi exploré que le voisement. Quoiqu'il en soit, ce manque de polyvalence combinatoire pourrait avoir une incidence directement pratique. En effet, si chaque trait a ses propres indices et chaque indice ses propres propriétés, on arrive vite, par quelques rapides estimations, à 360 propriétés différentes (en partant des 12 traits "universels", à raison d'une moyenne de 6 indices par trait et de 5 propriétés par indice). Pour le français, on descend au mieux à 140 (en prenant les 28 indices de DELATTRE[1] avec 5 propriétés par indice). Cette rapide évaluation laisse l'ingénieur de la parole avec la perspective peu réjouissante de devoir "construire" plus de 100 détecteurs de propriétés pour une langue comme le français qui n'a guère plus de 30 phonèmes. Après un stade de grande prolifération des découvertes phonétiques, il reste donc à tenter un effort sérieux de reprise synthétique des indices et des propriétés, de recherche d'invariant, de polyvalence et, en attendant, d'étude des contraintes proprement linguistiques, phonologiques[15], qui peuvent augmenter grandement les chances de détection.

DIFFERENCES DE NIVEAU

On est amené à poser une telle différence entre deux ensembles d'unités linguistiques chaque fois qu'il n'y a pas correspondance bi-univoque. Le phonème a différents allophones ([p] dans *cape*, [p$^{\neg}$] dans *apte*, [p$_{\vee}$] dans *cape bleue*, etc.). Ainsi le trait aura différents allo-traits. Le trait de voisement est représenté dans les différents contextes par différents groupements et indices. A l'initiale, la durée de la tenue et celle de la voyelle préconsonantique n'interviennent évidemment pas; dans *trompe perdue*, [p$^{\neg}$] n'ayant pas de détente, l'indice durée et intensité de l'explosion est neutralisé, etc... Différentes études ont montré qu'un indice pouvait avoir aussi des allo-indices. C'est le cas du V.O.T. qui varie entre autres en fonction de l'aperture des voyelles, du lieu des consonnes, des groupements de consonnes. C'est encore le cas des différents groupements de propriétés qui représentent l'indice suivant les contextes (par exemple, la propriété (IV) n'entre pas dans la composition des sibilantes). De même une propriété comme la périodicité peut avoir différentes allo-propriétés, ne serait-ce que le nombre de raies signifcativement plus élevé à la finale qu'à l'initiale, en anglo-américain.

TRAIT, INDICE ET PROPRIETE DANS UN MODELE STATIFICATIONNEL

Issu pour ainsi dire des diagrammes de HOCKETT qui combinent les relations de composition (différence de rang) et celles de représentation (niveau), le M.S. a été développé en phonologie essentiellement par LAMB. L'exemple de la relation entre morphèmes, allomorphes et phonèmes a servi à illustrer cette double distinction. Un morphème n'est, en effet, pas composé de phonèmes qui sont, non seulement de rang mais de niveau différent; c'est l'allomorphe qui est de même niveau que le phonème qui en est composé. Les morphèmes sont alors composés de morphophonèmes de même niveau. L'habitude de rechercher une forme de base en phonologie (ou en morphographie) générative pour les allomorphes de *grand* (grã, grãt, grãd), par exemple, peut aider à comprendre ce qu'est un morphophonème par rapport à un phonème. Le *d* de la représentation de base de *grand* est l'unité qui permet par ses règles de prévoir le passage aux phonèmes /t/ ou /d/ des allomorphes; c'est donc une unité qui n'est pas itentifiable au phonème : il est de même rang, mais pas de même niveau.

Les M.S. tels qu'ils ont été élaborés ne descendent pas eux non plus

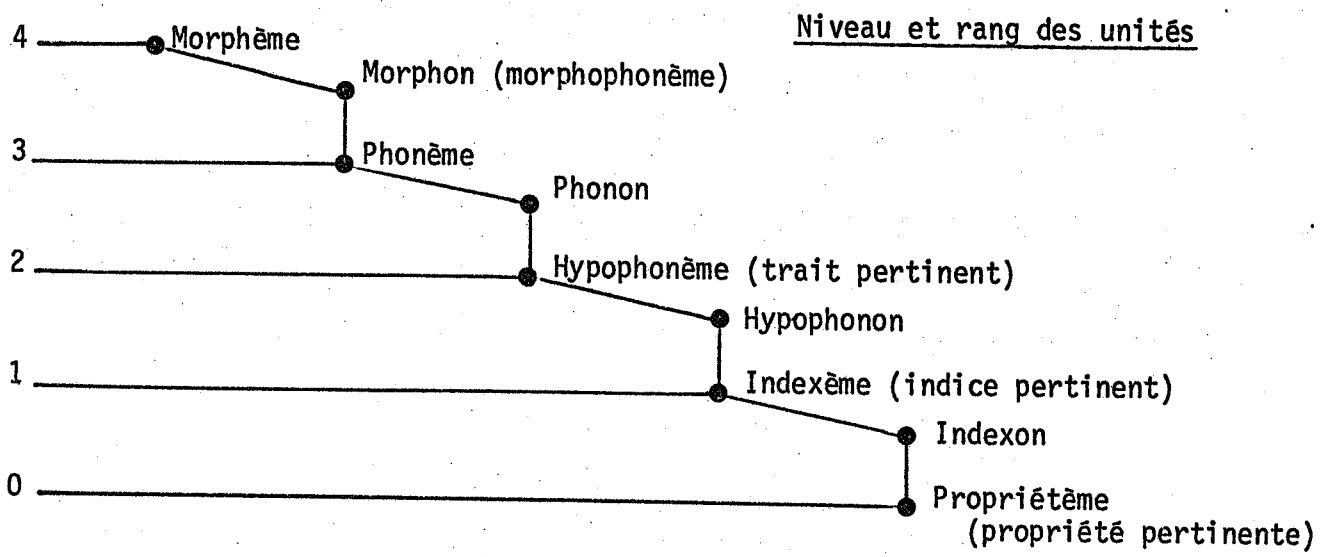
au-dessous d'unités de la taille du trait, pour la simple raison qu'ils ignorent l'existence d'unités inférieures. LAMB ne s'est pas complètement désintéressé de cette manifestation, mais elle ne peut présenter, selon lui (il avoue ne pas en être sûr) qu'un type de structuration différent de l'organisation linguistique. Ce que nous proposons en réponse à sa dernière question : "Could it be another stratum ?" n'est pas de passer du trait à sa manifestation, ceci a déjà été tenté mieux que nous ne pourrions le faire[16] , mais bien plutôt de reculer les limites de la structuration linguistique jusqu'à la propriété dans la mesure où il peut être montré que cette dernière a une pertinence linguistique.

Le diagramme montre l'économie du système, du morphème à la propriété. Chaque noeud est étiqueté dans la terminologie un peu luxuriante de LAMB. Les relations obliques sont des relations de composition (rang) les verticales de représentation (niveau). Ainsi un morphème comme *grand* est composé de morphons (ou morphophonèmes) *g r ã d*, par exemple, et le morphon *d* est représenté au niveau inférieur par les phonèmes /t/, /d/ ou par zéro, de même rang. Ces phonèmes sont composés de phonons de même niveau mais de rang inférieur, qui sont représentés à leur tour par des hypophonèmes (trait pertinent) de niveau inférieur. Qu'est-ce alors que le phonon ? C'est un trait pertinent qui est neutralisé dans certaines conditions de composition des phonèmes. C'est par exemple, dans le cas du trait de voisement, sa neutralisation quand il entre dans la composition des nasales en français (indifférentes du point de vue du voisement). Si l'on dépasse le terminus descendant du M.S. cela donne en structurant sur le même modèle les indices et les propriétés : des hypophonons qui composent le trait pertinent et qui sont représentés par des indexèmes (indice pertinent à son niveau). Un exemple d'hypophonon sera celui d'un indice neutralisé lorsqu'il entre en composition avec d'autres indices représentant un trait pertinent. Le V.O.T. des fricatives sourdes (délai entre la fin du bruit de friction et le voisement vocalique) qui existe en américain mais aussi en français, semble un indice non utilisé. Inversement l'indice des joints hauts relevé et testé à la synthèse par DELATTRE[1] pour les fricatives n'est pas utilisé pour les occlusives.

Les indexèmes à leur tour sont composés d'indexons qui sont représentés par les propriétèmes (propriétés pertinentes). Un exemple sera fourni par une propriété comme IV qui est neutralisée lorsqu'elle entre en composition avec le trait de sibilation de /s, z, ʃ, ʒ/[9].

Voici donc le modèle présenté sous sa forme statique. Deux mots de son fonctionnement génératif. Les passages de rang à rang et de niveau à niveau sont effectués grâce aux composantes tactiques. Ce sont elles qui donnent les contraintes sur les structures de composition acceptables et les conditions sur les règles de réalisation. Ces règles peuvent être exprimées sous forme graphique propre à LAMB ou sous une forme algébrique plus courante.

Outre que le M.S. a l'avantage de fournir une méthodologie propre à éviter certaines confusions de notions, il offre la possibilité d'appliquer aux indices l'analyse des contraintes phonologiques qui a déjà fait ses preuves pour les traits.



REFERENCES

[1]. DELATTRE, P. (1967). 6th Int. Congr. Phonetic Sci., 35-47.

[2]. LISKER, L. & ABRAMSON, A.S. (1964). Word 20, 384-422.

[3]. SUMMERFIELD, A.Q. & HAGGARD, M.P. (1974). J. Phonetics 2, 279-285.

[4]. WAJSKOP, M. & SWEERTS, J. (1973). J. Phonetics 1, 121-130.
SERNICLAES, W. (1975). R.A. Inst. Phonet., Bruxelles 10/1, 83-103.

[5]. COOPER, W.E. (1974). J. Phonetics 2, 303-313.

[6]. BAILEY, P.J. (1974). S.C.S. Stockholm, preprints 3/4, 47-51.

[7]. SERNICLAES, W. (1975). 6° J.E. Parole. GALF 1, 246-255.

. BEECKMANS, R. & SERNICLAES, W. (1975). 8th Int. Congr. Phon. Sci.(à paraître).

[8]. PIROGOV, A.A. (1974). Vokodernaja Telefonja. Metody i problemy. Ed. Svajaz'.

[9]. ABRY, C., BOË, L.J. & ZURCHER, J.F. (1975). 6° J.E. Parole. GALF 1, 228-245.

[10]. RABINER, L.R., CHENG, M.J., ROSENBERG, A.A. & Mc GONEGAL, C.A. (1976). IEEE Trans. ASSP - 24, 399-418.

[11]. LAMB, S.M. (1966). Outline of stratificational Grammar. Georgetown Univ. Press.

[12]. LAMB, S.M. (1966). Language 42, 536-573.

[13]. LIBERMAN, A.S. (1974). Trav. Inst. Phon. Strasbourg 6, 13-22.

[14]. SGALL, P. & al. (1969). A Functional Approach to Syntax in Generative Description of Language. Elsevier, New-York.

[15]. ROSSI, M. (1975) 6° J.E. Parole. GALF 1, 160-190.

[16]. LINDBLOM, B. (1971). 7th Int. Congr. Phonetic Sci., 63-97.

VIII èmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

LIGNE MELODIQUE et "QUE" COMPLETIF EN FRANCAIS DE MONTREAL

Louise Dagenais et Alain Marchal

• Résumé :

Cette communication traite de la corrélation entre le niveau syntaxique et l'intonation. Nous avons examiné en particulier le problème de la complétive en que dans les cas où le que se réalise et ceux où il disparaît. L'effacement du /que/ en français de Montréal permet d'observer le phénomène de récupération de l'information syntaxique par la courbe mélodique. Les cas observés permettent d'envisager une généralisation applicable à la reconnaissance automatique.

• Summary :

This paper deals with the interrelation between syntax and intonation. We examine the complementation phrase in que with and without que. The deletion of que in French spoken in Montreal is a good example of the way that syntaxic information can be restated by the Fo contour. Cases where que appears and those whereit is deleted permit a generalization that could be used for automatic speech recognition.

LIGNE MELODIQUE et "QUE" COMPLETIF EN FRANCAIS DE MONTREAL

Louise Dagenais et Alain Marchal
Laboratoire de Phonétique Expérimentale
Université de Montréal

Quelques études récentes (Connors, 1975) ont suscité notre intérêt pour un phénomène courant en français de Montréal: la chute du "que" complétif. La théorie générative standard définit la complémentation en "que", comme sous-catégorisant chaque verbe - du moins est-ce là une analyse possible. Tout verbe sous-catégorisé [+ Que], est dit complétif. Par exemple: sembler, penser, dire, paraître (3ième pers. sing. seulement), falloir, vouloir, se souvenir, trouver, prévenir, accepter, savoir, se rappeler, "étant donné", etc... (N.B. Cette liste constitue la liste exhaustive des verbes relevés dans la partie du corpus analysé).

Notre étude a voulu vérifier une hypothèse qui veut que dans les cas d'effacement du "que", l'information soit récupérée par la courbe mélodique. Nous avons minutieusement examiné le contour mélodique dessiné sur la dernière syllabe de la principale et sur le mot qui lui est adjacent à droite, afin de détecter toute perturbation potentiellement liée à la perte de "que".

Constitution du corpus

Nous avons utilisé le corpus Sankoff-Cedergren[1]. Retenant les enregistrements de 17 locuteurs féminins, de tout âge et tout statut socio-économique, nous avons recueilli et analysé sur sonagrammes: 27 phrases où le "que" est absent, 32 où le "que" apparaît phonétiquement plein, i.e. sous la forme /kd/ et 15 où la consonne /k/ apparaît, liée à la voyelle qui suit (v.g. les pronoms on, "i",...) ou formant un groupe consonantique (v.g. que c'est [ksE]).

Analyse

Nous avons procédé à l'analyse sonagraphique, en bande large (300 Hz) pour le découpage des segments, et en bande étroite (45 Hz) pour le tracage de la courbe mélodique.

---

1. Le corpus Sankoff-Cedergren est composé de 120 enregistrements représentatifs du français parlé à Montréal.

Résultats

Syllabe précédant la complétive	1ère syllabe de la complétive	QUE = Ø 27	QUE = Kd 32	QUE = K + V = K + C + V 15
1. montée	descente	23	19	8
2. montée	Fo légèrement plus élevé - descente	3	3	1
3. plat	montée ou ton H		2	1
4. plat (H) (x)	plat (M) (x-)	1	5	2
5. plat (x)	reprise à égalité		3	3

On entend ici par montée, un écart bien dessiné du contour mélodique au-dessus de la syllabe, par rapport à la syllabe qui précède, ou par rapport à la partie initiale de la syllabe considérée. La valeur de l'écart (Seuil - 30 Hz), bien que souvent importante, n'a pas de signification en soi. Par descente ou chute, on doit comprendre le mouvement bien dessiné de la chute du Fo dans les mêmes proportions que la montée. Il existe des cas pour lesquels il n'est pas possible de relever de montée ou de descente, au sens où nous l'entendons: c'est-à-dire qu'il n'y a pas de contour. On observe néanmoins une différence de niveau entre le fondamental des voyelles à la frontière de la principale et de la complétive. H, haut: signifie qu'à ce niveau la fréquence fondamentale est supérieure à la fréquence laryngienne moyenne (dans les exemples traités: maximum absolu). M, moyen vaut pour les cas où le niveau de Fo lui correspond. Nous avons noté (x-) et (x+) pour les cas où la valeur de Fo sur la syllabe initiale de la complétive s'avérait à peine plus élevée ou plus basse que celle de principale qui la précédait; x est une variable sur H ou M (moyen).

Analyse linguistique

Nos résultats indiquent pour une majorité de cas, qu'une complétive en que peut être identifiée par une montée sur la syllabe finale de la principale. Cependant, les trois variantes de la complétive en que que nous avons analysées, ne répondent pas dans les mêmes proportions à ce critère.

En examinant de plus près les résultats, on observe que le comportement du contour mélodique, de la variante où le "que" complétif est absent est beaucoup plus régulier. Pour 24 des occurences (les cas 1, et 4), il est manifeste

que le Fo atteint une valeur nettement plus élevée sur la syllabe finale de la principale. Les trois occurences du cas 2, où la première syllabe de la complétive est coiffée d'un contour qui au début est légèrement plus élevé que le contour ascendant de la voyelle qui précède, sont un peu plus délicates à traiter. Remarquons immédiatement que cette différence est minime et que d'autre part la syllabe de la complétive est caractérisée par une chute du Fo. Ceci nous ramène en quelque sorte au cas numéro un. Si on voulait par exemple que nos résultats trouvent quelque application en reconnaissance de la parole, voilà ce que nous proposons: il faudrait concevoir, 1) un découpage en segments, puis un découpage syllabique; 2) un décodage du Fo inscrit au-dessus de ces syllabes qui tienne compte du fait que sur le premier segment de la complétive, la chute du Fo s'amorce assitôt, et départager ainsi la complétive de la principale.

Les résultats moins systématiques obtenus pour les variantes avec que, s'expliquent aisément par le fait de la présence d'un ou de deux segments phonétiques, selon la variante utilisée, peuvent suffir à eux seuls à l'analyse en S + Complétive. En dépit des variations de contours, on peut observer que les comportements de contours 1, 4, -- $\begin{bmatrix} H \\ x \end{bmatrix} + \begin{bmatrix} M \\ x- \end{bmatrix}$ ou montée - chute, i.e. les cas où le Fo est plus élevé sur la principale, constituent 24 des 32 cas pour la variante en [kə] et 10/15 pour la variante en $\begin{bmatrix} K+V \\ K+C+V \end{bmatrix}$; ces résultats corroborent le comportement du Fo dans les cas de chute de que.

Si, comme nous le concevons, on peut inclure le cas 2 dans les cas réguliers (à cause de la chute qui se dessine sur le [kə] ou $\begin{bmatrix} K+Voy. \\ +C+V \end{bmatrix}$ qui suit, nos résultats gagnent en régularité: 27/32 pour la variante en [kə] et 11/15 en $\begin{bmatrix} K+V \\ K+C+V \end{bmatrix}$.

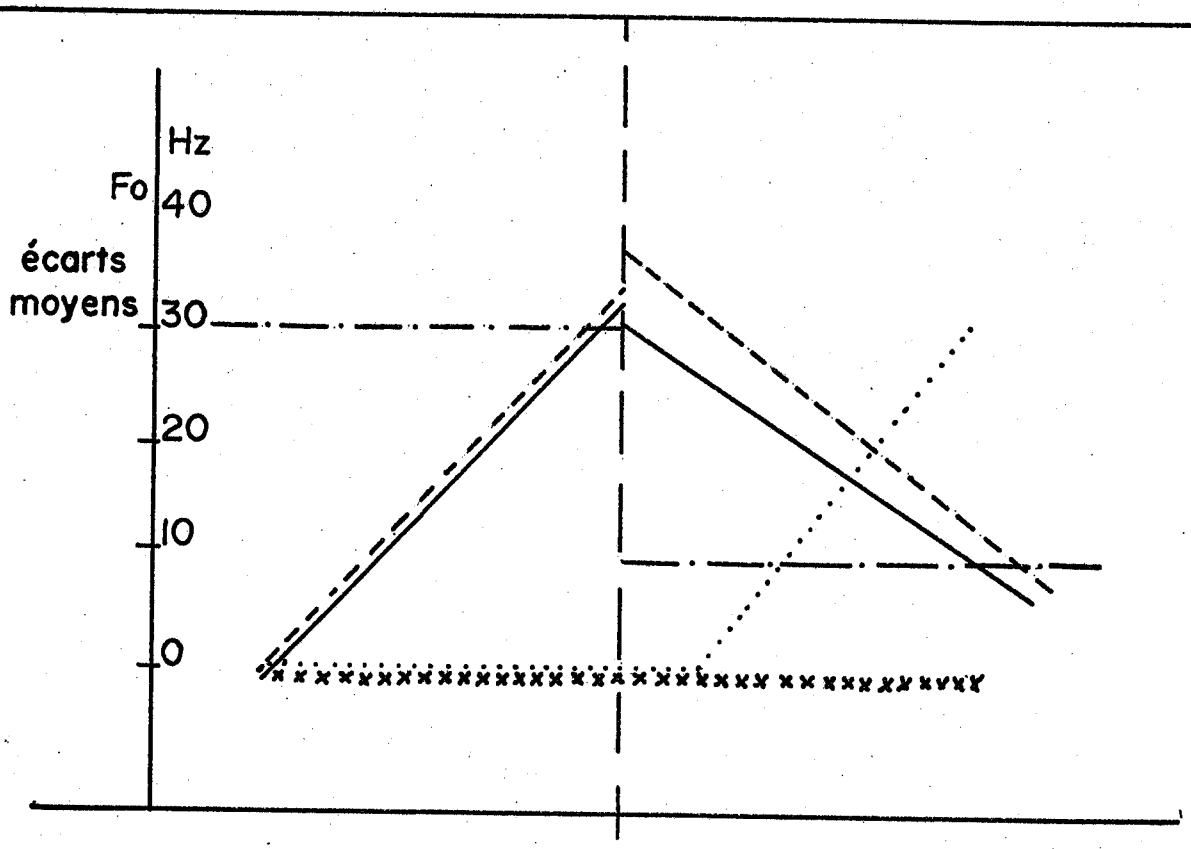
En conclusion, on peut caractériser très nettement la variante complétive sans que, par une montée du Fo sur la dernière syllabe de la principale suivi d'une descente sur la syllabe de la complétive. Les variantes où le que apparaît sont moins systématiquement caractérisées par cette montée H - ou un niveau Haut. Néanmoins et particulièrement dans le cas de la variante en [kə], des résultats réguliers à 84.3% nous semblent suffisants pour une généralisation significative.

Référence:

CONNORS, K.; "L'effacement du que - Règle syntaxique" - Recherches linguistiques à Montréal 4: 17-34 (1975)

DERNIERE SYLLABE de la PRINCIPALE

PREMIERE SYLLABE de la COMPLETIVE



——— (1) = 50 occurrences } 57/74

- - - - - - - - (2) = 7 occurrences

..... (3) = 3 occurrences

—·—·—·— (4) = 8 occurrences } 8/74

×××××××× (5) = 6 occurrences

Les variations du Fo. entre la PRINCIPALE et la COMPLETIVE

VIII èmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

EXPERIENCES PERCEPTIVES SUR L'INTONATION

BERTRAND DELGUTTE

RESEARCH LABORATORY OF ELECTRONICS

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

• Résumé :

Un groupe d'auditeurs a formulé des jugements quantitatifs sur le naturel de phrases dans lesquelles on a fait varier systématiquement les contours de F_0 à l'aide d'un synthétiseur. On a appliqué ce paradigme à l'étude de phrases courtes contenant des mots susceptibles d'être mis en relief par la modalité sémantique. L'influence de la syntaxe et de la modalité sur les contours de F_0 se manifeste plus souvent par des tendances que par des règles strictes. Au niveau phonétique, les contours de F_0 sont représentés par une succession de contours-types en petit nombre. La validité de cette représentation est discutée à la lumière des jugements perceptifs.

• Summary :

A group of listeners made quantitative judgements about the naturalness of sentences for which the F_0 contour has been systematically varied with the help of a speech synthesizer. This paradigm was applied to the study of short sentences containing words likely to be emphasized by semantic modality. Syntax and modality tend to influence the F_0 contours, but their influence does not strictly follow a set of rules. At the phonetic level, the F_0 contours were described as sequences of a small number of patterns. The consistency of this description with the perceptual judgements is evaluated.

* This work was supported in part by grant number NS-04332 from the National Institutes of Health.

EXPERIENCES PERCEPTIVES SUR L'INTONATION

Préliminaires méthodologiques

Peu d'études de l'intonation ont utilisé les techniques expérimentales perceptives au niveau de la phrase. Quelques expériences ont porté sur les aspects prédicatifs de l'intonation ou sur certains cas d'ambiguités levées par l'intonation. Seule l'école hollandaise d'intonation[2,8] a appliqué les techniques perceptives aux aspects les plus généraux de l'intonation. Ces techniques souvent fondées sur des jugements de similarité entre contours de F_0, ont permis de définir avec précision les éléments phonétiques qui constituent l'intonation hollandaise, les contraintes qui régissent leurs combinaisons, et comment les auditeurs intègrent ces éléments dans des formes intonationnelles abstraites.

La présente communication explore l'hypothèse qu'on peut déduire certaines contraintes linguistiques ou phonétiques sur les paramètres prosodiques (par exemple les contours de F_0) à partir de comparaisons entre le naturel de plusieurs valeurs de ces paramètres. On peut voir le naturel comme étant analogue à la notion de grammaticalité. Pour que l'association d'un jugement de naturel à un paramètre prosodique ait un sens, il faut (1) que le stimulus (valeur du paramètre) soit contrôlé avec précision, (2) que le jugement ne dépende pas d'un seul individu, et (3) que ce jugement soit quantitatif. En pratique, on a réalisé les stimuli à l'aide d'un synthétiseur en faisant varier sytématiquement le contour de F_0 d'une phrase dont tous les autres paramètres de synthèse étaient fixés. Ces stimuli ont été ensuite présentés à un groupe d'une vingtaine d'auditeurs représentatifs du français général. Pour certaines expériences (5, 6) on avait enregistré les stimuli sur bande magnétique. Les auditeurs exprimaient par un nombre le degré de naturel de chaque stimulus par rapport à un étalon. On a présenté chaque stimulus au moins deux fois pour vérifier la consistance des auditeurs. Pour d'autres expériences (1,2,3,4), les stimuli étaient enregistrés sur carte "Language Master" ce qui permettait aux auditeurs de comparer les stimuli ad libitum. Les auditeurs classaient les cartes par ordre de naturel. Apres l'expérience, on a regroupé pour chaque stimulus les réponses des différents auditeurs, et on a testé statistiquement les différences entre les moyennes des distributions de reponses. Ces tests donnent un sens a l'expression: "stimulus A est plus/moins/aussi naturel que stimulus B".

On applique cette technique expérimentale à l'étude de l'influence de la syntaxe et de la modalité sur les contours de F_0. Le mot modalité regroupe commodément les éléments linguistiques qui expriment l'attitude ou l'opinion du locuteur vis-à-vis de son énoncé. Plus précisément, l'étude porte sur des phrases du type [sujet-verbe-objet] dans lesquelles on a introduit systématiquement des mots ou constructions qui peuvent exprimer l'attitude du locuteur: verbes modaux (pouvoir, devoir, vouloir), constructions impersonnelles, négatifs, adverbes, quantificateurs. Les expériences perceptives ont été précédées par une analyse des contours de F_0 d'un corpus lu par trois locuteurs. Ce corpus comprenait une soixantaine de phrases du type décrit ci-dessus.

Etant donné que le nombre de contours de F_0 que l'on peut distinguer dans une phrase est très élevé, il est important dans ce type d'expérience d'avoir une description schématique des contours de F_0 qui restreint a priori le nombre de contours possibles. On peut utiliser les techniques perceptives pour affiner ou tester ce système de description phonétique. Nous avons adopté avec certaines modifications le système de Vaissière[9]. Ce système représente le contour de F_0

d'une phrase par une succession de contours-types surimposés sur une ligne de base descendant en pente douce. Chaque contour-type est une succession de mouvements élémentaires de F_0 (Fig. 1). Grosso-modo, un contour-type est associé à un mot lexical de 3 syllabes ou plus. Les mots courts se regroupent sous un même contour type, ou reçoivent un contour-type dont certains mouvements sont supprimés (voir les mouvements entre parenthèses, Fig. 1). Un test perceptif où les auditeurs jugeaient quelles syllabes étaient proéminentes dans plusieurs phrases, a montré que seules les syllabes contenant P ou R_p étaient souvent jugées proéminentes. Parmi les 160 syllabes du test, tous les mouvements sauf R_f etaient représentés. Ce résultat suggère d'appeler P et R_p "mouvements proéminents". Comme les contours-types P3, P4 et PN n'ont pas de mouvement proéminent, on les désignera par "contours-types inaccentués". Un énoncé déclaratif se termine normalement par un P4 ou PN. Le contour-type P1 (continuation) marque la fin d'un groupe de souffle non final et n'est pas étudié dans nos expériences qui portent sur des phrases contenant un seul groupe de souffle. Le mouvement P du contour-type P2 réalise une frontière prosodique à l'intérieur du groupe de souffle. La "règle d'accent"[9] prévoit que cette frontière a lieu de préférence a la fin d'un constituant syntaxique majeur. Le contour-type PE dont la première syllabe est proéminente (ou parfois P2 dont la dernière syllabe est proéminente) permet de mettre en relief le mot sur lequel il se produit. Pour les mots monosyllabiques, on ne peut distinguer a priori P2 et PE, et le contour-type correspondant se note P2/E.

Syntaxe, modalité et intonation

Pour les phrases du types [sujet-(auxiliaire) - (modal)-verbe-objet], l'analyse du corpus a montré qu'une grande variété de configurations de contours-types sont possibles. Le plus souvent, le sujet reçoit un P2, et le verbe un P2 ou un inaccentué. Ce résultat est en accord avec ceux de Di Cristo[5] selon lesquels le groupe nominal sujet se termine par une frontière prosodique majeure, et la base verbale par une frontière prosodique mineure facultative. S'il est présent, l'auxiliaire est toujours inaccentué, et le modal souvent inaccentué. Dans certains cas, le modal peut être mis en relief par un PE ou même un P2. Ceci se produit particulièrement quand la phrase est insérée dans un contexte qui suggère un sens particulier pour le modal (par exemple probabilité pour le verbe "pouvoir"). Plus rarement, le verbe peut être rendu emphatique par un PE. L'expérience 1 (voir les stimuli Fig. 2) confirme les résultats de l'analyse du corpus: la configuration de contours-types la plus fréquente est la plus naturelle. En particulier, un sujet inaccentué est moins naturel qu'un P2, ce qui suggère que la présence d'une frontière prodosique entre groupe sujet et groupe verbal est importante pour la perception de la phrase. Certaines autres configurations observées dans le corpus, par exemple la variante de modalité avec mise en relief du modal par un PE, sont aussi acceptables. Par contre, les configurations de contours-types observées plus rarement sont moins naturelles. Par exemple, un P2 sur le modal n'est pas naturel, peut-être parce que la réalisation d'une frontière entre le modal et le verbe rompt l'homogénéité du groupe de sens [modal-verbe]. De même, une emphase sur le verbe "manger" par un PE -surprenante dans des conditions expérimentales - est peu naturelle.

Pour toutes les phrases du corpus, l'introduction d'une négation produit des changements importants sur le contour de F_0: le négatif "pas" reçoit un P2/E, le verbe devient inaccentué, et le modal ne reçoit plus de P2 ni de PE. Par conséquent, on n'observe qu'une seule configuration de contours-types. L'expérience 2 (voir Fig. 3) confirme partiellement ces résultats: la configuration de contours-types la plus naturelle est celle qui a été observée dans le corpus. En particulier, un P2 au lieu d'un inaccentué sur le modal ou sur le verbe cause une forte diminution du naturel. Cependant, l'expérience perceptive

et l'analyse du corpus ne coïncident pas sur deux points. Premièrement, le négatif peut être inaccentué sans affecter le naturel, ce qui suggère que, lors de la lecture du corpus, les locuteurs ont eu tendance à mettre en relief le négatif qui était un mot nouveau. Deuxièmement, une variante de modalité, avec mise en relief du modal par un PE, est tout à fait naturelle bien qu'elle n'ait pas été observée dans le corpus. Le fait que cette mise en relief ne puisse être réalisée par un P2 suggère l'interprétation suivante: P2 introduit une frontière prosodique en violation de la syntaxe entre le modal et son négatif, tandis que PE ne crée pas de frontière prosodique. En d'autres termes, PE a une fonction purement contrastive, mais P2 a une fonction à la fois démarcative et contrastive.

L'analyse du corpus a aussi montré que les adverbes, les verbes impersonnels (falloir, se pouvoir), et les quantificateurs reçoivent généralement un contour-type P2 ou PE. Pour ces phrases, on n' a observé qu'une ou deux configurations de contours-types. Cependant, dans l'expérience 3, portant sur la construction impersonnelle "Il fallait que Brigitte mange des tartines", toutes les configurations de contours-types étaient également naturelles. En particulier, un P2 sur l'impersonnel n'est pas plus naturel qu'un inaccentué. L'expérience 4, portant sur la phrase "Beaucoup de garçons mangent des tartines" a donné des résultats similaires: tous les stimuli avaient le même naturel, y compris ceux avec un quantificateur inaccentué. Quelles que soient les causes de ces désaccords entre les expériences perceptives et l'analyse du corpus, ce résultat suggère qu'il est imprudent de fonder la formulation de règles d'intonation exclusivement sur un corpus lu.

En conclusion, l'influence de la structure syntaxique sur les contours de F_0 se traduit par une tendance à marquer la fin d'un constituant syntaxique majeur par une frontière prosodique. La violation de cette tendance par la création d'une frontière prosodique en contradiction avec la syntaxe cause parfois une diminution du naturel. Cependant, omettre la frontière prosodique à la fin d'un constituant affecte rarement le naturel. Par ailleurs, les parties de l'énoncé encadrées par deux frontières prododiques semblent parfois être des "groupes de sens" plutôt que des constituants syntaxiques au sens strict. Pour cette raison, l'expression "aspect démarcatif de l'intonation" semble plus appropriée que "influence de la syntaxe sur l'intonation." L'influence de la modalité se traduit par une tendance à mettre en relief certaines parties de l'énoncé (adverbes, négatifs, quantificateur , ...). L'emphase sur un mot inapproprié, ou dans un contexte inattendu, peut causer une diminution du naturel. Cependant, ne pas mettre en relief l'un de ces mots (adverbes ...) ne semble pas affecter le naturel. Comme le relief des différentes parties de l'énoncé dépend d'autres facteurs que de la modalité (par exemple le contexte), nous préférons parler d'"aspect contrastif de l'intonation". De nombreux auteurs ont abouti à des conclusions similaires [1,3,6,7].

Expériences perceptives et systèmes phonétiques

Dans cette section, nous montrons par deux exemples comment les techniques perceptives permettent d'améliorer le système phonétique. Jusqu'ici, nous avons sans justification regroupé les contours-types P3, P4 et PN sous la dénomination "inaccentué". L'expérience 5 (Fig. 4) étudie les relations entre ces trois contours-types dans un cas particulier. Les stimuli S2 et S5, qui sont des exemples de P4, sont les plus naturels. Les auditeurs n'ont pas pu les distinguer bien qu'ils diffèrent d'environ 1,3 ton. S3 (un exemple de P3), et S4 (un P4 approchant un PE) sont assez naturels. S7 (un P2 très schématique) est peu naturel dans ce contexte(voir expérience 1). S1 (un exemple de PN), et S6

qui n'est pas un contour-type sont les moins naturels, sans doute à cause des sauts abrupts de F_0 au début ou à la fin du modal. Sl serait peut-être plus naturel s'il y avait une pause entre le sujet et le modal. Ces résultats ne sont pas incompatibles avec l'hypothèse que P3, P4 et PN sont des variantes contextuelles d'un contour-type inaccentué sous-jacent qui serait caractérisé par une absence des mouvements de F_0 nécessitant un effort (montées, descentes abruptes, sauts). Il reste à voir si on peut généraliser cette hypothèse à d'autres contextes phonétiques.

Le fait que P2 et PE permettent tous deux de mettre un mot en relief suggère d'étudier si, à partir de ces deux contours-types, un locuteur peut produire une proéminence avec la même facilité. Cette question est abordée dans l'expérience 6, où l'on a demandé aux auditeurs de juger le degré de proéminence du mot "volontiers" dans la phrase "Brigitte mange volontiers des tartines". Les stimuli diffèrent en ce que l'adverbe est réalisé avec un P2 ou un PE, chacun à trois niveaux différents séparés par des intervalles d'un ton (Fig. 5). Les résultats montrent que dans le domaine de variation étudié, la proéminence croît avec la hauteur et que, dans ce contexte particulier où ils ont la même probabilité d'apparition, P2 et PE créent la même proeminence quand leurs pics ont la même hauteur. En conclusion, on peut produire une même proéminence à partir de P2 ou PE, avec des efforts comparables. La possibilité pour le locuteur de donner plus ou moins de relief à un mot en faisant varier la hauteur d'un pic de F_0 n'est pas prévue dans notre système phonétique. Notons aussi que, si dans cette expérience un intervalle de 1 ton entre deux mouvements proéminents est perceptible, dans l'expérience 5 un intervalle de 1,3 ton entre deux mouvements non proéminents n'était pas perceptible. Si de tels phénomènes étaient fréquents, ils mettraient en doute la validité de la procédure qui consiste à appliquer à la phrase les valeurs de seuil de F_0 pour les syllabes isolées[5].

En conclusion, le système phonétique de Vaissière semble compatible avec les jugements de naturel, mais il lui manque une autre dimension (hauteur des mouvements proéminents) pour qu'il puisse rendre compte des phénomènes de proéminence. A ce point, on peut se demander dans quelle mesure nos résultats dépendent du choix d'un système phonétique particulier, et il est intéressant de comparer plusieurs systèmes. Comme les systèmes de Delattre[4], de Di Cristo[5] et de Martin[6], le système de Vaissiere[9] permet un découpage hiérarchique de l'énoncé en plusieurs niveaux d'unités: l'unité de niveau 1, terminée par Pl, est le groupe de souffle; l'unité de niveau 2, terminée par P2, pourrait s'appeler "groupe d'accent". Mais il est possible qu'il y ait en realité un plus grand nombre de niveaux: (1) P3 pourrait réaliser une frontière prosodique de niveau inférieur a P2[9]; (2) puisque les auditeurs sont capables de distinguer plusieurs hauteurs du mouvement P (expérience 6), des P2 ayant des pics de hauteurs différentes pourraient marquer la fin d'unités de niveaux différents. Comme dans le système de Di Cristo[5], les oppositions entre frontières prosodiques de differents niveaux dans le système de Vaissière[9] ne sont pas seulement réalisées par la pente et par l'ampleur de la dernière syllabe de l'unité, mais aussi par la forme du contour de F_0 sur les syllabes précédentes. Nous pensons que beaucoup de divergences entre systèmes phonétiques pourraient êtres éclaircies par des expériences perceptives semblables à celles illustrées dans cette communication. Le naturel n'est pas le seul attribut qui puisse faire objet de jugements.

REFERENCES

(1) Allen, J. and O'Shaughnessy D. (1975) "Fundamental frequency contours of auxiliary phrases in Enlgish", 6èmes Journées d' Etude sur la Parole, II, 7-47.

(2) Collier, R. (1972) "From pitch to intonation", unpublished Doctoral Dissertation, Catholic University of Leuven.

(3) Collier, R. and t'Hart, J. (1975) "The role of intonation in speech perception", in "Structure and Process in Speech Perception", A. Cohen and S.C. Noteboom (eds), Springer-Verlag, New York.

(4) Delattre, P. (1966) "Les Dix Intonations du Base du Français", French Review, 40, 1-14.

(5) Di Cristo, A. (1975) "Recherches sur la structuration prosodique de la phrase française", 6èmes Journées d'Etude sur la Parole, I, 94-116.

(6) Maeda, S. (1976) "A characterization of American English Intonation", Unpublished Doctoral Dissertation, MIT.

(7) Martin, P. (1975) "Analyse phonologique de la phrase française", Linguistics, 146, 36-67.

(8) t'Hart, J. and Collier, R. (1975) "Integrating different levels of intonation analysis", Journal of Phonetic, 3, 235-255.

(9) Vaissière, J. (1975) "Caractérisation des variations de la fréquence du fondamental dans les phrases françaises", 6èmes Journées d'Etude sur la Parole, I, 39-50.

FIGURES

Figure 1.: Contours-types et mouvements élémentaires

Les mouvements entre parenthèses réalisés ou non selon le contexte phonétique.

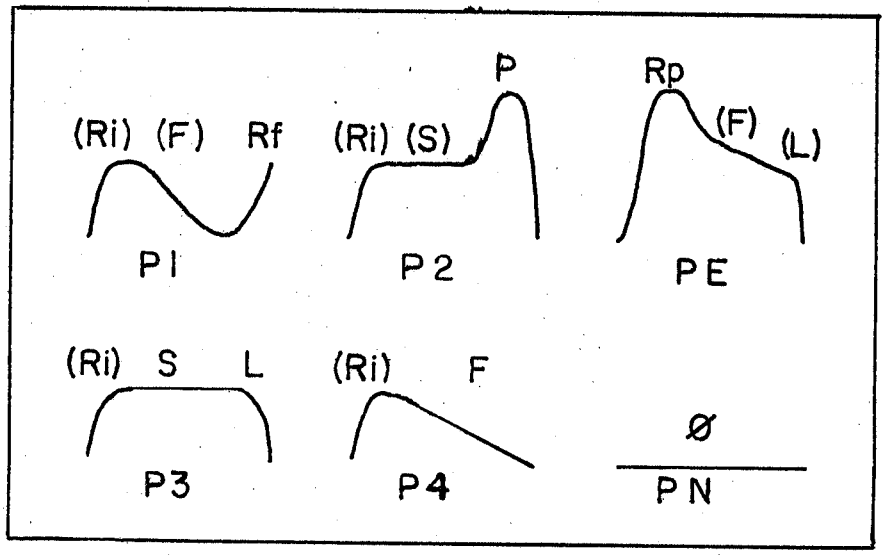


Fig. 2.: Stimuli de l'expérience 1

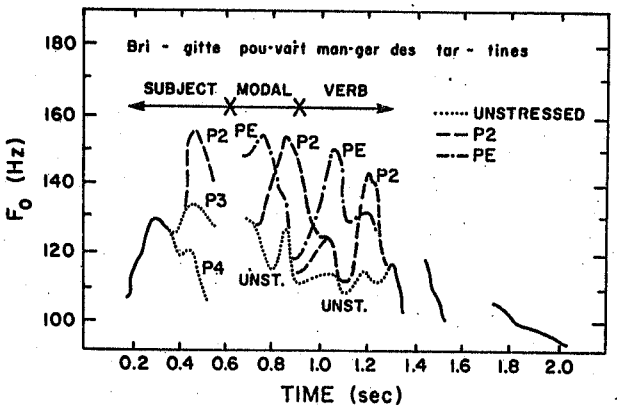


Fig. 3.: Stimuli de l'expérience 2

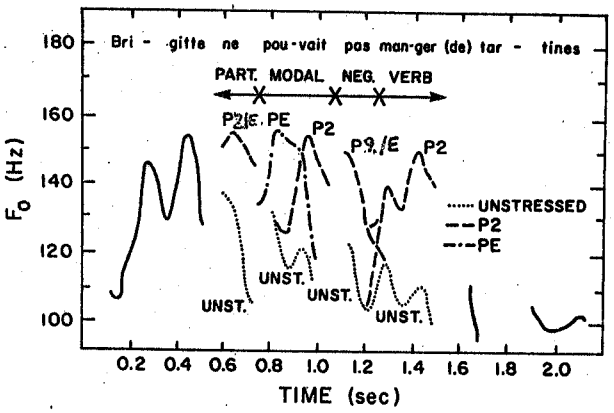


Fig. 4.: Stimuli de l'expérience 5

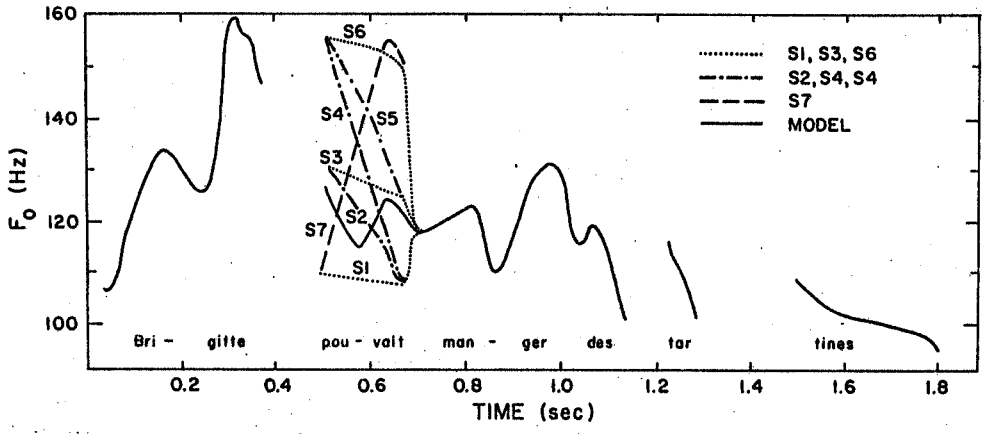
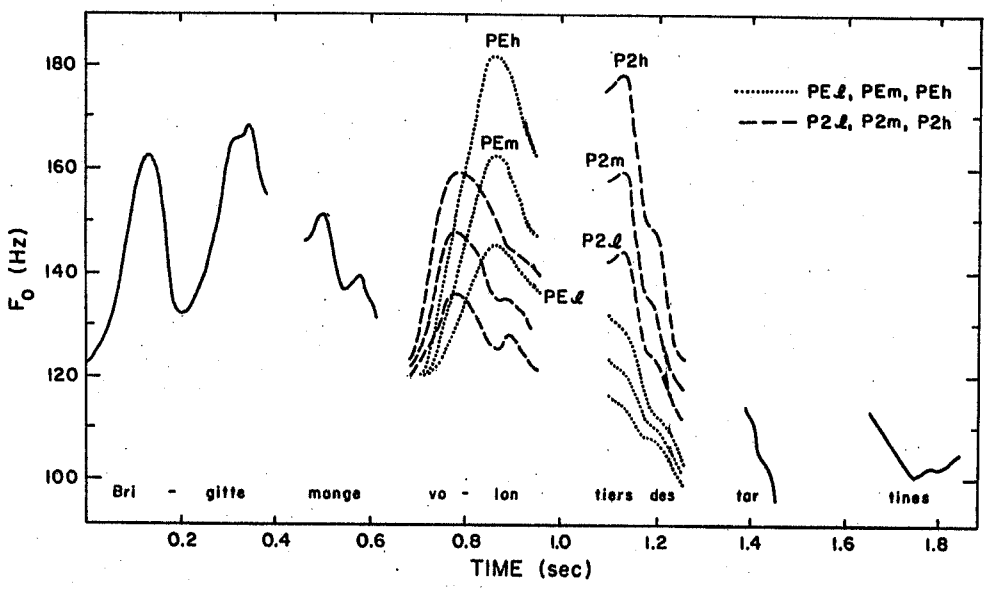


Fig. 5.: Stimuli de l'expérience 6



VIII èmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

POUR UNE DETERMINATION DES FONCTIONS GRAMMATICALES DE L'INTONATION : ANALYSE DE QUELQUES TESTS D'ENCODAGE ET DE DECODAGE.

M.C. HAZAEL-MASSIEUX

• Résumé :

La recherche de critères permettant une meilleure définition des fonctions grammaticales de l'intonation peut être effectuée au moyen d'un certain nombre de tests d'encodage et de décodage présentés et exploités ici. Le rôle de la mélodie dans la transmission d'informations grammaticales est ainsi mis en évidence. On peut distinguer plus nettement le grammatical de l'émotif en analysant le consensus des sujets qui reçoivent le message et en comparant l'information visée par l'encodeur et l'information reçue par le décodeur. La multiplication de travaux de cet ordre ouvre des perspectives pour l'établissement de grammaires de l'intonation de chaque langue qui préciseraient les fonctions morphologiques et syntaxiques de l'intonation.

• Summary :

The investigation of criteria leading to a better definition of the grammatical functions of intonation may be carried out using a certain number of coding and decoding tests that are presented and exploited here. The part played by melody in the transmission of grammatical information is thus emphasized. What is grammatical can be distinguished more clearly from what is emotive by analysing the consensus of the subjects that receive the message and by comparing the information provided by the coder and those received by the decoder. The multiplication of such studies opens up new perspectives for establishing intonative grammar for every language that would define the morphological and syntactical functions of intonation.

POUR UNE DETERMINATION DES FONCTIONS GRAMMATICALES DE L'INTONATION : ANALYSE DE QUELQUES TESTS D'ENCODAGE ET DE DECODAGE.

Les tests d'intonation peuvent permettre de préciser la façon dont l'intonation est perçue, ou interprétée par des auditeurs. On les utilise également à l'encodage pour voir comment un sujet est susceptible de produire une phrase conformément à des consignes qui peuvent lui être données plus ou moins explicitement. Ils sont indispensables pour déterminer par exemple ce qui est pertinent dans une courbe mélodique (que ce soit pour la reconnaissance automatique de la parole ou pour la synthèse), et ce qui est fonctionnel dans l'intonation d'une langue, c'est-à-dire ce qu'une grammaire de l'intonation doit retenir.

Dans la perspective de la recherche des fonctions grammaticales de l'intonation, et plus précisément de la détermination des courbes mélodiques fonctionnelles d'une langue, on a tenté de recourir à divers tests d'encodage et de décodage. On a entrepris, par ce moyen, de voir comment des sujets (en tout 52 personnes ont été testées, soient 21 + 31) interprètent des phrases qui leur ont été soumises. Ces phrases intonées au nombre de 14 sont constituées par la structure "Il a acheté une voiture" qui reçoit des intonations différentes.

Contrairement à ce qui est généralement pratiqué dans ce domaine, on n'a pas demandé à un acteur ou à un locuteur talentueux d'enregistrer ces phrases, mais à 4 sujets à qui l'on donnait des consignes de réalisation (avec colère, réponse naturelle, etc.) et qui n'avaient reçu aucune formation théâtrale, ou qui n'avaient aucune pratique de l'art dramatique.

Il ne s'agit donc pas de voir si l'encodage est bon (en le contrôlant par le décodage) ou si le décodage est bon (en le contrôlant par l'intention), ce qui est à juste titre une procédure contestable. Il s'agit de préciser :

- quelles sont les phrases pour lesquelles la transmission de l'information par l'intonation (puisque aucun autre procédé n'est laissé à la disposition des sujets testés) est bonne - sans préjuger de l'information en question (1). Ces phrases sont signalées par le pourcentage élevé de réponses semblables données par les sujets chargés d'interpréter les phrases (consensus des réponses élevé) ;

- quelles sont les phrases pour lesquelles la transmission de l'information par l'intonation est mauvaise - toujours sans préjuger de l'information en question : dans ce cas précisément, il pourra être parfois très difficile de définir une information précise, quand le trop grand nombre de réponses différentes ne laisse pas la possibilité d'en retenir une comme probante. C'est ici la dispersion des réponses (importante) qui permettra de déterminer les phrases dont l'interprétation semble particulièrement difficile aux sujets testés. De façon annexe, l'élévation du chiffre des réponses nulles ou ininterprétables confirmera pour ces phrases les difficultés au décodage.

---

(1) Bien qu'une consigne soit donnée à l'encodeur, les décodeurs peuvent déclarer recevoir une information tout autre.

Dans le tableau ci-dessous, on ne s'étonnera pas de trouver 18 phrases au lieu des 14 annoncées : on a voulu présenter certaines phrases deux fois afin de voir si elles étaient décodées chaque fois de la même façon, quand la phrase qui les précède immédiatement est différente, et quand il y a seconde audition. Il s'agit bien évidemment de la même phrase reportée deux fois sur la bande enregistrée et non pas de la même phrase enregistrée deux fois. On constate dans tous les cas, que pour chacune des phrases doublées, la réponse demeure la même, avec toutefois une légère diminution du nombre des réponses lors de la deuxième reproduction. Tout se passe comme si les décodeurs - pourtant avertis du fait qu'une même phrase pouvait se retrouver plusieurs fois-,avaient hésité à interpréter comme une phrase déjà rencontrée une phrase re-donnée. D'où le refus de se prononcer (réponses nulles) ou la volonté de trouver une autre interprétation que l'on peut constater pour quelques uns des décodeurs.

Notons enfin que si nous produisons ici les chiffres obtenus lors du test I (31 personnes testées sur le locuteur I) et lors du test II (21 personnes testées sur le locuteur II), nous avons aussi effectué des tests d'interprétation sur les locuteurs III et IV. Nous avons été amené à sélectionner les deux premiers tests à l'exclusion des autres pour les raisons suivantes :

Les réalisations des phrases 1, 5, 13, 9, 12, 16, 17, 3, 14 (1) par le locuteur III sont pratiquement semblables à celle de la phrase 2 et ne sont dès lors pas interprétées par les auditeurs autrement que comme "réponse" ou "ton indifférent" (avec tout au long de la présentation du test un accroissement des réponses nulles, les décodeurs ennuyés de leur "manque d'imagination" tentant de s'en excuser en refusant de répondre au lieu de donner toujours la même réponse).

Quant aux tests de décodage portant sur le locuteur IV, nous avons dû aussi les éliminer pour une autre raison : pour arriver à produire les phrases demandées, ce locuteur a dû recourir à des bruits expressifs divers : "brr..." pour l'effroi, "oh" et "ah" d'impatience ou de contentement, soupir pour le regret, etc. Ces bruits ne se retrouvent que pour les phrases 1, 13, 9, 16, 17, 3, 14, mais rendaient inutilisables une fois de plus les résultats obtenus par ces phrases lors du décodage, le consensus souvent important pouvant être mis sur le compte des onomatopées et bruits divers et non porté au crédit de l'intonation.

Pour les locuteurs III et IV, notons que les résultats obtenus pour les phrases 10, 15, 2, 11, 18, 7, 4, 6, 8 restent comparables à ceux qui ont été obtenus par les locuteurs I et II : fort consensus, interprétation conforme à l'intention du locuteur (cf. ci-dessous).

Les 14 phrases ont été enregistrées par les sujets encodeurs dans l'ordre suivant (on a par la suite modifié l'ordre et supprimé les contextes destinés à faciliter l'endocage) :

---

(1) Ceci tend à renforcer les résultats obtenus en montrant que l'encodage de ces phrases est beaucoup plus difficile que celui des autres phrases.

a - Il a acheté une voiture ?
b - Il a acheté une voiture ? (avec colère)
Il a gagné une grosse somme. Qu'a-t-il fait de tout cet argent ?
c - Il a acheté une voiture. (réponse naturelle)
d - Il a acheté une voiture. (évidence)
e - Il a acheté une voiture. (avec colère)
f - Il a acheté une voiture. (avec contentement)
g - Il a acheté une voiture. (avec hésitation)
h - Il a acheté une voiture. (avec effroi)
i - Il a acheté une voiture. (avec insistance)
j - Il a acheté une voiture. (avec regret)
k - Il a acheté une voiture. (avec impatience)
l - Qu'a-t-il fait de tout cet argent ? Il a acheté une voiture ?
m - Il a acheté une voiture. (triomphe claironnant)
n - Il a acheté une voiture. (interpellation)

Elles ont été soumises à l'interprétation des sujets décodeurs dans l'ordre suivant (18 phrases à interpréter) :

1) Il a acheté une voiture (f)
2) Il a acheté une voiture (c)
3) Il a acheté une voiture (e)
4) Il a acheté une voiture (d)
5) Il a acheté une voiture (g)
6) Il a acheté une voiture (b)
7) Il a acheté une voiture (m)
8) Il a acheté une voiture (d)
9) Il a acheté une voiture (h)
10) Il a acheté une voiture (a)
11) Il a acheté une voiture (c)
12) Il a acheté une voiture (i)
13) Il a acheté une voiture (j)
14) Il a acheté une voiture (e)
15) Il a acheté une voiture (a)
16) Il a acheté une voiture (f)
17) Il a acheté une voiture (k)
18) Il a acheté une voiture (n)

On constate que certaines phrases entraînent un consensus important au décodage (avec corrélativement une faible dispersion des réponses), alors que d'autres suscitent des réponses très dispersées, voire apparemment contradictoires.

TABLEAU

Phrase N°	Intention	Nombre de décodeurs ayant donné l'interprétation prioritaire		
		Test I (sur 31)	Test II (sur 21)	Total (sur 52)
10	interrogation	30 : interrogation	21 : interrogation	51
15	interrogation	28 : interrogation	20 : interrogation	48
2	assertion	30 : assertion	16 : assertion	46
11	assertion	28 : assertion	16 : assertion	44
18	interpellation	----------	15 : interpellation	----
7	triomphe	----------	13 : triomphe	----
4	évidence	25 : évidence	9 : évidence	34
6	interrogation + colère	11 : interrogation	21 : interrogation + étonnement	32
8	évidence	19 : évidence	9 : évidence	28
1	contentement	15 : envie	13 : étonnement	28
5	hésitation	12 : hésitation	16 : indifférence	28
13	regret	13 : regret	7 : indifférence	20
9	effroi	11 : étonnement	8 : évidence	19
12	insistance	10 : insistance	8 : insistance	18
16	contentement	9 : envie	8 : étonnement	17
17	évidence + impatience	7 : évidence	9 : indignation	16
3	colère	7 : étonnement	7 : insistance	14
14	colère	3 : étonnement	8 : insistance	11

Les premières phrases du tableau (10, 15, 2, 11, 18, 7, 4, 6, 8), caractérisées par le consensus important des décodeurs, sont également celles pour lesquelles il y a transmission effective de l'information : l'interprétation donnée par le plus grand nombre des décodeurs est conforme à l'intention de l'encodeur. Ce n'est pas le cas pour les phrases suivantes dans le tableau qui, décodées avec hésitation (accroissement du nombre des réponses nulles ou ininterprétables) et sans nette convergence des décodeurs, sont interprétées le plus souvent non conformément à l'intention du locuteur. L'information visée ne passe pas (si ce n'est éventuellement pour un nombre très faible de locuteurs - généralement très minoritaire : 1 ou 2 décodeurs-) et c'est une autre qui est reçue.

Les premières phrases sont en outre remarquables par le fait que chaque phrase a une courbe mélodique caractéristique qui l'oppose à toutes les autres et que les réalisations mélodiques de ces phrases ne varient pas sensiblement d'un locuteur à l'autre (locuteurs I, II, III, IV). En revanche, les autres phrases (1 à 14 dans le tableau) sont à peu près toutes produites sur une courbe mélodique semblable à celle de la phrase 2 (assertion ou "réponse naturelle"), que nous pourrons appeler "courbe neutre" ou "courbe non marquée" ; c'est par la dureté des attaques, les pauses, le renforcement de l'intensité que, tant

bien que mal, les encodeurs I et II essayent de faire passer l'information demandée (mais l'échec est presque total).

On concluera donc provisoirement :

- au rôle important de la mélodie dans la bonne transmission de l'information. On voit ici fonctionner un véritable code, susceptible d'un apprentissage (n'importe quel locuteur de la langue est capable de produire ou re-produire correctement les courbes mélodiques caractéristiques des phrases 10 à 4 dans le tableau et de transmettre ainsi avec succès l'information demandée).

- que les autres moyens acoustiques (intensité, durée, timbre, qualité...) sont utilisés avec plus ou moins de talent selon les locuteurs, mais le résultat (la bonne transmission de l'information) reste toujours incertain : l'imitation elle-même est difficile ; il n'y a pas ici de code constitué, si ce n'est peut-être au théâtre à propos duquel on pourrait éventuellement discuter de l'existence d'un "code second" organisant des indices en système conventionnel susceptible d'apprentissage.

- que ce sont des informations grammaticales (par opposition à des indices émotifs) que la mélodie est chargée de transmettre et transmet effectivement. On pourrait préférer dans une autre situation recourir au lexique ou à la morpho-syntaxe pour la transmission de ces informations . Les diverses émotions seront en revanche éventuellement exprimées (on devrait dire suggérées) en recourant à des phénomènes physiologiques, voire "métaphoriques", sans doute assez généralisés dans l'espèce humaine, et qui débordent le cadre d'une langue, d'une culture. Il n'y a pour ces phénomènes guère d'apprentissage possible : on peut tout au plus conseiller à celui qui veut au moyen de sa voix suggérer la colère, d'essayer de se dégager des interdits de sa société qui apprend à contrôler ses émotions, ou bien de faire de l'expression théâtrale : dans ce cadre il apprendra à encoder des émotions, mais de façon souvent outrée.

- selon le niveau de langue choisi, les courbes mélodiques que l'on utilisera pour la transmission d'informations grammaticales ne seront pas toutes retenues : on préfèrera ainsi parfois recourir à une marque morpho-syntaxique d'interrogation (par exemple "Est-ce que Jacques est venu ?" ou bien "Jacques est-il venu ?") plutôt qu'à une marque mélodique : "Jacques est venu ?" produit avec une courbe montante. La courbe caractéristique du "triomphe", normalement utilisée par les enfants scolarisés en maternelle, dont on peut dire qu'elle appartient véritablement à leur langue, qu'elle doit être décrite dans la grammaire de leur langue, n'est plus utilisée par l'adulte que pour imiter cette langue enfantine. La liste des courbes mélodiques chargées de la transmission d'informations grammaticales ne s'arrête pas à la trop célèbre courbe interrogative. Mais établir la grammaire intonative d'une langue, dresser la liste des courbes ayant une fonction grammaticale n'est pas suffisant : il convient encore de préciser les règles d'usage de ces courbes, d'associer une grammaire de l'intonation à une grammaire situationnelle.

On aurait souhaité encore souligner et discuter bien des points qui ressortent de l'analyse des courbes des encodeurs au détecteur de mélodie ou bien de l'exploitation des fiches de réponses des décodeurs (caractéristiques des courbes mélodiques pour les phrases chargées de transmettre une information grammaticale, comparaison des courbes d'un locuteur à l'autre et mise en correspondance avec les pourcentages de réponses identiques obtenus pour chaque locuteur lors du décodage, importance de l'ordre des phrases dans le test lors

du décodage, perturbations entraînées par la volonté de l'encodeur de faire passer une information secondaire - d'ordre émotif - en plus de l'information principale - grammaticale -, éventuelles constantes non-mélodiques dans la réalisation des phrases chargées d'exprimer une émotion, etc.) On se contentera ici de noter les perspectives générales ouvertes par des travaux de cet ordre.

Pour une recherche théorique, on tentera d'affiner les critères de détermination du grammatical par opposition à l'émotif, et l'on pourra ainsi espérer parvenir un jour à l'établissement d'une véritable grammaire de l'intonation dans laquelle on ne confondrait plus les courbes mélodiques ayant une fonction grammaticale avec les indices acoustiques fort divers servant éventuellement à la transmission d'émotions. On cherchera également à établir une liste exhaustive des courbes grammaticales pour chaque langue en pensant qu'à côté de l'assertion, la jussion et l'interrogation, la mélodie peut-être, diversement selon les langues, chargée de transmettre des modalités de ces trois attitudes fondamentales (1) : par exemple l'évidence, modalité de l'assertion, l'appel, modalité de la jussion, l'interrogation méta-linguistique, modalité de l'interrogation, etc.

La linguistique appliquée tirera tout profit de cette recherche en s'interrogeant sur les points suivants : Faut-il apprendre à des étrangers ou à des individus souffrant de troubles du langage n'imp orte quelle phrase intonée quelle qu'y soit la fonction de l'intonation ? Faut-il à cette fin sélectionner en priorité les phrases que l'on remarque le plus dans le discours pour leur intonation, c'est-à-dire les plus expressives, au détriment des phrases où l'intonation a une fonction grammaticale ? Ces dernières phrases étant tellement engagées dans le processus de la transmission des informations grammaticales avec le segmental ne sont le plus souvent même pas signalées dans le discours comme "phrases où l'intonation joue un rôle important". Au-delà des fameuses phrases interrogatives que l'on retient toujours pour l'apprentissage aux étrangers (et que l'on retient souvent mal en oubliant que seules sont obligatoirement montantes les courbes mélodiques des phrases où l'interrogation n'est pas marquée dans la morpho-syntaxe) quelles sont les autres courbes mélodiques ayant une fonction grammaticale que l'on doit apprendre ? Comment les décrire, comment les enseigner, comment en commenter l'usage afin de reconnaître le rôle fondamental de l'intonation dans le discours (la langue écrite ne saurait s'en passer puisqu'elle doit noter par tiellement l'intonation au moyen d'un certain nombre de signes de ponctuation, blancs, alinéas..) ? A côté des fonctions morphologiques de l'intonation qui intègre le message en situation en informant sur l'attitude du locuteur, il faudra aussi un jour s'attacher à ses fonctions syntaxiques : intégration de chaque phrase dans un contexte par l'intonation et types d'intégration possibles.

---

(1) Cf. Benveniste : Problèmes de Linguistique Générale, t. I, Gallimard, 1966, p. 130.

TESTON Bernard
10, Butte des 3 Moulins
13. - AIX-EN-PROVENCE

VIIIèmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

L'INTONATION GENERATIVE

D. J. HIRST
UNIVERSITE DE PROVENCE

• Résumé :

La grammaire générative ne s'est préoccupé que très peu de l'intonation. Seuls quelques auteurs ont proposé une matrice de traits prosodiques pour rendre compte de variations mélodiques distinctives des langues à tons. Une nouvelle matrice est proposée pour rendre compte à la fois des tons et de l'intonation. S'il est possible dans une assez large mesure de prédire les traits prosodiques à partir de l'indicateur syntagmatique sousjacent, la description est énormément simplifiée si l'on postule une matrice de traits intonatifs déterminés par la structure syntaxique et qui, à leur tour, déterminent les traits prosodiques.

• Summary :

Generative grammar has not paid much attention to questions of intonation. A few authors have proposed a matrix of prosodic features to account for significant pitch variations in tone-languages. A new matrix is proposed which can account not only for tones but also for intonation. While it is possible to a great extent to predict the prosodic features of a sentence from the underlying phrase-marker, the description is considerably simplified if a matrix of intonative features is made use of, which are determined by syntactic structure and which, in turn, determine the prosodic features.

L'INTONATION GENERATIVE

D. J. HIRST
UNIVERSITE DE PROVENCE

L'étudiant de l'intonation anglaise a la chance de pouvoir baser son travail sur un grand nombre de recherches systématiques et importantes dont je ne mentionnerai que O'CONNOR & ARNOLD, 1961, FAURE, 1962, CRYSTAL, 1969. L'intonation de l'anglais britannique a été décrite et bien décrite. Si l'intonation pose encore des problèmes pour le traitement automatique, ce n'est donc pas par manque de description. L'apport spécifique de la linguistique associée au nom de CHOMSKY a été de ne plus prendre comme modèle les sciences dites "exactes", "descriptives" ou encore "taxonomiques", mais d'emprunter aux mathématiques un modèle logistique de formalisation qui, au lieu de partir d'un corpus pour aboutir à une description, tente plutôt de rendre compte des réalisations concrètes en termes de formes abstraites sousjacentes et de règles de dérivation.

Par son caractère explicite, une théorie générative est particulièrement adaptée à un traitement automatique, qu'il s'agisse de synthèse ou de reconnaissance.

La linguistique générative ne s'est préoccupée que très peu de l'intonation. La tentation a été trop grande de s'occuper d'abord des questions de phénomènes segmentaux, laissant ainsi l'intonation pour une étape ultérieure de la recherche. Certains auteurs, cependant, ont proposé une matrice de traits prosodiques pour rendre compte des variations mélodiques dans des langues à tons. Ainsi W S-Y WANG (1967, 1972, 1975) a proposé une matrice comportant les traits : HIGH, CENTRAL, MID, CONTOUR, RISE, FALL, CONVEX, qui sont, d'après WANG, suffisants pour rendre compte de toutes les variations significatives rencontrées dans "plusieurs centaines de sys-

tèmes tonaux" (1972, p. 489). Le système de WANG, cependant, avait l'inconvénient de représenter les tons statiques par des niveaux absolus. Or, WANG lui-même cite plusieurs cas de langues africaines et mexicaines où les variations pertinentes sont des variations relatives par rapport à la syllabe qui précède. D'autres exemples semblent indiquer qu'il faille postuler pour des tons statiques à la fois des niveaux absolus (pour chaque locuteur) et des niveaux relatifs. J'ai donc proposé une matrice de six traits prosodiques (HIRST, 1977b) qui sont (HIGH) (LOW) (CONTOUR) (UP) (DOWN) et (CONCAVE). Ces traits sont assortis des conventions suivantes : (HIGH) et (LOW) indiquent des niveaux absolus (pour chaque locuteur) qui se réfèrent au début de chaque syllabe : (CONTOUR) indique une variation tonale perceptible pendant la durée de la syllabe, (UP) et (DOWN) indique une variation perceptible respectivement vers le haut et vers le bas, entre la fin de la syllabe et le début de la syllabe (associés à + CONTOUR) : [+ CONCAVE] enfin indique une valeur tonale au milieu de la syllabe, inférieure à celle du début et de la fin. Pour une syllabe marquée (- HIGH ; - LOW ; - UP ; - DOWN), le début est au même niveau que la fin de la syllabe précédente.

L'intérêt de cette matrice, outre le fait qu'elle peut rendre compte des faits qui échappaient à la matrice de WANG, réside dans le fait que la même matrice peut, sans problème, représenter des variations mélodiques dans des langues autres que des langues à tons. (Un exemple de l'application de cette matrice à l'intonation de l'anglais est donné dans HIRST, 1977c). Ceci paraît d'ailleurs la conséquence logique d'un système qui se veut universel. Comme le fait remarquer WANG (1967, pp. 99-100), il est très difficile de distinguer, à partir des seuls documents acoustiques, des énoncés dans des langues à tons (comme le mandarin) des énoncés dans des langues sans tons (comme l'anglais). En effet, les mêmes variations tonales qui en chinois moderne servent à distinguer :

a) bā b) bá c) bǎ et d) bà (avec les sens : a) huit b) soulever c) prendre et d) cesser, se retrouvent, en anglais

sur le mot "bar" dans les énoncés suivants :

a) The bar was almost empty.

b) Is he in the bar ?

c) Surely you don't mean the bar !

d) He's in the bar.

Si l'on admet que la "sortie" de la composante phonologique caractérisée par une matrice de traits tels que ceux que je viens de présenter, il faudrait s'interroger sur la nature de l'entrée de cette composante. Dans quelle mesure, autrement dit, doit-on représenter l'intonation par des constituants de la phrase, et dans quelle mesure peut-on prédire cette intonation à partir d'autres informations (syntaxiques) déjà présentes dans la phrase.

Si l'on compare l'opposition en anglais entre :

1a) It's a green house (c'est une maison verte)

1b) It's a green-house (c'est une serre),

on trouvera respectivement :

2a) It's a green house
(Ø) (Ø) (H) (C.D)

2b) It's a green-house
(Ø) (Ø) (C.D) (L)

(le ton de la syllabe est représenté par les traits qui ont la valeur (+) dans la représentation prosodique. (C) = CONTOUR, (H) = HIGH, (L) = LOW, (U) = UP, (D) = DOWN, (Cc) = CONCAVE. (Ø), qui représente une syllabe où tous les traits ont une valeur négative, ne sera plus marqué dans les exemples suivants).

D'une façon ou d'une autre, il faudrait démontrer que cette représentation est déterminée par les indicateurs syntagmatiques :

3a)... $({}_{NP}({}_{Det}\ a)_{Det} + ({}_{Adj}\ green)_{Adj} + ({}_{N}\ house)_{N})_{NP}$...

3b)... $({}_{NP}({}_{Det}\ a)_{Det} + ({}_{N}({}_{Adj}\ green)_{Adj} + ({}_{N}\ house)_{N})_{N})_{NP}$...

Il ne serait bien sûr pas difficile de construire une règle ad hoc qui affecterait les traits de la représentation de 2a), 2b) à partir des indicateurs de 3). Si, par contre, nous regardons les phrases :

4a) The green house is empty
(H) (D) (C.D)

4b) The green-house is empty
(H) (C.D)

il n'est pas évident que la même règle (ou les mêmes règles) pourrait affecter les traits prosodiques indiqués à partir des mêmes indicateurs syntagmatiques de 3). Ceci est encore plus évident si nous considérons les phrases suivantes :

4a) The old green house is empty
(H) (D) (D) (C.D)

4b) The old green-house is empty
(H) (D) (C.D)

ou encore :

5a) Is it a green house ?
(H) (D) (C.U)

5b) Is it a green-house ?
(H) (L) (U)

ou encore :

6a) I know it's a green house
(H.C.D) (L) (L)

6b) I know it's a green-house
(H.C.D) (L)

Surtout lorsque l'on considère que dans des phrases comme 6), un locuteur peut distinguer les deux énoncés en jouant sur les indices comme l'intensité et la durée, mais le plus souvent les réalisations seront identiques, car le contexte a neutralisé les variations mélodiques.

Les exemples que je viens de donner montrent clairement qu'il n'est pas commode de dériver la représentation en traits prosodiques directement à partir de l'indicateur syntagmatique. Il faut donc un autre niveau de représentation de l'intonation, niveau plus abstrait qui sera déterminé par la structure syntaxique et qui à son tour, déterminera la représentation prosodique. J'ai déjà présenté une telle représentation de l'intonation de l'anglais (HIRST, 1977a) sous la forme d'une matrice de traits abstraits appelés traits intonatifs. Dans le cas des exemples que nous venons de voir, il s'agit chaque fois des suites

7a) a green house
(+Stress)(+Stress)

7b) a green-house
(+stress)

Il devient ensuite facile de formuler une règle qui affectera les représentations en 7) à partir des indicateurs syntagmatiques de 3), ainsi que des règles phonologiques qui dériveront la représentation prosodique à partir de la représentation intonative.

Si les exemples que je viens de donner relèvent clairement d'une différence de structure syntaxique, il n'est pas toujours le cas de paires de phrases qui ne diffèrent que par la représentation intonative. Il est possible, cependant, en faisant appel à un niveau d'analyse que j'ai appelé "structure de surface non-réduite" de rendre compte non seulement d'un très grand nombre d'ambiguïtés syntaxiques qui sont levées par l'intonation, mais aussi d'un certain nombre de contraintes, comme par exemple, le fait que la phrase :

8) Does he speak English, or not ?

ne peut pas avoir une frontière non-terminale (= intonation montante) à la fin. De la même façon, dans la phrase :

9) I wanted an apple and he gave me an apple.

Le centre d'intonation de la deuxième proposition doit être sur gave et non pas sur apple.

Avec une théorie d'intonation telle que celle que je viens d'évoquer, il devient envisageable de formuler des règles de synthèse de l'intonation à partir d'un texte écrit, qui affecteraient les phrases des schémas intonatifs les moins marqués du point de vue de la structure non-réduite impliquée par l'intonation, et de définir les cas où il serait nécessaire d'indiquer dans le texte (par des moyens typographiques tels que le soulignement) d'autres variantes intonatives.

REFERENCES

CRYSTAL, D. (1969)
Prosodic systems and Intonation in English. Cambridge University Press : Cambridge.

FAURE, G. (1962)
Recherches sur les caractères et le rôle des éléments musicaux dans la prononciation anglaise. DIDIER, Paris.

HIRST, D. J. (1977a)
Intonative features : a syntactic approach to English Intonation.(Janua Linguarum, Series Minor). Mouton, LA HAYE.

HIRST, D. J. (1977b)
Prosodic and intonative features in English intonation. A paraître.

HIRST, D.J. (1977c)
The transcription of English Intonation, in Mélanges offerts à G. FAURE, Studia Phonetica, Didier, Paris, Montréal.

O'CONNOR & ARNOLD,(1961)
Intonation of colloquial English, Longmans, LONDRES.

WANG, W.S-Y. (1967)
Phonological features of tone. International Journal of American Linguistics, Volume 33, n° 2, pp. 93-105.

WANG, W. S-Y,(1972)
The many uses of Fo, in Papers in Linguistics and Phonetics to the Memory of Pierre Delattre. Mouton, LA HAYE, pp. 487-503.

WANG, W.S-Y. (1975)
Prosodic features. Communication au VIII° Congrès International des Sciences Phonétiques, LEEDS, 1975.

VIII èmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

PROBLEMES DE NEUTRALISATION DES MARQUES PROSODIQUES

APPLICATION A LA RECONNAISSANCE AUTOMATIQUE

Philippe Martin
Experimental Phonetics Laboratory
University of Toronto.

• Résumé :

L'utilisation de la prosodie en reconnaissance automatique pose trois problèmes: 1) la localisation des contours, c'est-à-dire la reconnaissance des syllabes accentuées (en écartant l'accent d'insistance); 2) l'identification des contours étant donnée leur neutralisation possible dans leur fonction indicatrice de la structure prosodique; 3) le fait que les structures syntaxique et prosodique ne se correspondent pas nécessairement dans les phrases non ambigües.

Dans l'hypothèse où les problèmes 1) et 3) peuvent être résolus, on propose un algorithme d'analyse de la structure prosodique qui tienne compte des réalisations possibles des contours neutralisés.

• Summary :

The use of prosody in automatic speech recognition raises three questions: 1) the limitation of the melodic contours by identification of the stressed syllables (without consideration of emphatic stress); 2) the classification of the contours given their possible neutralisation in their role as markers of the prosodic structure of the sentence; 3) the fact the prosodic structure is not always equivalent to the syntactic structure in unambiguous sentences.

Assuming that solutions to the first and last of these problems can be found, an algorithm is given which analyses the prosodic structure taking into account the possible neutralisations of the melodic contours.

Problèmes de neutralisation des marques prosodiques
Application à la reconnaissance automatique

Philippe Martin

Structure prosodique

L'énoncé peut être analysé en quatre lignes parallèles, de nature sémantique, syntaxique, prosodique et syllabique, présentant chacune des éléments minimaux classés hiérarchiquement selon des structures indépendantes mais liées entre elles par des lois d'association déterminées.

Le rapport de la structure prosodique avec les structures syntaxiques et sémantiques peut être en particulier décrit par des règles d'accentuabilité des unités syntaxiques assorties d'un mécanisme de neutralisation sélective des contours prosodiques qui réalisent les accents (Martin 1975, 1976a, 1976b, 1977a). Ces règles rendent compte à la fois de l'hypothèse d'une liaison "forte" (Di Cristo, 1975), ou "faible" (Choppy et al., 1975; Vaissière, 1976) existant entre syntaxe et intonation.

Règle d'accentuabilité

La règle d'association prosodie-syntaxe exprime qu'une unité syntaxique minimale peut correspondre à une unité prosodique minimale et, partant, est porteuse d'un accent manifesté par un contour mélodique spécifique, si elle ne contracte ni un rapport de solidarité (i.e. de présupposition réciproque), ni un double rapport de sélection (i.e. de présupposition unilatérale) avec d'autre(s) unité(s). Dans ces deux derniers cas, l'unité considérée forme une unité minimale accentuable, soit avec le ou les autres unités dont elle est solidaire, soit avec l'unité la plus proche qu'elle sélectionne à droite (à moins que cette dernière relation ne soit suspendue par la présence d'une autre unité, en rapport de combinaison - i.e. de non-présupposition - avec les deux premières. L'unité considérée est alors accentuable).

Les accents effectivement réalisés, et manifestés par les marques de présupposition des unités prosodiques que sont les contours mélodiques, constituent les points d'ancrage entre les deux structures.

Désaccentuation

A cette règle d'accentuabilité s'ajoute une hypothèse sur la réalisation des marques prosodiques neutralisées, voulant que la désaccentuation des unités se fassent dans un ordre correspondant à la profondeur de leur position dans la structure prosodique. Donc, s'il y a dominance des structures syntaxiques et sémantiques, c'est-à-dire absence d'ambiguité, les éléments correspondant aux niveaux les plus bas de la structure prosodique seront les premiers désaccentués, tandis que les contours des unités prosodiques de niveaux supérieurs, bien que redondants, seront effectivement réalisés.

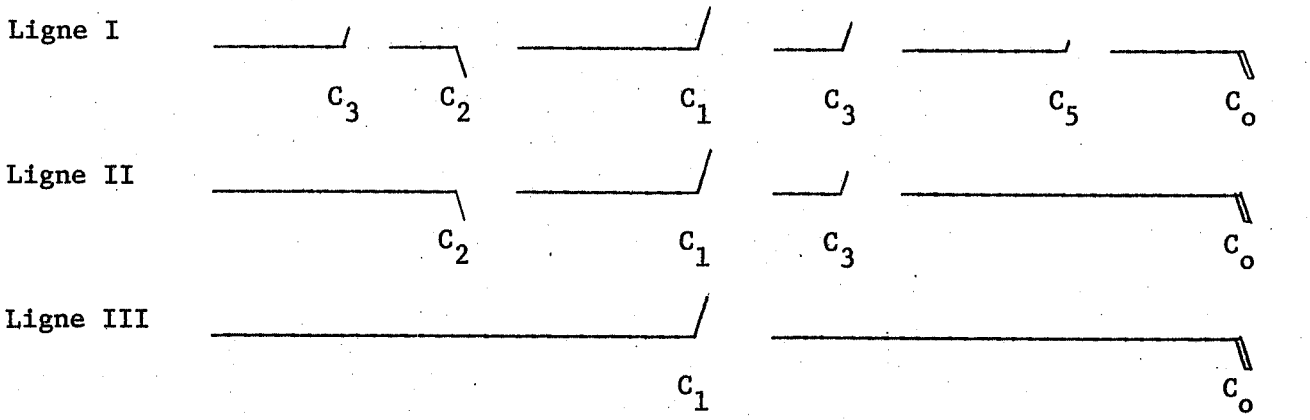
Il s'agit là d'un mécanisme sélectif d'élimination des marques redondantes, dans l'hypothèse où la structure syntaxique est complètement spécifiée par les relations de présupposition existant entre les unités syntaxiques et/ou sémantiques.

Un exemple de désaccentuation - donc de disparition de contours prosodiques - est donné par la séquence:

La grande soeur de Charlotte écrit des lettres parfumées

dans laquelle les contours des unités prosodiques minimales subissent une élimination sélective selon le niveau de neutralisation atteint dans la structure prosodique définie par le parenthésage: (pour les règles d'attribution des contours, voir ici même, Martin, 1977b)

(((La grande)(soeur))(de Charlotte))((écrit)((des lettres)(parfumées)))



Si toutes les unités accentuables sont effectivement accentuées, la ligne prosodique sera composée de 6 unités, présentant 6 contours (ligne I). Elle sera réduite à 4 unités si les unités syntaxiques correspondant au dernier niveau de la structure prosodique sont désaccentuées (ligne II), et à 2 unités si celles du deuxième niveau sont également désaccentuées (ligne III). Les unités prosodiques minimales, terminées par des contours mélodiques spécifiques placés sur la dernière unité syntaxique composante, correspondent alors à des unités de sens de plus en plus grandes.

Cas d'indépendance des structures syntaxique et prosodique

Dans l'exemple précédent, la structure prosodique correspondait à chaque fois à la structure syntaxique (présuppositionnelle) des unités minimales accentuées. Si cette dernière est spécifiée sans ambiguité, les deux structures pourront très bien être différentes, les marques prosodiques n'ayant pas à remplacer d'éventuelles marques syntaxiques déficientes dans leur rôle indicateur de la structure syntaxique. On pourrait par exemple avoir l'énoncé

((Elle)(écrit))(((des)(lettres))(parfumées))

pourvu de deux structures prosodiques distinctes, l'une (1) correspondant à la structure syntaxique définie plus haut, l'autre (2) étant différente:

(1) (Elle écrit)((des lettres)(parfumées))



(2) ((Elle écrit)(des lettres))(parfumées)



En effet, si la structure syntaxique est complètement définie par les marques syntaxiques et sémantiques des unités composantes, rien n'oblige le locuteur à privilégier l'une ou l'autre de ces structures prosodiques, si ce n'est le choix d'un style redondant dans la production de l'énoncé (discours formel, lecture, etc.). Les unités de sens délimitées par les contours réalisés ne correspondront pas nécessairement à ceux de la structure sémantique.

Réalisation des contours

Dans une structure prosodique, un contour déterminé ne doit s'opposer, dans sa réalisation, qu'aux contours qui peuvent apparaître à sa place. Ses caractéristiques de pente, durée, amplitude, hauteur pourront donc être quelconque pourvu que le contour qu'elles manifestent contraste avec tout autre contour qui aurait pu apparaître à cette place. Il pourra donc se réaliser sous la forme d'un autre contour dont la présence ne peut entraîner de changement dans la structure prosodique.

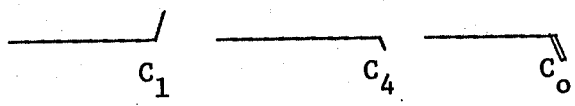
Ainsi, le contour C_3 de l'exemple vu plus haut:

(Elle écrit)((des lettres)(parfumées))



ne doit s'opposer qu'aux contours relatifs à des niveaux supérieurs C_o et C_1, et pourra par exemple se manifester avec une pente descendante, en se confondant alors avec un contour C_4 :

(Elle écrit)((des lettres)(parfumées))



Il s'agit là d'une conséquence connue de tout mécanisme de neutralisation.

Manifestation des contours

Un autre phénomène, placé cette fois sur le plan phonétique, explique la variabilité des contours observés, et porte sur la hiérarchie des paramètres acoustiques qui les manifestent. Leur description phonologique utilise, dans l'ordre, les traits de sens, durée et amplitude de variation mélodique. Leurs réalisations phonétiques font intervenir en plus les caractéristiques de pente et de fréquence de départ, selon une hiérarchie qui peut être variable selon les locuteurs (les contrastes de sens de variation étant toutefois les plus importants). Ainsi, une différence d'amplitude différentiant les contours C_1 et C_3, tous deux montants, pourra être accompagnée ou remplacée par une différence de durée.

A ces variations dans l'ordre d'utilisation des paramètres acoustiques s'ajoute celle des contrastes mis en oeuvre, contrastes qui, même exprimés sous forme relative (pourcentage de variation mélodique rapporté à la fréquence fondamentale moyenne, durée rapportée à la durée moyenne des voyelles accentuées, etc.) apparaissent différemment selon les locuteurs.

Localisation et reconnaissance

On se rend aisément compte à la lumière de ces observations que l'utilisation de la prosodie pour la reconnaissance automatique de la structure syntaxique ne peut être entreprise que dans des conditions bien définies. La seule reconnaissance de la structure prosodique pose déjà le problème de l'adaptation au locuteur, qui doit se faire soit par apprentissage préalable, soit par une exploration par niveaux ou par états de la structure de la phrase analysée (cf. Martin, 1975b). Dans ce dernier cas, outre la comparaison du sens de variation montant ou descendant, l'identité ou la différence d'ordre de deux contours sera établie par le calcul d'un index de type

AD = (a. Amplitude) + (b. Durée) + (c. F_o départ) + (d. Pente)

dont les coefficients a, b, c et d seront ajustés par rapport aux valeurs moyennes des paramètres qu'ils multiplient. Les différences dans la hiérarchie des traits acoustiques utilisés seront ainsi compensées. Ce calcul présuppose la possibilité d'isoler les contours, c'est-à-dire d'identifier les syllabes accentuées tout en éliminant les accents d'insistance. On peut se baser par exemple sur l'énergie du signal filtré de manière à ne garder que le premier et/ou le deuxième formant, la durée vocalique et l'amplitude de variation mélodique correspondante. D'autres procédés sont également envisageables (cf. Vaissière, 1976).

Un analyseur de la structure prosodique

La neutralisation des contours dans leur rôle indicateur de la structure prosodique implique, en reconnaissance, l'identification non pas de la classe à laquelle appartient une réalisation déterminée, mais plutôt de celle des classes auxquelles cette réalisation n'appartient pas.

Les contrastes que doit présenter un contour donné peuvent être exprimés par un graphe, dont les sommets correspondent aux unités obtenues par la règle générale de division:

$$C_j \longrightarrow C_{j+2k-1} \quad C_j$$

avec $j = o, 1, \ldots, 6$

et k = différence de niveau entre l'unité de C_j et l'unité de C_{j+2k-1}

et les arcs aux différentes oppositions à réaliser:

contraste de sens de variation montant - descendant

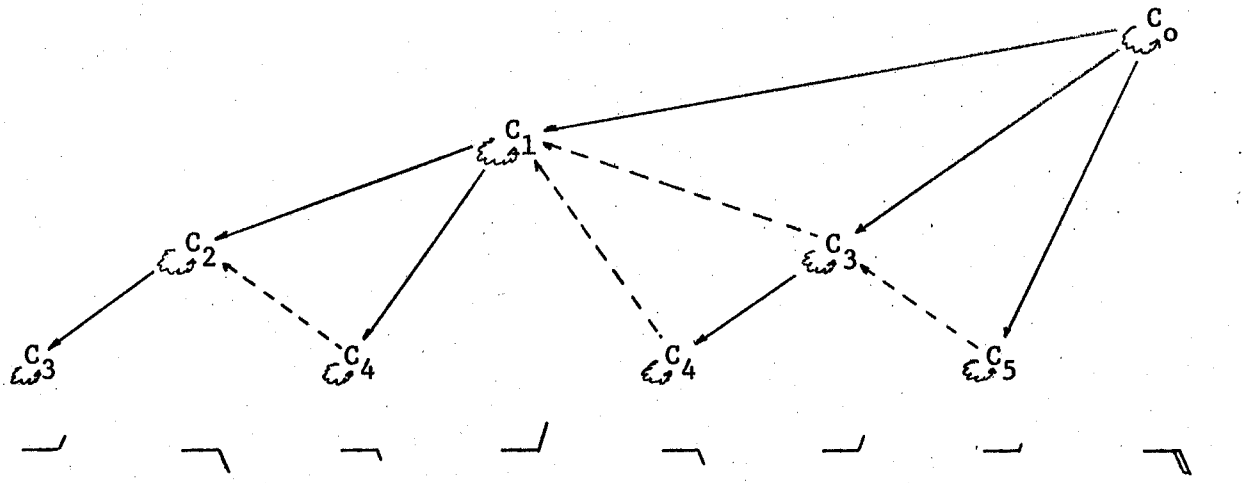
contraste d'amplitude, durée, niveau de départ, pente

contraste ϕ

Un contour doit alors se différencier par rapport à tous les contours rencontrés sur un trajet ascendant effectué dans ce graphe à partir du contour considéré, selon des caractéristiques correspondant à chaque chemin parcouru. Ainsi, C_4 doit s'opposer par le sens de pente à C_3 et C_o, et par l'amplitude à C_1.

Ce graphe peut être utilisé par un analyseur de la structure prosodique, qui intègre ainsi le mécanisme de neutralisation des contours dans son fonctionnement.

Cet analyseur opère de droite à gauche en commençant par le contour final C_o, supposé reconnu au départ. A partir d'une hypothèse C_x sur la nature d'un contour observé, il réalise différents tests de comparaison sur le contour immédiatement placé à gauche C_y, dont les résultats définissent les contours possibles pour C_y, étant donné l'hypothèse C_x, relativement au graphe.



Le cheminement de droite à gauche à effectuer selon les arcs ←, ←- ou ⟜ est alors:

1) C_y ← C_x si C_x et C_y sont de pente opposée (montant descendant ou descendant montant), et si la valeur de l'index AD pour C_y, AD(y), est inférieure ou égale à celle de C_x, à une constante K près, dépendant de la moyenne des variations de AD dans la phrase:
$$AD(y) < AD(x) + K;$$

2) C_y ←- C_x si $AD(y) > AD(x) + K$, quelles que soient les pentes des deux contours;

3) C_x ⟜ C_x (i.e. $C_y = C_x$), si les contours sont de même pente et que:
$$AD(y) - AD(x) < K.$$

Plusieurs chemins peuvent correspondre aux résultats de chaque test, si des lignes de même nature partent du contour hypothèse C_x. Ainsi, à partir de C_o, l'analyseur admettra comme premier essai pour le contour suivant C_1 (quelles que soient ses caractéristiques, puisque les contraintes de la grammaire font que ce contour doit toujours être montant, même s'il est réalisé descendant, comme c'est souvent le cas). A partir de cette nouvelle hypothèse, le contour placé immédiatement à gauche sera examiné, jusqu'à l'obtention de tous les contours ou d'une contradiction éventuelle (les résultats du test ne correspondent à aucun chemin possible). Dans ce cas, il y aura retour en arrière (par pas de 1 contour) et de nouveaux essais seront effectués à partir des chemins non encore utilisés.

Utilité relative d'un analyseur de la structure prosodique

La limitation la plus sérieuse de ce genre d'analyse est due à la non équivalence nécessaire entre les structures prosodiques et syntaxiques, qui ne se correspondront avec certitude que si toutes les autres marques syntaxiques, sémantiques, et syllabiques sont neutralisées (et que l'ambiguité résultant a effectivement été levée par le locuteur par l'utilisation de la prosodie). En

fait, même si ces conditions sont remplies, le dispositif de reconnaissance ne pourra les utiliser que si l'analyse lexicale montre qu'il y a effectivement ambiguité syntaxique, et que cette ambiguité n'est pas due à une mauvaise identification d'unités appartenant à la ligne syllabique.

Cependant, dans les cas d'indécisions irréductibles, les informations prosodiques pourront être utilisées soit globalement pour déterminer, parmi toutes les structures possibles, celle qui correspond le mieux à la structure prosodique trouvée, soit partiellement en utilisant l'analyse locale de la structure prosodique effectuée à partir d'un contour identifié d'une manière assez sûre. Le processus partiel, implanté dans une analyse syntaxique par états, pourrait incorporer au score de la meilleure structure d'état un coefficient relatif à l'application des règles de transitions prosodiques.

REFERENCES

Choppy, C., Lienard, J.S. et Teil, O. (1975) Un algorithme de prosodie automatique sans analyse syntaxique, Actes des 6èmes J.E.P., GALF, Toulouse, pp. 387-395.

Di Cristo, A. (1975) Recherches sur la structuration prosodique de la phrase française (essai d'analyse phonosyntaxique), Actes des 6èmes J.E.P., GALF, Toulouse, pp. 94-116.

Martin, Ph. (1975a) Eléments pour une théorie de l'intonation, R.A. de l'Institut de Phonétique U.L.B., Bruxelles, 9/1, pp. 97-126.

(1975b) Intonation et reconnaissance automatique de la structure syntaxique, Actes des 6èmes J.E.P., GALF, Toulouse, pp. 51-62.

(1976a) L'accentuation en français: théorie présuppositionnelle, R.A. de l'Institut de Phonétique U.L.B., Bruxelles, 10/1, pp. 75-82.

(1976b) Questions de dominance des faits prosodiques sur les marques syntaxiques, Acte del seminario sull' italiano parlato, Firenze, octobre 1976.

(1977a) Phonosyntaxe et phonosémantique vont en bateau. (Questions de phonosyntaxe et de phonosémantique en français), Linguisticae Investigationes, II (à paraître).

(1977b) Utilisation d'un modèle prosodique pour la synthèse par concaténation de mots en français et en anglais, Actes des 8èmes J.E.P., GALF, Aix-en-Provence.

Vaissière, J. (1976) Une procédure de segmentation automatique de la parole en mots prosodiques, en français, Actes des 7èmes J.E.P., GALF, Nancy, pp. 103-114.

Rôle de l'intonation dans la segmentation de l'énoncé et la reconnaissance de la parole

VIIIèmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

CORRELATIONS ENTRE MELODIE, INTENSITE ET DUREE DANS LA PHRASE FRANCAISE ENONCIATIVE

G. CAELEN et G. MAURAND

Section de Linguistique - Université de TOULOUSE-LE-MIRAIL

Résumé :

Notre étude porte sur une évaluation quantitative et qualitative des valeurs paramétriques (essentiellement énergie et fréquence) des groupes prosodiques en relation avec leur position dans la phrase et leur longueur. A l'aide de ces valeurs, nous pouvons tracer les courbes représentatives de chacun des groupes constitutifs de la phrase. Nous concluons sur les corrélations que l'on peut établir entre les paramètres.

Summary :

Our study bears upon a quantitative and qualitative analysis of the parameters (particulary energy and pitch) of the prosodic features in relation with their position in the sentence and their length. With these values, we can draw the curves of each group of the sentence. We conclude by defining the correlations which can be found between the parameters.

CORRELATIONS ENTRE MELODIE, INTENSITE ET DUREE DANS LA PHRASE FRANCAISE ENONCIATIVE

G. CAELEN et G. MAURAND
Section de Linguistique Université Toulouse-le-Mirail

I INTRODUCTION

Les études acoustiques et perceptives concernant l'intonation de la phrase française sont abondantes et les chercheurs prétendent maintenant connaître de manière satisfaisante la structure intonative de la phrase française. Les études sur les deux autres paramètres sont moins nombreuses, et rares sont celles qui intégrent l'étude simultanée des trois paramètres que sont l'énergie, la durée et la fréquence du fondamental. Dans cette étude nous nous proposons d'apporter une évaluation quantitative et qualitative des trois phénomènes prosodiques et d'étudier les corrélations qui peuvent s'établir entre eux.

II APPAREILLAGE ET ACQUISITION DES DONNEES

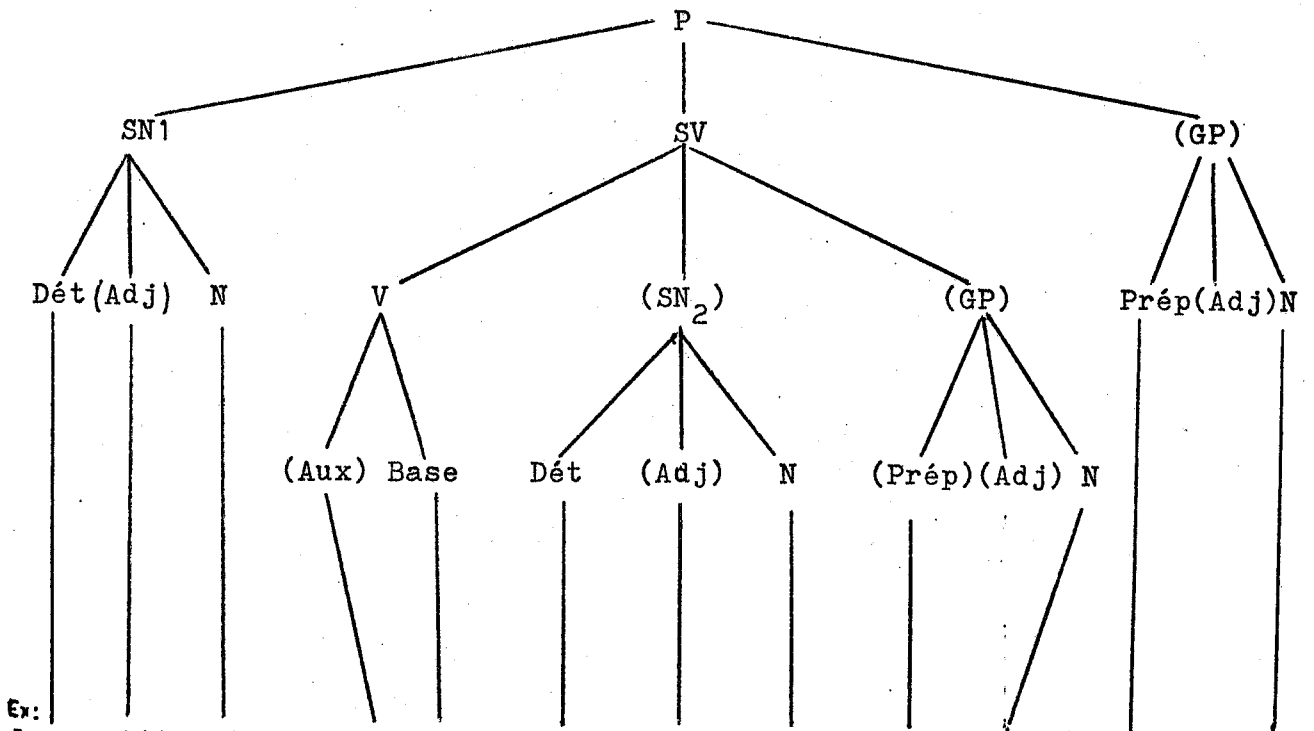
1. Enregistrement en chambre anéchoïde de 84 phrases, présentées en fin de texte sur fiche individuelle.

2. Analyse des enregistrements :
par mélodimètre et intensimètre de l'Institut de Phonétique d'Aix en Provence

par programme de détection du fondamental et de calcul de l'intensité mis au point au laboratoire C.E.R.F.I.A. (1)
Les données recueillies (concernant les trois paramètres) ont été minutieusement pondérées afin de heutraliser les phénomènes micro-prosodiques et les caractéristiques intrinsèques. Nous avons ensuite segmenté la phrase en groupes prosodiques (2),(3),(9), celui-ci pouvant se définir comme une unité acoustique supra-segmentale, délimitée par la coïncidence des frontières des groupes de durée, d'intensité et de fréquence fondamentale.

III Corpus

84 phrases lues par un seul locuteur. Elles comprennent 2 à 4 groupes prosodiques. Pour cette étude nous n'avons retenu que des phrases dont le SN_1 (syntagme nominal sujet) est un nom. Elles peuvent comprendre ou non un SN_2 (syntagme nominal complément d'objet) un groupe prépositionnel subordonné au syntagme verbal (SV), un groupe prépositionnel subordonné à P (phrase) comme l'indique l'arbre ci-dessous :



GP subordonné à P peut se trouver en début de phrase avant SN_1
Les groupes parenthésés ont une présence facultative.

III RESULTATS

1) Quantitatifs : voir tableau Annexe I

Soit F_0 la fréquence fondamentale, F_{oi} est la valeur initiale du groupe prosodique, F_{of} la valeur finale, F_{ofi} la valeur initiale de la tonique et F_{off} sa valeur finale.

Nous avons déterminé la moyenne de F_{of} sur l'ensemble des groupes initiaux, courts ou longs : 147 hz σ 11.4 hz, ainsi que la dynamique de base de F_0 (ou F_{ob}) selon la méthode prescrite par Rossi et Chafcouloff (4) F_{ob}=101 hz. Dans le cas de groupes longs, nous avons appelé F_01, F_02, F_03, F_04 les valeurs fréquentielles correspondant aux 4 limites de l'inflexion à double pente. Pour l'énergie nous avons adopté la même notation (E_i, E_f etc...). Nous avons déterminé également l'énergie de base du locuteur, soit E_b = 28.2 dB. Rappelons que le seuil différentiel de perception de l'intensité est de $\frac{\Delta I}{I}$=3 dB (5). Pour la durée, nous nous sommes contentés de chercher la durée vocalique des groupes ce qui est insuffisant pour décrire le comportement du groupe prosodique en matière de rythme. Nous avons cependant calculé la durée moyenne de la voyelle initiale de groupe et celle de la voyelle finale dans les groupes courts, D_i = 114 ms, D_f = 214 ms, puisque la durée ne semble pas être influencée par la position du groupe dans la phrase.

2) Qualitatifs :

Après étude de chaque groupe prosodique, nous avons relevé une forte similitude de comportement entre les groupes pouvant occuper une même position dans la structure de la phrase. Nous les avons donc regroupés, pour la commodité de l'exposé (ce qui n'exclut naturellement pas que le groupe syntaxique se comporte prosodiquement d'une manière spécifique) en 3 catégories correspondant aux 3 positions possibles des groupes dans la phrase : les groupes de finalité (G.prép SN2, V) les groupes non finaux (SN2, V, G.prép) et les groupes initiaux (G.prép, SN1) .

a) les groupes de finalité : (GF)

Ils se caractérisent, comme nous le savons, par, non seulement, une chute mélodique de la voix depuis les régions moyennes vers les régions basses du registre, mais également par une forte chute de l'intensité, depuis E_b jusqu'à un niveau très faible, inférieur à tout autre dans la phrase. Les pentes négatives de l'énergie des groupes longs,

prennent plus bas que celles des groupes courts, pour atteindre une énergie proportionnellement plus faible.

b) les groupes non finaux : (GNF)

Si le groupe non final est court et succède pour les V à un SN1 long, on constate une inflexion simple montante pour F_o et descendante pour E. Les écarts entre valeurs initiales et valeurs finales se maintiennent dans la moyenne, F_o s'élevant de 3.7 demi-tons au-dessus de F_{ob} et l'énergie diminuant de 3.6 dB depuis la zone d'énergie maximum. Si le SN est court, l'inflexion de l'énergie s'inverse et croit dans la proportion exactement inverse à 0.2 dB près. Si le groupe non final est long, on constate une double pente pour les deux paramètres, mais opposée : le premier élément de la paire se prononce de plus en plus bas à l'intérieur des régions moyennes, tandis que son énergie augmente (des régions moyennes jusqu'aux régions d'énergie élevée); le deuxième élément renverse le mouvement initial et s'élève depuis les régions basses vers les régions infra-aigües, alors qu'il atténue sa force progressivement, en passant de la zone d'énergie moyenne supérieure vers la zone d'énergie moyenne inférieure.

c) les groupes initiaux : (G I)

Ils sont caractérisés par une forte montée tonale (près de 6 demi-tons) accompagnée d'une forte chute d'énergie, inférieure cependant à la chute énergétique caractéristique de la fin de phrase. De même que F_o atteint les régions les plus hautes, de même E très élevée caractérise le début de ces groupes. Contrairement aux groupes non finaux, la première pente du groupe long, à double inflexion, se caractérise par une chute mélodique de la voix, des régions hautes vers les régions moyennes, tandis qu'augmente l'énergie depuis le niveau intermédiaire vers le niveau fort. La deuxième pente au contraire, s'élève vers l'aigu depuis les régions voisines de la F_{ob}, alors que l'énergie, initialement forte, décroît grandement, sans toutefois atteindre la valeur minimale des groupes de finalité.

IV CONCLUSION

Les recherches acoustiques menées sur la constitution des groupes prosodiques de la phrase énonciative française permettent, d'une part de confirmer les études antérieures sur l'intonation, et d'autre part,

d'avoir un aperçu sur le comportement global des autres paramètres :

1. Chaque groupe prosodique est caractérisé par un mouvement de l'énergie (soit croissant, soit décroissant) qui dépend à la fois de la longueur de la phrase (c'est à dire du nombre de groupes prosodiques) de la place du groupe dans la structure et de la longueur respective des autres groupes avec un invariant cependant, le chute de l'énergie en fin de phrase. La structure de la phrase se compose donc d'une suite alternée d'inflexions énergétiques, la pente finale étant négative.

2. Chaque groupe prosodique possède une mélodie et une énergie spécifiques. Elles s'inscrivent à l'intérieur de frontières bien délimitées, caractéristiques de la position du groupe dans la phrase. Leurs inflexions sont donc définissables en termes de niveaux: voir représentation schématique: Annexe II

3. Alors que le rythme du groupe ralentit, une montée ou une chute tonale importante s'accompagne toujours d'une forte décroissance énergétique.

V REFERENCES

(1) CAELEN J. CAZENAVE P. Mesure du fondamental par filtrage variable 8° J.E.P. Aix-en-Provence 1977

(2) FAURE Contribution à l'étude du statut phonologique des structures prosodématiques Studia Ponetica 3 Didier 93-107 (1970)

(3) DI CRISTO A. Recherches sur les structures prosodiques de la phrase française. Essai d'analyse phonosyntaxique. 6° J.E.P. Toulouse 1975

(4) ROSSI M.CHAFCOULOFF M. Les niveaux intonatifs Travaux de l'Institut de Phonétique d'Aix vol 1 1972

(5) ROSSI M. L'intensité spécifique des voyelles Phonetica n° 24 1971

(6) VAISSIERE J. Contribution à la synthèse par règles du français Thèse 1971

(7) BOE L.J. et CONTINI M. Contribution à l'étude quantitative de l'évolution de la fréquence laryngienne de la phrase énonciative en français. Bull. Inst. Phon. de Grenoble vol II 1973

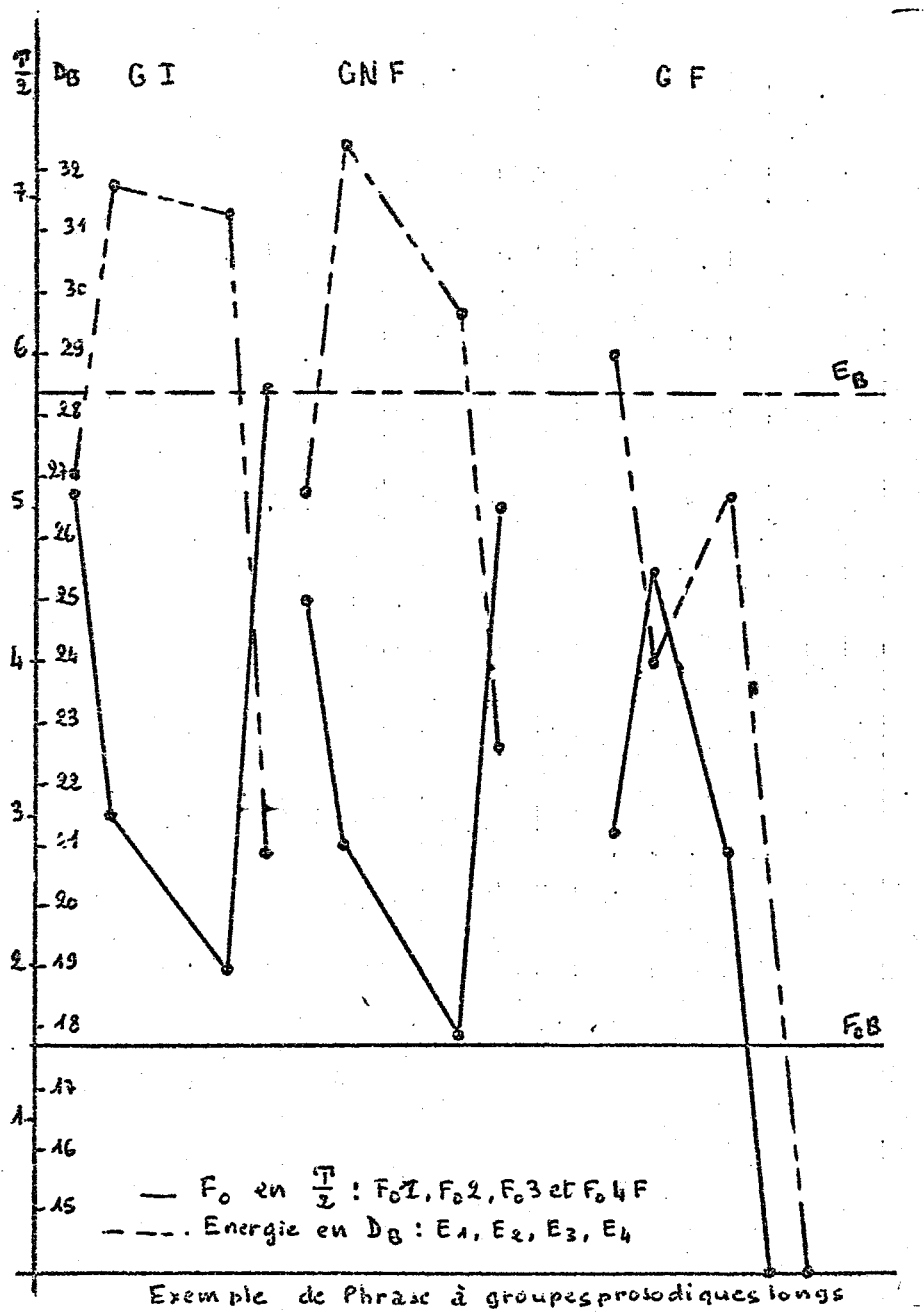
(8) CAELEN G. MAURAND G. Fréquence, intensité et durée : étude comparative des fonctions dans les phrases énonciatives simples et étendues 7° J.E.P. Nancy 1976

(9) VAISSIERE J. Une procédure de segmentation automatique de la parole en mots prosodiques , en français . 7° J.E.P. Nancy 1976

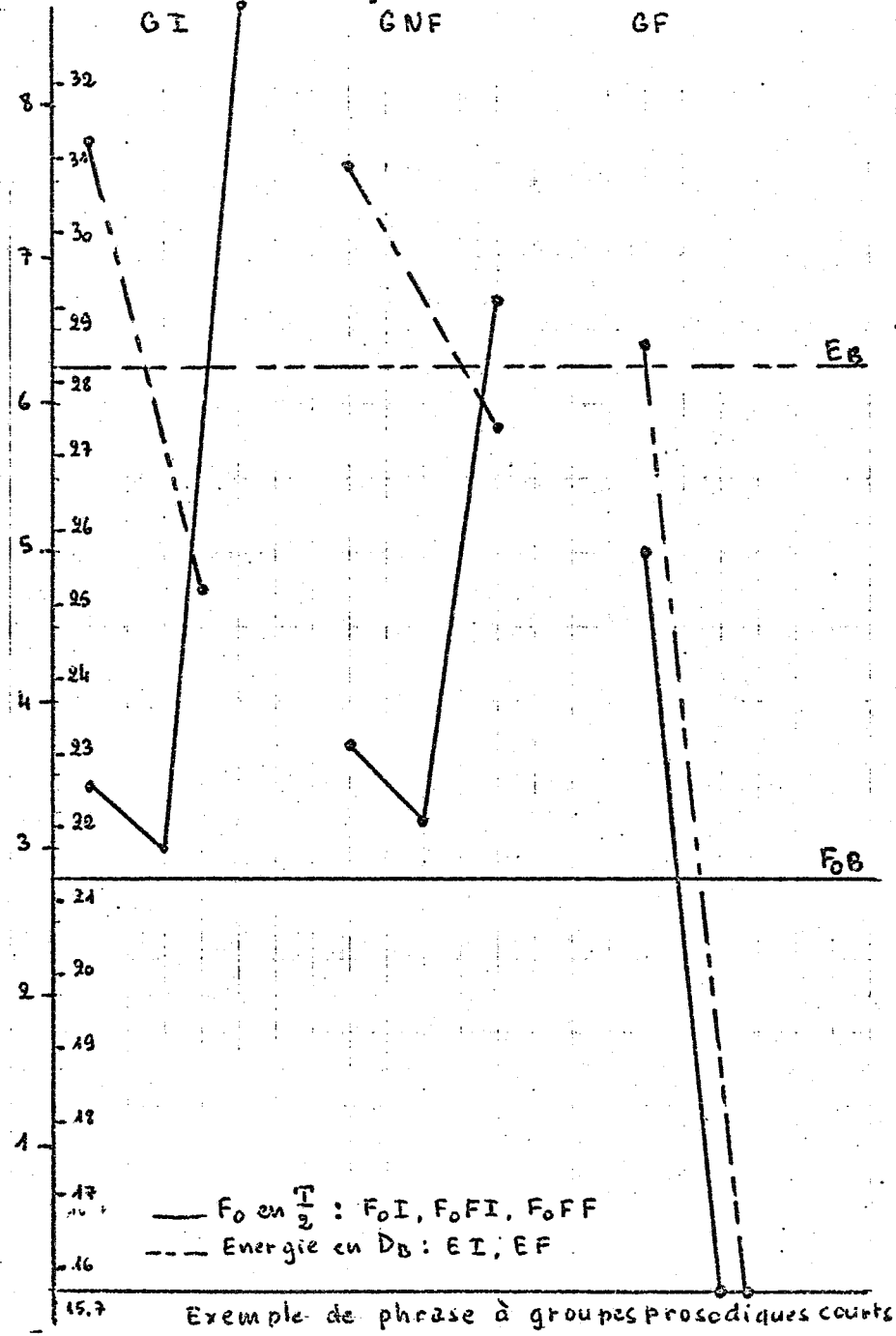
Annexe 1 : Tableau des valeurs paramétriques des groupes prosodiques

Position des groupes dans la phrase	longueur	paramètre prosodique	schema des inflexions	valeurs initiales	Ecart-type σ±	valeurs finales	Ecart-type σ±	valeurs relatives en 1/2 tons($\frac{T}{2}$) et DB
GI	courts ∠ 6 syllabes	Fo	↗	FoI = 106.2	1.3	FoFI = 103.2 FoFF = 151.6	0.7 10.5	FoI - FoB = 0.6 $\frac{T}{2}$ FoFI - FoB = 0.2 $\frac{T}{2}$ FoFF - FoB = 6.1 $\frac{T}{2}$
		E	↘	EI = 31.2	0.2	EF = 25.2	0.3	EI - EB = 3 DB EF - EB = - 3 DB
	longs ou à 2 éléments semantiques	Fo	1 ↘ 2 3 ↗ 4	Fo1 = 131.5 Fo2 = 113.1	3.2 1.8	Fo3 = 105.8 Fo4F = 136.9	2 8.6	FO1 - FO2 = 2.2 $\frac{T}{2}$ FO4 - FO3 = 3.7 $\frac{T}{2}$
		E	1 ↗ 2 3 ↘ 4	E1 = 27 E2 = 31.7	1 1.2	E3 = 31.3 E4 = 21	0.3 1.5	E2 - E1 = 4.7 DB E3 - E4 = 10.3 DB
GNF	courts ∠ 6 syllabes	Fo	↗	FoI = 108.5	2.4	FoFI = 104.3 FoFF = 133.1	1.8 2.6	FoI - FoB = 0.9 $\frac{T}{2}$ FoFI - FoB = 0.4 $\frac{T}{2}$ FoFF - FoB = 3.9
		E	SN2, Gprep, cr ↘ V si SN1 > 4 syll. V si SN1 ≤ 4 syll. ↗	EI = 31	0.5	EF = 27.3	1.3	EI - EB = 2.7 DB EF - EB = - 0.9 DB
				EI = 27.3	0.9	EF = 30.7	1.6	EI - EB = - 0.9 DB EF - EB = 2.5 DB
	longs ou à 2 elements sémantiques	Fo	↘ ↗	Fo1 = 125.3 Fo2 = 112	7.8 0.7	Fo3 = 107.3 Fo4I = 101.7 Fo4F = 130.3	10.4 2.9 8.9	Fo1 - Fo2 = 1.6 $\frac{T}{2}$ Fo4I - Fo4F = 3.4 $\frac{T}{2}$
		E	↗ ↘	E1 = 26.7 E2 = 32.3	1.6 0.9	E3 = 29.7 E4 = 22.6	1.1 4.9	E2 - E1 = 5.6 DB E3 - E4 = 7.1 DB
GF	courts ∠ 6 syllabes	Fo	↘	FoI = 118.9	1.6	Fo F = 89.3	3.8	FoI - FoB = 2.2 $\frac{T}{2}$ FoB - FoF = 2.8 $\frac{T}{2}$
		E	↘	EI = 28.5	0.4	EF = 15.7	1.9	EI - EB = 0.3 DB EB - EF = 12.5 DB
	longs ou à 2 éléments sémantique	Fo	↗ ↘	Fo1 = 113.1 Fo2I = 106.5 Fo2F = 127	4.4 5.8 0.5	Fo3 = 111.5 Fo4 = 94.9	0 2.2	Fo2F - Fo2I = 2.5 $\frac{T}{2}$ Fo3 - Fo4 = 2.7 $\frac{T}{2}$
		E	1er élément ↗ ou ↘ ↘ 1 2 1 2	E1 = 25.5 E2 = 30 E1 = 29 E2 = 24	5.5 3. 1 3	E3 = 26.8 E4 = 13.8	0.3 2.2	E2 - E1 = 4.5 DB E1 - E2 = 5 DB E3 - E4 = 13 DB

Annexe II : Contour simplifié de la variation de la fréquence fondamentale et de l'énergie des groupes prosodiques selon leur position et leur longueur



Exemple de Phrase à groupes prosodiques longs



Exemple de phrase à groupes prosodiques courts

VIII èmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

PROPOSITIONS POUR UN MODELE D'ANALYSE DE L'INTONATION

A. DI CRISTO et M. ROSSI
Institut de Phonétique d'AIX-EN-PROVENCE

• Résumé :

On propose une méthode d'analyse automatique susceptible de s'appliquer aux niveaux acoustique, perceptuel et fonctionnel. On définit une technique permettant d'éliminer les effets de la coarticulation, et de dégager des indices de segmentation. On prévoit 4 ensembles de règles qui conduisent :

1) à corriger les effets microprosodiques,
2) à interpréter perceptuellement la courbe en fonction du seuil de glissando, de la modalité d'intégration de la hauteur ou de la sonie, du seuil différentiel de Fo, de l'interaction de Fo et de I,
3) à dégager la structure constituante de la phrase (règles intonosyntaxiques),
4) à préciser la réalisation phonétique des traits prosodiques en fonction du contexte (règles intonotactiques).

• Summary :

A TENTATIVE MODEL FOR THE ANALYSIS OF INTONATION

A method of automatic analysis to be applied to the acoustic, perceptual and functional levels is presented.
This procedure allows the elimination of coarticulation phenomena and brings out segmentation cues.
4 sets of rules are taken into account - they are related to :

1) the neutralization of microprosodic effects
2) the perceptual interpretation of the prosodic contour as a function of
   a) the glissando threshold,
   b) the auditive integration of pitch and loudness,
   c) the threshold (of Fo),
   d) the interaction of Fo and intensity.
3) the constituant structure of the sentence (intonosyntactic rules
4) the phonetic actualization of prosodic features as a function of the context (Intonotactic rules).

PROPOSITIONS POUR UN MODELE D'ANALYSE DE L'INTONATION

M. ROSSI et A. DI CRISTO

1. INTRODUCTION

COLLIER critique à juste titre la plupart des études effectuées sur l'intonation (1). En effet, ces travaux reposent essentiellement sur un relevé des mesures objectives et leur interprétation directe.

Ces mesures, d'autre part, sont limitées aux points-clefs de la phrase considérés comme porteurs de la fonction et l'interprétation perceptuelle, si elle existe, est le plus souvent de type introspectif.

Nous voudrions élaborer un modèle d'analyse qui éviterait ces écueils et qui pourrait servir de cadre méthodologique aux recherches à venir sur l'intonation.

Il est nécessaire, dans un premier temps, de décrire la phrase et les unités qui la constituent au niveau subphonémique, c'est-à-dire de procéder à un relevé des variations prosodiques objectives qui incluent les variations microprosodiques. On pourrait définir ce niveau comme celui de la performance.

La nécessité de la description subphonémique devient évidente si l'on a présent à l'esprit le fait que la microprosodie introduit des variations dont on doit tenir compte dans l'interprétation des tracés acoustiques (2). Par conséquent, ce premier niveau d'analyse permet de définir un premier ensemble de règles de conversion essentielles.

Toutefois, si l'on en reste à cette étape, on ne dépasse pas le niveau de la performance, un niveau qui a subi un certain nombre de transformations et de distorsions (contraintes) et qui est, par là même, inapte à rendre compte de la compétence du locuteur, c'est-à-dire, de la relation de l'intonation à la syntaxe.

En effet, le continuum prosodique est, à l'audition et au décodage, interprété par des mécanismes perceptuels qui opèrent une sorte de filtrage inverse pour retrouver dans le signal les traits liés de façon univoque aux instructions linguistiques (3, 4). On doit donc simuler ce "filtrage inverse" à un deuxième niveau d'analyse qui serait un niveau interprétatif dont la sortie contiendrait la définition des règles susceptibles de reconnaître la structuration syntaxique de l'énoncé.

Ce modèle interprétatif est indispensable non seulement dans la perspective d'une analyse linguistique, mais aussi dans celle de la reconnaissance automatique.

On peut se demander d'ailleurs si les échecs relatifs dans ce dernier domaine ne sont pas dûs 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 (5).

Nous voulons présenter ici les grandes lignes de ce que pourrait être ce modèle interprétatif.

2. PRESENTATION DU MODELE D'ANALYSE

L'analyse objective, qui constitue la première étape, doit être de type pluriparamétrique et faire intervenir la mélodie (relief, pente, niveaux), la durée, l'intensité et éventuellement le timbre. On a trop longtemps confondu, en effet, mélodie et intonation ; or ces deux plans ne sont pas toujours en relation biunivoque.

La présence d'un relief mélodique et d'un glissando (figure 1) n'indique pas toujours une limite de groupe intonatif ; ces deux éléments peuvent être liés à la mécanique respiratoire ou à la présence d'un accent expressif. Dans ce cas précis, le relief mélodique se réalise sur la première syllabe du mot et conserve une fonction démarcative au niveau du mot. Mais, si le relief mélodique et le glissando sont accompagnés d'une augmentation de la durée et d'une sonie dominante, alors ils constituent deux des traits définitoires de l'intonème progrédient et assument une fonction démarcative (figure 1).

Le modèle interprétatif comprend :

1) des règles d'effacement microprosodique (R1)
2) des règles de conversion perceptuelle (R2)
3) des règles intonosyntaxiques (R3)
4) des règles intonotactiques (R4).

R1 et R2 supposent la reconnaissance préalable de certains traits phoniques afin que les conversions puissent être opérées correctement.

Ainsi, pour l'évaluation de l'augmentation de la durée - qui joue un rôle essentiel dans la définition acoustique de l'intonème (6, 7) et pour la correction en fonction de l'intensité spécifique, les traits suivants doivent avoir été identifiés :

1° Haut/non Haut et Nasal/non Nasal, pour les voyelles

Les voyelles hautes, en effet (ex. [i], [u]), sont plus brèves que les voyelles basses (ex. [a]) de l'ordre de 30 %. Les voyelles nasales sont plus longues que les orales basses de l'ordre de 50 %.

On sait, d'autre part, que le facteur de correction relatif à l'intensité spécifique des voyelles hautes et de l'ordre de 5 dB (8).

2° Voisé/non Voisé, pour la consonne (C1) qui précède la voyelle et Silence/non Silence, pour la consonne (C2) subséquente à la voyelle.

Si (C1) est non voisée, elle abrège (V) de 20 % environ. La voyelle accentuée ou finale de groupe progrédient a une durée à peu près identique si elle se trouve en syllabe ouverte (ex. chanta) ou en syllabe fermée par l'une des consonnes :
[b, d, g, m, n, l].

Si on considère la structure (C V) comme la plus fréquente en syllabe inaccentuée, il n'est donc pas nécessaire d'apporter de correction à la durée dans le cas où le trait voisé est détecté sur (C2), que la consonne appartienne ou non à la syllabe dont (V) est le centre.

On fait, dans ce cas, une économie considérable, puisque l'évaluation de la durée peut se faire sans avoir recours à la reconnaissance préalable de la limite syllabique.

Après la voyelle accentuée ou finale de groupe progrédient, un silence - dont la valeur seuil est à déterminer en fonction du corpus - peut être assimilé à la tenue d'une consonne non voisée qui abrège considérablement la voyelle (troncation de 50 % par rapport à la voyelle de syllabe ouverte) (2).

La question se pose de savoir comment identifier de façon automatique la fin du groupe intonatif.

On détecte au préalable les reliefs mélodiques caractérisés par une configuration statique ou montante.

Si, parallèlement à chaque relief, on observe une augmentation significative de la durée et de l'intensité par rapport à la voyelle subséquente (V2), alors on est en présence d'une fin de groupe intonatif ou accentuel.

L'évaluation de la durée doit être faite après les corrections sur (V1) et (V2) (figure 2), en fonction de la durée spécifique et du contexte (C1) et (C2). Si (V1) > (V2) de 30 %, on admet la présence d'une frontière. Si (V1) ≃ (V2) ou (V1) > (V2) d'une valeur inférieure au seuil de 30 %, on teste l'intensité que l'on corrige d'après la procédure définie par ROSSI (9).

Si I1 > I2 d'une valeur égale ou supérieure au seuil de 3 dB, alors on admet la présence d'une frontière, sinon la rupture mélodique n'a aucune valeur démarcative.

L'intérêt de cette méthode réside dans le fait qu'elle permet de ne pas prendre les variations microprosodiques ou un relief mélodique quelconque pour une frontière de constituant (5).

Dans une deuxième étape, on applique R1 et R2 à la mélodie, à l'intérieur de chaque groupe intonatif identifié.

a) Sur les voyelles supposées non accentuées ou désaccentuées, on prend systématiquement, pour définir la hauteur,(Fo) situé aux 2/3 de la durée vocalique (10) et on applique le facteur de pondération dû à la hauteur spécifique (2). Si on observe un nouveau relief mélodique à l'intérieur du groupe, on reprend la procédure définie ci-dessus (INIT, sur la figure 2).

b) Sur la voyelle de fin de groupe, il semble que la hauteur spécifique ne manifeste plus ses effets. En revanche, l'influence du contexte est fondamentale.

Si la perception de la parole procède selon un modèle d'analyse par synthèse (4, 11), on peut penser que les mécanismes perceptuels prennent en compte les contraintes physiologiques ou acoustiques pour évaluer les indices prosodiques. Ainsi, dans le cas de troncation (V1 + occlusive non voisée), la configuration montante n'est pas perceptible, si on lui applique le seuil de glissando(10). Comme cette configuration est l'image d'une tension physiologique et que dans les contextes sans troncation elle est perceptible, il est probable que l'oreille suit la modulation au-delà de sa durée réelle (4), comme elle le fait pour les transitions de formants. Par conséquent, on prolongera la modulation sur toute la durée corrigée.

D'autre part, si (C1) est voisée, on relèvera (Fo) de 20 % au début de la voyelle. On fait l'hypothèse que la trajectoire de ce point à la fin de la modulation prolongée représente la configuration perçue par l'auditeur.

La dernière étape, avant R3, consiste à calculer les niveaux intonatifs dans la tessiture du locuteur et à redéfinir la mélodie des groupes intonatifs en fonction de ces niveaux.

La procédure adoptée pour le traitement des reliefs mélodiques et des groupes permet de localiser les frontières intonatives et d'établir une hiérarchie parmi ces dernières.

Une analyse syntaxique conduite parallèlement permet de définir les correspondances entre ces frontières et les limites de constituants. L'étude de ces relations aboutit à l'élaboration d'un ensemble R3 de règles intonosyntaxiques (13). Les règles intonosyntaxiques peuvent être assimilées à des règles de base et à des générateurs élémentaires qui permettent d'engendrer les structures intonatives à partir de la hiérarchie des constituants.

A ce niveau, intervient un second type de règles intonosyntaxiques qui peuvent être assimilées à des règles transformationnelles et qui assurent le passage de l'intonation de base à l'intonation de surface.

L'intonation de surface est une structure intermédiaire dont la réalisation dépend de règles intonotactiques, qui appartiennent à la composante phonétique de la grammaire.

Le modèle interprétatif que nous proposons rappelle à plus d'un titre le modèle d'analyse par synthèse et peut connaître une double application dans les domaines de la reconnaissance automatique et de la synthèse de la parole.

R E F E R E N C E S

1. COLLIER, R. (1974) : Intonation from a structural viewpoint : a criticism, Linguistics, 129 : 5-28.
2. DI CRISTO, A. (1977) : Contribution à l'étude de la structuration de la phrase en français moderne : les faits microprosodiques, Thèse de Doctorat (3è cycle), Université d'AIX-EN-PROVENCE.
3. LIBERMAN, A.M., COOPER, F.S., SHANKWEILER, D.L. and STUDDERT-KENNEDY, M. (1967) : Perception of the speech code, Psychological Review, 74 : 431-61.
4. LIEBERMAN, P. (1967) : Intonation, Perception and Language, M.I.T. Press.
5. LEA, W.A. (1973) : An approach to syntactic Recognition without phonemics, IEEE Transactions on audio and electroacoustics, Au-21 (3) : 249-58.
6. KLATT, D. (1976) : The linguistic uses of segmental duration in English : Acoustic and Perceptual Evidence, J.A.S.A.,59, 1208-21.
7. de ROOIJ, J.J.(1976) : Perception of prosodic boundaries, IPO Annual Progress Report, 11 : 20-24.
8. ROSSI, M. (1971) : L'Intensité spécifique des voyelles, Phonetica, 24 : 129-61.
9. ROSSI, M. (1970) : Sur la hiérarchie des paramètres de l'accent, Proceedings of the 6th Int. Cong. Phon. Sci. : 779-85.
10. ROSSI, M. (1971) : Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole, Phonetica, 23 : 1-33.
11. HALLE, M. and STEVENS, K.N. (1964) : Speech recongition : a model and a program for research, in FODOR, J.A. and KATS,J.J., The Structure of Language : Readings in the Philosophy of Language.
12. ROSSI, M. and CHAFCOULOFF, M. (1972) : Les niveaux intonatifs, Travaux de l'Institut de Phonétique d'AIX, 1 : 167-78.
13. DI CRISTO, A. (1975) : Recherches sur la structuration prosodique de la phrase française, Actes des 6èmes Journées d'Etude sur la Parole : 94-116.

ANNEXE : EXEMPLE D'APPLICATION

Phrase : Il a perdu son chameau en Suisse (fig. 1)

Relief 1 : 148 Hz sur la syllabe per.
1. dV1 = 70 ms ; V1 → non haute ; V1 → non nasale ; C1 → non voisée : 70 x 1.2 = 84 ms ; C2 → voisée.
2. dV2 = 70 ms ; V2 → haute : 70 x 1.3 = 91 ms ; V2 → non nasale ; C1 → voisée ; C2 → non voisée ; C2 → non silence.
3. dV1 < dV2 → le relief mélodique n'est pas indicateur de fin de groupe.

Relief 2 : 150 Hz sur la syllabe meau.
1. dV1 = 90 ms ; V1 → haute : 90 x 1.2 = 108 ms ; V1 → non nasale ; C1 → voisée ; C2 → voisement.
2. dV2 = 120 ms ; V2 → non haute ; V2 → nasale : 120/1.5 = 80 ms ; C1 → voisement ; C2 → non voisée ; C2 → non silence.
3. dV1 > dV2 (dV1/dV2 = 1.35) → le relief est indicateur de fin de groupe.

Relief 3 : 140 Hz sur la syllabe Suisse : fin de phrase.

Hauteurs : # Il a perdu son c h a m e a u / en S u i s s e // #
100 100 140 114 108 112 120-150 108 140 - 84 //
Après l'identification des groupes majeurs et le calcul de la hauteur, on explore à nouveau chaque groupe ; on identifie un relief sur du qui est parallèle à un allongement de la durée et à un relief d'intensité significatif.

Configurations : # Il a perdu ↓ Statique son chameau ↓ Statique / en Suisse ↓ chute // #

Niveaux : # M M IA M M M IA M IA - G #

Règles intonosyntaxiques :
A. Règles de bases - RB1 : P → SN ↑↑ SV ↑↑ (SP) ↑↑
RB2 : SV → Aux V ↑(SN) etc...
B. Règles de transformation - RT1 : /↑↑/→ ∅ /$_P$(SN Pron. — X)$_P$ etc...

Règles intonotactiques - Rt1 : /↑↑/ → H + I A + C + I + L
Rt2 : /↑/ → H + M + (I) + L
Rt3 : /C/ → ∅ / X $\frac{}{X}$ X , ssi $x \leqslant 1$, où X = intonème et x = nombre de syllabes.

On a identifié une frontière mineure après perdu (↑ = H + M + I + L, Rt2), c'est-à-dire une limite de constituant non dominé par P. Après chameau, on a identifié une frontière majeure (↑↑= H + I A + I + L, Rt1 et Rt3), c'est-à-dire une limite de constituant dominée par P. Mais on ne sait pas si ce constituant est SN, SV ou SP. Le relief sur per n'indique pas une limite de constituant mais une limite de mot devant per. La structure syntaxique déduite de l'analyse est la suivante :
1ère scrutation : #$_P$((il a perdu son chameau) (en Suisse))$_P$#, où les deux interprétations les plus plausibles sont : (SN + SV) ou (SV + SP).
2ème scrutation : #$_X$((il a perdu) (son chameau))$_X$, où x est un constituant inférieur à P.
Si x = SN, on peut avoir par exemple Adj + GN ou Dét + GN ; si x = SV, SN1 est pronominalisé et on a par exemple V + SN, ou V + SP. Les 4 structures les plus plausibles seront par exemple : 1. ((Adj GN)(SV)), 2. ((Dét GN)(SV)), 3. ((SN Pron.≠Aux V≠SN)(SP)), 4.((SNPron.≠Aux V≠SP)(SP)).

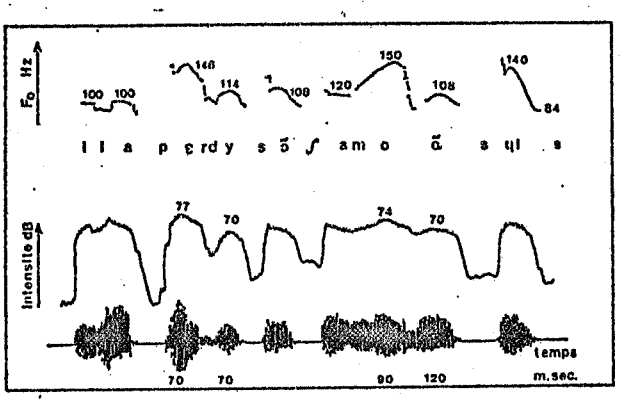


Fig. 1 Document utilisé pour l'illustration de la méthode proposée.

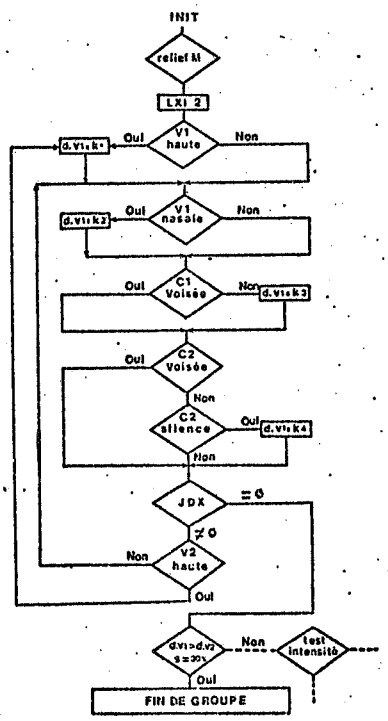


Fig. 2 Organigramme de la procédure utilisée pour la reconnaissance des frontières de groupes.

VIIIèmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

PRISE EN COMPTE D'INFORMATIONS PROSODIQUES DANS UN MODELE DE COMPREHENSION DU DISCOURS CONTINU

J.P. HATON, J.F. MARI, G. MESSENET, J.M. PIERREL, C. SANCHEZ

Centre de Recherche en Informatique de Nancy
Equipe de traitement du signal et reconnaissance de formes
C.O. 140 54037 - NANCY

• Résumé :

Les études menées récemment de façon isolée ont montré l'utilité des informations prosodiques en reconnaissance du discours parlé. On présente ici les grandes lignes d'un système intégrant la prosodie comme une source d'informations parmi d'autres. La prosodie intervient
1) dans la segmentation de la phrase en mots prosodiques et en syllabes et
2) dans le processus d'analyse syntaxico-sémantique avec un formalisme unique de réseau de transition procédural.

• Summary :

Recent works have led to the conclusion that prosody can be useful in speech understanding. We present here the basic ideas of a general system which incorporates prosody as one knowledge source among others. In this system prosody is used
1) during the segmentation into "prosodic words" and syllables of a sentence and
2) during the syntactic-semantic analysis (at this level a unified model of transition network with procedural nodes is used).

I - Introduction

La prosodie a fait l'objet de nombreuses études, en particulier, depuis plusieurs années, en ce qui concerne son rôle dans la compréhension du discours parlé.
La prosodie constituant, en quelque sorte, une forme de ponctuation du discours parlé, est utilisée aussi bien dans un but génératif (par exemple en synthèse de la parole) que dans un but analytique (par exemple en reconnaissance-compréhension de la parole). C'est l'aspect analytique, moins évident et plus difficile à mettre en oeuvre, qui nous intéresse ici.

Une mise en oeuvre des contraintes prosodiques nécessite une définition préalable des paramètres physiques contenant l'information prosodique et des méthodes d'extraction de ces paramètres. Nous utilisons comme paramètres le contour mélodique, l'intensité syllabique et les pauses. Le rythme d'émission serait aussi un paramètre pertinent mais difficile à détecter de façon fiable. L'un ou l'autre de ces paramètres ont été utilisés par différents chercheurs.

Ainsi, DI CRISTO [1] utilise essentiellement les pauses et la mélodie pour détecter divers types de frontières syntaxiques afin d'obtenir une segmentation du semi- continuum vocal.
J. VAISSIERE [2] [3] recherche des patrons intonatifs types dans le contour mélodique d'une phrase et en déduit une segmentation en mots prosodiques.
P. MARTIN [4] propose un algorithme qui, en partant également de contours-types, permet d'atteindre la structure syntaxique superficielle de la phrase. Sa méthode possède à la fois un aspect génératif et un aspect analytique. Le travail de O'MALLEY [5] se propose le même but, en utilisant le seul paramètre de durée et dans le cas trop simple et trop limité des expressions algébriques.
W.A. LEA [6] propose également un système permettant une segmentation en syntagmes de la phrase en s'appuyant sur un principe d'analyse par synthèse. Son approche revient à utiliser l'aspect génératif des règles prosodiques.

L'ensemble de ces travaux met en évidence l'utilité de la prosodie pour automatiser la compréhension. Mais ces études restent pour l'instant isolées, en dehors du contexte d'un système général de reconnaissance de la parole. Seul LEA [7] travaille actuellement à l'intégration de son système au système général développé par BBN ; ce problème est très délicat et, de plus, fonction de la langue.
Notre but est de concevoir un système de compréhension de phrases plus général que notre système actuel [8], dans lequel la prosodie ferait partie intégrante au même titre que les autres sources d'informations : acoustique, phonétique, syntaxique, sémantique, de façon à accroître l'efficacité du système et à obtenir une coopération optimale entre les divers niveaux.

D'autre part, les informations prosodiques sont caractérisées par un degré variable d'incertitude et de flou qui rendent leur utilisation délicate. Nous considérons ainsi, grâce à une structure de données et à un algorithme adaptés, ces informations comme un moyen -parmi d'autres- permettant d'émettre des hypothèses au cours de la segmentation et au cours de l'analyse syntaxico-sémantique. Nous allons brièvement passer en revue ces deux aspects sur un exemple.

II - Prosodie et segmentation

La prosodie, en tant qu'aide à la segmentation, intervient dans notre système à deux niveaux :

- mise en place dans la phrase de délimiteurs de mots prosodiques (éventuellement assortis de "taux de confiance"), par recherche de contours types (cf. [3]) à l'aide d'une méthode de reconnaissance de formes et de pauses dans la prononciation.
- utilisation de l'indice binaire de voisement pour définir dans la phrase des "îlots voisés" sur lesquels on applique un algorithme récursif de recherche de syllabes. Cet algorithme n'utilise dans sa recherche qu'un seul paramètre prosodique : l'intensité. Nous n'avons pas étudié la validité des variations mélodiques à ce niveau.

La figure 1-a indique les marqueurs de mots (M) et d'îlots voisés (V) passés aux niveaux supérieurs de segmentation et d'analyse lexicale et syntaxique (la mélodie est donnée par un appareil "Mélographe" du CNET-ETA connecté en temps réel à notre calculateur Mitra 125).

III - Prosodie et analyse syntaxico-sémantique

Dans le cadre d'un système général de compréhension de la parole utilisant des contraintes syntaxico-sémantiques, les indices prosodiques donnent des informations intéressantes utilisables à deux niveaux : d'une part, pour déterminer le point de départ du processus de reconnaissance, d'autre part, pour limiter le nombre d'hypothèses à prendre en compte à chaque pas du processus.

a) La prosodie comme critère de choix d'un point de départ pour l'analyseur syntaxique

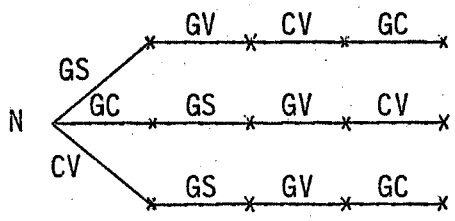
Il est intéressant d'utiliser pour le discours continu une analyse syntaxique qui puisse débuter en plusieurs points d'une phrase [9] et pas seulement au début de cette phrase, comme on le trouve classiquement. La détermination du mot-clé de départ fait intervenir des paramètres phonétiques, syntaxiques et sémantiques car ce mot doit posséder des propriétés bien particulières. L'utilisation d'indices prosodiques (accentuation et marqueurs de mots prosodiques) s'avère efficace pour aider à la localisation de ces mots-clés.

b) La prosodie comme aide pour lever les ambiguïtés

Le système MYRTILLE que nous développons actuellement est essentiellement basé sur un processus d'hypothèse et test. Le traitement syntaxico-sémantique a pour rôle principal d'émettre des hypothèses sur les prochains mots à reconnaître. Une étude préalable sur une représentation syntaxico-sémantique d'un sous-ensemble assez vaste du français parlé nous a montré que la plupart des ambiguïtés sur la structure de la phrase à choisir pouvaient être levées grâce à un test sur un seul groupe d'hypothèses représentant les diverses possibilités acceptées pour le mot suivant. Néanmoins, certaines phrases françaises entraînent à un moment, lors de l'analyse syntaxico-sémantique, de très fortes ambiguïtés sur la structure sous-jacente : ce sont, en particulier, les phrases comportant soit une incise soit une transformation introduite pour provoquer une insistance sur un mot particulier.

A titre d'exemple, la fig. 1 montre trois constructions possibles pour la phrase : "les chercheurs nancéiens étudient la prosodie cette année".

Si l'on appelle GS le groupe sujet, GV le groupe verbal, CV le complément du verbe et GC le complément circonstanciel, on a l'arbre syntaxique suivant pour représenter ces trois constructions :



Il faut donc, au noeud N, choisir l'une des trois branches possibles ; or, chaque branche commence par un groupe nominal (article, nom), des considérations purement syntaxiques ne permettent donc pas de lever l'ambiguïté existant en ce point. Pourtant, il semble raisonnable de penser que des indices prosodiques permettront de distinguer chacune de ces trois constructions. Il faut, pour cela, pouvoir caractériser prosodiquement les incises et les insistances et savoir les détecter. Ceci paraît possible : les premières sont caractérisées par des pauses significatives encadrant le groupe de mots en incise, quant aux secondes, elles correspondent à un patron intonatif particulier suivi d'une pause, comme on le constate sur les fig. 1-a et 1-b.

IV - Mise en oeuvre dans un système de compréhension automatique

Comme nous l'avons vu, la prosodie intervient à deux niveaux de la reconnaissance :

a) au niveau phonémique comme point de repère pour la segmentation en syllabes.

b) au niveau syntaxico-sémantique. On peut encore ici distinguer deux types d'utilisation :

(i) d'une part, pour pondérer les choix à traiter (cf. paragraphe II-b). A ce niveau, prosodie, syntaxe et sémantique seront pris en compte par un processus unique travaillant par exemple sur un réseau de transition possédant des noeuds procéduraux (c'est le choix que nous faisons actuellement). Dans un tel réseau, les transitions rendent compte des informations syntaxiques alors que les noeuds font référence à des procédures utilisant la sémantique et la prosodie pour classer par plausibilité croissante les différentes sorties de ce noeud. Il ne s'agit en aucun cas de supprimer certaines possibilités mais d'élaborer une heuristique particulière qui doit permettre à chaque instant de traiter en premier lieu la structure la plus probable en fonction du contexte sémantique et des traits prosodiques.

(ii) d'autre part, pour aider à la validation des hypothèses émises. Cela correspond à l'utilisation des marqueurs de segmentation en mots prosodiques lors de la reconnaissance phonémique des mots. Sans vouloir s'appuyer toujours sur ces marqueurs, il semble qu'ils doivent intervenir pour pondérer le score de réussite fourni après la reconnaissance : pour une reconnaissance phonémique équivalente, il faudra privilégier un mot débutant et finissant sur un marqueur de segmentation par rapport à un mot possédant en son centre un tel marqueur.

V - Conclusion

Nous avons très brièvement esquissé ici l'intégration d'informations proso-

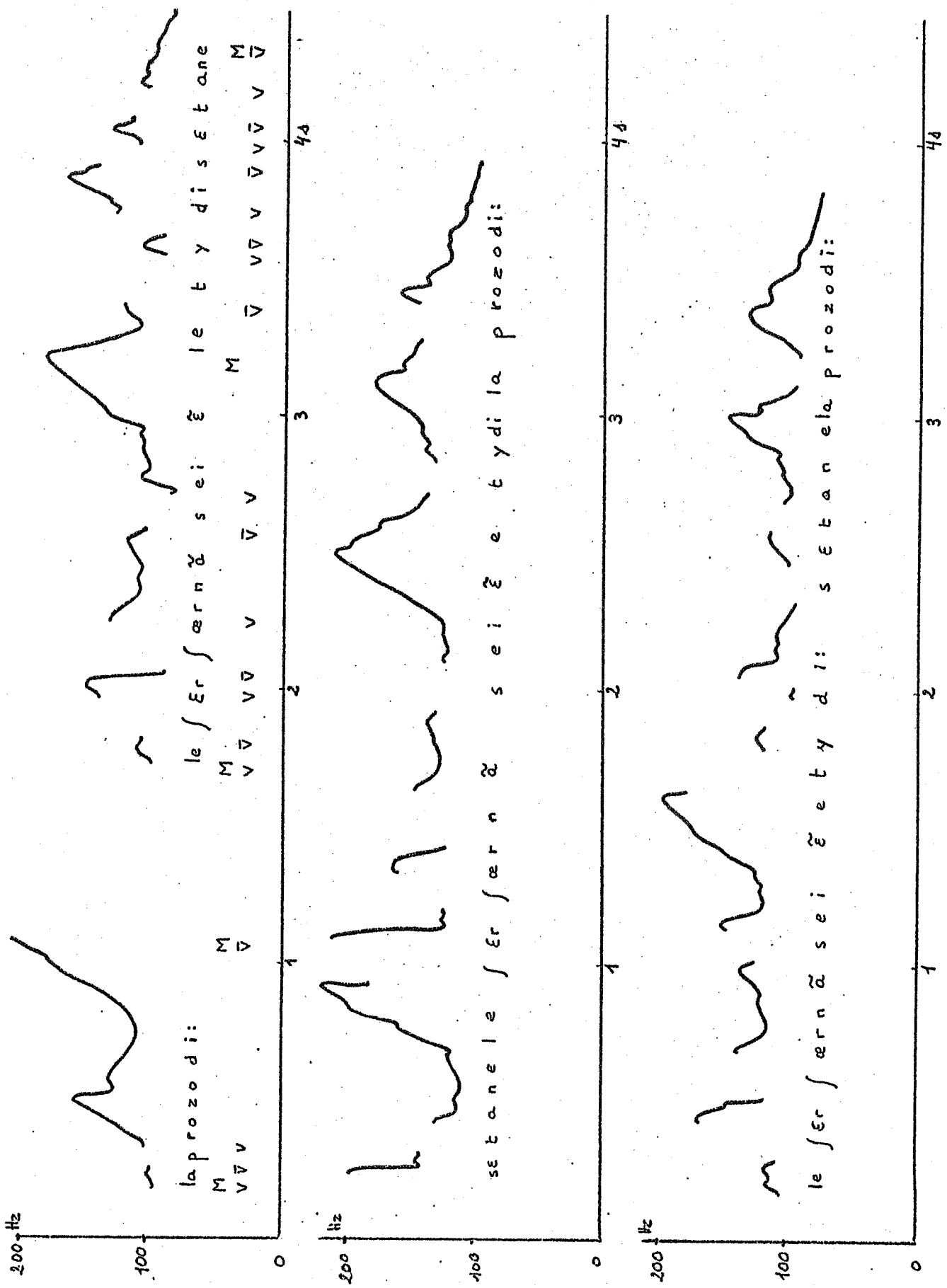


Figure 1 - contour mélodique et marqueurs prosodiques.

diques dans un système général du discours parlé, d'une part, au niveau de la segmentation et d'autre part, au niveau de l'analyse syntaxico-sémantique. Le caractère flou et variable de ces informations rend leur prise en compte délicate dans un système par ailleurs très indéterministe. Néanmoins, l'amélioration apportée est certaine et, dans une optique d'Intelligence Artificielle, la prosodie apparaît comme une des sources d'informations nécessaires pour parvenir à la compréhension d'une phrase.

BIBLIOGRAPHIE

[1] A. DI CRISTO - "Recherches sur la structuration prosodique de la phrase française" 6èmes journées d'Etude sur la Parole (JEP), Toulouse, mai 1975.

[2] J. VAISSIERE - "Caractérisation des variations de la fréquence du fondamental dans les phrases françaises" 6èmes JEP, Toulouse, mai 1975.

[3] J. VAISSIERE - "Une procédure de segmentation automatique de la parole en mots prosodiques, en français" 7èmes JEP, Nancy, mai 1976.

[4] P. MARTIN - "Intonation et reconnaissance automatique de la structure syntaxique" 6èmes JEP, Toulouse, mai 1975.

[5] M.H. O'MALLEY et al. - "Recovering parentheses from spoken algebraic expressions" IEEE Trans., AU-21, n° 3, pp. 217-220, 1973.

[6] W.A. LEA et al. "A prosodically guided speech understanding strategy" IEEE Trans., ASSP-23, n° 1, pp. 30-38, 1975.

[7] W.A. LEA - "Use of intonational phrase boundaries to select syntactic hypotheses in a speech understanding system" 92nd Meeting ASA, San Diego, Nov. 1976.

[8] J.P. HATON, J.M. PIERREL - "Organization and operation of a connected speech understanding system at lexical, syntactic and semantic levels" Int. Conf. ASSP, Philadelphia, April 1976.

[9] R. MOHR, J.P. HATON - "A parsing algorithm for imperfect patterns and its application to speech and image recognition" 3rd Int. J. Conf. on Pattern Recognition, San-Diego, USA, nov. 1976.

VIIIèmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

LES PAUSES ET LES GROUPES RYTHMIQUES : LEUR DUREE ET DISTRIBUTION EN EN FONCTION DE LA VITESSE D'ELOCUTION

Monique SAINT-BONNET . et JEAN-Louis BOE

E.N.S.E.R. Grenoble Inst. de Phonétique de Grenoble

• Résumé :

Cette étude a pour but de mettre en évidence la part respective occupée par la phonation et les pauses dans un énoncé. Un texte de 80 secondes environ a été lu deux fois à vitesse naturelle par 10 locuteurs (5 hom. et 5 fem.). Les résultats montrent que la durée de phonation varie très peu en fonction du locuteur, de son sexe et du débit. On montre ainsi que ce sont les pauses qui jouent un role important dans la vitesse d'élocution. La deuxième partie du travail a consisté à étudier la répartition, la fréquence et la durée des pauses en relation avec la structure syntaxique et le débit.

• Summary :

Precisions concerning the respective part of phonation time and pause duration are stated for oral reading. A passage of ZOLA (80 sec.) has been read twice at natural rate by 5 men and 5 women. Data stow that phonation time is almost independant of the speaker, the sex and the rate of reading. Pause duration plays an important part for the rate of speaking. A second part of this study is devoted to the distribution, the duration and the number of pauses in relation to the syntactic structure and the rate of speaker.

INTRODUCTION :

Cette étude a pour but de mettre en évidence les différents paramètres qui interviennent au niveau du débit de la parole, c'est à dire, pour l'essentiel, la durée de phonation et celle des pauses.

Pour simplifier les problèmes nous avons choisi la lecture d'un seul et même texte, par 10 locuteurs (5 hommes et 5 femmes), effectuée deux fois à des intervalles de temps espacés.

Il s'agit d'une page de Germinal de ZOLA, de 234 mots, soit environ 340 syllabes (selon la prononciation des 16 [ə]);elle est constituée de 10 phrases dont la plus courte comprend 16 syllabes.

Les enregistrements ont été effectués en chambre sourde et il a été seulement demandé aux locuteurs de lire à leur vitesse habituelle : il s'agit donc d'une étude entre différents débits naturels.

On a mesuré :
- la durée totale de la lecture : simplement chronométrée,
- la durée de phonation : automatiquement à l'aide d'une procédure décrite par ailleurs [1,2] nous nous sommes fixés une constante d'intégration de 20 ms et une dynamique de 30 dB.
- la durée de chaque pause, de chaque segment de parole compris entre deux pauses le total des pauses, le nombre des pauses : sur les tracés d'intensité délimités à l'aide d'un oscillogramme.
Pour la détermination des pauses, nous avons choisi un critère subjectif : toute séquence perçue comme une interruption de parole.
- la fréquence laryngienne moyenne F_ℓ : automatiquement à l'aide de la durée du voisement et du nombre de périodes détectées [1,2].

		Durée totale (s)	% de phonation	Nombre de syllab/mn	Nombre de pauses	Durée moy. de pauses ms.	Ph *	F_ℓ
H	m	83,2	70,0	248	39	644	1486	130
	$(\frac{\sigma}{m})\%$	11	8,5	11	18	18	16	8
F	m	79,3	74,5	258	38	542	1549	204
	$(\frac{\sigma}{m})\%$	7	4	6	16	20	14	7

* Durée moyenne des groupes de phonation (en ms).

TABLEAU I : Durée moyenne de la durée des pauses, des groupes de phonation, du débit, de la fréquence laryngienne pour 5 hommes et 5 femmes.

RESULTATS

Le tableau I présente la moyenne des mesures pour les 10 locuteurs groupés par sexe et leurs 2 lectures.

L'analyse des résultats sera faite en comparant :

. les différences entre sexes,
. les variations entre locuteurs pour chaque groupe,
. les variations pour un même locuteur, d'une lecture à l'autre.

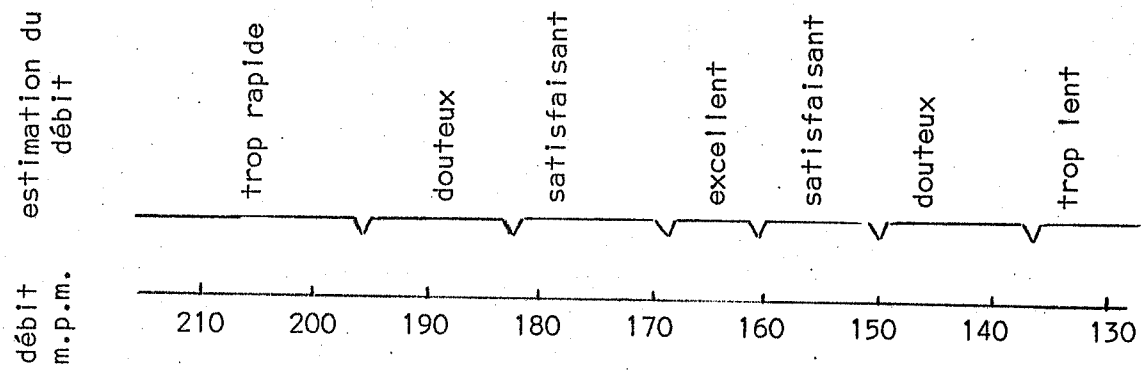


TABLEAU II - D'après Fairbanks (1960)

1. La durée totale et la durée des pauses

En moyenne la durée totale correspond à 169 mots par minute pour les hommes et à 177 m.p.m pour les femmes. C'est un débit qualifié d'excellent selon Faibanks [4] (tableau II). Les débits extrêmes relevés (143 m.p.m. et 198 m.p.m.) seraient considérés dans cette échelle comme la limite du trop lent et du trop rapide. On peut remarquer aussi que d'une lecture à l'autre, la vitesse a varié de 13% au plus pour un locuteur.

- comparaison homme-femme : La durée totale moyenne est supérieure de 5% chez les hommes. Or la durée de phonation moyenne ne varie pas avec le sexe : cette différence est donc essentiellement due à une diminution du temps de pause chez les femmes.

Chez les hommes, le temps de lecture se décompose en 70% de durée de phonation et 30% de durée des pauses, tandis que chez les femmes la phonation occupe 74,5% et les pauses 25,5% (Tableau III).

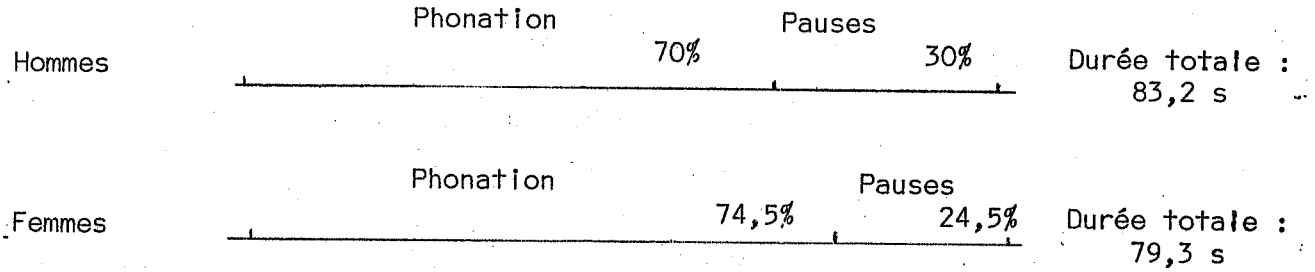


TABLEAU III

Cette différence n'est pas très significative ici, compte tenu du petit nombre de locuteurs, mais elle a été relevée comme telle sur de plus vastes échantillons [2].

- les variations entre locuteurs: si on observe maintenant les plages de variation de chacune de ces durées on observe que :

la durée de phonation varie dans les mêmes proportions chez les hommes et les femmes. L'écart type ramené à la moyenne $(\frac{\sigma}{m})\%$ est faible, 5% pour chaque groupe alors que pour la durée totale il correspond respectivement à 11% et 7%.

TABLEAU IV

	totale	phonation	pauses
H	11	5	30
F	7	5	18

$(\frac{\sigma}{m})\%$ pour les durées

Ce sont donc aussi les seules variations de la durée des pauses qui expliquent la différence de variation de la durée totale entre les hommes et les femmes ; en effet, alors que la durée des pauses varie de 18% chez les femmes, elle varie de 30% chez les hommes.

On peut donc avancer que le facteur essentiel de variation du débit est constitué par la durée des pauses aussi bien pour les hommes que pour les femmes. D'autre part, ce facteur temporel et ses variations sont plus faibles chez les femmes que chez les hommes ; ce qui explique chez celles-ci, à la fois un débit plus rapide et des variations de ce débit plus faibles.

Le coefficient de corrélation entre durée totale et durée des pauses est de 0,94 chez les hommes et 0,84 chez les femmes (il est significatif à partir de 0,55 pour un seuil de confiance de 95%).

- les variations pour un même locuteur : la variation individuelle du débit entre deux lectures est en moyenne de 6,8% pour les hommes et 5,3% pour les femmes, alors que la variation individuelle moyenne de la durée de phonation est identique et très faible chez les uns et les autres.

La différence dans la part respective des pauses et de la parole entre les hommes et les femmes se retrouve aussi au niveau de chaque locuteur.

	totale	phonation	pauses
H	6,8	3,4	13
F	5,3	3,5	7,5

TABLEAU V : variation des durées entre deux lectures pour un même locuteur (exprimée en %).

2. Le nombre des pauses

. Comparaison homme-femme :

Il faut se demander maintenant comment se traduit la diminution de la durée des pauses chez les femmes (ou son accroissement chez les hommes). S'agit-il d'une réduction de leur nombre ou de leur durée respective?

Le nombre des pauses est pratiquement le même pour les deux catégories de locuteur (39 et 38) : c'est donc leur durée qui est à l'origine de la disparité (16% de différence).

	Nombre de pauses		Durée d'une pause (ms)	
	m	$(\frac{\sigma}{m})\%$	m	$(\frac{\sigma}{m})\%$
H	39	18	644	18
F	38	16	542	20

TABLEAU VI

On constate que, pour chacun des deux groupes, le débit (exprimé en syllabes/minute) est en corrélation beaucoup plus étroite avec le nombre de pauses qu'avec leur durée moyenne.

R	débit	
	nombre de pauses	durée des pauses
H	- 0,84	- 0,67
F	- 0,60	*

TABLEAU VII

Coefficient de corrélation R

*non significatif

Nous avons déjà vu que les corrélations sont plus fortes pour les locuteurs masculins entre les pauses et leur débit, car elles sont plus longues.

Pour chaque catégorie de locuteurs (H et F) l'accélération du débit se fait préférentiellement par la réduction du nombre de pauses alors que la différence de débit entre les deux sexes n'est due qu'à la différence de la durée moyenne des pauses.

. pour un même locuteur : il s'agit de vérifier si un même individu réduit plutôt le nombre que la durée de ses pauses quand il parle plus vite . Dans 5 cas sur 10 c'est le nombre qui diminue alors que la durée moyenne augmente simultanément ou reste stable. Dans 3 cas c'est le contraire ; dans 2 cas la durée et le nombre de pauses diminuent simultanément.

Il semble donc que chaque locuteur ait tendance à réduire plutôt le nombre des pauses que leur durée. Mais le faible nombre de sujets ne permet pas de trancher, d'autant plus que la syntaxe du texte est relativement complexe (nombreuses propositions relatives, coordinations, appositions etc...). D'autre part dans un texte écrit la présence des virgules rend leur réalisation presque systématique. Or, ici, les virgules et les points représentent 78% des pauses possibles et sur ces pauses possibles, 71% ont été réalisées. Ce pourcentage est élevé, compte tenu du fait que la réalisation de toutes les pauses rendait la lecture peu intelligible. On observe d'ailleurs que, mis à part les 9 points (limite de phrases) qui n'ont jamais été omis(100% de réalisations sur 180 cas), seules 6 pauses (matérialisées par des virgules) n'ont jamais été supprimées. Donc 15 pauses seulement (soit 28% du total) étaient suffisantes pour assurer une bonne réalisation du texte et cependant aucun locuteur n'en a réalisé moins de 51%. On peut se demander si ce pourcentage représente la proportion de pauses qu'il faut réaliser pour répondre aux exigences de l'intelligibilité, du naturel et de la respiration phonatoire?

Il faut rappeler ici que les pauses ont des fonctions multiples, l'intelligibilité de la construction syntaxique ne justifie la présence que d'une faible proportion d'entre elles (c'est-à-dire celles qui délimitent les phrases). Mais l'interruption de la parole est nécessaire au lecteur comme à l'auditeur pour assimiler progressivement le texte [8].

3. la durée moyenne des pauses

Les pauses durent en moyenne 100 ms de plus pour les hommes mais les dispersions pour les deux sexes sont les mêmes (environ 19%).

Pour un même locuteur, les variations de durée des pauses sont très étendues puisqu'elles peuvent varier de 30 ms (entre deux voyelles par exemple dans la séquence "et/en sortant", "l'après-midi/il") à 2 secondes. La durée moyenne des pauses correspondant aux fins de phrases (marquées par un point) est presque deux fois plus importante que celle des pauses en général (1,1 s pour les deux sexes.

4. les groupes de phonation

La durée moyenne des groupes de phonation (séquence située entre deux pauses) est sensiblement la même en moyenne pour les deux catégories de locuteurs; d'autre part elle est en corrélation très étroite avec le nombre de pauses ($R = -0,97$).

Mais on observe qu'entre les deux lectures d'un même locuteur, la variation en % du nombre de pauses est deux fois plus importante que celle de la durée moyenne des groupes de phonation. Ceci indique que la réduction de la durée d'un groupe de phonation due aux nombres de pauses est légèrement atténuée par un allongement de la durée de phonation.

Le nombre de syllabes allongées précédant une pause explique vraisemblablement ce phénomène.

5. La hiérarchie de suppression des pauses

Est-ce que les pauses qui sont supprimées ou ajoutées sont situées à des places que l'on peut caractériser sur le plan linguistique? Sur les 43 pauses matérialisées, 15 ne sont jamais supprimées : les 9 points et 6 autres que nous avons cherchés à spécifier dans un niveau d'analyse syntaxique. Cette approche n'a pas abouti dans la mesure où le corpus utilisé est trop limité.

On observe que les pauses supprimées, d'une part suivent des groupes de sens courts (1 à 3 mots) et d'autre part qu'elles le sont plus facilement devant ou après un groupe comportant une marque syntaxique (dans, comme, où...) permettant d'identifier clairement ses limites.

Si l'on fait intervenir un facteur de distance entre le déterminé et le déterminant (ou le thème et le propos) à l'échelle de la phrase : dans une phrase où le sujet et le verbe sont séparés par une incise qui dure ici plus de 4 s, les pauses qui les séparent sont supprimées dans 53% des cas alors que les autres pauses ne disparaissent que dans 26% des lectures.

Il semble donc que l'on puisse avancer une hiérarchie dans la résistance des pauses mais il faudrait pouvoir donner une formulation quantitative de la distance (ou de la cohésion) entre deux groupes de sens, ce qui dépasse notre propos.

6. La fréquence laryngienne moyenne et le débit

Enfin on n'observe aucune corrélation, chez un même locuteur, entre sa fréquence laryngienne moyenne F_ℓ et son débit. Cependant chez les hommes le coefficient de corrélation (tous locuteurs confondus) entre le débit et F_ℓ est de - 0,85 (donc largement significatif) et chez les femmes de - 0,67.

Donc la corrélation est négative contrairement à ce à quoi on s'attendrait si on extrapolait la comparaison entre les deux sexes. Ce résultat demande à être confirmé sur de plus vastes échantillons.

7. CONCLUSION

Le débit moyen apparaît comme la résultante d'un système faisant intervenir trois facteurs indépendants qui peuvent se cumuler : le nombre de pauses, la durée de chaque pause et la vitesse d'élocution dans chaque groupe de phonation; l'importance relative de ces trois facteurs est décroissant dans ce même ordre. La suppression des pauses se fait selon une hiérarchie qu'il faut spécifier en termes linguistiques (distance et profondeur des groupes dans l'arbre syntaxique).

REFERENCES

[1] *BOE J.L.*
"Etude des vibrations des cordes vocales dans la parole : méthodes, résultats et applications"
Revue d'Acoustique *37*, 105-112 (1976)

[2] *BOE J.L., CONTINI M., RAKOTOFININGA H.*
"Etude statique de la fréquence laryngienne. Application à l'analyse et à la synthèse des faits prosodiques du français".
Phonetica *32*, 1-23 (1975)

[3] *BOE J.L., RAKOTOFIRINGA H.*
"A statistical analysis of laryngeal frequency : its relationship to intensity level and duration".
Language and speech *18*, 1-13 (1975).

[4] *FAIRBANKS G.*
Voice and articulation drillbook
2 ed. Harper & Row, Pub. New York, Evanston , London (1960)

[5] *FONAGY I., MAGDICS K.*
"Speed of Utterance in Phrases of Different Lengths".
Language and Speech *3*, 179-192 (1960).

[6] *GOLDMAN-EISLER F.*
"The distribution of pause durations in speech".
Language and speech *4*, 232-237 (1961).

[7] *GOLDMAN-EISLER F.*
Psycholinguistics - Experiments in spontaneous speech
Academic press, London , New-York (1968)

[8] *GOLDMAN-EISLER F.*
"Pauses, clauses, sentences".
Language and speech *15*, 103-113 (1972)

[9] *GROSJEAN F., DESCHAMPS A.*
"Analyse des variables temporelles du français spontané".
Phonetica *28*, 191-226 (1973).

[10] *LUCCI V.*
"Etude phonostylistique du rythme et de la variabilité de la longueur en français parlé et français lu".
Bulletin de l'Institut de Phonétique de Grenoble *2*, 139-161 (1973).

[11] *LUCCI V.*
"Rythme et longueur du message parlé. La conversation".
Bulletin de l'Institut de Phonétique *3*, 139-152 (1974).

[12] *MALECOT A., JOHNSTON R., KIZZIAR P.A.*
"Syllabic rate utterance length in French".
Phonetica *26*, 235-251 (1972).

[13] *SNIDECOR J.C.*
"A comparative study of the pitch and duration characteristics of impromptu speaking and oral reading".
Speech Monographs *10*, 50-56 (1943).

VIIIèmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

PREMIERS ESSAIS D'UTILISATION DE LA DUREE POUR LA SEGMENTATION EN MOTS DANS UN SYSTEME DE RECONNAISSANCE

Jacqueline Vaissière

C.N.E.T.

Lannion

Résumé :

Cette communication illustre l'utilisation possible de la durée pour la reconnaissance des frontières de certains mots dans la phrase. L'expérience se limite à la voix d'un seul locuteur, et sur un corpus limité. Après une division du continuum en pseudo-syllabes et en voyelles (voir Fig.1 et Ref.1), le programme calcule pour chaque segment détecté sa longueur (voir Fig.4). Les segments les plus longs sont interprétés comme appartenant à la dernière ou à la première syllabe d'un mot (voir Fig.3). Ce programme devra être modifié pour pouvoir s'adapter à différents rythmes d'élocution, et à divers locuteurs. Des corrections devront être apportées, tenant compte de la classe des phonèmes sous-jacents (sourds ou sonores, plosives ou fricatives, et les autres).

Summary :

This paper represents an attempt to use the information contained in the lengthening of certains syllabes or vowels in a sentence, as an help to word boundaries location in a speech recognition system (Keal system). After an automatic segmentation of the continuum (see Fig.1 and Ref.1) into pseudo-syllables and vocalic segments, is completed, a program computes and compares the length of the successive segments (see output of the program on Fig.4). The longer segments are interpreted as belonging to the last or to the first syllable of a word. Results for one speaker are discussed. Further studies are needed to adjust the parameters to various speakers and different speeds of elocution, and to integrate some relevant influences from the identities of the underlying phonemes (voiced-non-voiced, plosives-fricatives, and the remainings).

PREMIERS ESSAIS D'UTILISATION DE LA DUREE POUR LA SEGMENTATION EN MOTS DANS UN SYSTEME DE RECONNAISSANCE

L'évolution du fondamental sur les voyelles, de la durée des phonèmes en séquence et de leur intensité concourent au même rôle de démarquation et de marquage syntaxique dans les phrases françaises.

Des essais d'intégration de la fréquence du fondamental dans le système de reconnaissance du CNET, pour la segmentation en mots dans les phrases lues sans insistance sont en cours. L'information syntaxique contenue dans l'évolution du fondamental sur la dernière syllabe des mots n'est pas encore utilisée, du fait même que les mots susceptibles d'avoir été prononcés sont "découverts" en fin de chaine, par l'analyseur syntaxique lui-même. L'intégration de cette information est donc complexe.

Cette communication s'intéresse à l'intégration du second facteur prosodique, l'évolution de la durée des phonèmes au cours de la phrase. Le premier programme de la chaine de reconnaissance, l'analyseur phonétique (SONIADEL), opère une segmentation du continuum en pseudo-syllabes (fondée sur l'évolution de l'énergie globale), en voyelles et en consonnes.[1] Bien que cette segmentation ne corresponde pas toujours à la segmentation qui serait faite manuellement (certains phonèmes peuvent être rajoutés, d'autres soudés entre eux), elle est en général satisfaisante ce qui nous a permis d'envisager d'utiliser les durées des segments découverts comme aide à la segmentation en mots (voir un exemple de segmentation à la figure 1).

Nous reverrons succintement dans une première partie les principaux facteurs connus intervenant sur la durée des phonèmes lors de leur actualisation dans les phrases. Dans la seconde partie, nous proposerons un algorithme très simple pour l'utilisation de la durée et nous présenterons les résultats obtenus par l'application de cet algorithme sur un corpus de phrases prononcées par un locuteur.

I) Facteurs intervenant sur la durée des phonèmes dans la phrase.

En premier lieu, tous les phonèmes n'ont pas une durée intrinsèque identique et certains phonèmes ont tendance à être plus longs que d'autres.

La figure 2a illustre les résultats bien connus de House et Fairbanks[2] sur la durée intrinsèque des voyelles anglaises. On remarque sur cette figure que les voyelles fermées (/i/et/u/) ont tendance à avoir une durée intrinsèque inférieure aux voyelles ouvertes (/a/). Une telle tendance a été vérifiée pour d'autres langues que l'anglais, telles que l'allemand, le danois, le suédois, etc... (Pour un résumé et une bibliographie à ce sujet, voir Lehiste[3]) En ce qui concerne le français, Pierre Fouché[4] distinguait déjà des voyelles longues et des voyelles brèves et notait que la durée semblait être corrélée avec leur degré d'ouverture. Delattre[5] donne pour exemples les paires de mots suivantes :
/ɔ/, /o:/: pomme, paume et /oe/,/ø:/: jeune, jeûne
Delattre note également que les voyelles nasales ont tendance, en syllabes fermées, à être plus longues que les voyelles orales qui leur correspondent :
/ɛ/,/ɛ̃/: messe, mince, /ɔ/,/ɔ̃:/: mode, monde et /a/,/ã:/: bac, banque.

Moins d'études concernent la durée intrinsèque des consonnes. Celle-ci semble être fonction à la fois du point d'articulation et du mode d'articulation de la consonne. Mais les conclusions des différents auteurs ne convergent pas (voir Lehiste[3] pour une bibliographie).

En second lieu, la durée de chaque phonème semble être influencée par la nature des phonèmes environnants.

L'influence des consonnes sur les voyelles a été beaucoup étudiée (voir Lehiste[3]). La figure 2b résume les résultats de House et Fairbanks[2] sur l'anglais.

Le mode et le point d'articulation de la consonne ont une influence sur la durée de la voyelle qui précède : les consonnes voisées plosives et fricatives) ont une influence allongeante par rapport aux consonnes non voisées qui leur correspondent (comparez la durée moyenne des voyelles dans l'environnement /p, t, k,/ et /b, d; g,/ d'une part, et /f, s/ et /v, z/ d'autre part). Les fricatives ont également une influence allongeante par rapport aux plosives. Le point d'articulation de la consonne semble influencer la longueur de la voyelle, mais les résultats sont moins nets (différence entre /b/et/g/ d'une part, et /d/, d'autre part).

En troisième lieu, la place du phonème dans la syllabe et le nombre de phonèmes dans la syllabe semblent avoir une influence. Les syllabes ouvertes ont tendance à être plus courtes que les syllabes fermées (Delattre note une différence de 150 ms pour les syllabes en position finale dans le groupe). La durée de la voyelle n'est pas simplement influencée par la nature des consonnes adjacentes, mais encore par le nombre de consonnes qui suivent et précèdent cette voyelle.

En quatrième lieu, la place de la syllabe dans le mot et le nombre de syllabes dans le mot ont une influence sur la durée des phonèmes qui les composent. Il est bien connu que la dernière syllabe des mots français est la plus longue (voir les travaux de Armstrong et de Coustenoble, [6] de Spang-Thomsen, [7] entre autres). Delattre note que "la durée est le seul des trois éléments acoustiques qui soit toujours par sa proéminence un facteur de l'accent". (Delattre, 1938, pg 145). Beauchemin [8] a comparé la durée des voyelles en syllabes accentuées et en syllabes non accentuées (333 ms en moyenne pour le premier type et 145 ms pour le second type). Zichova [9] découvre dans le cadre de ses expériences que la limite inférieure des voyelles inaccentuées est approximativement de 60 ms et celle des voyelles accentuées de 130 ms. Dans les mots longs, nous avons remarqué que la première syllabe a également tendance à être légèrement plus longue, toutes choses égales ailleurs, que les syllabes suivantes (excepté la dernière), grâce à une allongation de la première consonne du mot, en général (cette allongation de la première syllabe est concommitente à la réalisation de l'éventuelle montée de Fo durant la consonne). Cette tendance ne semble pas être particulière au français, et elle a été découverte pour d'autres langues, tel que l'anglais et le suédois (voir les travaux de Lindblom [10] et de ses collègues pour le suédois, et ceux de Barnwell [11] pour l'anglais). L'influence du nombre de syllabes dans un mot est également bien connue (voir les travaux des auteurs précédents, par exemple, qui ont fait l'étude systématique de ce phonème) : plus le mot est long, et plus les syllabes qui le composent ont tendance à se raccourcir. Cette affirmation doit être nuancée par le fait que dans les mots de plus de quatre syllabes, on peut observer chez certains locuteurs français une tendance à alterner syllabes longues et syllabes brèves. Certaines infuences de la décomposition du mot en morphes ont été également observées par Mr Morand et Mme Caelen. (12)

En cinquième lieu, la place du mot par rapport aux pauses de la phrase et le rythme avec lequel la phrase a été prononcée joue un rôle. Le phénomène de l'allongement de la dernière syllabe avant une pause est bien connu, et il n'est pas particulier au français.

Toutes les influences précédemment citées interfèrent de façon complexe et il ne serait pas aisé de prédire par règles le profil temporel des phrases, tout au moins de façon fine (notons que le degré d'insistance sur tel ou tel mot, qui est également un facteur déterminant pour l'attribution de durées, ne peut être prédit à priori). Certains auteurs ont noté des phénomènes de compensation de durée d'une syllabe à l'autre (Kozhevnikov et Chistovitch [13] Huggins [14]. Néanmoins certains chercheurs ont entrepris d'établir des modèles de génération de la

durée. (Klatt, O'Shaughnessy et Oller pour l'anglais, Lindblom, Carlstrom et leurs collègues pour le suédois).

II) Algorithme et résultats

La figure3 représente le modèle théorique pour le profil temporel des mots en français, modèle que nous avonsdérivé des observations précédentes et des résultats obtenus lors de l'étude de la durée dans les phrases prononcées par un locuteur professionnel. Ce modèle ne tient compte que des influences énoncées en quatrième lieu, c'est-à-dire des influences au niveau du mot et devra être corrigé par la suite en fonction de l'influence des phonèmes.

A partir de ce modèle extrêmement simplifié, nous avons déduit l'algorithme de reconnaissance suivant : toute syllabe de plus de 240 ms est considéréecomme la première ou la dernière syllabe d'un mot (condition I); toute voyelle de plus de 120 ms est considéréecomme la première ou la dernière voyelle d'un mot (condition 2), même si la durée de la syllabe est inférieure à 240 ms.

Le corpus de référence (annexe) a été enregistré par un certain nombre de locuteurs, puis analysé par le programme d'analyse phonétique de Guy Mercier. Ce programme calcule la durée des syllabes (ou pseudo-syllabes) qu'il segmente, ainsi que la durée des noyaux vocaliques. Ces données sont reprises par un second programme, écrit par Christian Le Corre, qui extrait les données qui lui sont nécessaires pour le programme prosodique dans son ensemble (évolution du fondamental d'une voyelle à l'autre, durée des voyelles et des syllabes reconnues, intensité maxima sur chaque voyelle). La figure 4 illustre une sortie de ce programme. L'algorithme a été testé manuellement en relevant sur ces sorties le numéro des syllabes dont la durée est supérieure ou égale à 240 ms (ce qui équivaut à un nombre supérieur ou égal à 18 dans la colonne indiquant la durée syllabique, la fréquence d'échantillons étant de 13,3 ms), et le numéro de celles dont le noyau vocalique a une durée supérieure ou égale à 120 ms (ce qui équivaut à un nombre supérieur ou égal à 9 dans la colonne indiquant la durée de la voyelle détectée).

Nous ne reportons dans cette communication que les résultats concernant un seul locuteur, car les échantillons d'un seul locuteur ont pu être analysés à la date d'aujourd'hui (15 Mars). Nous espérons pouvoir présenter les résultats concernant les autres locuteurs lors des Journées d'Etudes.

Dans le corpus de référence, le programme de segmentation a découvert 75 syllabes répondant à la condition I (être de durée supérieure ou égale à 240 ms). Parmi ces 75 syllabes, 51 correspondent effectivement à la dernière syllabe prononcée de mots lexicaux, 15 à la première syllabe de mots lexicaux, 4 à des mots monosyllabiques grammaticaux, et 3 à des syllabes dont le noyau vocalique est un e muet prononcé à la fin de mots lexicaux. Dans 73 cas (sur 75), une durée de plus de 240 ms ou de 240 ms indique bien la première ou la dernière syllabe de mots. Dans deux cas cependant, il s'agit de syllabes intercalaires (ni la première, ni la dernière), les syllabes "ver" et "pé" des mots "université" et "supérieures" ont été trouvées plus longues que 240 ms. Le nombre des erreurs, 2 sur 75 détections est relativement limité, mais il révèle cependant que l'algorithme n'est pas totalement fiable : une application stricte de l'algorithme aurait éliminé parmi les candidats les mots "université" et "supérieure" et aurait empêché une bonne reconnaissance des phrases 5 et 7. Les résultats sont illustrés sur le tableau suivant (à droite du tableau sont représentés les résultats pour les syllabes remplissant la condition 1 ou 2) :

Condition I : la syllabe doit avoir une durée égale ou supérieure à 240 ms
Condition II : la voyelle doit avoir une durée égale ou supérieure à 120 ms

NOMBRE DE SYLLABES	Satisfaisant à la Condition I	Satisfaisant à la Condition I ou II
	75	81
Dernière syllabe de mots lexicaux : - prononcées	51	55
- e muet	3	3
Premières syllabes de mots lexicaux :	15	16
Mots grammaticaux	4	4
Syllabes intercalaires	2	3

III) Conclusion

Les résultats semblent assez satisfaisants pour le locuteur étudié. Nous prévoyons néanmoins, que l'algorithme de départ se révèlera insuffisant pour l'ensemble des locuteurs et il faudra introduire une correction en fonction de la classe des phonèmes reconnus : plosives sourdes, fricatives sourdes, consonnes voisées et voyelles.

Le seuil devra être calculé en fonction de la vitesse d'élocution (nombre de syllabes par unité de temps), qui peut être dérivé automatiquement par le programme Soniadel. Il est évident que, si les locuteurs parlent vite, le seuil de 240 ms par la durée des syllabes sera trop élevé.

BIBLIOGRAPHIES et NOTES

(1) Mercier Guy, Segmentation et reconnaissance de la parole en temps réel et différé, J.E.P., Aix, 1971.
(2) House, A.S. and Fairbanks, G., The Influence of Consonant Environment upon the Secondary Acoustical Characteristics of Vowels, J.A.S.A., 25, 1953, pp 105-13.
(3) Lehiste I., Suprasegmentals, The MIT Press, 1970.
(4) Fouché P., Traité de Prononciation Française, Klincksieck, Paris, 1959.
(5) Delattre P., L'accent final en français : accent d'intensité, accent de hauteur, accent de durée, French Review, 12 (2), pg 145-5, 1938
(6) Coustenoble et Armstrong, Studies in French Intonation, Heffer, Cambridge, 1934.
(7) Spang-Thomsen, L'accent en français moderne, Orbis Litterarum, suppl. 3, 1963, pg 92-102.
(8) Beauchemin, N., Correlation des durées sous l'accent en Français, Proceedings of the 7th Intern. Congress of Phonetics Sciences, Montréal, 1971, pg 860-3.
(9) Zichova, M., Contribution à l'étude de la durée des voyelles françaises, Acta Carolinae, Philologica 6, Phonetica Pragensia, 1967, pg 113-24.
(10) Lindblom, B., and Rapp, K., Some temporal Regularities of Spoken Swedish, Symposium on Auditory Analysis and Perception of Speech, Leningrad, 1973.
(11) Barnwell, T.P., An Algorithm for Segment Durations in a Reading machine context, Technical Report, Research Laboratory of Electronics, MIT, 1971.
(12) Au cours de discussions.
(13) Kozhevnikov, V.A., and Chistovitch, L.A., Speech : Articulation and Perception, Traduction, 1965.
(14) Huggins A.W.F., The Perception of Timing in Natural Speech : compensation within the syllable, Language and Speech, Vol. 11, Part 1, 1968, pg 1-11.
(15) Le locuteur professionnel de la Compagnie IBM, à la Gaude.

ANNEXE

1. Le Centre National d'Etudes des Télécommunications
2. L'Institut de Phonétique de l'Université de Provence
3. Le Centre de Recherches de la Compagnie Générale d'Electricité
4. L'Institut de Phonétique de l'Université Libre de Bruxelles
5. Le Laboratoire de Phonétique de l'Université des Sciences Humaines de Strasbourg
6. Le Laboratoire de la Communication Parlée et de l'Instrumentation de Mesure de Grenoble
7. L'Ecole Nationale Supérieure des Télécommunications
8. Le Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur du C.N.R.S.
9. Le Laboratoire de Reconnaissance des Formes de la Faculté des Sciences de Paris
10. Le Laboratoire d'Electricité et d'Automatisme de l'Université de Nancy
11. Le Centre d'Etudes et de Recherches de la Compagnie I.B.M. à la Gaude
12. L'Institut de Phonétique de l'Université de Besançon
13. L'Institut de Phonétique de l'Université des Langues et Lettres de Grenoble

FIGURES (voir pages suivantes)

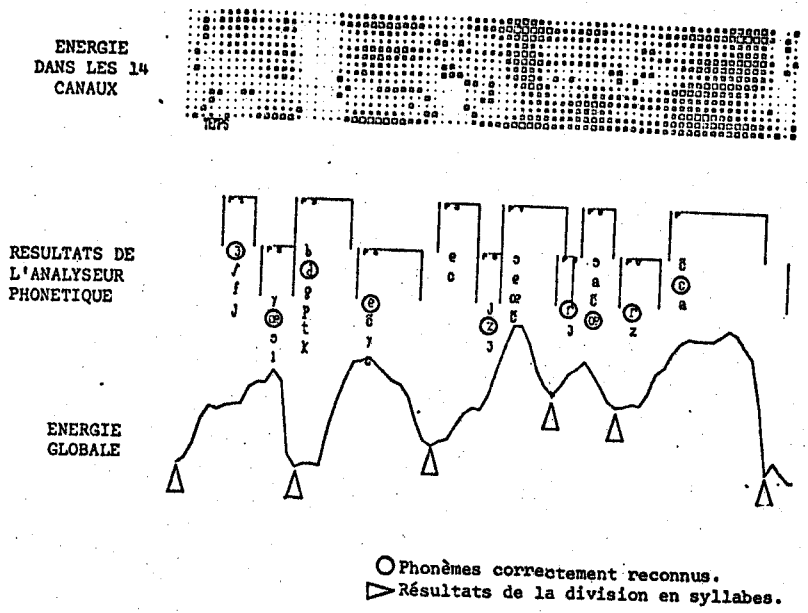


Figure 1: Résultats de la segmentation automatique du continuum en syllabes. Les frontières de syllabes correspondent à un minimun d'énergie entre deux noyaux vocaliques. La division en consonnes et en voyelles (qui ne correspond pas obligatoirement à la division en syllabes) est indiquée par des traits verticaux.

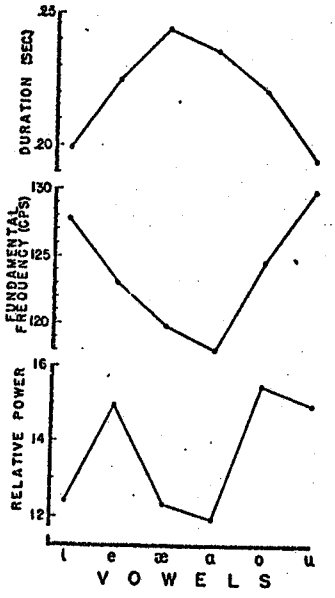


FIG. 4. Mean duration, fundamental frequency, and relative power of vowels. Consonant environments pooled.

Figure 2a: Résultats de House et Fairbanks[2] sur les valeurs intrinsèques de durée, de fréquence du fondamental et d'intensité de 6 voyelles anglaises

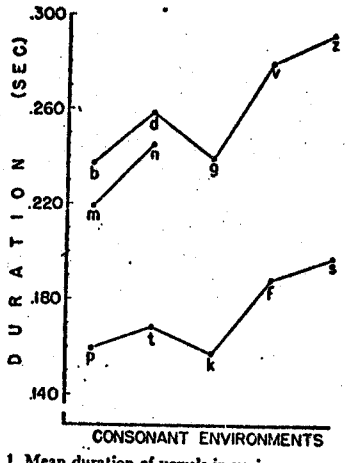


FIG. 1. Mean duration of vowels in various consonant environments. Vowels pooled.

Figure 2b: Influence des consonnes sur la durée des voyelles.

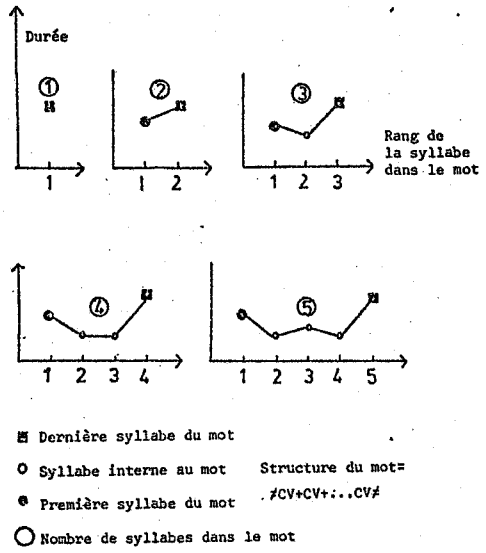


Figure 3: Durée théorique des syllabes successives dans des mots composés de une à cinq syllabes. La dernière syllabe a tendance à être la plus longue. L'allongement éventuel de la première syllabe peut être du à une allongation de la première consonne du mot. Lorsque le mot devient plus long, sa composition morphologique peut avoir une influence.

DONNEES POUR L'ETUDE DE LA PROSODIE

LOCALISATION DES SYLLABES				FREQUENCE FONDAMENTALE				DUREE DES SYLLABES			DUREE DES VOYELLES			ENERGIE DE LA SYL
NO SYL	DEB SYL	FIN SYL	MAXI	VAL MAX	ECART	% RELAT	LONG SYL	ECART	% REL	L-VOY	ECART	% RELATIF	ENERG DE LA SYL
1	4	27	12	93	0	0	24	0	0	4	0	0	1594
2	28	38	35	91	-2	-2	11	-13	-54	2	-2	-50	376
3	39	57	54	94	3	3	19	8	72	4	4	200	783
4	58	66	62	120	28	27	9	-10	-52	5	-1	-16	619
5	67	83	81	97	-19	-19	17	8	88	7	2	40	1295
6	84	94	89	99	2	2	11	-6	-35	5	-2	-28	757
7	95	118	108	99	0	0	24	13	118	5	0	0	765
8	119	133	129	139	40	40	15	-9	-37	4	-1	-20	579
9	134	142	140	103	-25	-25	9	-6	-40	4	0	0	518
10	143	152	148	95	-8	-7	10	1	11	8	4	100	657
11	153	165	157	102	7	7	13	3	30	4	-4	-50	1068
11	153	165	157	102	0	0	13	0	0	4	0	0	1068
12	166	176	173	113	11	10	11	-2	-15	1	-3	-75	416
13	177	195	188	115	2	1	19	8	72	6	5	500	1113
14	196	204	200	124	9	7	9	-10	-52	4	-2	-33	541
15	205	218	215	136	12	9	14	5	58	8	4	100	503
16	219	255	227	132	-4	-2	37	23	164	12	6	75	1536

Résultats obtenus par l'analyse automatique de la valeur des trois paramètres prosodiques (fréquence du fondamental, durée et intensité) dans la phrase" L'institut de Phonètique de l'université de Provence.".
L'écart du fondamental (en colonne 6: f=1/64.v) permet de repérer la fin des mots "institut" et"phonétique", et le début des mots "phonétique" et "université".
Les durées des syllabes (en colonne 8:la valeur est à multiplier par 13,3 ms) permet de repérer la fin des 4 mots lexicaux. Voir texte.

VIIIèmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

UTILISATION, POUR LA RECONNAISSANCE DE LA PAROLE CONTINUE, DE MARQUEURS PROSODIQUES EXTRAITS DE LA FREQUENCE DU FONDAMENTAL.

VIVES R., LE CORRE C., MERCIER G. et VAISSIERE J.

C.N.E.T., LANNION

• Résumé :

L'information fournie par un programme de segmentation automatique de la parole continue, à partir de l'évolution de la fréquence du fondamental, peut être utile à différents niveaux des systèmes de reconnaissance de la parole. Les tests ont été effectués sur 13 phrases prononcées par un locuteur et une évaluation des gains de la méthode est donnée sur les réponses des niveaux lexical et syntaxique du système complet de reconnaissance KEAL.

• Summary :

The information contained in the fundamental frequency (Fo) fluctuations along the sentences can be used with profit at different levels of recognition system. The results have been tested for the recognition by the System Keal of 13 sentences spoken by 1 speaker, and the improvements for the lexical and syntactic levels are evaluated.

UTILISATION, POUR LA RECONNAISSANCE DE LA PAROLE CONTINUE, DE MARQUEURS PROSODIQUES EXTRAITS DE LA FREQUENCE DU FONDAMENTAL.

VIVES R., LE CORRE C., MERCIER G. et VAISSIERE J.

C.N.E.T., LANNION

I. - INTRODUCTION.

Cette communication concerne l'utilisation dans le système de reconnaissance du CNET, Keal, (1) d'une partie de l'information contenue dans les variations de la fréquence du fondamental (Fo) d'une voyelle à l'autre dans la phrase déclarative lue par un locuteur, en vue de la réduction du nombre des traitements lexical et syntaxique.

Dans un article précédent, une méthode d'utilisation des variations de Fo avait été décrite (2). L'algorithme a été depuis lors légèrement modifié. Les modifications concernent premièrement le choix des valeurs de Fo : une référence à l'énergie est désormais introduite, afin de ne sélectionner qu'une seule valeur de Fo par syllabe, sur la voyelle et de remédier à certaines catégories d'erreur de la première version qui ne référait qu'à la courbe de Fo pour repérer les valeurs de Fo sur les voyelles. Deuxièmement, le groupe prosodique est maintenant décomposé en partie initiale, finale et médiane pour séparer les larges mouvements de Fo dûs à la réalisation globale des groupes prosodiques, de ceux qui sont porteurs d'indications sur les mots.

II. - RAPPEL D'ELEMENTS DE LA THEORIE SOUS-JACENTE ET ALGORITHME.

II.-1 - Décomposition de la phrase en groupe prosodique -

Une phrase peut être prononcée en un seul groupe de souffle, ou encore en plusieurs groupes prosodiques, séparés par des pauses.

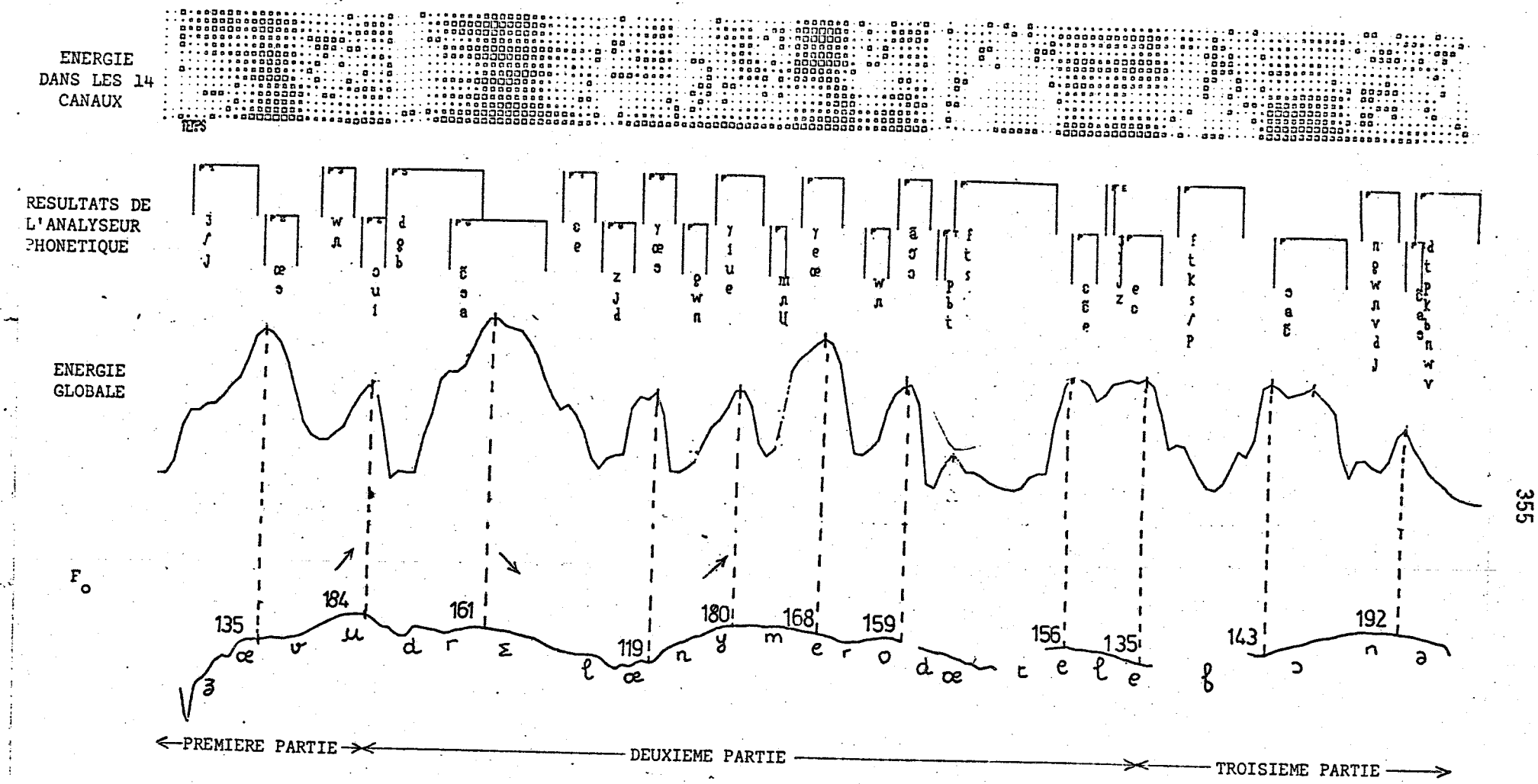
Le premier rôle du programme sera de détecter les pauses (d'une durée égale au moins à 140 ms de silence) et de délimiter ainsi les groupes prosodiques.

II.-2 - Décomposition des groupes prosodiques en deux ou trois parties :

A chaque groupe prosodique se superpose l'un des deux schémas prosodiques caractéristiques ∼ et ∧. Chaque schéma global peut se décomposer en deux ou trois parties :
une partie initiale montante (Partie 1), une partie finale (montante ou descendante) (Partie 3) et une partie médiane, entre les deux autres parties (Partie 2). Dans les phrases courtes, la partie finale peut succéder immédiatement à la partie initiale, et il n'y a pas de partie médiane.

Sur chaque groupe prosodique délimité, on utilise un programme de segmentation en syllabes (réf. 3 et réf. 4) qui permet de localiser les sommets vocaliques ; on repère la valeur de Fo coïncidant avec le maximum d'énergie (voir sur Fig. 1, la position des valeurs de Fo sélectionnées). Pour mesurer les variations de Fo d'une syllabe à l'autre, on calcule la différence entre les deux valeurs prises par Fo sur les deux sommets vocaliques consécutifs.

Une fois les groupes prosodiques délimités, le rôle du programme est de décomposer le groupe prosodique précédemment délimité (par l'existence de pauses) en deux ou trois parties : la première variation négative de Fo indique la fin de la première partie du groupe prosodique (Voir partie I sur Fig. 1) ; le début de la dernière partie est indiqué par le dernier changement de signe (avant la



Premier groupe prosodique de la phrase:" Je voudrais le numéro de téléphone de Monsieur Mercier." Les valeurs de fréquence du fondamental par le programme de prosodie correspondent à des sommets d'énergie. Le groupe prosodique a été séparé en trois parties par le programme: la partie initiale montante (jusqu'à la première voyelle de 'voudrais'), une partie intermédiaire, et une partie finale montante (sur la dernière syllabe du groupe dans cet exemple).

Avec les seuils s1=20Hz et s2=22, on trouve un marquage de type 1 (↗) sur le /u/ de 'voudrais' et le /y/ de 'numéro', ainsi qu'un marquage de type 2 (↘) sur le /ɛ/ de 'voudrais'.

Fig.1

pause) de la variation de Fo. La zone médiane est délimitée par la fin de la première partie et le début de la dernière partie (voir Partie 2 sur Fig. 1).

II.-3 - Interprétation des mouvements de Fo :

Les variations importantes de Fo dans la partie médiane caractérisent deux types d'indications sur les mots (5). Une montée significative entre deux voyelles (type 1), indique que la seconde de ces voyelles est, soit la première voyelle d'un mot (voir les mots "numéro" et "téléphone" à la fig. 1), soit la dernière syllabe d'un mot. Une chute rapide entre deux syllabes successives (type 2), peut indiquer soit, une frontière de mot entre ces deux syllabes (voir la fin du mot "voudrais" sur la Fig. 1), soit la première syllabe d'un mot de deux ou trois syllabes.

Le troisième rôle du programme est donc de repérer les montées et descentes importantes. Une variation positive et supérieure à un seuil s_1 entre deux voyelles consécutives entraîne un marquage de type 1 sur la seconde voyelle. Une variation négative supérieure en valeur absolue à un seuil S_2 entraîne un marquage de type 2 sur la première voyelle (voir Fig. 1), dans la partie médiane, du groupe (Partie 2). En cas de conflit de marquage, le type 1 l'emporte sur le type 2.

S'il y a une variation positive et une seule, supérieure au seuil S_1, dans la partie 1, elle reçoit également un marquage de type 1, car la montée initiale du groupe se réalise, en général, au début du premier mot lexical de ce groupe. S'il y a plusieurs montées supérieures dans la partie 1, elles ne sont pas interprétées.

Dans la partie 3, les variations de Fo ne sont pas interprétées.

III. - UTILISATION DES MARQUEURS PAR LE PROGRAMME D'ANALYSE LEXICALE.

Chaque marqueur élaboré est positionné, en principe, sur un segment voyelle d'une syllabe. Ceci n'est que théorique car 2 difficultés font que l'analyseur phonétique peut mélanger parmi les étiquettes voyelles, l'étiquette du phonème R : premièrement, dans des mots comme "pour" ou "par", il est souvent impossible de séparer la voyelle du R. Deuxièmement, dans les contextes gre, bre, etc il est très fréquent que la segmentation fournisse deux syllabes, là où il n'y en a qu'une, le R ressemblant beaucoup à une voyelle.

L'utilisation des marqueurs au niveau lexical du processus de reconnaissance, va consister à vérifier l'information prosodique sur les cadrages possibles de mots dans la phrase.

Les cadrages ne vérifiant pas les hypothèses prosodiques sont pénalisés ou éliminés (6). Dans un premier temps, nous avons pensé à bonifier les cadrages en accord avec l'information prosodique. Le score (entre 0 et 1000) obtenu pour un cadrage particulier est bonifié, dans la limite de 1000, par la quantité :

$$25 \times N \times M$$

où N est le nombre de phonèmes du mot de référence et M le nombre de coïncidences (0, 1 ou 2) où l'on a vérifié l'information prosodique pour ce cadrage. Tous les mots ne sont pas concernés par l'élimination ou la pénalisation.

La marque de type 1 peut éliminer les mots de plus de 2 syllabes dont la voyelle d'une syllabe médiane se trouve cadrée en première ou dernière syllabe. Elle élimine également tous les mots contenant un R cadré sur le R d'un segment voyelle marqué.

La marque de type 2 est plus discriminante : elle élimine ou pénalise les

mots de plus de trois syllabes dont la voyelle de la première syllabe est cadrée sur un segment marqué de type 2.

Les phrases qui nous ont servi de support pour les tests utilisent un vocabulaire de 49 mots dont : 34 possèdent le critère :"plus de deux syllabes ou contient au moins 1 R".

IV. - PREMIERS RESULTATS ET DISCUSSION.

Nous ne pouvons qu'être prudents, quant à l'interprétation des résultats car nous n'avons testé nos programmes que sur treize phrases, prononcées par 1 locuteur. Le locuteur n'est pas le locuteur sur lequel on a pratiqué l'apprentissage phonétique. Les phrases sont des intitulés de centres de recherches où il existe des équipes travaillant sur la parole. Sur chaque phrase, nous avons passé les programmes de reconnaissance phonémique, prosodique, lexicale et syntaxique. La table 1 montre des résultats comparatifs d'expériences ne tenant pas compte de la prosodie puis faisant intervenir l'information prosodique réelle et idéale. Les marqueurs prosodiques idéaux ont été placés "à la main" en complément des marqueurs réellement trouvés dans la phrase : ils ne correspondent pas à la façon "idéale" de prononcer la phrase, mais à une manière qu'on peut qualifier de riche mélodiquement. Une série d'expériences montrent l'influence des bonifications au niveau de la syntaxe.

a) Dans les expériences que nous avons menées, nous n'avons pas noté d'erreur dans le marquage des syllabes qui aurait entraîné l'élimination de bonnes détections.

b) Analyse des résultats obtenus au niveau lexical.

Chaque mot du dictionnaire a été recherché dans chacune des treize phrases. Chaque cadrage de mot obtenant au moins le score 400 est appelé détection. Une détection supprimable au sens prosodique est une détection d'un mot du dictionnaire possédant le critère :"plus de deux syllabes ou contient au moins 1 R". Ces détections représentent en moyenne 60 % du nombre total de détections. Les deux premières séries d'expériences qui ont été menées, montrent, qu'en moyenne, 35 % des détections supprimables sont éliminées grâce aux marques prosodiques réelles et 47 % le sont grâce aux marques prosodiques idéales. Il est à remarquer que si une marque prosodique réelle est à l'origine, en moyenne, de 7 suppressions, les marques ajoutées pour simuler un marquage prosodique idéal, ne provoquent en moyenne plus que 3,3 suppressions chacune. Cette chute d'efficacité s'explique par la redondance des marques idéales qui sont ajoutées en complément des marques prosodiques réelles : une marque idéale ajoutée, par exemple, en fin de mot, ne fait souvent que confirmer la marque réelle de début du mot suivant.

c) Analyse des résultats obtenus au niveau syntaxique.

Il n'y a pas eu d'améliorations dans le nombre de phrases correctement analysées. Le coefficient syntaxique qui apparaît dans la Table 1 pour évaluer les performances de l'analyseur syntaxique, est proportionnel à l'encombrement des piles développées et au nombre de calculs effectués. Il permet de noter une certaine détérioration des résultats dans les expériences où les mots bien cadrés prosodiquement ont été bonifiés. Les bonifications ont favorisé non seulement les détections de mots effectivement prononcés, mais aussi environ entre 10 à 20 % des autres détections. Ces dernières entrent en concurrence avec les bonnes détections, pendant le développement des solutions et ont pour conséquences les dégradations en temps et en place, obtenues sur certaines phrases. Les expériences poursuivies en n'appliquant plus la technique des bonifications montrent un allègement appréciable du traitement syntaxique.

Résultats du niveau Lexical numéro de phrase	1	2	3	4	5	6	7	8	9	10	11	12	13	Totaux
nombre de phonèmes	38	39	40	45	54	52	67	34	57	52	48	48	40	614
nombre de détections sans prosodie	91	112	102	174	183	174	234	98	170	186	177	151	163	2015
nombre de détections supprimables	52	59	51	105	104	92	149	61	105	104	105	100	107	1194
nombre de marques réelles	3	4	3	5	4	6	6	3	5	6	4	5	6	60
nb. de suppressions	11	23	9	52	18	45	39	22	24	28	32	51	62	416
nombre de marques idéales	7	7	7	8	9	9	10	5	9	9	7	10	7	104
nb. de suppressions	18	30	19	57	48	54	51	25	36	41	46	67	69	561
Résultats ~~du niveau~~ Syntaxique														
coeff. syntaxique sans prosodie	137	22	67	37	4900	54	-	7	-	369	-	-	-	
coeff. syntaxique avec pros. réelle et bonifications	335	22	96	31	2200	50	-	7	-	282	-	-	-	
coeff. syntaxique avec pros. idéale et bonifications	19	23	55	39	98	48	-	7	-	280	-	-	-	
coeff. syntaxique avec pros. réelle sans bonifications	52	22	64	31	1500	50	-	7	-	282	-	-	-	

TABLE 1 :Résultats. Les phrases marquées '-' ne sont pas reconnues par l'analyse syntaxique.

BIBLIOGRAPHIE.

(1) Compte rendu Final - Convention SESORI n° 74-80, "Deux Langages Parlés pour un Dialogue Intelligent entre l'homme et la machine", Rapport RP/CEI/CSI 17, RP/TMA/ETA/39, Novembre 1976, 112 pg.

(2) Vaissière J., Le Corre C.

Premiers essais de segmentation automatique de la parole continue en mots à partir des variations du fondamental dans la phrase, Recherches Acoustiques, Vol. III, 1976, p. 193 - 208.

(3) Mercier G., Segmentation de la parole et reconnaissance des syllabes à l'intérieur des mots - Automatisme tome XVII n° 3 - mars 1972 - p. 69 - 75.

(4) Gresser J-Y., Mercier G., Automatic sygmentation of speech into syllable and phonemic units. Application to french words an utterances.
Auditory - Analysis and perception of speech, (Fant and Tatham)
Academic Press, Londres, New-York, San Francisco 1975 p. 349 - 359.

(5) Vaissière J.

Further Note or French Prosody, Research Laboratory of Electronics, MIT, PR n° 115, 1975, PP 251 - 61.

(6) "Ils ne mouraient pas tous, mais tous étaient frappés".

Intonation et communication parlée

VIIIèmes JOURNEES D'ETUDE SUR LA PAROLE

Aix-en-Provence, 25-27 Mai 1977

ANALYSE ET REEDUCATION DES PARAMETRES PROSODIQUES CHEZ L'ENFANT SOURD

Marie-Christine HATON et Jean-Paul HATON

Centre de Recherche en Informatique de Nancy
Equipe de Traitement du Signal et Reconnaissance des Formes
C.O. 140 54037 - NANCY

Résumé :

Les spécialistes s'accordent à dire que les paramètres prosodiques sont les premiers paramètres à rééduquer chez l'enfant non-entendant, car ils jouent un rôle fondamental dans l'émission et la compréhension de la parole.

Cet article concerne les recherches menées en ce domaine dans le cadre du projet SIRENE. Dans une première partie nous présentons les méthodes d'étude des paramètres mélodie, intensité, rythme, ainsi que quelques résultats partiels obtenus à partir d'enregistrements d'enfants sourds. Dans une deuxième partie nous rappelons brièvement les algorithmes développés pour tenter la rééducation et la maitrise par l'enfant des paramètres prosodiques.

Summary :

Prosody plays a major role in speech emission and understanding therefore, prosodic parameters are among the first to be reeducated with deaf children.

This paper is concerned with research carried out in this area in our SIRENE project. In a first part we present a method for studying pitch, intensity and rhythm parameters. Some preliminary results are given. In a second part we briefly present algorithms which make possible the reeducation of these prosodic parameters.

ANALYSE ET REEDUCATION DES PARAMETRES PROSODIQUES CHEZ L'ENFANT SOURD

I. INTRODUCTION.

Très tôt dans la vie de l'enfant, et même avant qu'il ne prononce des mots, on peut distinguer les traits prosodiques caractéristiques de sa langue maternelle. Un fait important est que ces formes s'enracinent si profondément qu'il sera pour lui difficile de les modifier suffisamment pour parler une langue étrangère comme les sujets natifs. Bien sûr les qualités des voix d'individus parlant la même langue peuvent présenter de grandes variations. De plus, chaque individu présente une voix de couleur différente suivant son degré de fatigue, son humeur... Il reste que les paramètres prosodiques sont fondamentaux pour l'intelligibilité de la parole. Dans la définition de son concept de "clarté phonétique", J.C. Lafon (1) cite l'intensité et la durée comme les principaux facteurs influents en plus de la distortion et des effets de masquage du bruit.

Dans l'optique de la réalisation de notre système de rééducation vocale des enfants non-entendants (SIRENE (2)), nous avons développé un certain nombre de programmes d'étude des voix de ces enfants, en particulier de leurs traits prosodiques. Nous présentons ici quelques résultats obtenus à partir de corpus de données pour l'instant trop limités pour permettre une généralisation ou une analyse statistique mais choisis pour leur caractère significatif pour l'auditeur. Ces données concernent d'une part deux enfants entendants et d'autre part sept enfants mal-entendants enregistrés à l'I.N.J.S. de Paris (*).

II. DEFAUTS DE PROSODIE CHEZ L'ENFANT SOURD.

Il existe deux systèmes d'innervation du larynx, phonatoire et respiratoire, auxquels s'ajoute une innervation végétative qui peut parasiter les précédentes (encas d'émotion par exemple). Tous les problèmes rencontrés par l'enfant au niveau de la production du son glottal peuvent être attribués à un défaut de commande des muscles du larynx et/ou à un défaut de commande du système expiratoire.

La voix de l'enfant sourd présente une qualité très particulière, difficile à décrire ou à définir mais assez caractéristique. Divers qualificatifs lui sont appliqués suivant les auteurs = voix raide, plate, soufflée, rauque ou même gutturale, voix de fausset. La cause principale en est une mauvaise tension et un mauvais ajustement des cordes vocales. Le fondamental de la voix est anormalement élevé (c'est le cas le plus fréquent, le larynx est alors haut et trop contracté) ou au contraire trop bas, un manque de tonicité provoquant une détente excessive du larynx. La voix est peu expressive et l'intonation mal respectée dans la phrase.

La fréquence fondamentale de vibration des cordes vocales F_0 est influencée par l'intensité du flot d'air qui les traverse : les variations de hau-

(*) Enregistrements réalisés sous la responsabilité de M.R. DISSOUBRAY que nous tenons à remercier pour son aide et l'intérêt qu'il porte à notre étude.

teur suivent souvent les variations d'intensité. F_0 augmente par exemple pour les phonèmes qui demandent une énergie importante, au passage des consonnes ouvertes vers les consonnes fermées.

A l'attaque de la voix, on observe certains défauts par manque de coordination entre la pression sous-glottique et la mise en vibration des cordes vocales. Une pression exagérée provoque un coup de glotte, à l'émission de /k/ ou /g/ par exemple. Une fermeture insuffisante de la glotte crée l'apparition d'une zone de bruit avant le son voisé... L'intensité augmente souvent en fin de phrase ou pour des sons difficiles à prononcer. Les articles sont faibles ou même inaudibles.

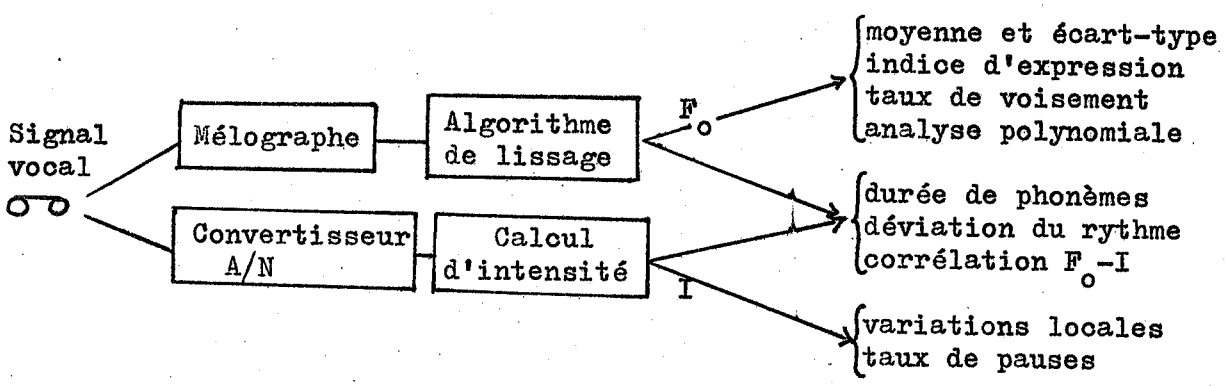
Le rythme et la durée des segments de parole sont affectés par un défaut de coordination respiratoire. La phrase est souvent attaquée rapidement puis les pauses deviennent plus fréquentes et plus longues. Ou bien la parole est prolongée jusqu'à épuisement de l'air. De façon générale, les cordes vocales laissent passer un flot d'air trop important pendant les sons voisés, ce qui réclame deux fois plus d'inspirations.

III. ETUDE QUANTITATIVE DES VOIX.

Nous avons constitué par chacun des enfants trois fichiers contenant un signal temporel de 3,4 s échantillonné à 8kHz, le contour mélodique et le contour d'intensité en synchronisation.

1. Fondamental de la voix et intonation.

La fréquence fondamentale F_0 est obtenue tous les 75èmes de seconde grâce à un Mélographe CNET/ETA. Un algorithmede lissage permet d'éliminer automatiquement les valeurs parasites comme celles dues à la non-détection d'une impulsion glottale dans un segment voisé. Des programmes de calcul permettent ensuite de déterminer la valeur moyenne $\overline{F_0}$ et l'écart-type σ_{F_0}, un indice d'expression que nous définissons comme $\frac{F_0 \text{ max.} - F_0 \text{ min.}}{\overline{F_0}}$, le taux de voisement (rapport des durées de portions voisées à la durée de parole, pauses exclues), le coefficient b caractéristique de la projection du contour de F_0 sur une base de polynômes de Tchebycheff (cf (3)). Le schéma ci-dessous résume ces possibilités.



Le tableau I donne un exemple de résultats obtenus pour deux enfants non-entendants S_1 et S_2 et deux enfants entendants E_1 et E_2. On peut remarquer l'intérêt de l'analyse polynomiale jointe au calcul de σ_{F_0} = en effet σ_{F_0} dépend de l'indice d'expression comme des variations locales rapides alors que notre coefficient b chiffre le caractère erratique du contour mélodique. On a ajouté une colonne d'avis subjectif sur la voix du sujet.

2. Intensité.

Le signal vocal est échantillonnéici à 8 kHz grâce à un convertisseur analogique-numérique. On peut approcher la valeur de l'intensité I par l'enveloppe du signal temporel ou la calculer en intégrant l'amplitude carrée du signal sur un intervalle de temps fixe, ce que nous faisons ici. Nous avons choisi une fenêtre temporelle de 8ms avec déplacement toutes les 4ms (la largeur de la fenêtre correspond ainsi à deux périodes d'un fondamental de 250 Hz).

La figure 1 montre l'enveloppe du signal vocal, l'intensité et le contour mélodique pour un enregistrement du sujet S2.

Le calcul de I permet d'étudier le taux des pauses-rapport des portions où l'intensité est inférieure à un seuil préfixé à la durée totale (cf tableau I), et les variations locales d'intensité. Un programme étudié au départ pour la rééducation vocale permet de mettre en évidence la corrélation entre F_0 et I dans la production de sons qui demandent un effort particulier.

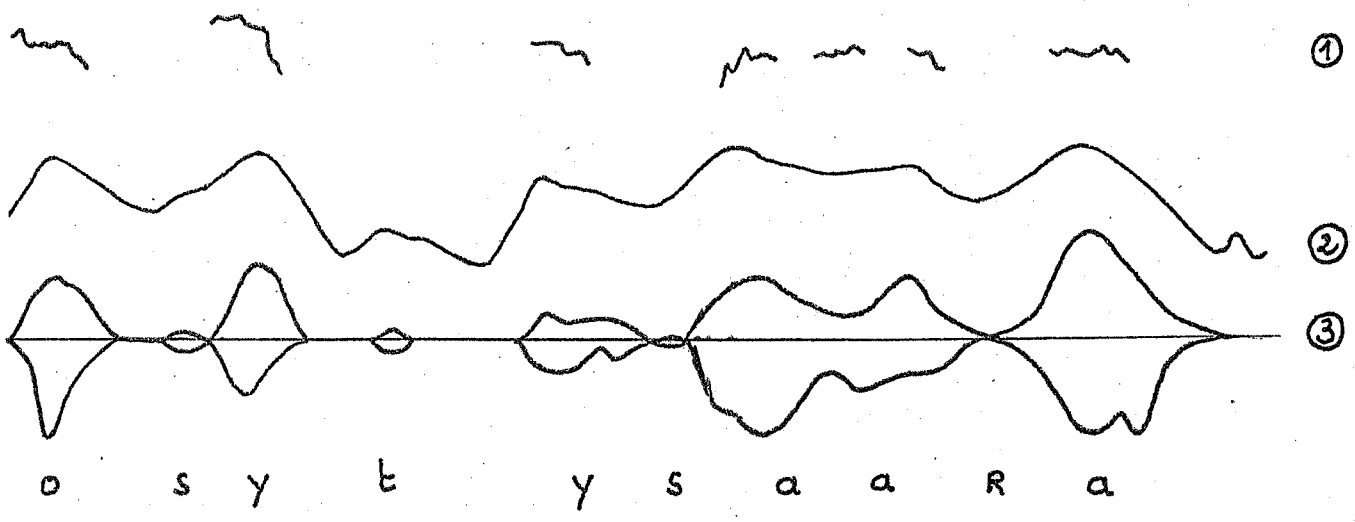


Figure 1.

"Au Sud du Sahara" prononcé par le sujet S2.
1 - Contour mélodique 2 - Contour d'intensité 3 - Enveloppe du signal

	$\overline{F_o}$ Hz	σ_{F_o} Hz	an. pol. coeff. b	indice d' expression	taux de voisement	taux de pauses	qualité de la voix (subjectif)
S1	247	15	2.6	0.11	47 °/₀	20 °/₀	monotone, peu compréhensible
S2	269	46	2.0	0.27	48	25	hachée, meilleure intelligibilité que S1
E1	260	35	1.5	0.42	79	5	bonne
E2	263	33	1.6	0.47	85	8	bonne

Tableau I

3. Rythme et durée.

L'étude des déviations du rythme se fait grâce à la comparaison des courbes d'intensité et de mélodie à des courbes moyennes d'enfants entendants. Nous pensons dans un deuxième temps adapter un programme de segmentation du signal pour l'étude des durées respectives des différents phonèmes (en plus de l'étude des problèmes de substitution, d'insertion et de suppression de phonèmes qui demandent l'intervention d'une procédure de classification de phonèmes).

4. Conclusion.

Les programmes cités auparavant présentent pour nous un triple intérêt :

- tout d'abord pour une étude plus systématisée des voix des enfants sourds et la recherche de méthodes d'évaluation subjective de la qualité de ces voix. L'étude faite sur une population plus importante doit permettre de juger de la signification des différences par rapport aux voix non pathologiques, grâce au test de Student en particulier. On trouve dans la thèse de G. Voiron (4) dirigée par G. Roinsol une étude de ce type faite grâce à un gros travail de compilation manuelle, travail que nous tentons d'automatiser complètement. Nous pensons de plus rechercher les corrélations possibles entre les paramètres physiques dont il est question ici, les paramètres physiologiques (sexe, âge, perte auditive) et des paramètres subjectifs d'évaluation de la parole (intelligibilité, âge équivalent pour le développement du langage...),

- ensuite pour l'étude des voix en cours de rééducation vocale en vue de la constitution d'un fichier de paramètres pour chacun des sujets,

- enfin pour une adaptation audéveloppement de procédures de rééducation des paramètres prosodiques.

IV. REEDUCATION DES PARAMETRES PROSODIQUES.

1. En ce qui concerne la perception de la parole, une meilleure

compréhension du discours continu serait obtenue, selon R. Dissoubray, grâce à l'indication du voisement et l'indication de durée pour les colonnes, en complément de la lecture labiale.

2. Pour ce qui est de la rééducation vocale, un système géré par ordinateur présente de grandes facilités, tant dans la souplesse d'utilisation que dans les réglages (détermination de seuils, etc) et dans la possibilité d'autoapprentissage pour l'élève.

C'est ainsi que nous développons des jeux visuels dans lesquels l'enfant doit par exemple maintenir son fondamental dans une zone préfixée par le rééducateur d'après la valeur qui lui convient le mieux du point de vue esthétique et d'après ses possibilités, faire passer F_0 d'un niveau bas à un niveau plus élevé ou inversement, tenter de reproduire, dans la phrase cette fois, un contour intonatif type, etc...

La visualisation de la courbe temporelle d'intensité permet de mettre en évidence certains caractères comme le coup de glotte ou la voix "soufflée". Il est connu que l'enfant a dans les premiers essais tendance à élever le niveau d'intensité quand on lui demande de monter vers les aigus. Nous visualisons d'une part, F_0 par rapport à I (dans une échelle logarithmique), d'autre part, les variations de l'un par rapport à celles de l'autre dans des coordonnées calculées pour mettre le mieux en évidence les différences par rapport à la voix normale. Le deuxième graphe est utilisé pour les essais de stabilité des cordes vocales, le premier pour les essais à F_0 ou I constant, l'autre paramètre pouvant évoluer.

Le rythme de la phrase et le débit sont évalués et peuvent être comparé à des schémas types de façon à conduire vers une parole de meilleure intelligibilité.

V. CONCLUSION.

Les paramètres prosodiques essentiels : mélodie, intensité, rythme comptent parmi les plus mal assimilés par les mal-entendants et les plus difficiles à rééduquer.

Nous avons présenté dans cet article des algorithmes permettant une étude quantitative automatique de ces paramètres. Ces algorithmes ont été appliqués à des voix d'enfants sourds, en nombre encore insuffisant pour pouvoir en tirer des conclusions définitives. Ces buts que nous poursuivons sont d'une part une systématisation de l'étude des voix pathologiques et d'autre part pour un suivi de l'élève en cours de rééducation et une adaptation des procédures de rééducation.

Dans une dernière partie nous esquissons simplement le travail mené dans la conception de ces procédures de rééducation assistée par ordinateur. Ces dernières utilisent essentiellement le sens visuel de l'élève et mettent en oeuvre, entre autres, des techniques de reconnaissance automatique de la parole.

REFERENCES.

(1) J.C. LAFON "Auditory basis of phonetics"
in Manual of phonetics, B. MALMBERG ed.

(2) M.C. HATON, J.P. HATON "SIRENE : un projet de système interactif pour la rééducation vocale des enfants non-entendants"
Rev. Gale Ens. Déficients Auditifs, pp. 203-209, 67, n°4,1975.

(3) M.C. HATON, J.P. HATON "Essai de classification des voix d'enfants sourds par analyse polynomiale de la mélodie" 6èmesjournées d'Etude sur la Parole, Toulouse, 1975.

(4) G. VOIRON "Contribution à l'étude objective de la voix des déficients auditifs moyens et profonds avec l'utilisation d'un indicateur de mélodie". Thèse INJS, Paris, 1974.

VIII èmes JOURNEES D'ETUDE SUR LA PAROLE

Aix - en - Provence, 25 - 27 Mai 1977

DE L'ANALYSE PHONOLOGIQUE A L'ANALYSE PHONETIQUE DES ENONCIATIVES EN FRANCAIS

L. THERRIEN-ROY, D. JACQUEMIN & L.J. BOË
Institut de Phonétique de Grenoble

• Résumé :

Cette étude des contours intonatifs de la phrase énonciative française a été menée à partir de la théorie de l'intonation proposée par Ph. MARTIN. Les auteurs tentent de confronter cette théorie à des réalisations et étudient particulièrement l'existence de contrastes entre les contours montants et descendants qui marquent le découpage, par le locuteur, en unités de signification minimales. Quelques types de contours sont précisés de façon quantitative dans leur amplitude, durée et position par rapport à la fréquence laryngienne moyenne de la phrase.

• Summary :

This study of the pitch pattern of the declarative sentence is conducted on the basis of a theory of intonation put forward by Ph. MARTIN. The authors attempt to apply the above mentionned theory to actual utterances, and discuss particularly the existence of contrasts between rising and falling patterns, used for dividing the sentence into minimal sense units by the speaker himself. Some patterns are quantitatively described mainly as to their amplitude, duration and position in relation to the mean laryngeal frequency.

DE L'ANALYSE PHONOLOGIQUE A L'ANALYSE PHONETIQUE DES ENONCIATIVES EN FRANCAIS

L. THERRIEN-ROY, D. JACQUEMIN, L.J. BOË

I - INTRODUCTION

Cette étude s'inscrit dans le cadre d'une analyse systématique des réalisations de la phrase énonciative française; elle est menée à partir de la théorie de l'intonation proposée par MARTIN[1]. Sa "grammaire" des contours intonatifs est décrite en un système d'oppositions binaires, appliqué à des unités minimales de signification. La délimitation résulte du découpage, de la structure de surface, par le locuteur lui-même. Chaque contour ne s'applique qu'à la syllabe accentuée (en général, la dernière) de l'unité de signification. La succession des contours d'une phrase énonciative est ainsi fixée à partir de 5 contours types : C_1, C_2, C_3, C_4 et C_d fonctionnant par contraste (tableau I).

	C_1	C_2	C_3	C_4	C_d
					
montant	+	-	+	-	-
ample	+	+	-	-	-
long	-	-	-	-	+
extrême	-	-	-	-	+

Tableau I

Les traits *long*, *montant* et *ample* permettent de préciser la pente du contour, tandis que le trait *extrême* situe son milieu par rapport à la fréquence laryngienne moyenne $\overline{F_0}$ de la phrase; une mesure systématique de $\overline{F_0}$ a été décrite par ailleurs[2].

Nous nous proposons :
- de vérifier le principe du fonctionnement du système des contours surtout en ce qui concerne l'opposition +/- *montant*
- de spécifier quantitativement l'*amplitude* et la *longueur* de certains contours et de les situer par rapport à $\overline{F_0}$.

II - PROCEDURE

Nous nous sommes servis d'un corpus qui avait déjà été enregistré à l'occasion d'une étude précédente sur la phrase énonciative[3]. Il est composé de 45 phrases à expansions successives portant sur les syntagmes nominal et verbal. Nous avons retenu 8 locuteurs (5 femmes et 3 hommes), la lecture des

360 tracés sur une échelle logarithmique de la F_ℓ[4] a été faite par référence aux différentes possibilités prévues par la théorie. Les mesures de F_ℓ tiennent compte des caractéristiques intrinsèques des consonnes voisées[5]; en ce qui concerne les sourdes, l'hypothèse d'une interpolation fictive du tracé de F_ℓ semble justifiée. Les résultats sont donnés en tons et centisecondes. Dans un premier temps ont été situés les contrastes résultant du premier niveau de découpage de la phrase; pour des niveaux ultérieurs nous avons considéré plus particulièrement les contours *+ montant* quant à leur réalisation effective et leur rapport de contraste.

III - RESULTATS ET DISCUSSION

1. Contours montants et de fin de phrase.

En ce qui concerne les contours montants et finals, tous les tracés du corpus présentent au moins un contraste entre une courbe montante de grande amplitude, qui correspond à la valeur maximale de F_ℓ pour la phrase et le contour descendant final. En outre, 31 des 45 phrases sont réalisées par au moins deux locuteurs avec plus d'un maximum nettement marqué. Une deuxième catégorie de contours montants, de moindre amplitude, apparaît dans 89 % des 60 tracés qui étaient susceptibles de présenter des C_3. Le tableau II donne les résultats moyens qui ont servi à définir quantitativement les trois types de contours que nous pouvons assimiler à la réalisation de C_1, C_3 et C_d

	C_1 ⟋	C_3 ⟋	C_d ⟍
amplitude (tons)	3	1,8	2,2
longueur (cs)	23	13	27
position du milieu du contour par rapport à F_ℓ (tons)	+ 1,6	+ 0,6	- 1,8

Tableau II

Dans le système C_d s'oppose à C_1 par les traits *extrême*, *long* et *ample*. Dans les réalisations que nous observons les tracés correspondants sont situés à peu près symétriquement de part et d'autre de F_ℓ (- 1,8 et 1,6 ton). Si la réalisation de C_d est *+ extrême* il semble bien que celle de C_1 le soit aussi.

Les mesures sur la dutée ne doivent pas être considérées comme absolument significatives, compte tenu de la variété des syllabes qu'elles recouvrent, mais il semble aussi que C_1 et C_d soient réalisés tous les deux *+ long*.

Par contre le trait *ample* semble conforme aux prévisions et par comparaison avec les valeurs des deux contours précédents C_3 est bien *- extrême*, *- long* et *- ample*.

Par ailleurs, il nous faut signaler que les contours *+ montant, + ample* ne recouvrent pas obligatoirement le thème par opposition au propos, mais procèdent d'un découpage en grandes unités qui peuvent se situer à plusieurs endroits de la phrase.

2. Courbes descendantes

En appliquant toutes les possibilités de découpage prévues par la théorie, 25 des 45 phrases du corpus sont susceptibles de présenter des contours *- montant* (C_2 ou C_4), soit un total de 200 pour les 8 locuteurs. Or nous n'avons observé que 24 tracés descendants soit 12%. La répartition par locuteur et par phrase est la suivante : 8 cas isolés, 2 cas × 2 locuteurs et 2 cas × 6 locuteurs. De tous ces tracés, 19 sont en position de contraste avec ceux du niveau de découpage supérieur et donc conformes à l'analyse de MARTIN. Les 5 autres se situent sur ce qui est, théoriquement, la dernière unité de signification minimale précédant le contour final; mais, à notre avis, on ne peut l'en extraire par contraste, la pente restant pratiquement constante jusqu'à la fin de la phrase. Ces 5 cas font partie des 8 réalisations isolées.

Les 19 contours descendants, confirmant la théorie, portent, en grande partie, sur deux types d'unités illustrés par les deux phrases suivantes avec les tracés que nous avons obtenus :

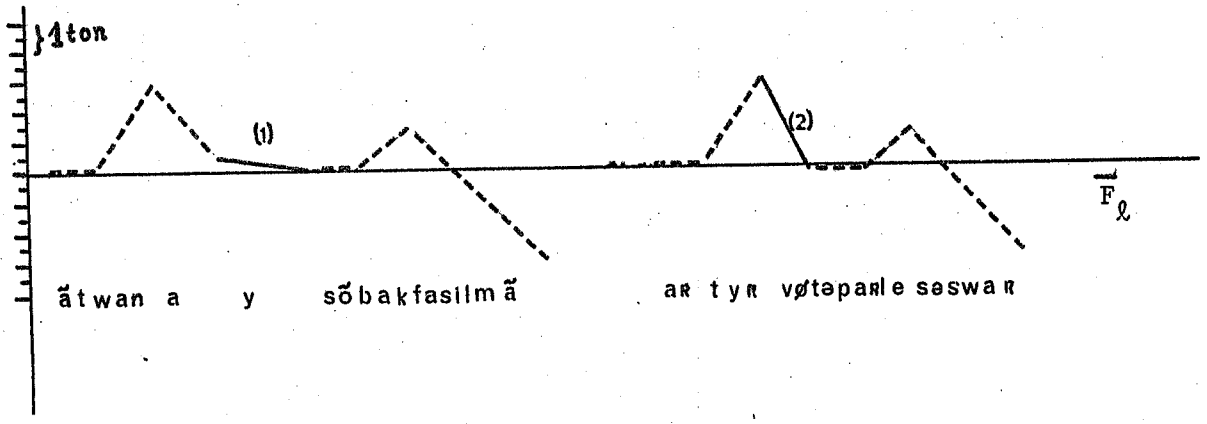


Antoine (a eu son bac) facilement.



Arthur (veut te parler) ce soir.

A partir des valeurs moyennes, nous avons établi leur position par rapport à $\overline{F_\ell}$ (figure 1).



En fait ces 19 contours *- montant* ne confirment que dans une proportion de 11% la théorie des contrastes entre unités de signification de niveaux différents, les réalisations de ces mêmes contours étant plates ou montantes respectivement dans 26% et 63% des autres cas. Ces résultats, très défavorables

à la théorie sur ce point, ont été à la base d'une comparaison effectuée entre le type d'unité segmentable dans lequel ces contours descendants s'inscrivent et d'autres unités non segmentables sur le plan de la signification; par exemple :

Annie, (la veinarde), (a gagné) le gros lot

(réalisé sans pause, comme d'ailleurs tous les autres exemples cités).

Cette analyse montre que, sur la syllabe [la] F_0 décroît de 3,5 tons, [vɛ] se situe à -0,5 ton par rapport à F_0 et [naRd] a un tracé montant de 2,5 tons. Les valeurs sont similaires pour [a] dont le chute est de 2,5 tons [ga] évoluant dans une bande de 1 ton autour de F_0 et [ne] remontant jusqu'à +1,5 ton. On observe donc, pour ces unités non segmentables la même évolution de F_0 que pour les unités segmentables dans lesquelles s'inscrivent les contours descendants dont nous venons de parler (figure 1).

Suite à ces constatations, nous croyons qu'il est justifié de mettre en doute l'existence de contraste entre contours *+ montant / - montant* à des niveaux de découpage inférieurs au premier. Le découpage auquel procède MARTIN, en ce qui concerne les contours *- montant* , détermine des unités qui n'apparaissent pas au niveau de la réalisation. Ce qui est proposé comme C_2 ou C_4 nous semble être une évolution de F_0 non significative (dans le cadre de la théorie) correspondant à la liaison entre deux contours *+ montant* . Cette remarque a aussi été faite par EMERARD .
D'ailleurs cette interprétation des tracés semble être confirmée : à partir du tableau II on peut voir qu'il existe un intervalle de 0,7 ton sous F_0 qui constitue la zone de retombée et / ou d'évolution de F_0 entre les contours correspondant aux unités de signification. En outre, le grand pourcentage de tracés montants obtenu tend à montrer que, lorsqu'il y a découpage de la part du locuteur, celui-ci se traduit par un contour montant.

IV - CONCLUSION

Le mérite de la théorie de MARTIN est d'avoir proposé des unités pour lesquelles il est possible de prévoir un contour intonatif, à partir d'un découpage appliqué à la structure de surface : c'est une méthode opérationnelle.

. L'analyse des réalisations confirme, telle que prévue, l'existence de contours *+ montant* et *+ ample* conséquentes à un premier niveau de découpage contrastant avec le contour final *- montant* et *- ample*. Le fait que plusieurs de ces contours *+ montant* peuvent apparaître dans une même phrase montre bien qu'ils ne reflètent pas seulement le cas de l'opposition reconnue entre thème et propos mais qu'ils résultent d'un découpage, en grandes unités de signification, opéré par le locuteur. Contrairement aux prévisions, ces contours C_1 et C_d sont tous les deux *+ long* et *+ extrême*.

. On relève également sur le tracé des courbes croissantes d'amplitude moindre que les précédentes que MARTIN étiquette C_3 et qui sont bien *- ample*, *- long* et *- extrême*.

. Cependant en ce qui concerne les niveaux de découpage inférieur, il semble que l'on puisse mettre en doute l'existence de contour - *montant* (C_2 ou C_4) contrastant avec les contours + *montant*. Leur rareté ainsi que les types de tracés dans lesquels ils s'insèrent laissent plutôt croire qu'ils sont la suite d'un contour + *montant* suivi d'une chute, d'un retour en-dessous ou aux environs de F_ℓ et d'une nouvelle montée sur la dernière syllabe accentuée de l'unité de signification suivante.

A notre avis, lorsque dans l'esprit du locuteur, unité de signification il y a, c'est un contour + *montant* qui en résulte comme en fait foi l'analyse du corpus.

REFERENCES

[1]. MARTIN, Ph. (1975). Analyse phonologique de la phrase française. Linguistics 146, pp. 35-67.

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

[3]. CONTINI, M., BOË, L.J. (1973). Contribution à l'étude quantitative de l'évolution de la fréquence laryngienne dans la phrase énonciative en français. Bulletin de l'Institut de Phonétique de Grenoble 2, pp. 77-92.

[4]. BOË, L.J. (1976). Etude des vibrations des cordes vocales dans la parole : méthodes, résultats et applications. Revue d'Acoustique 37, pp. 105-112.

[5]. LARREUR, D., BOË, L.J. (1973). Les caractéristiques intrinsèques des consonnes voisées du français dans la parole continue. Bulletin de l'Institut de Phonétique de Grenoble 3, pp. 25-29.

[6]. EMERARD, F. (1977). Synthèse par diphones et traitement de la prosodie. Thèse de 3° Cycle, Linguistique, Grenoble III.

BIBLIOTHÈQUE
Institut de Phonétique
AIX-EN-PROVENCE

Les 8èmes Journées d'Etude sur la Parole ont été organisées, sous l'égide du G.A.L.F., avec le concours du C.N.R.S., à l'Université de Provence (AIX-MARSEILLE I), par l'Institut de Phonétique (L.A. au C.N.R.S., n° 261), du 25 au 27 mai 1977.

Les thèmes de travail de ces Journées étaient les suivants :

I. La fréquence fondamentale : études physio-acoustiques et instrumentales.

Président : R. CARRÉ
Rapporteurs : L.J. BOE, M. CARTIER

- Variations de la périodicité
- Régulation de la vibration des C.V.
- La détection de Fo et du voisement.

II. La mélodie : études perceptives et synthèse de la parole.

Président : M. ROSSI
Rapporteurs : J. GENIN, M. WAJSKOP

- La perception de la hauteur et de ses variations (micromélodie)
- Corrélations entre l'intensité, la durée et la fréquence fondamentale
- Rôle des traits prosodiques dans la transmission et la synthèse de la parole.

III. L'intonation : aspects linguistiques et reconnaissance de la parole.

Président : G. FAURE
Rapporteurs : A. DI CRISTO, J.S. LIENARD, Ph. MARTIN

- Intonation et syntaxe
- Rôle de l'intonation dans la segmentation de l'énoncé et la reconnaissance de la parole
- Intonation et communication parlée.

On trouvera, dans ce volume, les textes des communications retenues par le Bureau du Groupe Communication Parlée.

Afin d'éviter les inconvénients trop connus des congrès et de donner à ce colloque le caractère de journées de travail et de réflexion sur un thème bien défini, le Bureau du Groupe a prévu une organisation originale des séances de travail. Il a désigné des rapporteurs qui ont été chargés :

1) de faire un exposé de synthèse sur la base des communications proposées,

2) de susciter la discussion en demandant aux auteurs des éclaircissements ou des compléments d'information, et en critiquant les thèses ou les méthodes développées dans les communications.

Les auteurs ont été conviés, sans avoir à répéter le texte publié dans les Actes, à compléter, illustrer ou justifier leur méthode ou leurs thèses.

La dernière demi-journée a été consacrée aux tables rondes :

1) Fonctionnement des cordes vocales,
2) Modèles linguistiques de l'intonation,
3) La terminologie des études prosodiques peut-elle être unifiée ?

Les textes des tables rondes et des discussions sont publiés dans le volume II des Actes des Journées.

Professeur Mario ROSSI
Directeur de
L'Institut de Phonétique