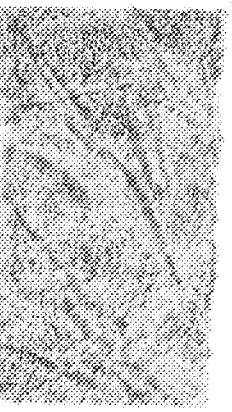
GROUPEMENT DES ACOUSTICIENS DE LANGUE FRANÇAISE

GROUPE

COMMUNICATION PARLEE

3e JOURNEES D'ETUDES

SUR LA PAROLE



GROUPEMENT DES ACOUSTICIENS DE LANGUE FRANCAISE

GROUPE "COMMUNICATION PARLEE"

COMPTE RENDU DES JOURNEES D'ETUDES SUR LA PAROLE

ORGANISEES LES 31 MAI, 1er ET 2 JUIN 1972

AU C.N.E.T. A LANNION

GROUPEMENT DES ACOUSTICIENS DE LANGUE FRANÇAISE

GROUPE

COMMUNICATION PARLEE

JOURNEES D'ETUDES

SUR LA PAROLE

PROGRAMME DES JOURNEES D'ETUDES

Mercredi 31 Mai 1972

Matinée :

Après-midi :

Jeudi 1er Juin 1972

Matinée :

Après-midi :

Vendredi 2 Juin 1972

Matinée :

Après-midi :

A L L O C U T I O N D E B I E N V E N U E

prononcée par Monsieur E. JULIER
Directeur du C.N.E.T. - LANNION

C'est un réel plaisir pour moi d'offrir l'hospitalité du C.N.E.T. - LANNION aux Troisièmes Journées d'Etudes sur la Parole, organisées par le Groupe "Communication Parlée" du GALF, en collaboration avec l'A.F.C.E.T. et la Section Bretagne de la S.E.E.

Les Journées d'Etudes organisées à GRENOBLE et à AIX EN PROVENCE avaient été, au dire des participants, couronnées d'un franc succès. Je souhaite donc que cette troisième manifestation soit digne des précédentes et soit l'annonce d'autres colloques aussi fructueux dans les années à venir.

Parmi les 116 inscrits à ces journées d'études, j'aimerais saluer tout particulièrement :

- Monsieur PIMONOW, Président du GALF et les participants venant des pays étrangers lointains.

- Monsieur LINDBLOM, de Stockolm,

- Monsieur SANTERRE, de Montréal,

et aussi de nombreux voisins immédiats de Belgique, d'Italie et d'Espagne.

Je souhaite enfin la bienvenue aux représentants de nombreuses universités françaises, de nos grandes écoles et de divers organismes publics et privés.

Une telle diversité d'origine parmi les participants à ces journées d'études prouve toute l'importance que les laboratoires attachent, à l'heure actuelle, aux problèmes liés à la parole. Au moment où les Etats-Unis d'Amérique semblent décidés à intensifier les programmes

de recherche, en particulier en reconnaissance automatique de la parole, il est réconfortant de constater que l'Europe est tout aussi intéressée par ce sujet.

Le Centre National d'Etudes des Télécommunications, pour sa part, et dans la mesure de ses moyens, essaie de faire progresser les connaissances dans le domaine de l'analyse et de la synthèse de la parole. La parole est en effet, la principale matière première que les Télécommunications doivent transmettre ; tout progrès dans la connaissance des caractéristiques du signal de parole peut être utile pour l'étude des systèmes de transmission ; tout progrès en reconnaissance et en synthèse de la parole permettra d'envisager la construction de centres de renseignements téléphoniques automatiques. J'espère que les visites des laboratoires prévues au cours des jours qui viennent vous permettront de vous faire une juste idée des travaux déjà accomplis.

Je souhaite que tous, ici, vous tirerez de réels bénéfices de ces trois jours d'échanges scientifiques à Lannion. Laissez-moi donc, en terminant, vous souhaiter sincèrement bon courage !

MODELES ARTICULATOIRES

APPROACHES TO ARTICULATORY MODELING

Björn E.F. Lindblom and J. Sundberg

Royal Institute of Technology (K.T.H.)
Département of speech communication
STOCKHOLM

Approaches to articulatory modeling

Table of contents

APPROACHES TO ARTICULATORY MODELING*

Björn E. F. Lindblom, Department of Phonetics, Stockholm University, Fack, 104 05 Stockholm 50 and Department of Speech Communication, Royal Institute of Technology (KTH), 100 44 Stockholm 70, Sweden
and
Johan Sundberg, Department of Speech Communication, Royal Institute of Technology (KTH), 100 44 Stockholm 70, Sweden

Abstract

Since human speech production is a highly complex and only partly understood process present attempts to model this process must by necessity involve many simplifications and approximations. The question thus arises how simplifications should be made and what criteria of evaluation should be used. In approaching the task of articulatory modeling an investigator may restrict his interests in a number of ways depending upon the short-term or long-term character of his objectives. Is his goal to construct a model primarily for the synthesis of an acoustic speech signal? If so, is it his intention to produce hi-fi, highly natural speech or speech of high intelligibility but with a strong computer-accent? How far "upstream" in the direction of the brain should the modeling proceed? Should the model accommodate the facts of normal speech as well as those of pathological or deviant articulatory behaviors (stuttering, Parkinson speech etc.) or those of non-speech uses of the speech organs such as singing, matrication, and deglutition? Should the model be complete enough to provide a theoretical phonetic basis for the elucidation of the facts of language structure?

---

* paper presented during "Journées d'études sur la parole" organized by G.A.L.F. at Lannion, France, June 1972.

Approaches to articulatory modeling

The purpose of the present paper is to illustrate some of the problems that arise in current work on articulatory modeling and to discuss solutions to these problems proposed by various researchers. A major part of this presentation will depart from Lindblom and Sundberg: "Acoustical Consequences of Lip, Tongue, Jaw and Larynx Movement", J. Acoust. Soc. Am. 50 (1971), pp. 1166-1179. Since the appearance of this article rules have been formulated describing the movements of the tongue tip. The acoustic correlates of such movements have been studied. These developments have made possible some further applications of the model. For instance, a new interpretation of consonant-vowel coarticulation will be suggested.

Although the model presented represents a gross simplification of the speech production process it will be demonstrated that it has many features useful for explaining a number of phenomena associated with speech and language structure. We conclude that it appears possible to find a balance between faithfulness to detail and the degree of insight that a model may offer. Consequently here as in other fields of science, it is the explanatory power of a model that should serve as the primary guideline in determining how it should be developed and evaluated.

Approaches to articulatory modeling

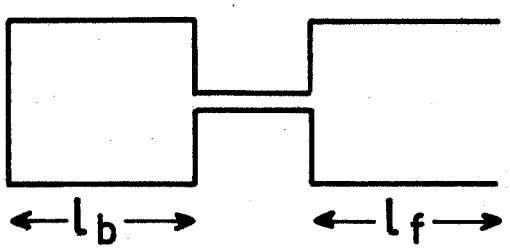
1. Introductory remarks

The purpose of this paper is to review the field of articulatory modeling. This review will show that during the last two decades there has emerged a tendency for investigators of speech production to carry their analysis further and further upstream, that is, further and further towards the centers in the brain from which the neural activity associated with speech emanates. This development can be examined from a practical point of view. We can ask ourselves: What methods are available to-day for such upstream exploration of the speech processes? How far upstream is it possible to go? Or, the tendency can be viewed as a matter of research strategy: How far upstream is it necessary to go? It is primarily on this latter question that our presentation will bear and we shall return to it when making our final conclusions.

2. From vocal-tract shape to sound

2.1 Some basic formant-cavity relations

When teaching students of phonetics the elements of acoustic phonetics it may sometimes be convenient to resort to a simplified description of the relation between formants and cavities for certain classes of speech sounds. The shape of the vocal tract may then be approximated as follows:



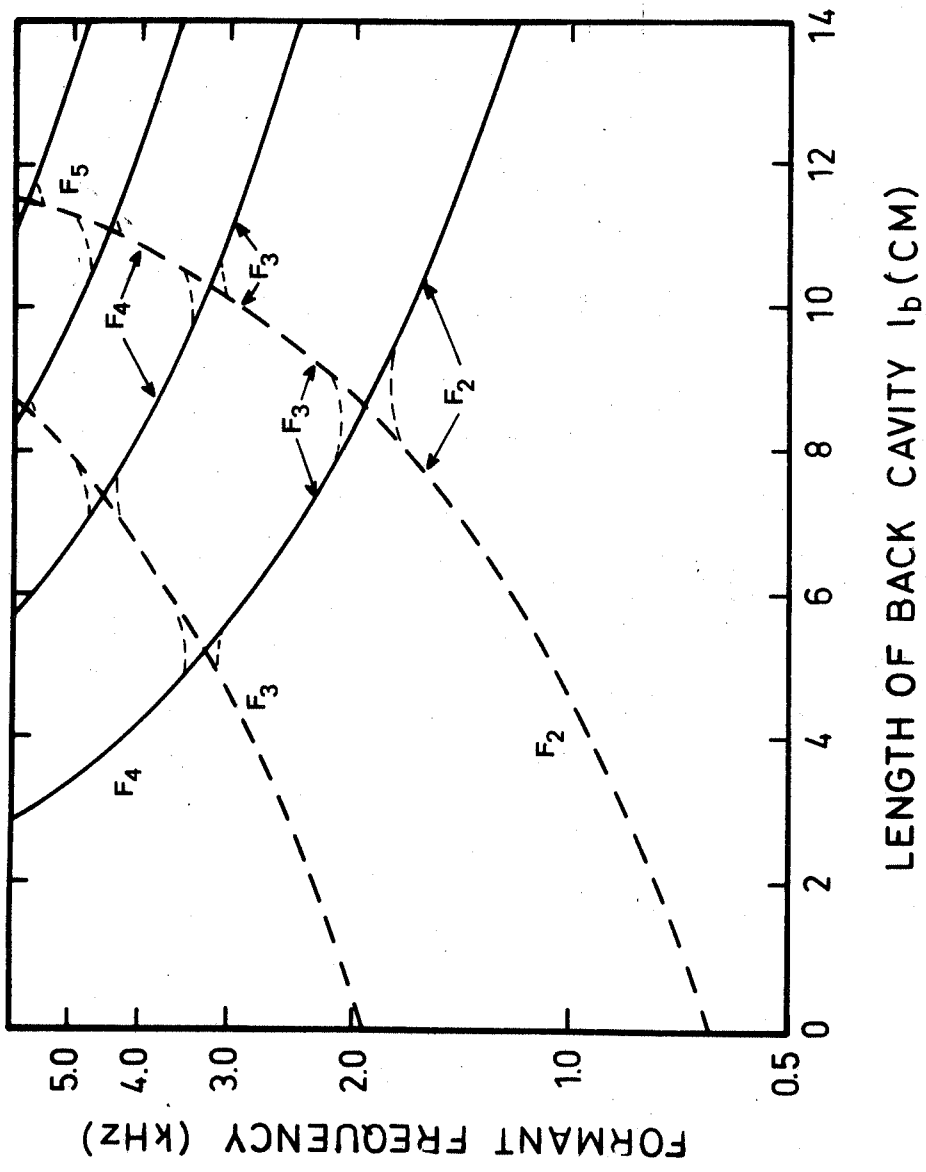

FORMANT FREQUENCY (kHz)
5.0
4.0
3.0
2.0
1.0
0.5
LENGTH OF BACK CAVITY l_b(CM)
0
2
4
6
8
10
12
14
F_5
F_4
F_3
F_3
F_2
F_4
F_3
F_2

Fig. 1

Approaches to articulatory modeling

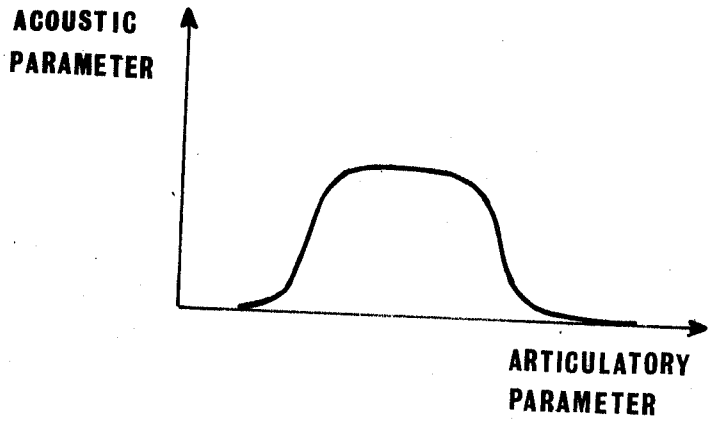
We assume that the vocal tract is a cylindrical uniform tube of a certain length. By introducing a constriction at some point along the tube we obtain a back cavity and a front cavity. Suppose furthermore that the cross-sectional area of this constriction is so small that we can regard the coupling between the two cavities as negligible. Under such conditions acoustic theory states that the front cavity will exhibit resonances at frequencies corresponding to odd multiples of $\frac{c}{4l_f}$, where c = speed of sound and l_f = length of front cavity, and that the back cavity will resonate at multiples of $\frac{c}{2l_b}$, where l_b = length of back cavity. Plotting resonance or formant frequencies as a function of the length of the back cavity we obtain the diagram shown in Fig. 1.

Since we have assumed that a radical constriction is present the first formant will always be close to zero for this idealized case. Thus the lowest formant seen in this nomogram is the second formant. It can be seen to be dependent on the longer cavity of the vocal tract; when the constriction is close to the glottal end it is a front-cavity resonance; when it is close to the labial end it is a back-cavity resonance.

2.2 "The quantal nature of speech"

It is of course true that the idealization of the vocal tract discussed here represents a form of articulatory modeling which can be justified only with regard to certain limited pedagogical purposes. Nevertheless nomograms of the sort shown on Fig. 1 were recently used by K.N. Stevens (1968) when he first presented his views on the quantal nature of speech (cf. also 1967). Basically the hypothesis of Stevens can be illustrated using a diagram of the following sort:

Approaches to articulatory modeling



The point is here that a change in an articulatory parameter may or may not lead to similar changes of acoustic attributes. As a result there are plateau-like areas where articulatory imprecision has very little acoustic effect. Stevens suggests that languages tend to seek out such regions and locate speech sounds at such plateaus. He argues that the nomogram of the first figure provides evidence in favor of his hypothesis. Here the plateaus are located at the points of intersection between formants. In view of inevitable coupling between cavities formants never coincide but curve off smoothly as indicated by the dashed lines. In the present case it is of interest to note that the proximity regions are found by Stevens to correspond closely to the terms for place of articulation that phoneticians normally use in describing consonants such as /kg/ which would be characterized by a coincidence of F_3 and F_4 would correspond to "the alveolar consonants /š ž č ǰ/ and the retroflex consonants such as /ḍ ṣ / which occur in many languages. The most anterior of the three points of constriction is where F_4 and F_5 coincide, and would, perhaps, correspond to the point of articulation for /s z t d/."

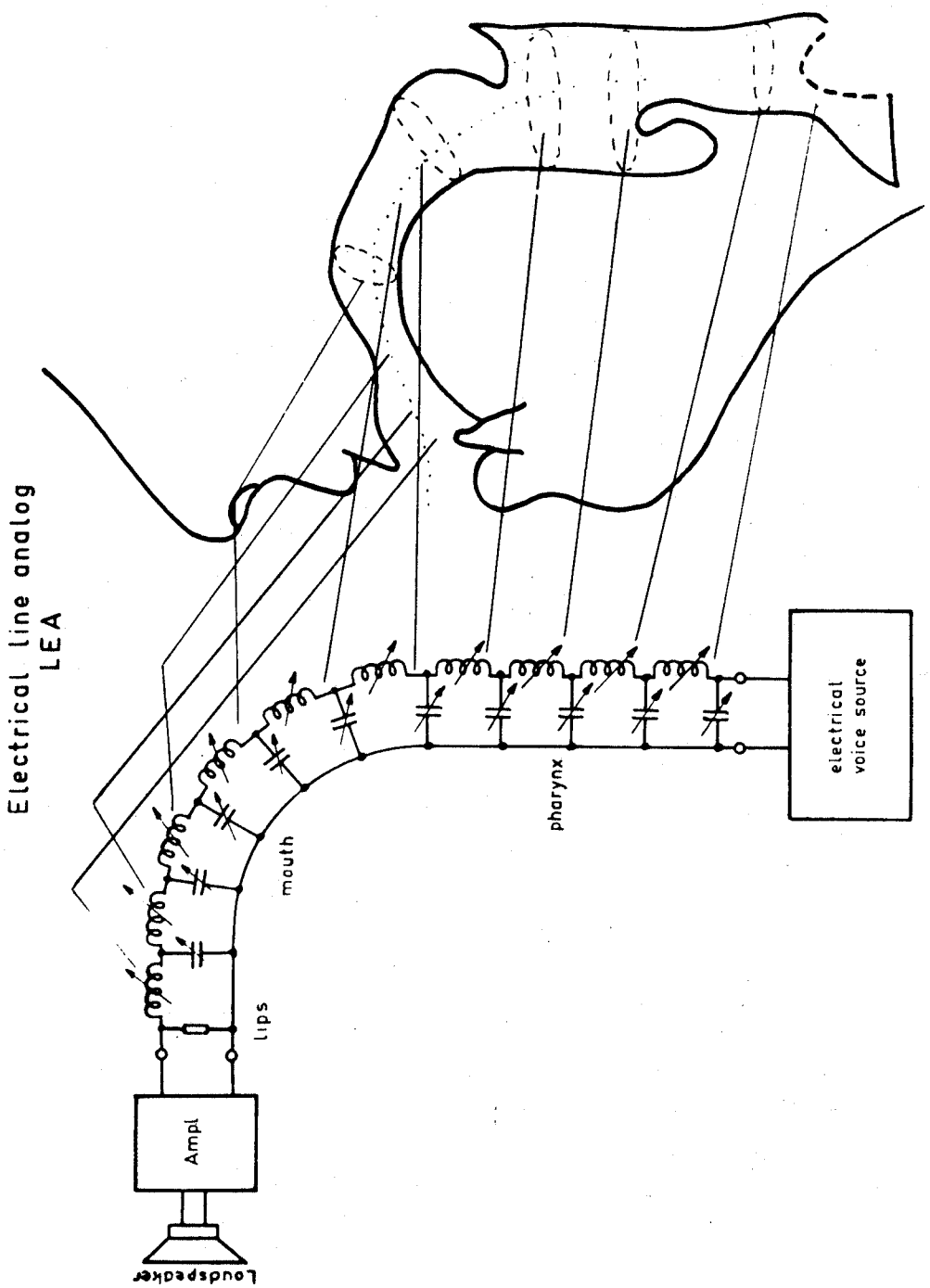

Electrical line analog
LEA
Loudspeaker
Ampl
lips
mouth
pharynx
electrical
voice source

Fig. 2

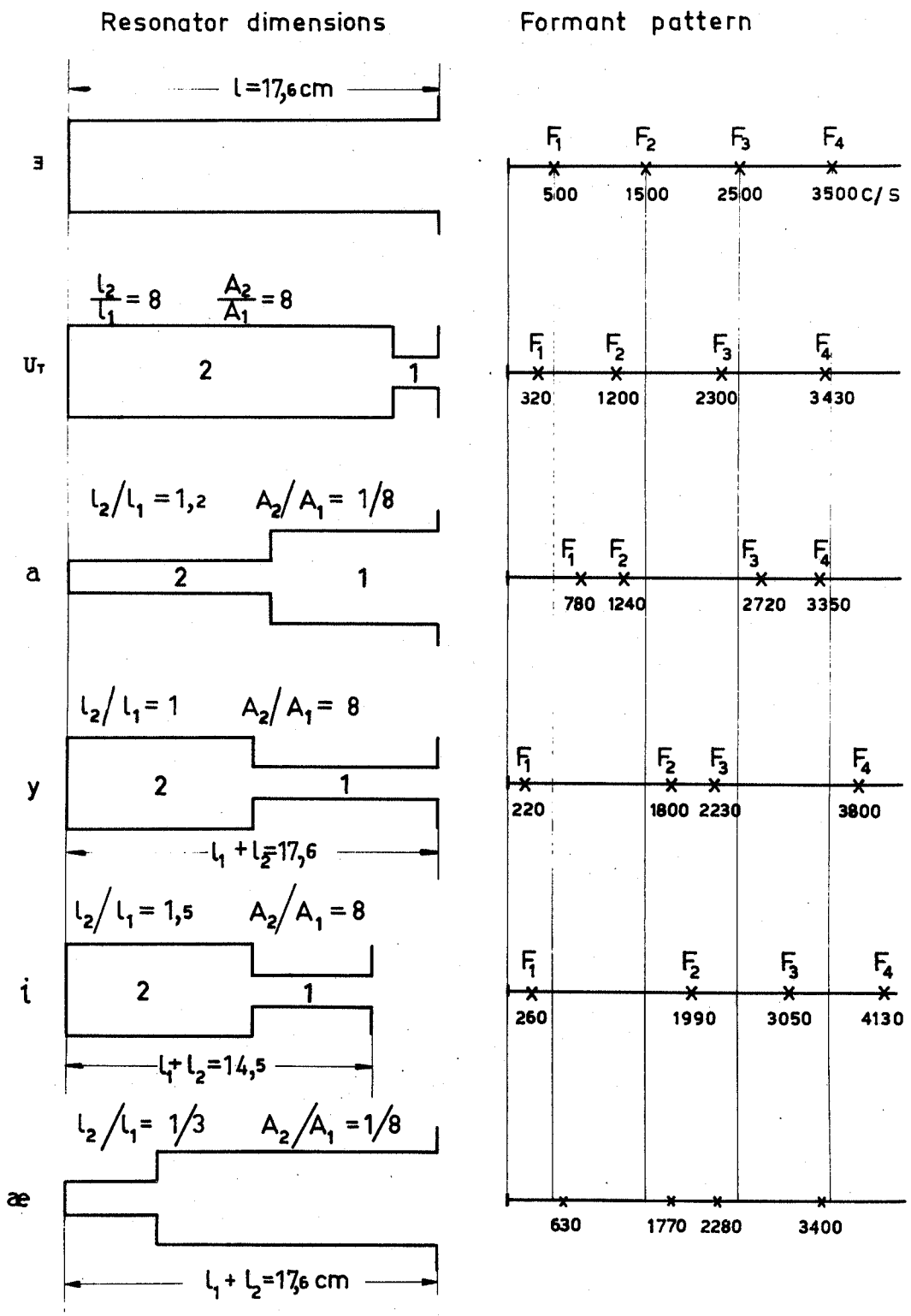

Resonator dimensions
Formant pattern
l=17,6cm
ɜ
$\frac{l_2}{l_1}=8$ $\frac{A_2}{A_1}=8$
ʊ
2
1
$l_2/l_1=1,2$ $A_2/A_1=1/8$
a
2
1
$l_2/l_1=1$ $A_2/A_1=8$
y
2
1
$l_1+l_2=17,6$
$l_2/l_1=1,5$ $A_2/A_1=8$
i
2
1
$l_1+l_2=14,5$
$l_2/l_1=1/3$ $A_2/A_1=1/8$
æ
$l_1+l_2=17,6$ cm
F1 F2 F3 F4
500 1500 2500 3500 c/s
F1 F2 F3 F4
320 1200 2300 3430
F1 F2 F3 F4
780 1240 2720 3350
F1 F2 F3 F4
220 1800 2230 3800
F1 F2 F3 F4
260 1990 3050 4130
630 1770 2280 3400

Fig. 3

2.3 Twin-tube models of vowels

Going back a few decades we find that attempts were made to analyze the vocal tract as consisting of a number of cylindrical sections. Assuming plane-wave propagation investigators chose to represent the acoustic properties of such sections in terms of equivalent electrical networks. For a given length of a cross-section the impedance elements of its network representation depend only on the cross-sectional area.

An example of an electrical analogue of the vocal tract is shown in Fig. 2. This is the model LEA which Fant used for most of his work reported in the Acoustic Theory of Speech Production (1960). Alternatively to derive the set of formant frequencies associated with a given vocal-tract shape researchers can to-day use digital computers to solve numerically Webster's horn equation. Or they may attempt to simulate a transmission-line equivalent as suggested by Flanagan (1965). Flanagan's model includes representations of the vocal tract, the nasal tract, the vocal cords as well as the subglottal system. Consequently, it is capable of simulating both phonation and respiration.

Using his electrical-line analogue LEA Fant demonstrated that twin-tube approximations of some simple vowel articulations could be successfully obtained. Fig. 3 shows that two tubes differing in terms of length and cross-sectional area are sufficient for attaining formant patterns that come close to those of human vowels such as [ʊ̟, a, y, i, æ]. This result is important because it helps to clarify the complex formant-cavity relations that characterize natural vowel productions. However, it was considered desirable to devise articulatory models that were more realistic and general from an articulatory point of view.

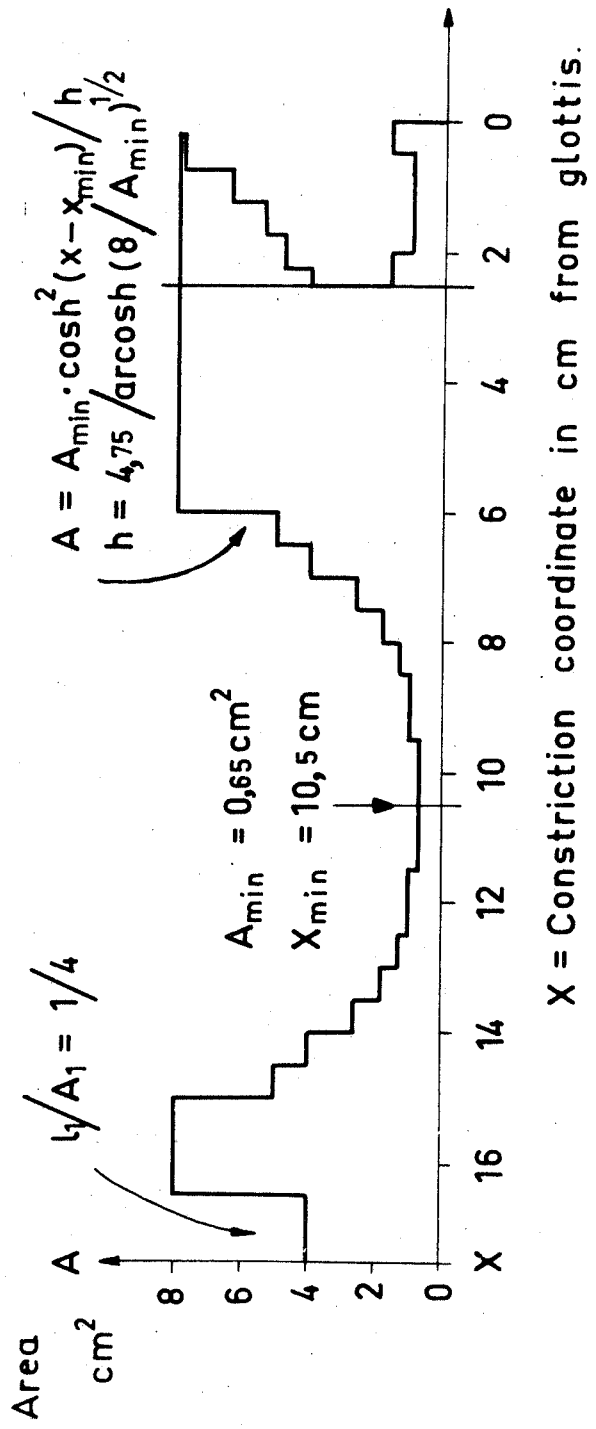

Area
A
cm²
8
6
4
2
0
X
16
14
12
10
8
6
4
2
0
$l_1/A_1 = 1/4$
$A_{min} = 0{,}65\,cm^2$
$X_{min} = 10{,}5\,cm$
$A = A_{min} \cdot \cosh^2(x - x_{min}) / h$
$h = 4{,}75 / \operatorname{arcosh}(8/A_{min})^{1/2}$
X = Constriction coordinate in cm from glottis.

Fig. 4

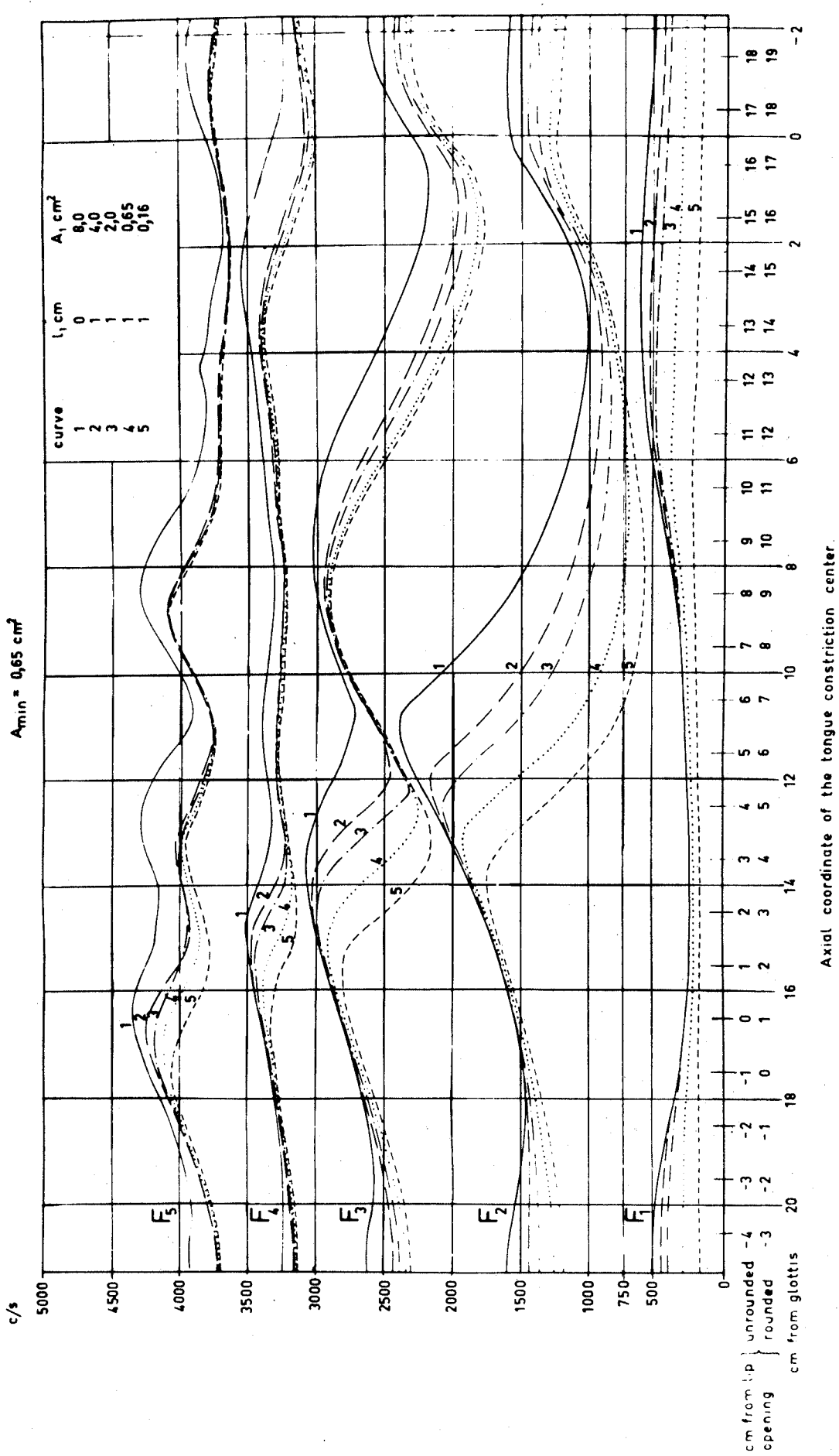

c/s
A_{min} = 0,65 cm²
curve l, cm A_1 cm²
1 0 8,0
2 1 4,0
3 1 2,0
4 1 0,65
5 1 0,16
F_5
F_4
F_3
F_2
F_1
5000
4500
4000
3500
3000
2500
2000
1500
1000
750
500
0
cm from lip opening
unrounded -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
rounded -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
cm from glottis
Axial coordinate of the tongue constriction center

Fig. 5

2.4 The three-parameter model

Following Dunn (1950), Stevens and House (1955) at MIT and Fant (1960) in Stockholm devised their versions of the so-called three-parameter models. The parameters of these models are closely related to the vocal-tract area function. The three numbers represent the place of oral constriction, the degree of this constriction and the length/area ratio for the lip section. These numbers can be said to delimit a class of area functions that approximate those occurring in human speech.

Fig. 4 illustrates an area function generated with the aid of Fant's three-parameter model. Note that once the values for the place and the degree of constriction have been selected the cross-sectional areas of adjacent points are uniquely determined according to a rule saying that these areas are a hyperbolic function of their distance from the place of constriction and the degree of constriction.

Fant undertook a systematic study of the acoustic properties of his model. Some of the results are shown in Fig. 5. From this nomogram we can infer the effects on formant frequencies of varying the place of constriction, the degree of constriction and "lip-rounding". With the aid of the three-parameter models it became possible to arrive at still deeper insights into the complex relations between articulation and acoustic aspects.

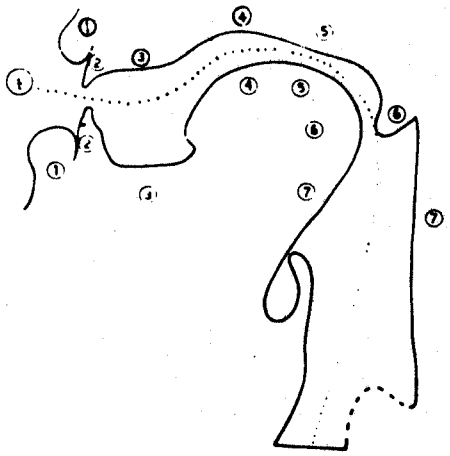
2.5 Deriving the vocal-tract area function

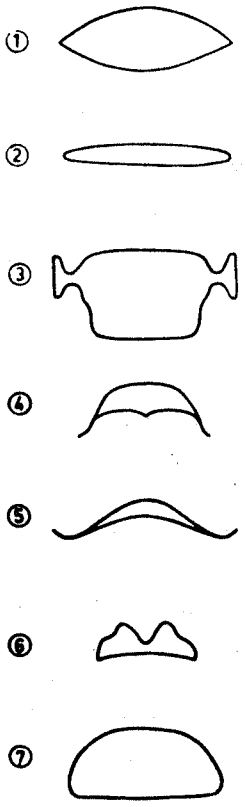
Attempts have also been made to derive formant frequencies from information obtained from X-ray pictures. Lateral profiles as well as tomographic cross-sections have been used to estimate the area functions associated with certain consonant and vowel articulations. Fig. 6 shows the procedure used by Fant to estimate

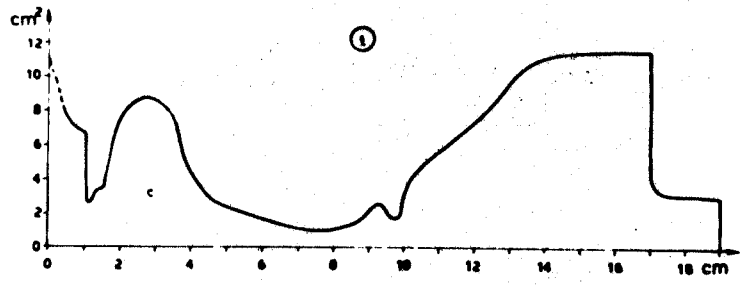
the cross-sectional areas normal to the vocal-tract midline. The continuous area function estimate is quantized so as to permit analogue computation of formant frequencies. In general the results indicate that fairly accurate predictions of formant frequencies can be made using measurements from X-ray pictures as input. This is the conclusion arrived at also by Heinz and Stevens (1964) who devoted special attention to the problems associated with going from lateral X-ray profiles to the acoustic output. They made measurements of casts of the palate and tried to formulate quantitative rules describing the conversion of cross-distances into cross-sectional areas.

Fig. 7 shows a plot of cross-sectional area against the distance (d) between the tongue contour and the palate. It can be seen that on this log-log plot the relationship between area and cross-distance is approximately linear. This result holds primarily in the region of the hard palate. In the pharynx other rules must be used. Our own work is based on similar results (1969) but we have had to revise the assumption of a flat tongue surface in the region near the velum.

Special problems arise below the velum. Direct observation of the pharyngeal cavities is difficult. Ladefoged using himself as subject (1971) had casts made of his pharynx under various special conditions. Lindqvist and Sundberg (1971) used fiber optics to obtain qualitative data on the shape of the pharynx during various sustained vowels. Fig. 8 is an illustration of the type of picture obtained with the fiber-scope. The bundle of fibers is introduced through the nose and positioned as shown in the profile tracing. They also attempted to make a quantitative treatment of their data and found good agreement with estimations of pharyngeal dimensions reported in other investigations.






cm²
0
2
4
6
8
10
12
14
16
18 cm

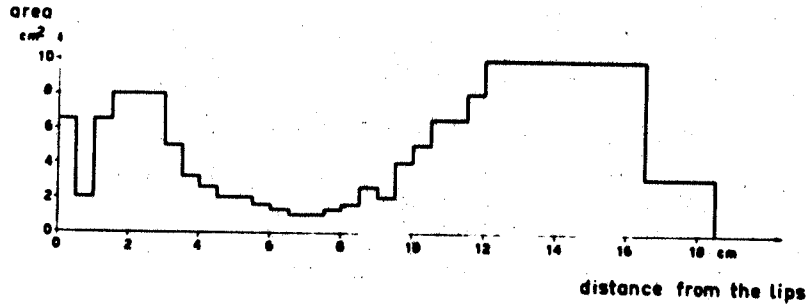
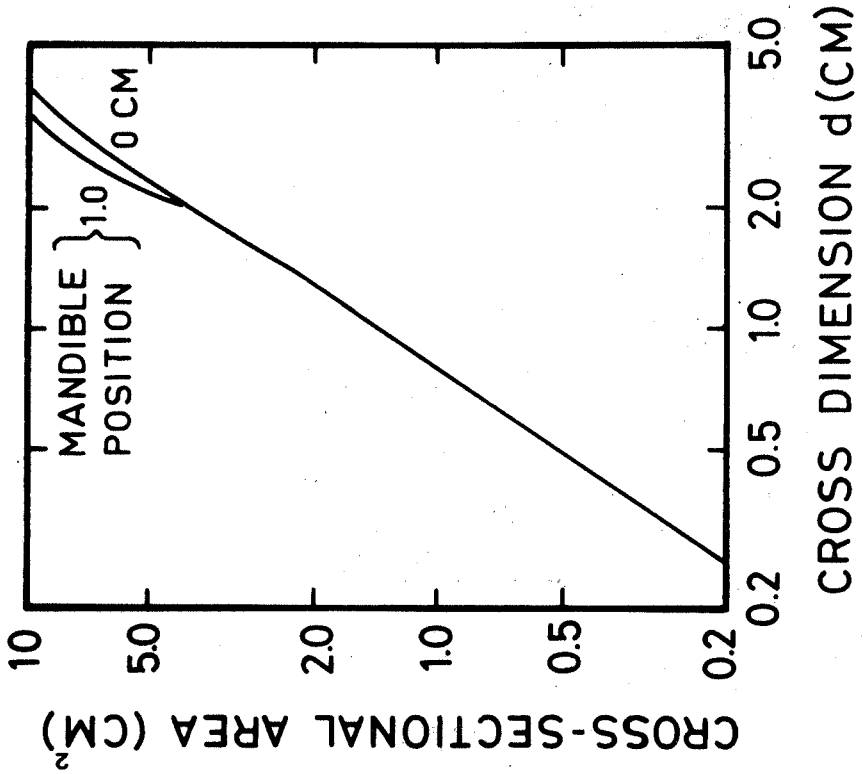

area
cm²
0
2
4
6
8
10
12
14
16
18 cm
distance from the lips

Fig. 6


CROSS-SECTIONAL AREA (CM²)
10
5.0
2.0
1.0
0.5
0.2
MANDIBLE POSITION
1.0
0 CM
0.2
0.5
1.0
2.0
5.0
CROSS DIMENSION d(CM)

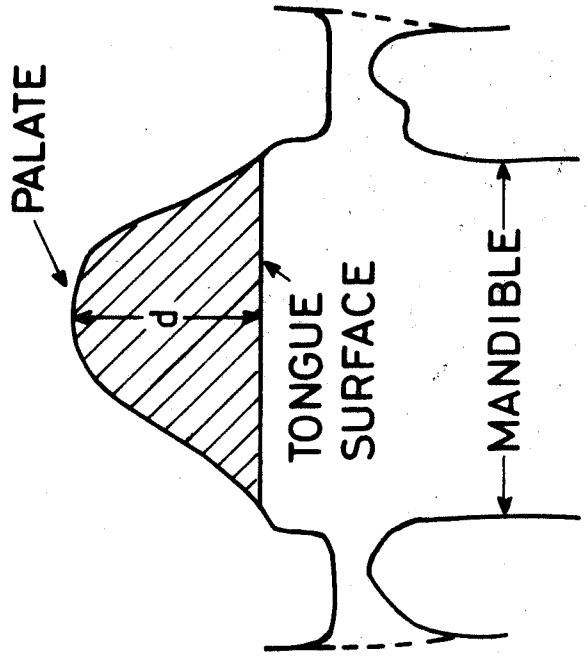

PALATE
d
TONGUE SURFACE
MANDIBLE

Fig. 7

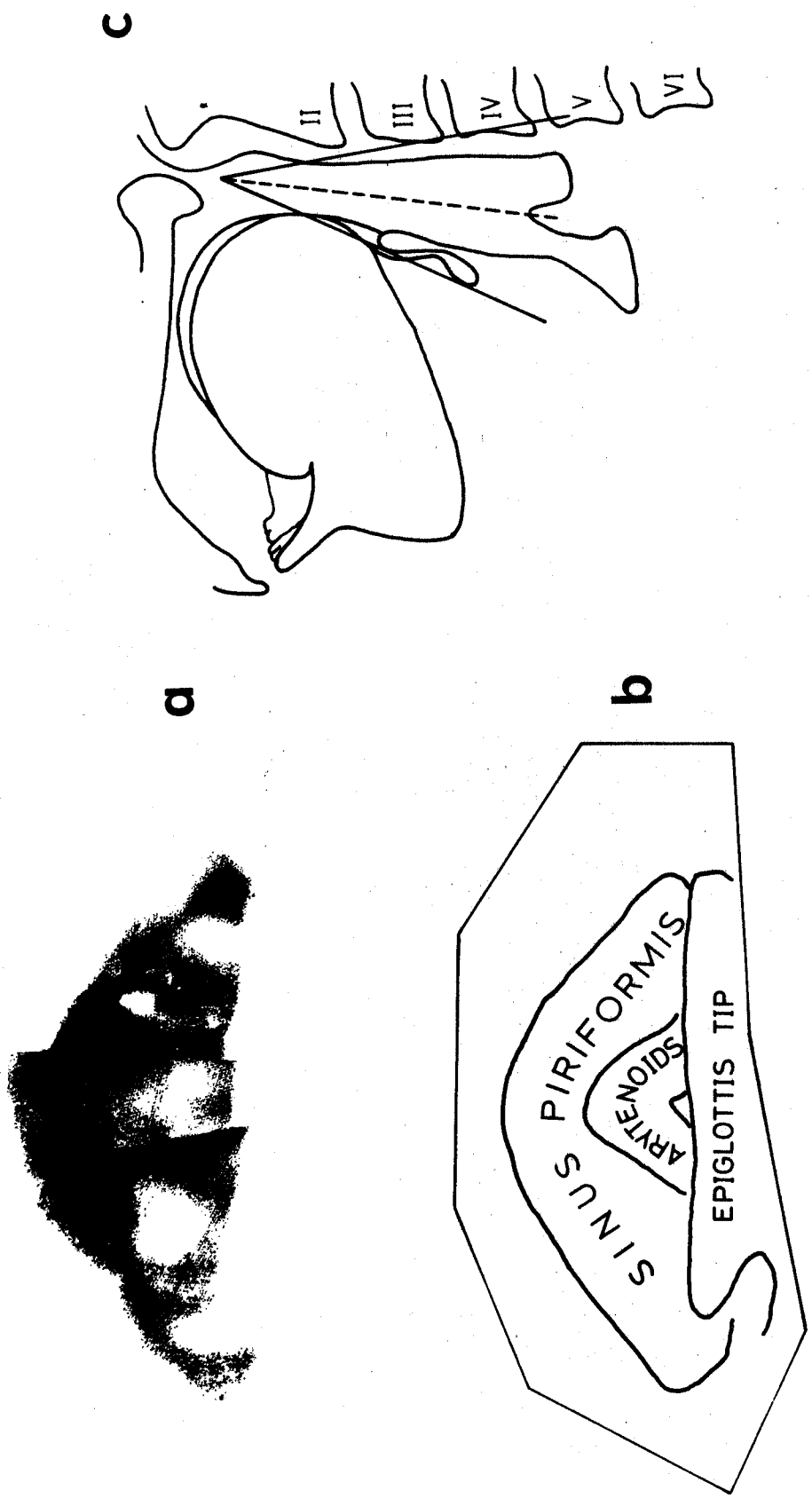

a
b
SINUS PIRIFORMIS
ARYTENOIDS
EPIGLOTTIS TIP
c
II
III
IV
V
VI

Fig. 8

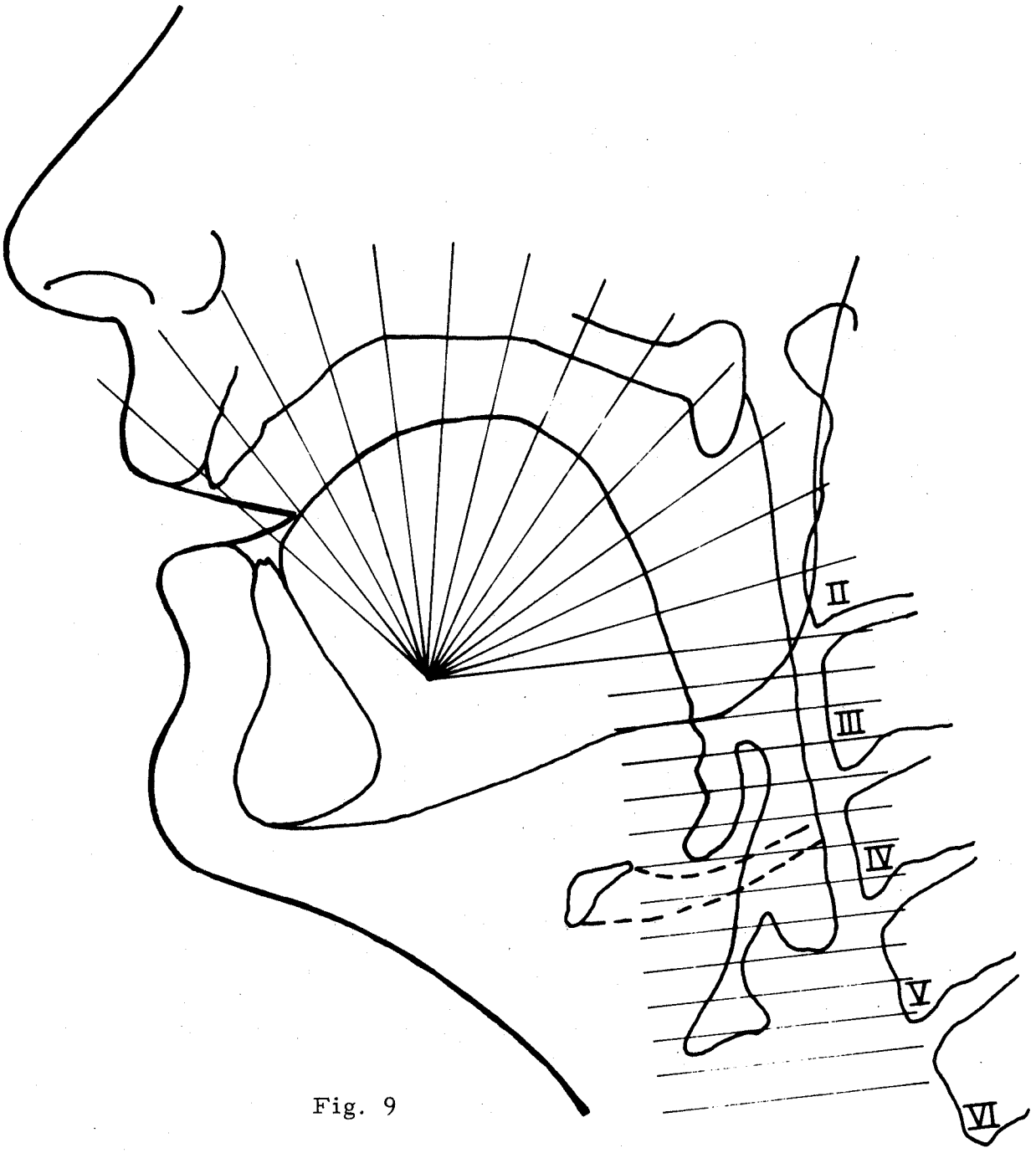
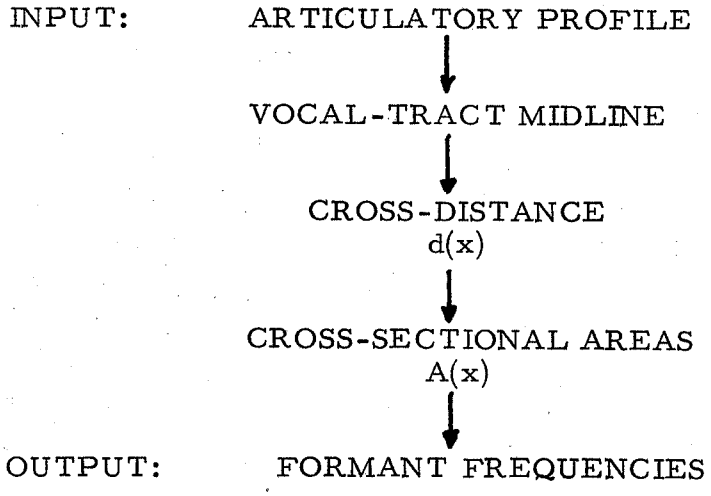


Fig. 9

To conclude our review of the procedures for deriving the vocal-tract area function from articulatory profiles we should add a few comments on how the cross-distances are obtained. Following Heinz and Stevens again investigators including ourselves have adopted the coordinate system shown in Fig. 9. Polar coordinates are used in the mouth region and Cartesian coordinates in the pharynx. The system is defined relative to fixed landmarks on the maxilla and other fixed structures. The coordinate segments delimited by the vocal-tract walls are halved and a contour of the vocal-tract midline is drawn so that it passes through the half-points. The cross-dimensions of the vocal-tract normal to this midline can then be measured at specific points along the tract.

We have now reached a point where it would be appropriate to summarize what we have said so far. These are the different stages reviewed so far:



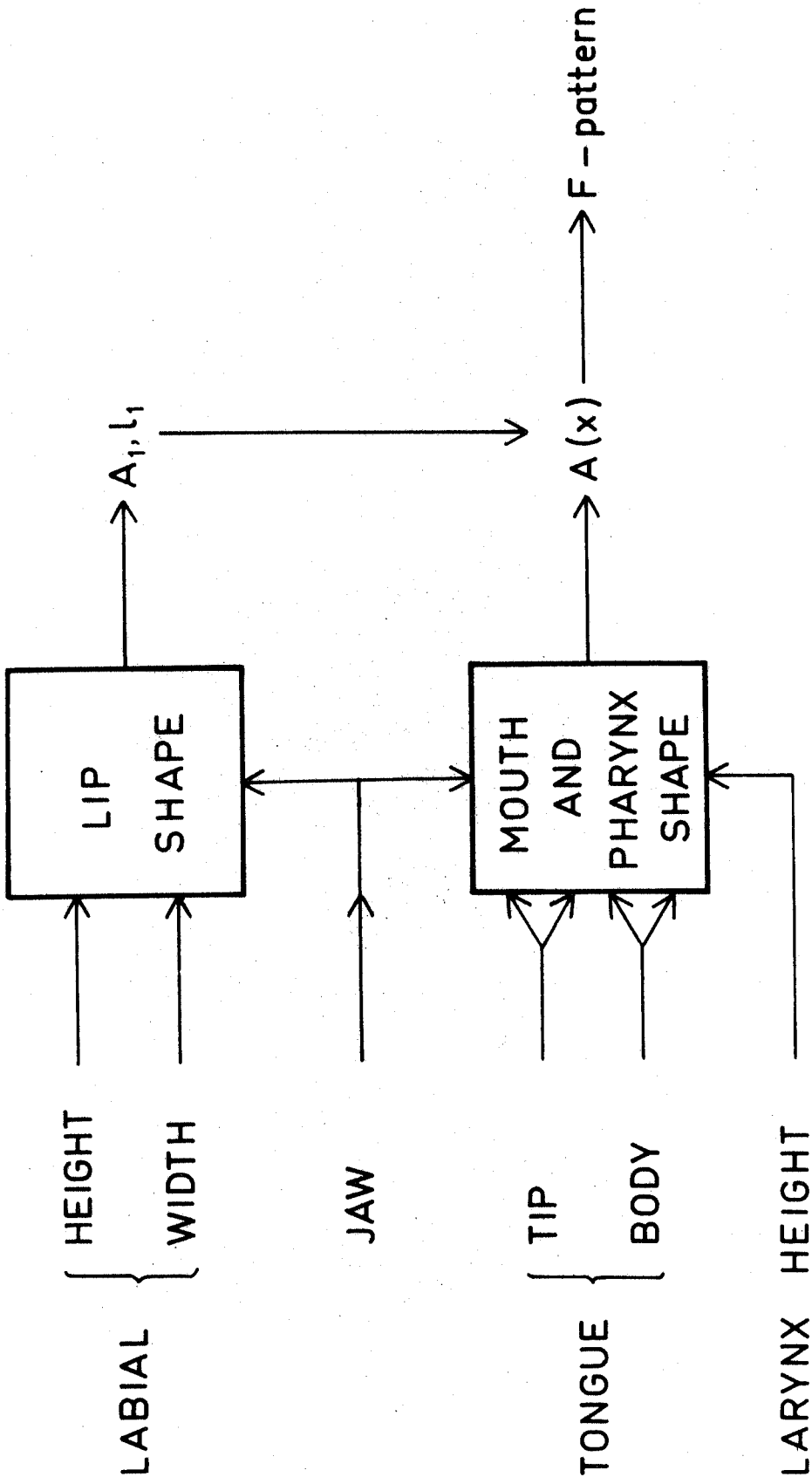
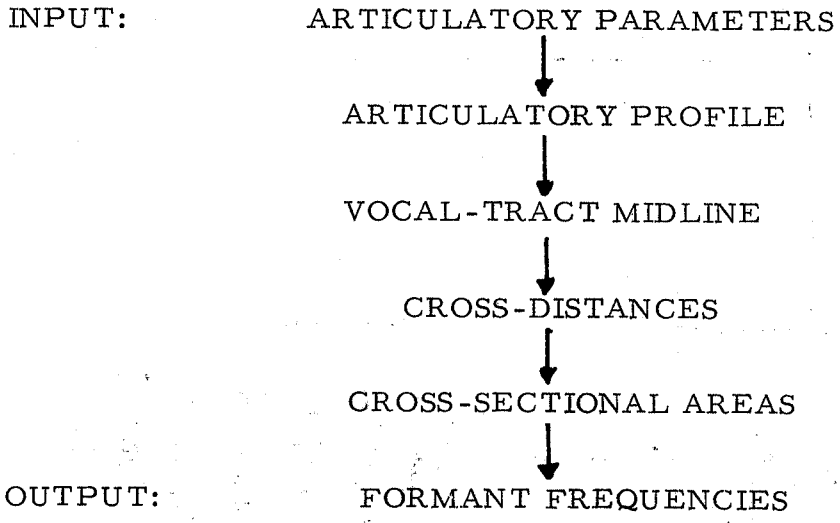

F - pattern
A(x)
A_1, L_1
LIP SHAPE
MOUTH AND PHARYNX SHAPE
LABIAL
HEIGHT
WIDTH
JAW
TONGUE
TIP
BODY
LARYNX HEIGHT

Fig. 10

3. From articulatory parameters to sound

We have reviewed the various steps involved in relating information contained in an articulatory profile to some of its potential acoustic correlates. We are now ready to go one step further and ask: how do we derive correctly the class of relevant articulatory profiles? Thus we add one more stage in our set of derivations:



This last step is the step that was recently taken by our own research group. We reported some aspects of this work in the October issue of the Journal of the Acoustical Society of America from last year (1971). We shall now report some further developments.

In constructing the model we have tried to retain features characteristic of the three-parameter models viz., control of place and degree of constriction. On the other hand, certain novel features have been introduced.

As can be seen from Fig. 10 the lip parameters are controlled independently of the position of the jaw. On the other hand, labial shape as well as vocal-tract termination depends on both jaw position and the values of the lip parameters. The model generates a

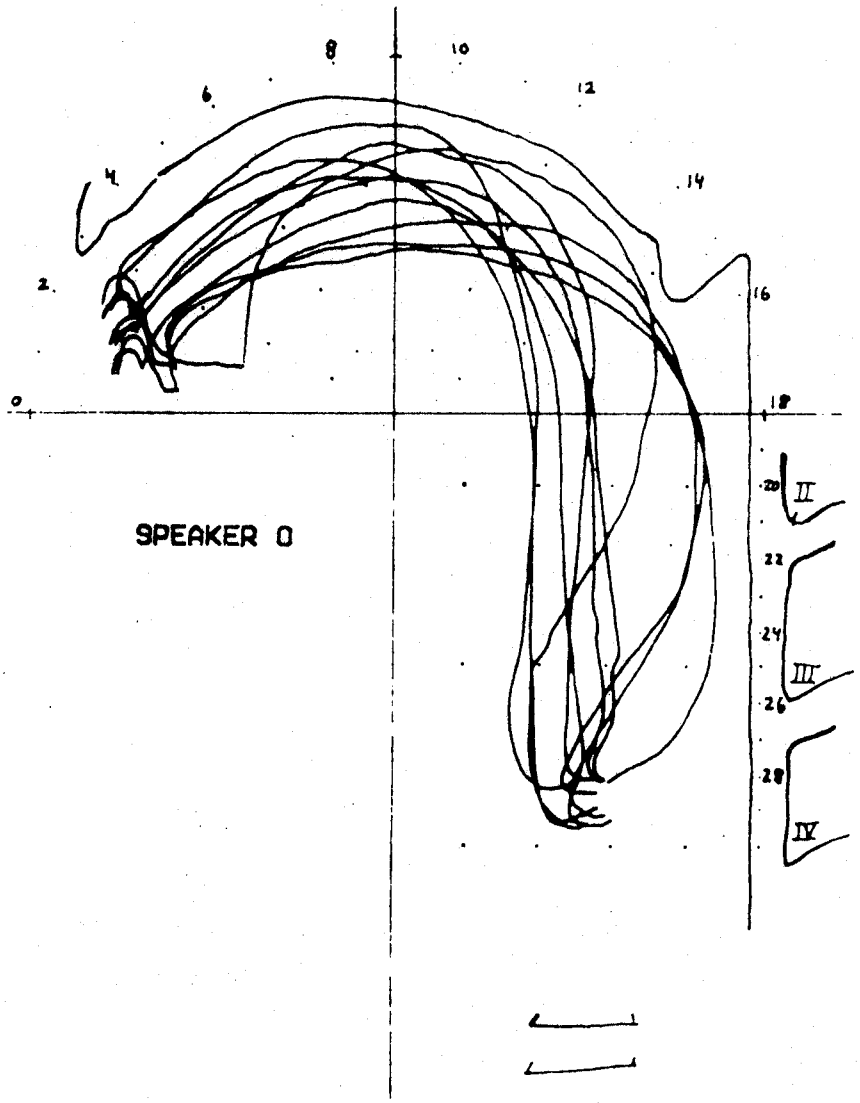


Fig. 11

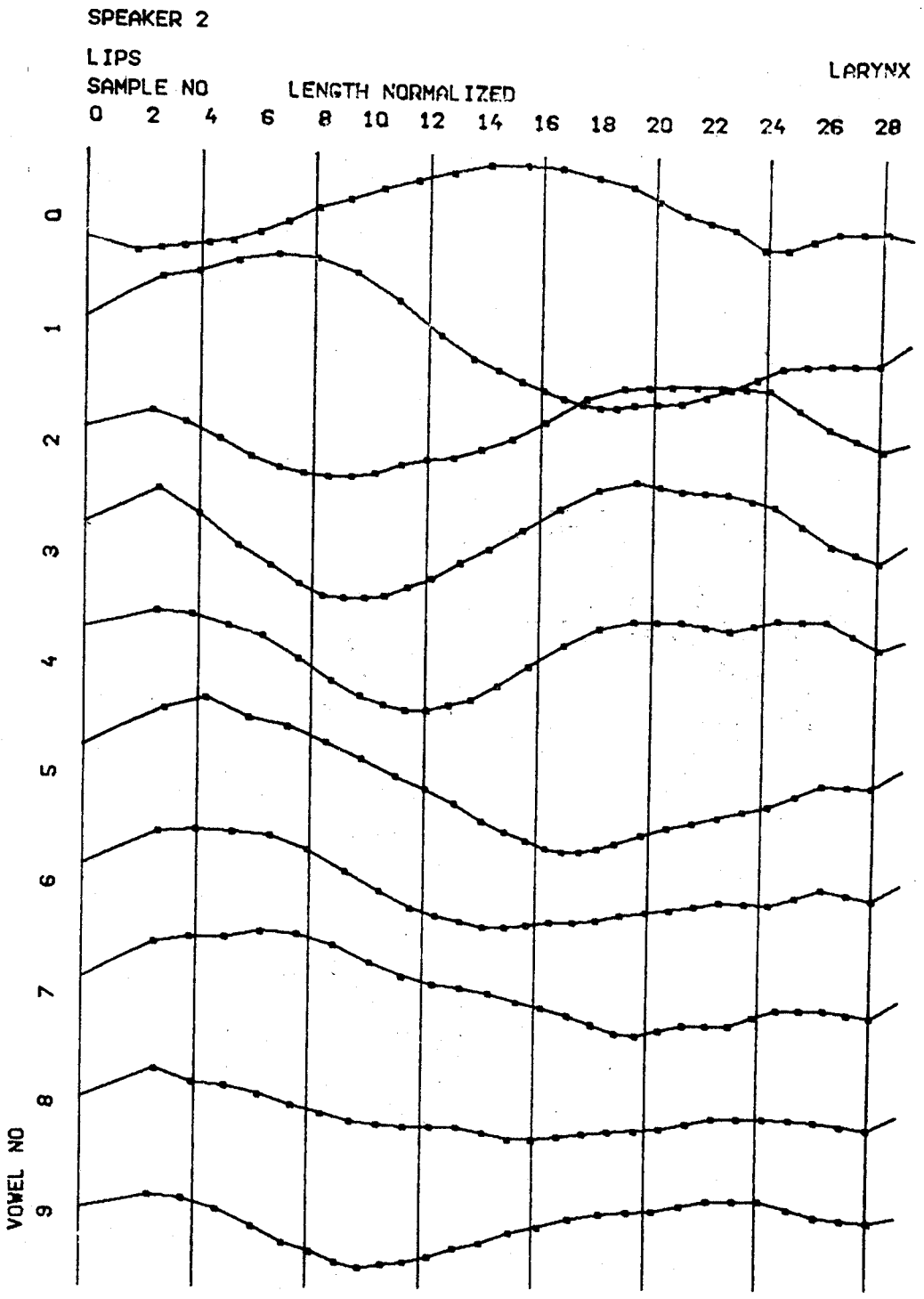

SPEAKER 2
LIPS
SAMPLE NO
LENGTH NORMALIZED
LARYNX
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28
VOWEL NO
0 1 2 3 4 5 6 7 8 9

Fig. 12

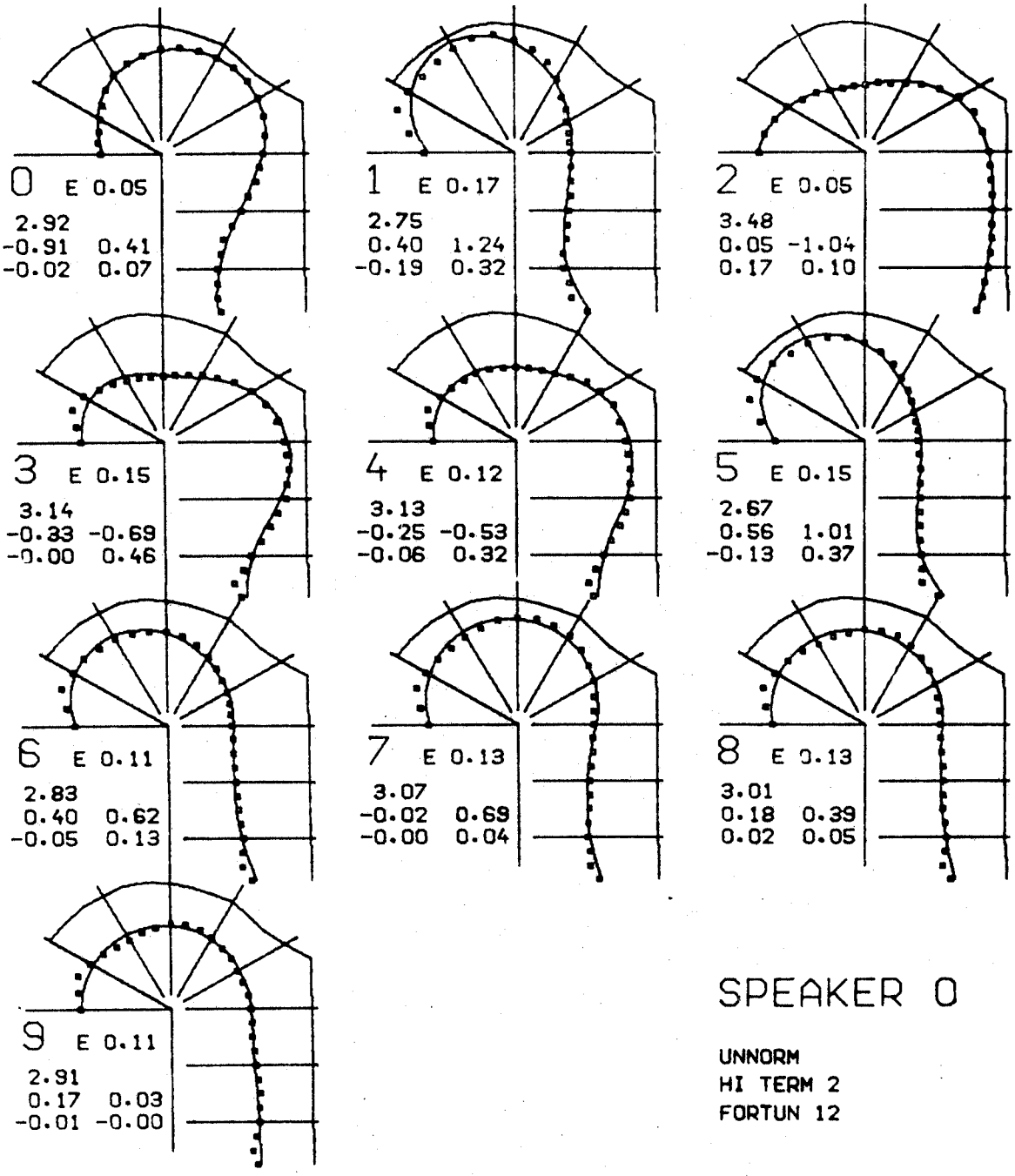

0 E 0.05
2.92
-0.91 0.41
-0.02 0.07
1 E 0.17
2.75
0.40 1.24
-0.19 0.32
2 E 0.05
3.48
0.05 -1.04
0.17 0.10
3 E 0.15
3.14
-0.33 -0.69
-0.00 0.46
4 E 0.12
3.13
-0.25 -0.53
-0.06 0.32
5 E 0.15
2.67
0.56 1.01
-0.13 0.37
6 E 0.11
2.83
0.40 0.62
-0.05 0.13
7 E 0.13
3.07
-0.02 0.69
-0.00 0.04
8 E 0.13
3.01
0.18 0.39
0.02 0.05
9 E 0.11
2.91
0.17 0.03
-0.01 -0.00
SPEAKER 0
UNNORM
HI TERM 2
FORTUN 12

Fig. 13

jaw-based tongue contour. A tongue body shape can be numerically derived provided that a specification is given of how much the tongue is deformed from its neutral shape and of the direction in which the tongue is displaced. Analogously, the tongue tip parameters are related to place and degree of apical elevation. They answer the questions of where? and how much? Finally it is also possible to vary the position of the larynx within certain limits.

3.1 The problem of tongue contour specification

An interesting contribution to the problem of tongue contour specification has come from our colleague Johan Liljencrants (1971).

Raw data obtained from X-ray pictures are presented in Fig. 11 for different vowels. A given shape is generally specified in terms of a table listing the coordinate lengths that are observed using a semipolar coordinate system similar to the one mentioned earlier.

These tables can be plotted as shown in Fig. 12 for example. These cruves show a certain resemblance to segments of waveforms. Thus Liljencrants suggests that the tongue contours be subjected to a regular Fourier analysis.

Fig. 13 shows that a very good agreement can be obtained between the observed contour and the Fourier model using only two harmonics and the DC term. Liljencrants interprets the phase of the "spatial fundamental" as indicative of the place of articulation.

3.2 The need for independent control of jaw position

During the course of our own work various alternative choices of parameters have suggested themselves. The question arises to what extent arguments in favor of the ones selected so far can be presented. Let us begin by challenging the incorporation of the jaw. Is it really necessary to have the jaw as a separate independent parameter? Our colleagues at Bell Laboratories have demonstrated

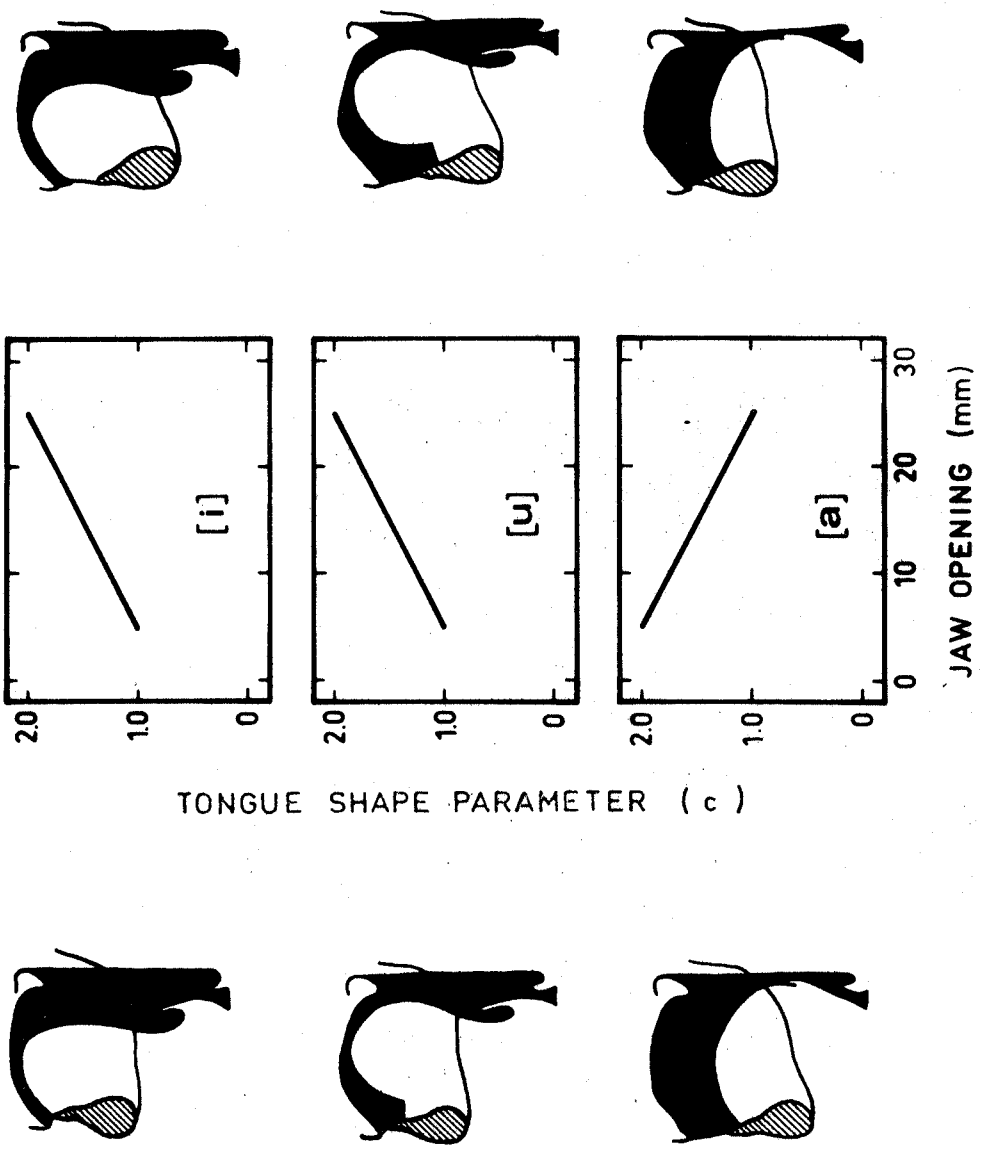


Fig. 14

that speech may be synthesized from articulatory parameters without using the jaw as an independent parameter(Tape demonstration of articulatory synthesis from a manual input). Listen to the following sample.

The intelligibility may not be a 100 % but if it is less than that it probably has very little to do with the treatment of the jaw. What are the arguments then?

The diagrams in the middle of Fig. 14 pertain to conditions under which our model will produce area functions appropriate for [i, a] and [u]. It is possible to produce an [i] either with the lower jaw raised and a moderate degree of palatalization. It is also possible to produce an [i] with a lowered jaw but with a larger degree of palatalization. In the left column of profiles the tongue and the jaw cooperate. In the right column they work against each other. What jaw positions would entail minimal displacement of the tongue from its neutral position for [i], [u], and [a]? From these diagrams we can infer that a jaw position that would be optimal according to this least effort criterion would be a raised one for [i] and [u] and an open or lowered one for [a]. We can conclude that the model being jaw-based provides a basis for explaining why [a] should be an open vowel and [i] and [u] close vowels.

3.3 Modeling universal phonetic constraints of vowel quality

A primary requirement on any model of articulation is that it be capable of producing most of the vowel qualities observed in the languages of the world. The present model delimits an acoustic vowel space of which we show an idealized representation in Fig. 15.

Suppose now that we place n vowels within this space and we ask a computer to maximize the distances between all pairs of vowels. What positions in the formant space are assigned to the vowels? Recently we made an attempt to answer this question (Liljencrants and Lindblom, forthcoming). The results of some preliminary

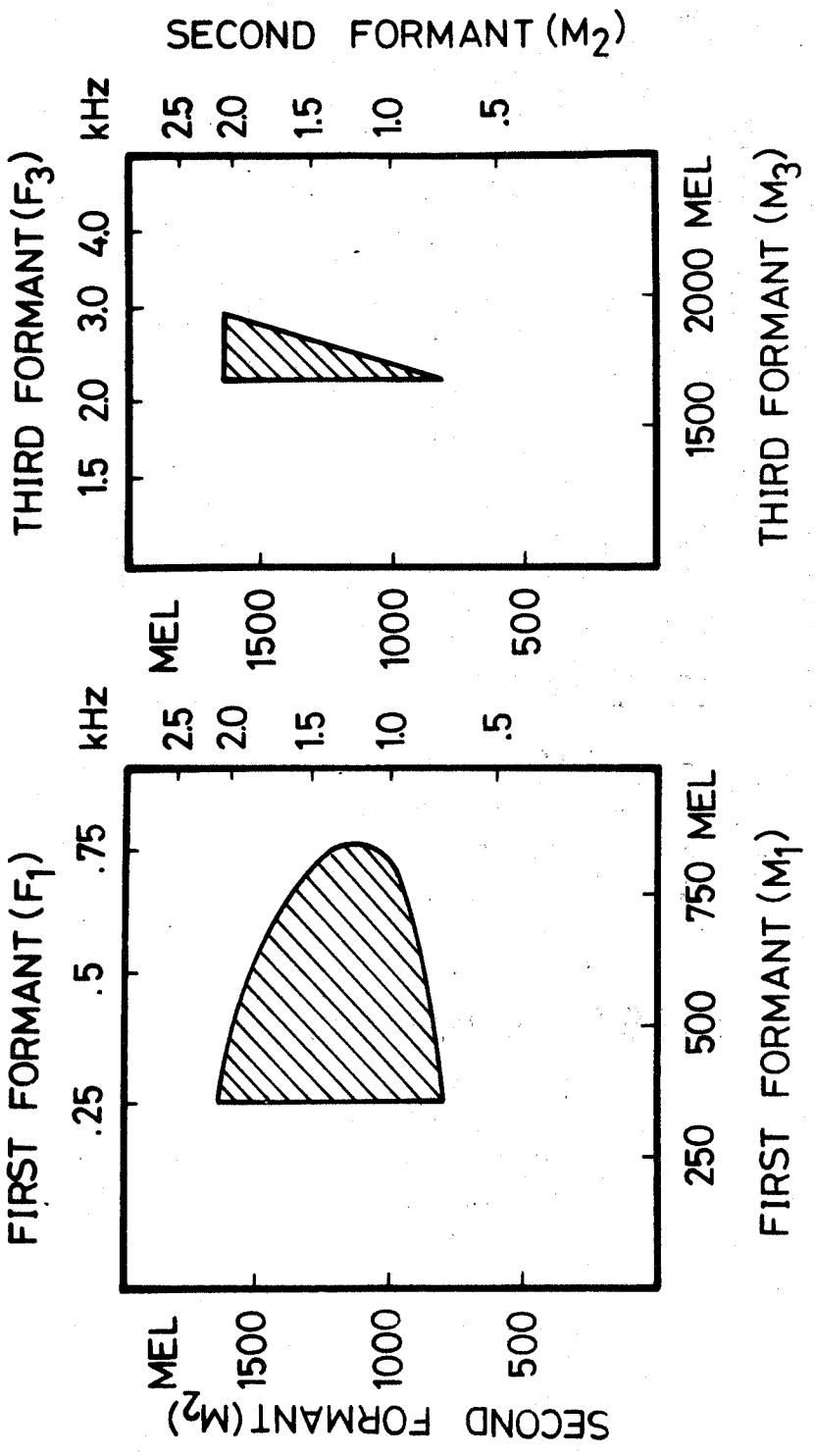

SECOND FORMANT (M_2)
kHz
2.5
2.0
1.5
1.0
.5
THIRD FORMANT (F_3)
1.5 2.0 3.0 4.0
MEL
1500
1000
500
1500 2000 MEL
THIRD FORMANT (M_3)
FIRST FORMANT (F_1)
.25 .5 .75
250 500 750 MEL
FIRST FORMANT (M_1)
SECOND FORMANT (M_2)

Fig. 15

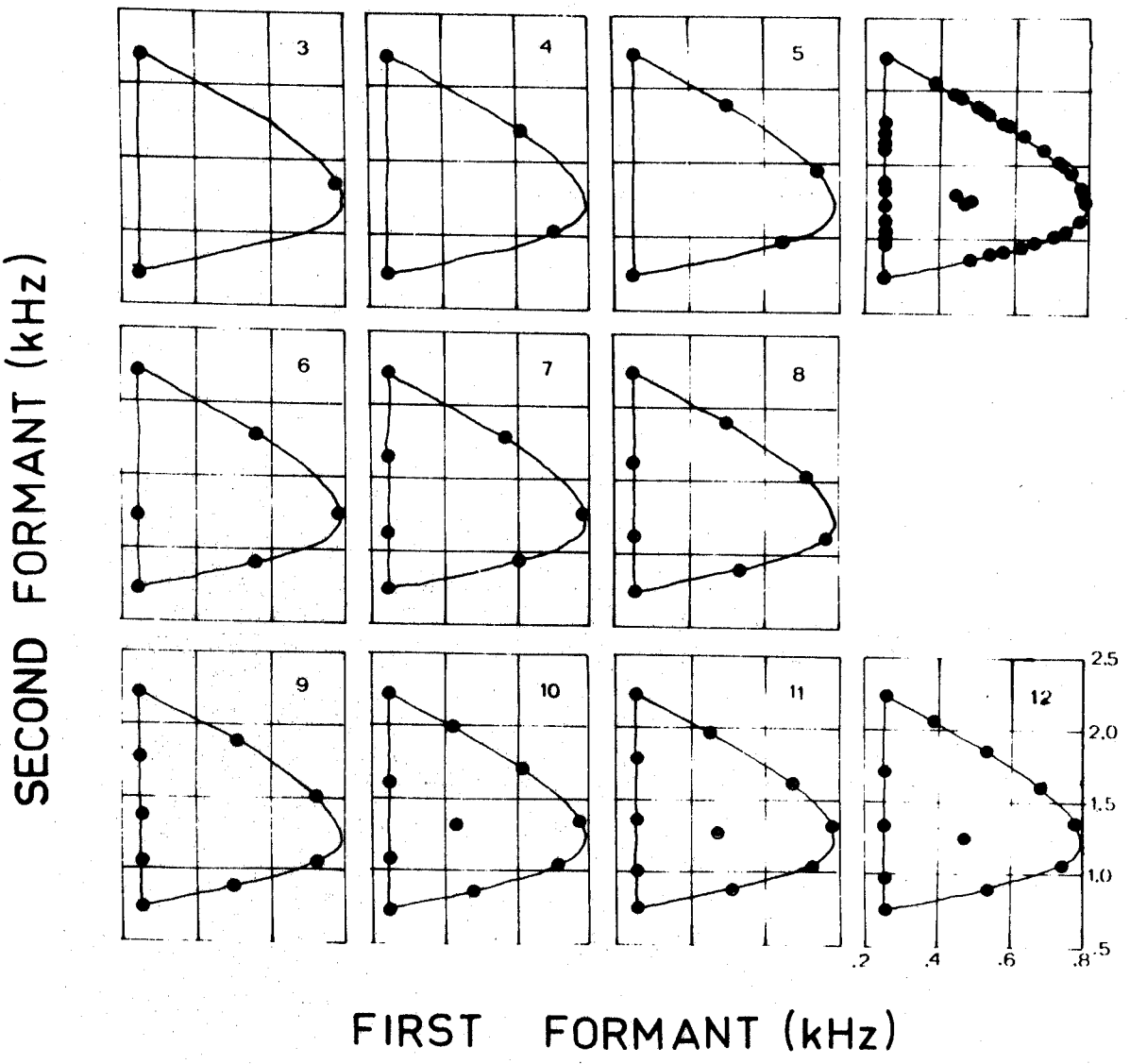

SECOND FORMANT (kHz)
3
4
5
6
7
8
9
10
11
12
2.5
2.0
1.5
1.0
.5
.2
.4
.6
.8
FIRST FORMANT (kHz)

Fig. 16

Phonetic values of predicted vowels

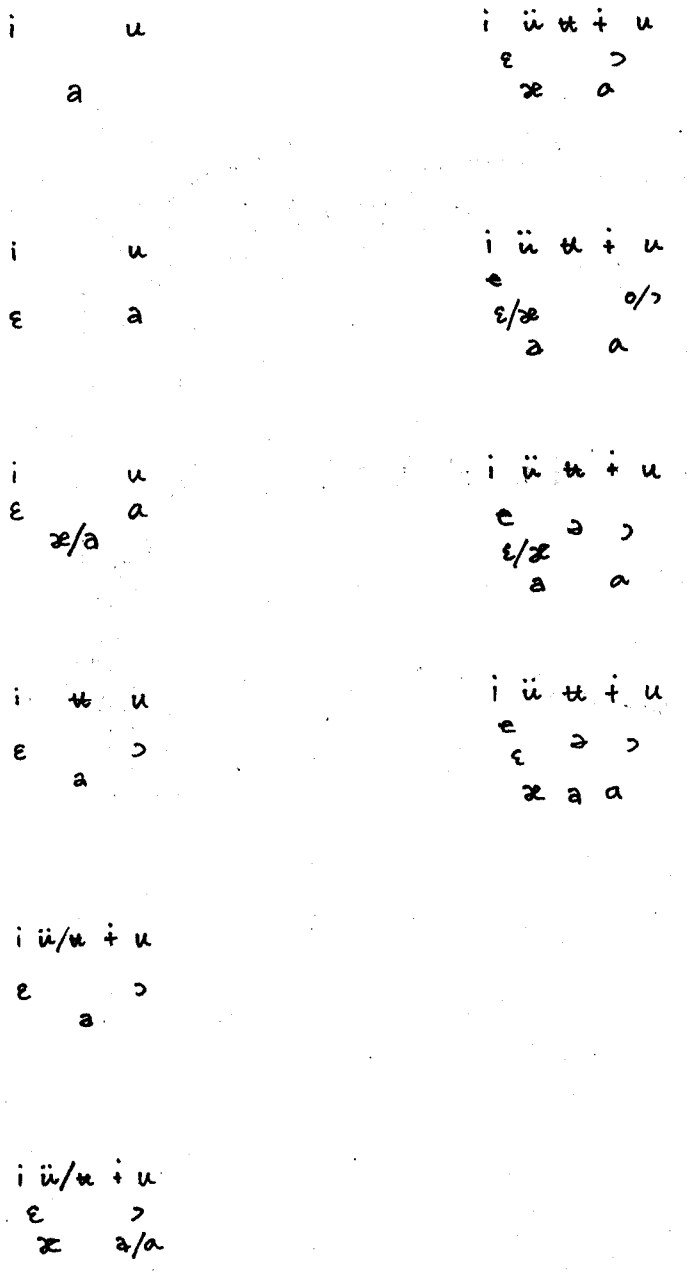


Fig. 17

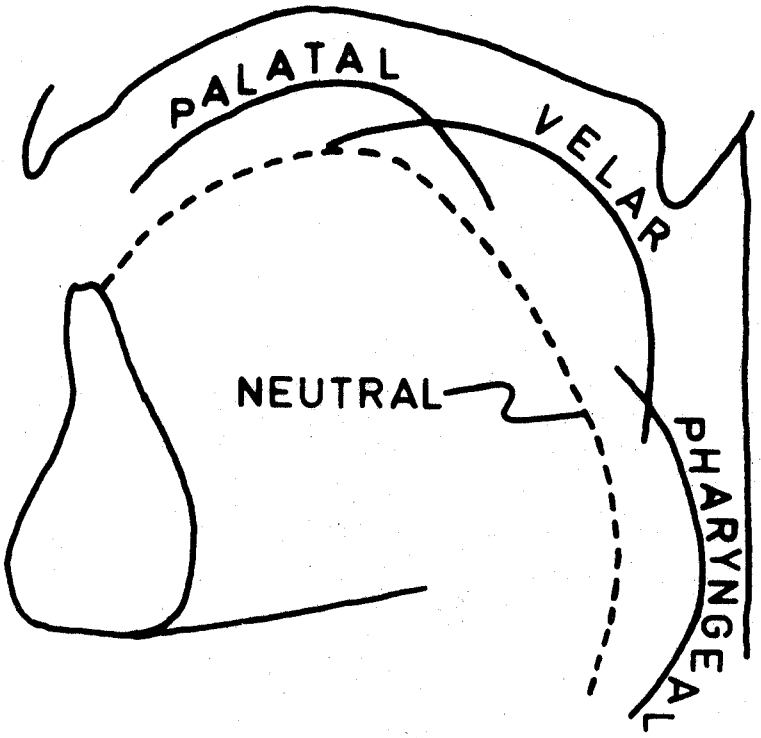


Fig. 18

calculations are shown in Fig. 16. Here you see the second formant frequency plotted against the first formant frequency for 3 through 12 vowels per system. The results come very close to actual vowel systems as can be seen from Fig. 17. Here the formant frequency values have been transcribed phonetically. We take this result as indicating that our model delimits the acoustic space for vowels in an approximately correct manner.

3.4 Neutralization

Our next example concerns the tongue body parameters. Remember the basic observation which we repeat in Fig. 18. We assume that the tongue body may be displaced from its neutral position in a palatal, velar, or pharyngeal direction. In an attempt to synthesize the long vowels of standard Swedish we once deliberately neglected to synchronize tongue movements with the lip and jaw movements. As a result phonation set in too early and the transition from a neutral tongue position to the target configurations for the different vowels became audible. The result is strikingly similar to the pronunciation of these vowels in a Southern Swedish dialect, Scanian. The following table gives the phonetic values of these vowels in the notation of Bruce (1970):

/i:/ → [ei]	/y:/ → [øy]	/u:/ → [eu]
/e:/ → [ɛe]	/ʉ:/ → [øʉ]	/o:/ → [ɛo]
/ɛ:/ → [æɛ]	/ø:/ → [œø]	/ɑ:/ → [æɑ]

In terms of the model parameters a natural description is obtained of these data. This is a case where the phonetic realism that can be attained using the notational devices developed within generative phonology is insufficient and where a description in terms of numerical speech production models seems preferable. For a discussion of the implications of articulatory models for phonological theory, see Lindblom (1971). We conclude from this example and from other

TONGUE BODY PARAMETERS

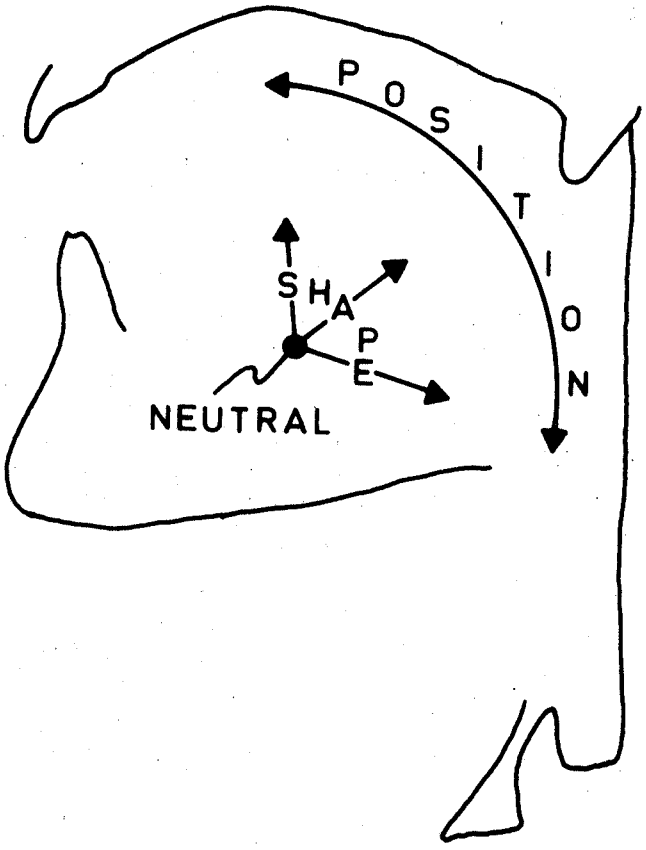


Fig. 19

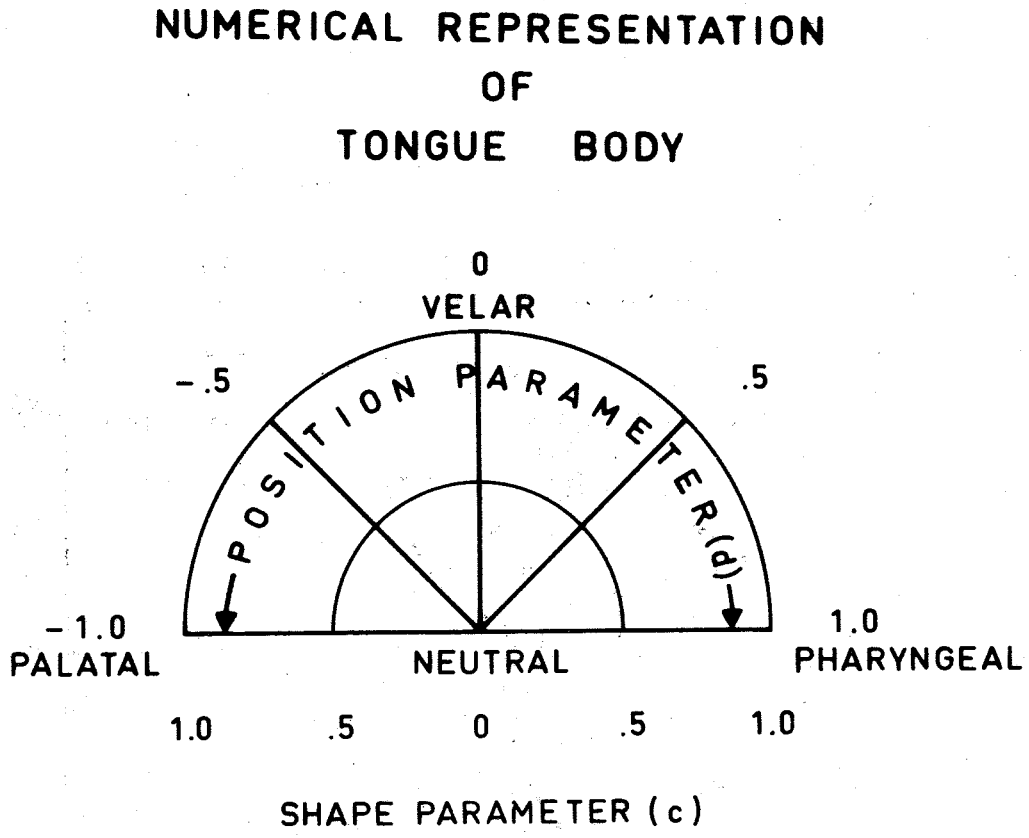

NUMERICAL REPRESENTATION
OF
TONGUE BODY
0
VELAR
-.5
.5
POSITION PARAMETER (d)
-1.0
PALATAL
NEUTRAL
1.0
PHARYNGEAL
1.0
.5
0
.5
1.0
SHAPE PARAMETER (c)

Fig. 20

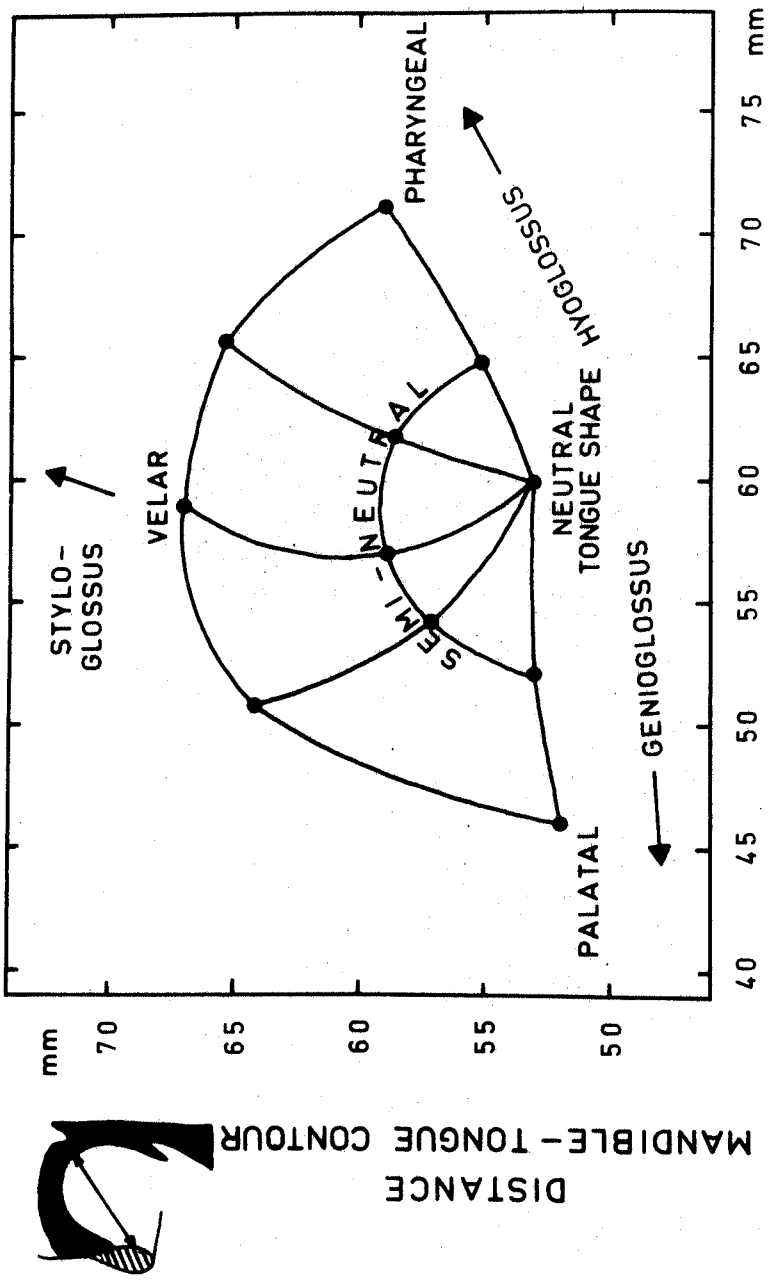

DISTANCE
MANDIBLE – TONGUE CONTOUR
mm
70
65
60
55
50
40
45
50
55
60
65
70
75 mm
STYLO-
GLOSSUS
VELAR
SEMI-NEUTRAL
PHARYNGEAL
HYOGLOSSUS
NEUTRAL
TONGUE SHAPE
GENIOGLOSSUS
PALATAL
TONGUE ROOT
DISPLACEMENT

Fig. 21

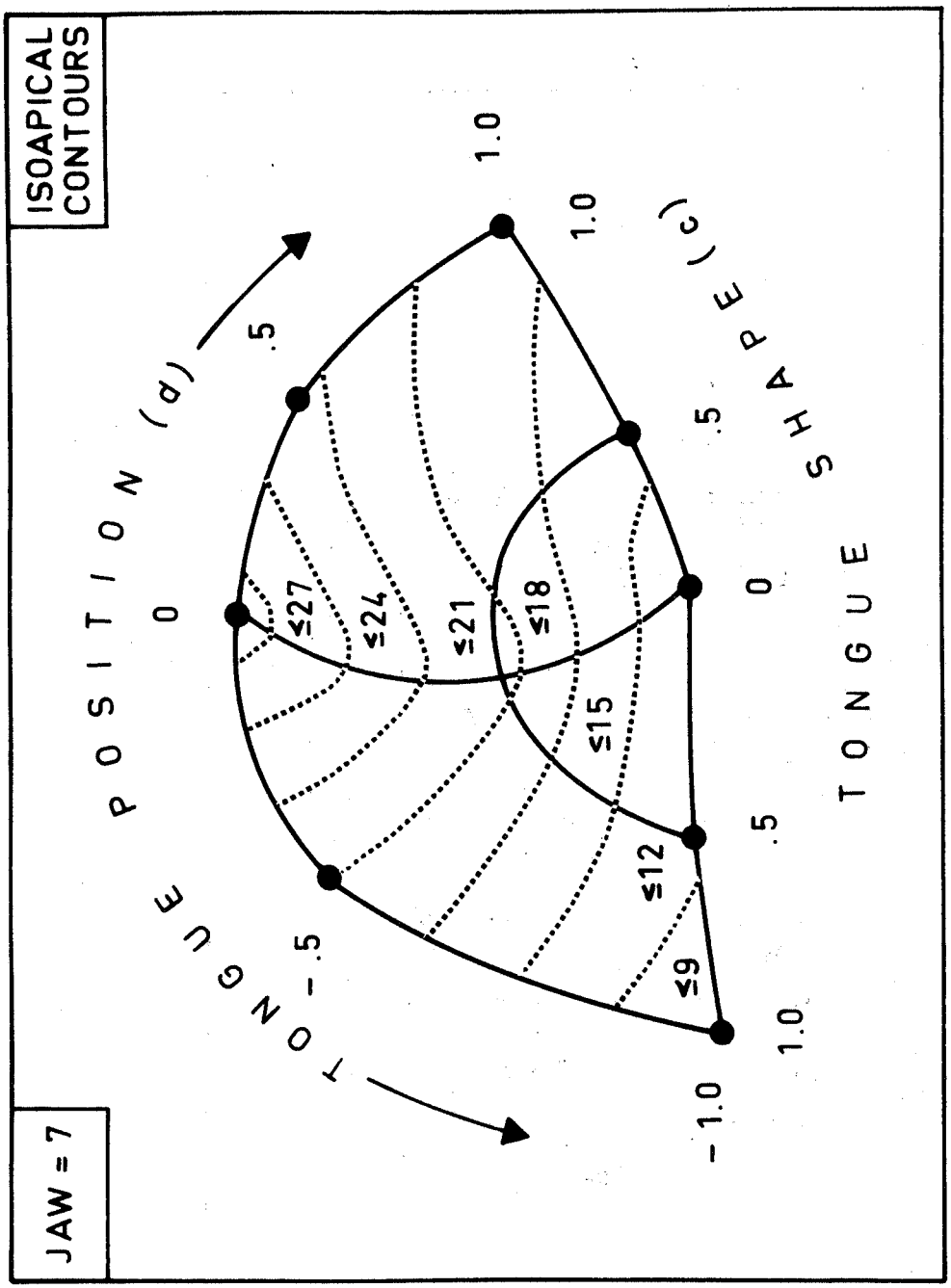

ISOAPICAL CONTOURS
JAW = 7
TONGUE POSITION (d)
TONGUE SHAPE (c)
1.0
.5
0
-.5
-1.0
≤27
≤24
≤21
≤18
≤15
≤12
≤9

Fig. 22

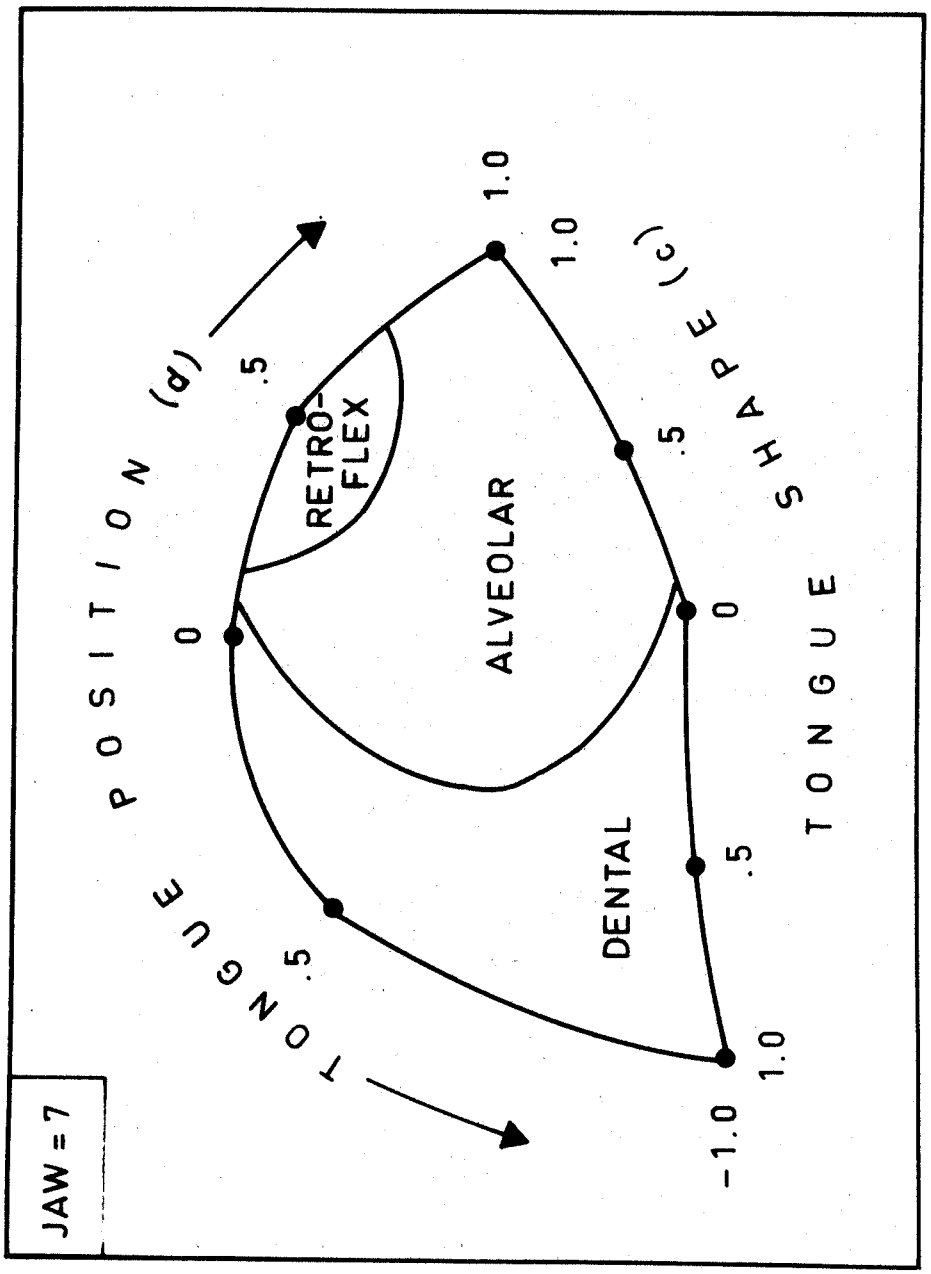

JAW = 7
TONGUE POSITION (α)
.5
0
-.5
RETRO-FLEX
ALVEOLAR
DENTAL
1.0
.5
0
-1.0
TONGUE SHAPE (c)
1.0
.5
0
.5
1.0

Fig. 23

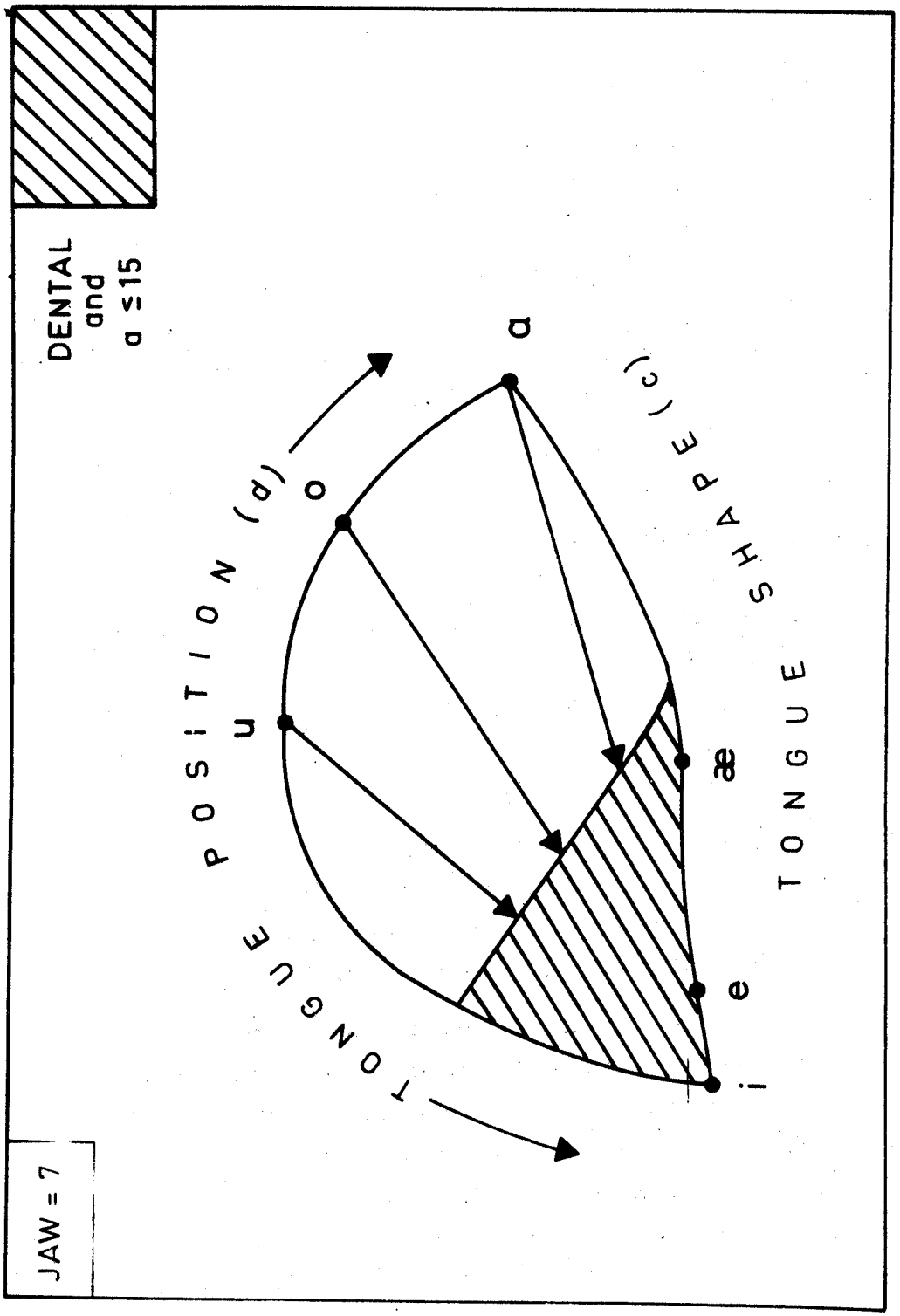


Fig. 24

evidence that an articulatory model should in all probability be equipped with a parameter representing the degree of tongue shape neutralization.

3.5 Tongue tip movement and coarticulation

Our final example concerns tongue tip movement and coarticulation. Again we remind you how we have defined our tongue body parameters. In Fig. 19 shape and position are independently controlled. A convenient way of representing these parameters numerically is shown in Fig. 20. Since for neutral or zero shape deformation there can be no position specification the articulatory space of the tongue body parameter takes the form of a semi-circle.

Fig. 21 shows that this semi-circle can be interpreted in terms of muscle actions. Along the x-axis we have tongue root displacement or the action of the genioglossus versus that of the hyoglossus. Along the y-axis we have displacement of the tongue in a velar direction or a pull from the styloglossus muscle. The model makes it possible for us to answer the following question: For any given position of the tongue body and the jaw by how much must the tongue tip be raised for a dental closure to be attained ? Will a dental closure be produced at all in the first place?

The answer is given in Fig. 22 showing isoapical contours, that is, positions of the tongue body compatible with apical closure for a jaw opening of 7 mm. However, dental closure is obtained only in certain areas of the semi-circle as is shown in Fig. 23.

The cross-hatched area in Fig. 24 represents an area corresponding to dental contact and a maximal elevation of the tongue tip of 15 mm. Now suppoes that in producing sequences such as [udu], [ada], [odo], [idi], [ede] etc. the speech production system uses a strategy of minimization of "effort". In producing a dental stop the tongue body does, as it were, half of the work and the apex does the other half. In quantitative terms, limit the elevation of the apex

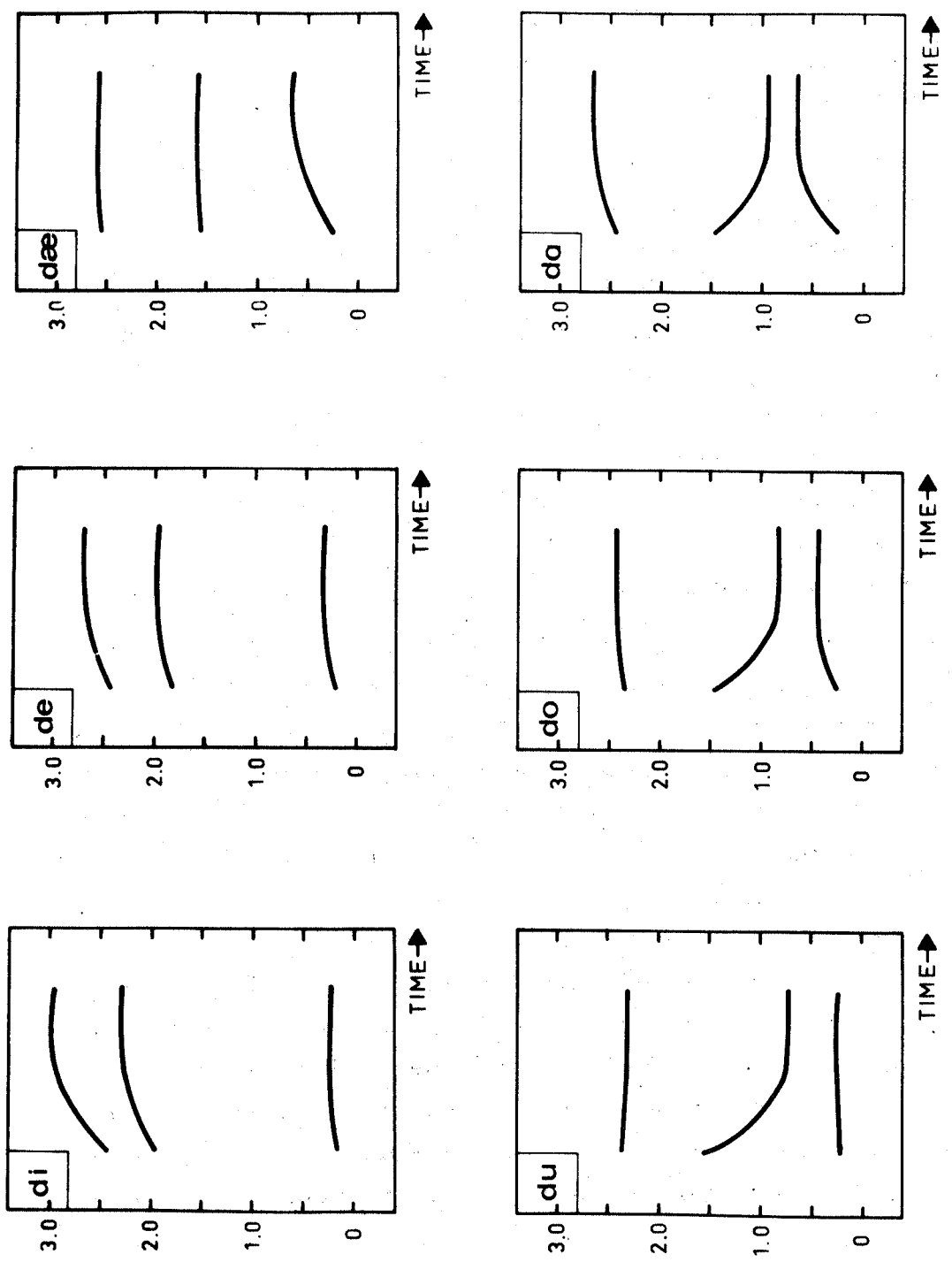


Fig. 25

to 15 mm and let the back of the tongue be displaced towards its target position but no further than necessary. What are the formant or locus frequencies that would be observed for these articulations? The predicted formant patterns are shown in Fig. 25.

When these diagrams are compared with spectrograms of the corresponding natural syllables it can be seen that there is good agreement with respect to F_1 and F_2 transitions and satisfactory agreement with respect to F_3. This result appears to lend further support to the selection of parameters for the model and seems to allow a new interpretation of coarticulation.

4. Concluding remarks

The last example brings us to the end of our presentation and it is time to raise again the question of the introduction. How far upstream is it necessary to go in modeling articulatory behavior?

Our answer must depend on the goals of our research. If we construct models of speech production in order to contribute to phonetic theory we should not hesitate to incorporate the jaw, neutralization as independent parameters. If on the other hand we construct models primarily to solve specific technological problems that arise in the application of phonetic theory, for instance to produce speaking machines for practical and commercial purposes, the answer may be different. However, these aspects must not be confused. How far upstream should we go then? A possible answer seems to be: as far as possible, as long as going upstream will increase the explanatory power of our models. This is the answer that we have tried to give in terms of the proposals presented in this paper. We hope that these suggestions might indicate directions in which better explanations of various phonetic facts may be sought.

Acknowledgments

This work was supported by the National Institutes of Health under Research Grant NS 04003.

References

Bruce, G. (1970): "Diphthongization in the Malmö dialect", Working Papers 3, Phonetics Laboratory, Lund University, pp. 1-19.

Dunn, H.K. (1950): "The calculation of vowel resonances, and an electrical vocal tract", J.Acoust.Soc.Am. 22, pp. 740-753.

Fant, G. (1960): Acoustic Theory of Speech Production, Mouton & Co., The Hague 1960 (2nd edition 1970).

Flanagan, J.L. (1965): Speech Analysis Synthesis and Perception, Academic Press, New York 1965 (2nd edition 1972).

Heinz, J.M. and Stevens, K.N. (1965): "On the relations between lateral cineradiographs, area functions, and acoustic spectra of speech", in Congress Reports (5th ICA, Liège), Vol. Ia, paper A44.

Ladefoged, P., Anthony, J.F.K., and Cordell, R. (1971): "Direct measurement of the vocal tract", Working Papers in Phonetics No. 19, UCLA, p. 4.

Liljencrants, J. (1971): "Fourier series description of the tongue profile", STL-QPSR No. 4, pp. 9-18.

Liljencrants, J. and Lindblom, B. (forthcoming): "Numerical simulation of vowel quality systems: The role of perceptual contrast", accepted for publication in Language.

Lindblom, B.E.F. (forthcoming): "Phonetics and the description of language", to appear in the Proc. of the VII International Congress of Phonetic Sciences, Montreal, Canada 1971 (to be publ. by Mouton & Co.).

Lindblom, B.E.F. and Sundberg, J. (1971): "Acoustical consequences of lip, tongue, jaw and larynx movement", J.Acoust.Soc. Am. 50, pp. 1166-1179.

Lindqvist, J. and Sundberg, J. (1971): "Pharyngeal constrictions", STL-QPSR No. 4, pp. 26-31.

Stevens, K.N. (1967): "The quantal nature of speech: evidence from articulatory-acoustic data", submitted manuscript as a chapter for Human Communication: A Unified View, eds. E.E. David, Jr. and P.B. Denes.

Stevens, K.N. (1968): "Acoustic correlates of place of articulation for stop and fricative consonants", Rep. No. 89, April, QPR, MIT, pp. 199-205.

Stevens, K.N. and House, A.S. (1955): "Development of a quantitative description of vowel articulation", J.Acoust.Soc.Am. 27, pp. 484-493.

Sundberg, J. (1969): "Articulatory differences between spoken and sung vowels in singers", STL-QPSR No. 1, pp. 33-46.

Questions J. S. LIENARD à M. LINDBLOM

Je souhaite poser deux questions à M. LINDBLOM. D'une part, le fait que la simulation du conduit vocal soit d'ordre électrique n'est-il pas une gêne pour rendre compte de certains phénomènes d'origine aérodynamique tels que les bruits de friction et d'explosion de certaines consonnes ? Que pensez-vous, d'autre part, de l'étude des différences entre locuteurs dans la mesure où la dimension du conduit vocal, sa configuration, et surtout ses mouvements varient sensiblement d'un individu à l'autre ?

Réponses

1°) The problems associated with the simulation of aerodynamic aspects of speech production are complex but in principle not insolvable.

2°) Differences between speakers provide interesting research topics. On constructing our articulatory model our aim has been to describe some of the degrees of Freedom of the articulatory system, we have not had the intention of creating a very general tool for the analysis of X-ray profiles from a large number of different talkers. It seems as if the Fourier model proposed by Johan Liljencrants for the description of tongue profiles is better suited for a quantitative analysis of this sort.

.../...

Questions R. CARRE à B. LINDBLOM

Est-il facile de déterminer les paramètres articulatoires à partir de la fonction d'aire ?

Les paramètres articulatoires peuvent-ils être plus intéressants dans des travaux de reconnaissance de parole que les formants par exemple ?

Réponses

1°) We have as yet rather limited experience of models based on articulatory parameters such as jaw, lips, larynx, etc. At least in our own group we have not undertaken any systematic studies of the problem to which you refer. However, we have done several experiments on compensatory articulation. The results of these studies indicate that in certain cases there appears to be no one-to-one relation between the area function and the articulatory parameters. What I have in mind is the finding that in spite of an abnormal externally controlled jaw position, subjects are capable of producing vowels with formant patterns corresponding very closely to those observed under normal conditions.

2°) In so far as the values of these parameters are more simply related to linguistic categories than to acoustic data an affirmative answer might perhaps be possible especially in the context of an analysis-by-synthesis scheme for recognition. Further research will be necessary however to make possible such "simplicity" considerations. This problem is associated with the question to what extent articulatory behavior is reorganized as a function of context in order to preserve the invariance of acoustic and perceptual events.

.../...

Question Mme LHOTE à M. LINDBLOM

Dans vos descriptions génératives, faites-vous intervenir les problèmes de coarticulation, et peut-on tenir compte de la coarticulation ?

Réponse

At present the quantitative phonetic description of coarticulation e.g., in terms of models of speech production, seems to be incompatible with the formalism developed with generative phonology for the description of phonological facts. Since it is well known that coarticulation is a characteristic not only of speech but also of phonological processes - synchronic as well as historical - it seems clear that an integration of the description of these different areas is an important desideratum in the study of language and speech.

TRANSITIONS ARTICULATOIRES ET TRANSITIONS ACOUSTIQUES DANS LA PAROLE REELLE

par

Laurent SANTERRE

Laboratoire de Phonétique

Université de Montréal

Canada

Avant d'aborder proprement l'étude des transitions acoustiques en rapport avec les transitions articulatoires, je vais établir une corrélation entre les mouvements articulatoires et les variations formantiques. Cette corrélation peut être symbolisée par le fonctionnement d'un modèle théorique simple de résonance acoustique et électronique qui nous permettra de prévoir les événements acoustiques de la parole réelle à partir de l'articulation, ou de déduire les mouvements articulatoires à partir de leurs effets acoustiques.

Cet exposé portera donc d'abord sur la synchronisation des films avec l'analyse sonographique, puis sur le fonctionnement du résonateur théorique modèle du canal buccal; nous pourrons ensuite étudier les rapports étroits qui existent dans la parole réelle entre les transitions articulatoires et leurs résultats acoustiques.

La corrélation dont je présente ici les grandes lignes a été établie à partir de l'analyse comparée sur les plans articulatoire et acoustique de 280 phrases de quatre syllabes prononcées par deux locuteurs francophones de Montréal.

Les radiofilms, de 48 images à la seconde en moyenne, sur le plan sagittal médian, ont été tracés image par image et mesurés à une dizaine de points à la fois. La prise de son a été faite au moyen d'un microphone protégé contre le bruit des appareils de prise de vue et elle a permis une analyse spectrale satisfaisante; le rapport signal-bruit y est de l'ordre de 40 dB.

La prise de vue et la prise de son se sont trouvées électroniquement synchronisées sur une bande magnétoscopique de circuit fermé de télévision. Cette synchronisation ne dispense pas de celle qu'il faut faire ensuite entre les séquences filmées et l'analyse spectrale. Cette dernière synchronisation consiste à faire coïncider étroitement la course du film avec celle des sonagrammes, de manière à faire correspondre chaque image du film avec la section sonagraphique qui représente la production acoustique à ce moment précis.

Pour réduire le coefficient d'erreur due aux fluctuations des vitesses de déroulements, j'ai pris comme points de repère des occlusives situées aux deux extrémités d'une même séquence sur le film et sur la bande enregistrée. Exemple: soit une séquence dont la première phrase est Donne un coup de râpe et la 10e, Les pêches sont vertes; entre l'explosion du [d] et celle du [p] de la dernière phrase, je compte 800 images sur le film et, au moyen d'un segmentateur, 317,5 cm sur la bande magnétique. La vitesse exacte de déroulement de bande du magnétophone est facile à préciser au moyen d'un chronographe; il suffit de mesurer la bande enregistrée en 100 ou 200 secondes et de diviser le nombre de millimètres par le nombre de centisecondes; dans l'exemple qui nous occupe, j'ai trouvé 190 mm/sec. Les 3175 mm de bande entre [d] et [p] ont été enregistrés en 16,71 secondes, ce qui donne une course de 1,33 mm d'enregistrement par centiseconde. Dans ce cas, le temps d'une image de film, soit 2,089 cs, correspond à 2,77 mm sur le sonagramme. Il faut donc faire correspondre chaque image du film avec des segments de 2,77 mm sur le sonagramme.

ORDRE DES MESURES ARTICULATOIRES

1. projection de la lèvre inférieure;

2. projection de la lèvre supérieure;

3. distance entre les lèvres;

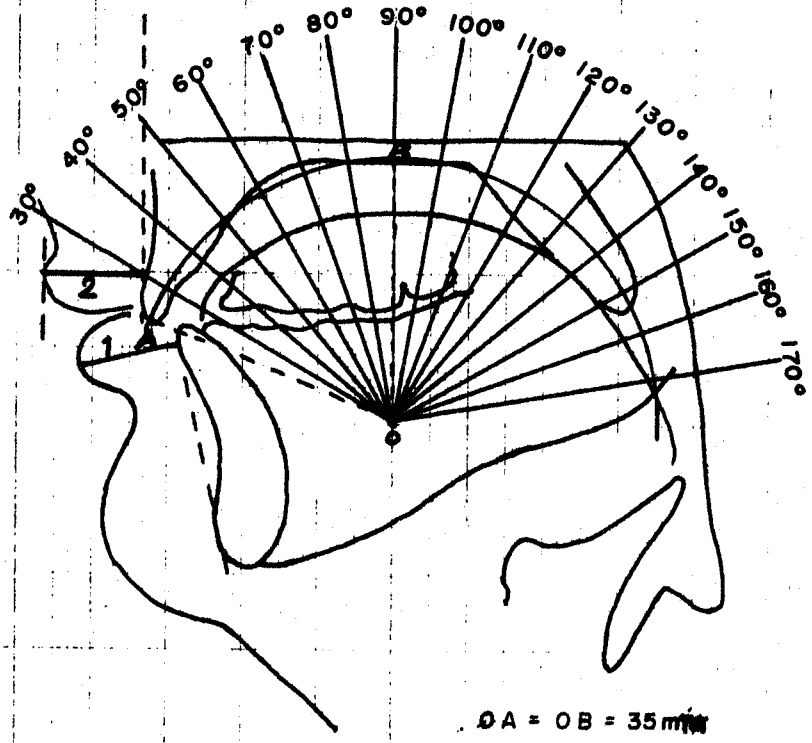
4. distance entre les incisives;

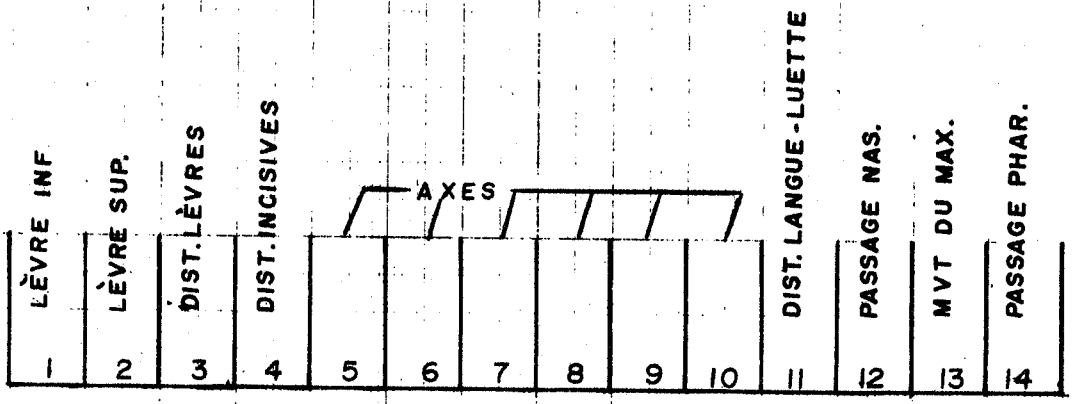
5 à 10. distances de la langue aux incisives, aux alvéoles, au palais dur, au voile du palais, sur des axes déterminés par leur angle;

11. distance de la langue à la luette;

12. ouverture des fosses nasales;

14. distance de la langue à la paroi pharyngale.


30°
40°
50°
60°
70°
80°
90°
100°
110°
120°
130°
140°
150°
160°
170°
1
2
3
OA = OB = 35 mm


AXES
1 LÈVRE INF
2 LÈVRE SUP.
3 DIST. LÈVRES
4 DIST. INCISIVES
5
6
7
8
9
10
11 DIST. LANGUE-LUETTE
12 PASSAGE NAS.
13 MVT DU MAX.
14 PASSAGE PHAR.

Les mesures sont exprimées en millimètres et correspondent aux 2/3 des dimensions réelles.

Dans la synchronisation de chaque sonagramme, on est guidé par les éléments transitoires comme les explosives; une fois fixée une image du film sur sa tranche spectrale, tout le reste de la phrase se trouve synchronisée. D'ordinaire d'un bout à l'autre de la phrase, les contraintes imposées par les distances entre les mouvements articulatoires et les transitoires acoustiques ne laissent pas plus de 1 cs de liberté dans la synchronisation d'une durée de 2 secondes.

Cette méthode permet de suivre point par point les mouvements articulatoires et les variations formantiques correspondantes. L'analyse de l'ensemble du corpus m'a permis de dégager les constantes de la corrélation articulatoire et acoustique, dont je me contenterai de rappeler plus loin les grandes lignes pour vérifier la validité d'un modèle théorique de résonance électroacoustique qui permet de schématiser la corrélation.

Résonateur mécanique

J'appelle résonateur mécanique tout résonateur dont le siège des vibrations est une colonne d'air, comme dans les tuyaux d'orgue ou la bouche, ou un support matériel, comme une corde d'instrument de musique. La fréquence du résonateur est déterminée par ses caractéristiques physiques.

La durée d'une oscillation est la période (T) et le nombre de périodes par seconde est la fréquence (F) exprimée en Hertz (Hz). La formule suivante rend compte des lois qui régissent le temps d'oscillation et partant, la fréquence d'un système mécanique.

$$T = 2\,\Pi\sqrt{\frac{m}{c}} \qquad F = \frac{1}{2\,\Pi\sqrt{\frac{m}{c}}}$$

Remarque: Cette formule très générale ne tient pas compte d'une caractéristique nécessaire de tout système oscillant non théorique, i.e. la résistance à l'oscillation. Dans le résonateur de Helmholtz, elle peut être représentée par le frottement de l'air à l'embouchure du résonateur. Cette résistance est d'autant plus grande que l'embouchure est longue et étroite, comme un goulot de bouteille. Didier (1) représente la conductivité (K) de l'embouchure - qui est l'inverse de la résistance (R) - par la formule suivante:

$$K = \frac{S}{L}$$

où S = section de l'embouchure et
L = longueur de l'embouchure.

Si la conductivité grandit, l'oscillation est plus libre de se produire et sa période est plus courte, donc la fréquence des oscillations augmente; au contraire, si K diminue, la fréquence s'abaisse.

1 Didier (A), Reproduction des sons et des images, Tome I, Masson, Paris 1964, p. 54.

On peut résumer les lois mécaniques des systèmes oscillants comme suit: la fréquence est directement proportionnelle à la racine carrée de l'énergie interne (c) qui l'alimente, inversement proportionnelle à la racine carrée de la masse (m) mise en vibration, directement proportionnelle à la section (S) de l'embouchure, et inversement proportionnelle à la longueur (L) de l'embouchure du résonateur.

Le rapprochement est facile à faire avec le conduit vocal: sa fréquence de résonance propre augmente quand le volume d'air qu'il contient diminue; elle augmente aussi quand l'énergie qui entretient l'oscillation augmente; elle s'abaisse, par contre, quand l'embouchure se referme et quand les lèvres s'avancent. Je reviendrai plus loin sur le parallèle entre le résonateur buccal et un résonateur théorique électrique.

Système oscillant électrique

Un circuit électrique oscillant obéit aux mêmes lois que le résonateur mécanique.

Sa fréquence de résonance est inversement proportionnelle à la racine carrée de la masse des électrons mis en oscillation, et directement proportionnelle à la racine carrée de l'énergie interne du système.

La formule suivante résume ces lois:

$$F = \frac{1}{2\,\Pi\sqrt{\frac{i}{V}}}$$

i = quantité de courant;

V = voltage ou pression.

(On reconnaît la formule précédente où m est l'équivalent de i, et c l'équivalent de V).

Cette formulation souligne le parallélisme entre les résonateurs mécaniques et les résonateurs électriques, mais elle cache les rapports que les éléments i et V entretiennent entre eux dans l'oscillation électrique. Les circuits oscillants électriques font appel, en général, à deux composantes: l'une est l'inductance (L) d'une bobine, et l'autre la capacitance (C) d'un condensateur. C'est dans ces deux composantes que se distribuent l'énergie et la masse (c et m, ou V et i) dont il vient d'être question.

La formule usuelle des circuits oscillants électriques est la suivante:

$$F = \frac{1}{2\,\Pi\sqrt{LC}}$$

L'inductance (L) est exprimée en henry (H), et la capacitance (C) l'est en farad (fd). Pour les généralités sur ces valeurs, je renvoie à l'aide-

mémoire Technor, pp. 45 à 52 (1) et au Dictionnaire de Piraux (2).

Pour poursuivre la comparaison avec le résonateur mécanique, on peut dire que la masse est ici constituée par la quantité d'électrons en mouvement dans L et C; l'énergie est représentée par le voltage qui fait passer alternativement ces électrons dans L et dans C.

On voit par la formule qu'on peut faire varier L et C sans changer la fréquence, pourvu que leur produit soit constant. (Il en sera ainsi pour des positions articulatoires différentes qui peuvent conserver la même fréquence par compensation.)

Ces formulations simplifiées des oscillations acoustiques et électriques ont fait ressortir leurs analogies, mais on ne peut s'en tenir là, si l'on veut appliquer les analogues physiques et électriques aux phénomènes de la parole. Il faut maintenant étudier de plus près ce qui se passe, aussi bien dans un résonateur acoustique que dans un résonateur électrique.

Le résonateur simple de Chiba et de Kajiyama

Fant (1968, p. 217) reprend à Chiba et Kajiyama la distribution des noeuds et des ventres de tension et de courant dans le canal buccal, selon qu'il résonne sur le premier, le deuxième ou le troisième formant.

On sait qu'un formant est un groupe de fréquences voisines, harmoniques ou partielles, qui sont amplifiées par la résonance propre du résonateur, tandis que les autres harmoniques sont absorbés ou affaiblis.

Chiba et Kajiyama en 1941, Fant en 1960, puis Flanagan en 1965, sont partis du fait que le canal vocal, d'une longueur approximative de 17,5 cm, résonne à des fréquences simultanées de 500 Hz, 1500 Hz et 2500 Hz environ, quand il est laissé entrouvert et en position de repos. Ce tuyau de 17,5 cm est bouché du côté de la glotte, et entrouvert, donc, du côté des lèvres, ce qui occasionne un maximum de courant aux lèvres et un minimum de courant à la glotte. En termes d'électronique, on dit que l'impédance est maximale à la glotte et minimale aux lèvres. On sait que là où le courant est maximal, le voltage est minimal, et que là où le courant est minimal, le voltage est maximal.

Un résonateur dont un bout est ouvert et l'autre fermé, résonne de telle sort que le minimum de pression se trouve du côté ouvert, là où l'air est libre et non retenu; à cet endroit, l'air à la pression atmosphérique a toute liberté de vibrer en amplitude; on dit qu'il y a maximum de courant; à l'autre extrémité du résonateur, l'air n'est pas libre, mais soumis à un maximum de pression, et le courant vibratoire y est à son point minimal.

---

1 M. Bibal et P. Heiny, Electronique appliquée, Delagrave, éd. de 1968.

2 Henry Piraux, Dictionnaire général d'Acoustique et d'Electroacoustique, Eyrolles, Paris, 1964.

Ce déphasage de 180° entre la pression, ou le voltage, et le courant se retrouve à tous les points du résonateur. Le conduit vocal est obligé de résonner à des fréquences inversement proportionnelles au carré du volume d'air qu'il contient, et de telle sorte qu'à tout moment il y ait aux lèvres un maximum de courant et un minimum de pression par rapport aux autres points.

Dans le cas qui nous occupe, i.e. celui du conduit vocal résonnant à ses fréquences propres de 500 Hz, 1500 Hz et 2500 Hz, le résonateur de 17,5 cm est parcouru par trois sortes d'ondes qui sont dans des rapports bien définis avec lui. On sait que la longueur d'onde est égale à la vitesse de propagation du son, divisée par la fréquence:

$$\lambda = \frac{V}{F} \qquad (V = \text{vitesse})$$

Pour 500 Hz, la longueur d'onde est:

$$\lambda = \frac{35\ 000 \text{ cm/sec}}{500} = 70 \text{ cm}$$

(La vitesse du son dans l'air est d'environ de 350 m/sec.) Le conduit de 17,5 cm ne représente qu'un quart de la longueur de l'onde (1/4λ).

Pour 1 500 Hz, λ = 23 1/3 cm; ici, le conduit vocal de 17,5 représente les 3/4 de λ;

Pour 2 500 Hz, λ = 14 cm; le conduit de 17,5 cm fait 5/4 de λ.

On peut voir que le résonateur a pour fréquences propres toutes celles pour lesquelles 17,5 cm représentent un nombre entier et impair de quarts de longueur d'onde. Un nombre pair de quarts de longueur d'onde ferait coïncider un maximum de pression avec le côté ouvert du résonateur, ce qui est impossible. Le conduit vocal théorique peut donc résonner sur toutes les fréquences pour lesquelles 17,5 cm égalent 1/4 ou 3/4 ou 5/4 ou 7/4 ou ... 9/4 de la longueur d'onde (λ). Il faut remarquer que cette régularité de rapport entre les fréquences propres d'un résonateur n'est pas nécessaire; elle tient ici à la régularité du modèle théorique. Les trois résonances considérées comme des formants feraient entendre un schwa à la synthèse. En pratique, le schwa, si on pouvait le mesurer, n'aurait pas nécessairement ces fréquences de 500 Hz et 1500 Hz; il peut tout de même servir de point de comparaison pour les autres fréquences de résonance et pour les autres positions articulatoires.

Le résonateur dont j'ai décrit le fonctionnement se comporte comme un tuyau d'orgue bouché. Tous les résonateurs que nous étudierons maintenant conserveront cette particularité d'être ouverts à un bout, mais présenteront les changements suivants: variation de l'ouverture, de la longueur, du volume d'air et surtout de la forme de la colonne d'air, selon les positions de la langue. Il faut donc voir dans quel sens toutes les modifications du résonateur font varier ses fréquences propres, i.e. les formants de voyelles.

Variations des caractéristiques du résonateur buccal

1. La masse ou le volume d'air

On a vu, par la formule $F = \frac{1}{2\pi\sqrt{\frac{m}{c}}}$, que la fréquence s'abaisse quand le volume augmente. A partir des films, on ne peut avoir qu'une approximation des variations de volume du canal buccal, puisqu'ils ne permettent pas des mesures sur les deux plans que comporte un volume. Cependant, il est possible d'avoir des précisions sur le volume à partir des autres variables, en raison de la fréquence mesurée. Je ne m'attache pas à faire la corrélation des fréquences de formants avec le volume du résonateur buccal, comme l'a fait Fant; je la ferai avec les fonctions articulatoires. Cette dernière démarche me paraît plus facile et plus sûre, en raison du grand nombre de réalisations filmées dont je dispose.

2. Longueur du résonateur

On a vu qu'une longueur de 17,5 cm conditionnait des fréquences de 500, 1500 et 2500 Hz. Voyons le cas d'un résonateur de 18 cm obtenu par projection labiale.

Si le quart de longueur d'onde est 18 cm,

$$\lambda = 72 \text{ cm}$$

$$72 = \frac{35\,000}{F}$$

$$F = 486 \text{ Hz}$$

la fréquence est donc inversement proportionnelle à la longueur du résonateur.

Dans l'articulation, la longueur du résonateur est mesurée par la distance entre la glotte et la sortie des lèvres. Je n'ai pas tenu compte des variations de longueur occasionnées par les positions du larynx; comme Fant, je m'en suis tenu à la projection labiale qui, elle, est une articulation volontaire et pertinente.

3. Variations à l'embouchure

Le fait d'entraver la libre circulation de l'air là où se produit un maximum de courant a pour effet d'abaisser la fréquence. Joos (1) a fait remarquer qu'une bouteille a une fréquence de résonance beaucoup plus basse qu'un tuyau d'orgue bouché de mêmes dimensions. Si l'on ferme complètement toute issue, la fréquence descend très bas, pour la même dépense d'énergie; la résonance devra même cesser, si la pression à l'intérieur empêche les molécules d'air de vibrer. On peut donc formuler une règle comme suit: la fréquence d'un résonateur buccal est [redacted] proportionnelle à la section de l'ouverture labiale.

---

1 Martin Joos, Acoustic Phonetics, Supplement to Language, vol. 24, no 2, 1948.

4. Energie

Selon la formule connue:

$$F = \frac{1}{2\pi\sqrt{\frac{m}{c}}}$$

la fréquence augmente en proportion de la racine carrée de l'énergie en jeu dans la vibration. Ochiai (1) a fait prononcer des voyelles à trois intensités différentes et a comparé les fréquences des formants. Il a constaté que le spectre était peu affecté par les différents niveaux d'intensité.

En effet, il ne faut pas confondre la force extérieure qui met en branle un système oscillant et entretient l'oscillation, avec l'énergie interne du phénomène. Cette dernière est la force qui se distribue, selon les positions de la masse (m) dans le cycle vibratoire, en énergie cinétique et en énergie potentielle. Dans un système mécanique, elle est représentée par la force du ressort qui ramène la masse déplacée de son point d'équilibre; la plus ou moins grande extension du ressort ne change pas la fréquence de la vibration, mais son amplitude. Il faudra revenir plus longuement sur cette force interne répartie alternativement en énergie cinétique et en énergie potentielle, pour expliquer le fonctionnement des systèmes électriques composés de bobines (L) et de condensateurs (C). Le fait de jouer fort d'un instrument de musique ne change pas la fréquence des notes. Même en l'absence d'énergie extérieure, un système conserve sa fréquence propre déterminée par l'élasticité de la matière vibrante. Les oscillations d'un diapason ne changent pas de fréquence pendant qu'elles s'amortissent; seule leur amplitude diminue.

5. Variations de fréquences causées par les déformations du résonateur

Il reste à voir comment les fréquences propres du résonateur buccal sont amenées à varier quand la langue quitte la position du repos du schwa que nous avons considérée.

Tout déplacement de la langue a des incidences sur la forme du résonateur et sur ses fréquences de résonance. Par analogie avec le système électrique, on parlera d'inductance et de capacitance, et par analogie avec le système mécanique, il s'agira d'énergie cinétique et d'énergie potentielle.

Fant (1968, p. 216) résume le phénomène en disant qu'une constriction en un point de courant maximal abaisse la fréquence propre du résonateur, tandis qu'une constriction en un point de voltage maximal a pour effet de relever la fréquence. On pourrait préciser, pour tous les points du résonateur théorique, les rapports du courant et de la pression ou du voltage, selon les fréquences; il sera plus intéressant de le faire sur les schémas articulatoires tirés de mes films. Pour l'instant il faut voir comment ces variations de courant et de pression sonique se trouvent symbolisées dans les termes du modèle électrique.

---

1 Yoshiejuki Ochiai, "Recherches sur les voyelles françaises au point de vue du timbre, phonétique et vocal", in Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung, Band 18, Heft 3, 1965, pp. 243-79.

Nous avons vu que la fréquence d'un système à inductance et à capacitance est soumise aux relations de la formule:

$$F = \frac{1}{2\pi\sqrt{LC}}$$

La fréquence est inversement proportionnelle à la racine carrée de la valeur LC, i.e. au produit de l'inductance par la capacitance. L'inductance, qui est le fait d'une bobine d'auto-induction parcourue par un courant variable, représente la vélocité de l'air en vibration dans le modèle acoustique; ses valeurs s'expriment en henry (H). La capacitance, qui est le propre d'un condensateur qui emmagasine des électrons sur ses plaques à des niveaux de potentiel différents, représente la pression dans le modèle acoustique; ses valeurs s'expriment en farad (fd) ou microfarad (mfd).

Comme pour le courant et le voltage dans un circuit, dont la résistance au phénomène est théoriquement sensée ne pas varier, l'inductance grandit en tout point quand la capacitance diminue; et vice versa, l'inductance diminue quand la capacitance grandit. Certains points sont le siège de grandes valeurs d'inductance et de petites valeurs de capacitance, d'autres sont le lieu de petites valeurs d'inductance et de grandes valeurs de capacitance; on les appelle points où prédomine l'inductance, et points où prédomine la capacitance.

Pour une fréquence donnée, le produit LC est toujours le même, c'est-à-dire que, si en un point les rapports sont de 20 L pour 5 C à un instant du cycle vibratoire, à l'instant suivant au même point, les valeurs pourront être de 10 L et 10 C; le produit LC égale toujours 100, mais on voit que la loi de la relation inverse est respectée, même quand l'inductance, ou la capacitance, est prédominante en un point.

Si la fréquence change, le produit LC n'est plus le même. Prenons un exemple: soit

$$L = 1 \text{ henry}$$

$$C = 1 \text{ microfarad (mfd)}$$

$$F = \frac{1}{2\,\pi\sqrt{.000001 \times 1}}$$

$$= \frac{.159}{.001} = 159 \text{ Hz}$$

Si L égale 100 fois moins que dans l'exemple précédent, soit .01:

$$F = \frac{.159}{.000001 \times .01}$$

$$= \frac{.159}{.0001} = 1590 \text{ Hz}$$

La fréquence est multipliée par 10 quand l'inductance est divisée par 100. Autre exemple:

Si L = .01 et C est quadruplé,

$$F = \frac{.159}{.000004 \times .01}$$

$$= \frac{.159}{.0002} = 795 \text{ Hz}$$

La fréquence est donc inversement proportionnelle à la racine carrée de l'inductance ou de la capacitance du système.

Il reste à voir comment L et C varient selon les constrictions du résonateur. Fant (1968, p. 317) résume les phénomènes en disant que toute constriction augmente L et diminue C, dans des proportions qui tiennent à la prédominance en énergie cinétique ou en énergie potentielle au point touché.

On peut expliciter cette règle comme suit:

1. Si cette constriction augmente beaucoup L et diminue peu C, la fréquence est amenée à baisser pour rétablir l'équilibre du système;

2. Si la constriction augmente peu L et diminue beaucoup C, la fréquence est amenée à monter pour rétablir l'équilibre;

3. Si la constriction augmente L dans la proportion où elle diminue C, la valeur de LC n'est pas changée, et la fréquence n'est pas affectée.

Autrement dit:

1. La constriction fait beaucoup varier L, i.e. l'augmente beaucoup, quand elle est faite en un point où le courant prédomine sur le voltage; alors C varie peu; la fréquence baisse;

2. La constriction fait beaucoup varier C, i.e. la diminue beaucoup, si elle est faite en un point où le voltage prédomine sur le courant; alors c'est L qui est peu affectée; la fréquence monte;

3. La constriction fait varier L et C en un point où le courant et le voltage sont de valeur égale; alors la fréquence ne change pas.

J'ai distingué plus haut entre l'énergie interne du système et l'énergie extérieure qui entretient l'oscillation et lui donne son amplitude sans affecter sa fréquence. J'ai alors dit que l'énergie (c), dans la formule $F = \frac{1}{2\pi\sqrt{\frac{m}{c}}}$ prenait la forme d'énergie cinétique et d'énergie potentielle au cours du cycle vibratoire. Le moment est venu d'expliciter ces facteurs et de voir comment ils conditionnent la fréquence du système.

L'énergie cinétique $w = \frac{1}{2} Li^2$; (w = énergie)

L'énergie potentielle $w = \frac{1}{2} CV^2$

On voit par ces formules que l'énergie cinétique est liée au courant (i) dans la bobine; elle est proportionnelle à la valeur d'inductance (L) et au carré du courant (i^2); l'énergie potentielle est liée au voltage (V) aux bornes du condensateur; elle est proportionnelle à la valeur de capacitance (C) et au carré du voltage (V^2).

Les variations de l'énergie cinétique et de l'énergie potentielle sont, comme celles du courant et du voltage, à tout moment et en tout point du résonateur, inversement proportionnelles et déphasées de 180°.

En certains points du résonateur, l'énergie cinétique ($\frac{1}{2} Li^2$) prédomine, comme le courant, sur l'énergie potentielle ($\frac{1}{2} CV^2$); en d'autres points, $\frac{1}{2} Li^2 = \frac{1}{2} CV^2$; en d'autres points enfin, l'énergie potentielle l'emporte sur l'énergie cinétique, comme le voltage sur le courant.

Règles:

1. Abaisser ou entraver le courant par constriction ou résistance équivaut à augmenter la valeur de L, puisque le fait de l'inductance est de s'opposer au courant; la formule de l'énergie cinétique en rend d'ailleurs compte:

$$w = \frac{1}{2} Li^2 ,$$

$$L = \frac{2w}{i^2} ;$$

si donc L grandit, la fréquence s'abaisse, selon la formule $F = \frac{1}{2 \pi \sqrt{LC}}$

2. Une constriction en un point où le voltage prédomine sur le courant, entrave très peu le courant, puisqu'il y est faible, mais ajoute à la pression déjà haute; augmenter le voltage équivaut à diminuer C selon la formule de l'énergie potentielle:

$$w = \frac{1}{2} CV^2 ,$$

$$C = \frac{2w}{V^2} ;$$

si donc C diminue, la fréquence augmente, comme le prévoit la formule.

A partir de ces considérations techniques, on peut prévoir l'effet de constriction de la langue dans le résonateur buccal, selon qu'elle se fera en un point où domine le courant ou le voltage, ou encore selon que le point touché comportera une prédominance en énergie cinétique ou en énergie potentielle.

Les avantages de cette comparaison entre les résonateurs mécanique, acoustique et électrique ont été d'expliciter les lois générales du fonctionnement de tout système oscillant et de faire apparaître le fondement des recherches actuelles effectuées au moyen des analogues.

Bien entendu, il ne peut être question d'assimiler le canal vocal à un tuyau de 17,5 cm de longueur; mais il n'est pas interdit de penser que les nombreux paramètres du profil du canal buccal que je tire des films puissent déterminer de façon assez précise les valeurs des variables LC d'un système électrique pour permettre éventuellement une reconstitution par synthèse des faits de parole que j'étudie.

Vérification du résonateur théorique

On peut maintenant faire le rapprochement entre le résonateur théorique (acoustique ou électrique) et le résonateur buccal.

La Fig. 1 représente un modèle, réduit aux deux tiers, du canal buccal de S.P. (sujet populaire) et un résonateur théorique de 12 cm (18 cm x 2/3). Il y a correspondance entre les axes, qui servent de points de repère pour les mesures articulatoires, et les numéros d'ordre sur le résonateur théorique. Ainsi on peut voir, par exemple, que 4 est une zone où, pour la résonance de F1, le courant est plus grand que la pression, et où, pour la résonance de F2, le courant est minimal et partant le voltage maximal. Pour tous les points de rétrécissement du canal buccal considéré comme résonateur théorique, on peut prévoir et expliquer les variations de fréquences pour F1 et F2.

Il est bien entendu que l'articulation réelle ne se limite pas au rétrécissement du canal buccal en un seul point et que plusieurs paramètres varient en même temps; il faut donc tenir compte de plusieurs points de constriction ou de déplacement en même temps.

L'orientation des flèches au bas de la Fig. 1 indique en tous points le sens des variations fréquentielles pour F1 et F2, au moment du rétrécissement du résonateur. On voit que F1 descend rapidement au point 1, moins rapidement à 4, ne bouge plus à 6, puis commence à remonter à 7, et monte enfin de plus en plus rapidement à mesure qu'on approche des points postérieurs extrêmes. F2 baisse à 1, est indifférent à 2, monte beaucoup à 4, est de nouveau indifférent à 6, puis redescend ensuite.

Les points neutres sont particulièrement intéressants pour comprendre que tout déplacement de la langue n'entraîne pas nécessairement des variations de résonance importantes et que, pour un mouvement, un formant peut varier tandis que l'autre ne change pas de fréquence.

Comme les mouvements articulatoires affectent toujours plusieurs zones à la fois, il y a lieu de voir si les variations s'additionnent ou s'annulent. Prenons comme exemple les zones touchées par l'articulation successive des voyelles antérieures de [ɛ] à [i]. La langue s'élève sur les axes de 70° et 80° et dans les zones adjacentes; une constriction sur ces axes fait monter F2 et descendre F1; d'un autre côté, l'agrandissement du canal pharyngal aux zones situées autour de l'axe 160° a pour effet de faire monter F2 et descendre F1: les variations articulatoires s'additionnent donc pour faire des voyelles antérieures des timbres à formants de plus en plus distants en passant de [ɛ] à [i]. (Fig. 2).

Voyons le cas des voyelles postérieures [ɔ], [o] et [ʊ] (Fig. 2). La masse de la langue se déplace vers des zones situées autour de l'axe 130°. F2 descend et F1 monte, par comparaison avec le schéma de [ə] pris comme point de

départ. Le dégagement sur les axes antérieurs fait aussi baisser F2 et monter F1; en s'élevant pour [o] et [ʊ], la langue libère le passage pharyngal inférieur, ce qui fait baisser F1; la constriction s'accentue sur les axes 120°, 130°, 140°, ce qui fait baisser F2 encore davantage. Cette tendance est fortement accentuée par l'action des lèvres qui, en s'allongeant et en se fermant, font baisser toutes les fréquences. F1 est plus abaissé par l'allongement du canal buccal et la fermeture antérieure qu'il n'est relevé par le rétrécissement vélaire. On aura l'occasion de le vérifier pour l'articulation de [R] non labialisé. D'ailleurs, autour de 130°, F1 commence à peine à manifester une tendance à s'élever; du côté des lèvres, les variations sont beaucoup plus marquées que du côté de la glotte où l'impédance ne peut pas varier beaucoup par ouverture et fermeture. Les ouvertures de la glotte sont minimes comparées à celles des lèvres, et la pression sous-glottique pour les voyelles est grande, comparée à celle de la cavité buccale.

Les [a] de S.P. s'expliquent par la grande ouverture; par rapport au point de comparaison [ə], les axes antérieurs sont un peu dégagés, ce qui fait monter F1 et baisser F2. De [a] à [ɑ], il y a recul de la langue sur les axes pharyngaux, d'où abaissement de F2 et relèvement léger de F1; mais ce relèvement de F1 est contratié et renversé par la fermeture et la projection labiales, (Fig. 2).

Corrélation entre les mouvements articulatoires et les variations formantiques

Comme je l'ai précisé plus haut, il ne s'agira pas d'une corrélation entre les volumes du résonateur buccal et les fréquences formantiques, mais d'une mise en rapport des variations formantiques avec les mouvements articulatoires mesurés sur le plan sagittal médian.

A partir d'un grand nombre de réalisations filmées et étudiées au sonagraphe, on devrait pouvoir dégager des constantes qui permettent de prévoir et d'expliquer les principaux traits de l'articulation et de la production acoustique de la parole réelle. On pourra ainsi mieux décrire un fait articulatoire en s'appuyant sur les données acoustiques correspondantes, ou au contraire, mieux identifier dans un spectre complexe ce qui doit être retenu comme formants, en se guidant cette fois sur les positions articulatoires. En raison de leur caractère complémentaire, les deux approches, synchronisées et éclairées l'une par l'autre, devraient permettre de combler une partie de leurs lacunes respectives.

Une corrélation doit être établie sur des ensembles assez importants, faute de quoi elle risquerait d'être une juxtaposition de faits sans valeur explicative; celle que je présente ici est le résultat de centaines de comparaisons; parmi les cas dont je fais état dans le texte, certains se trouvent illustrés dans la documentation (1).

Je souligne que cette corrélation n'a rien de théorique; elle repose sur l'observation de faits réels, articulatoires et acoustiques; il y aura lieu

---

1 Pour une documentation complète, voir Laurent Santerre, Les voyelles orales dans le français parlé à Montréal, Klincksieck, Paris 1972, (sous presse).

de voir si les constantes qui s'en dégagent sont de nature à infirmer ou à confirmer les données du résonateur théorique, acoustique ou électrique. Il est peu probable que la théorie elle-même des résonateurs soit prise en défaut; on est plutôt en droit de penser que les lois générales de la résonance s'appliquent partout, aussi bien dans la parole qu'en musique et en électronique; mais on pourra peut-être découvrir que le résonateur buccal est plus complexe qu'on n'avait cru, et que certaines données encore inconnues, par exemple celles des volumes, empêchent de voir comment s'applique la théorie dans les détails. Il est possible, par ailleurs, que la théorie attire l'attention sur des faits qui auraient pu échapper sans elle; des cas obscurs peu nombreux devraient rester sans explication; ils ne sauraient remettre en question toute la théorie et les moyens d'analyse et de mesure.

Du point de vue méthodologique, je dois dire que la démarche que j'adopte pour la présentation est l'inverse de celle que j'ai suivie dans la recherche. J'ai d'abord dégagé la corrélation, et c'est ensuite seulement que je l'ai vérifiée au moyen de résonateurs théoriques. Il s'est trouvé que les deux méthodes concordaient sans susciter de difficultés. Pour la présentation des résultats, j'ai cru qu'il y avait avantage à partir de la théorie, dont les grandes lignes sont simples, pour aller vers les faits réels, beaucoup plus complexes.

L'ordre de présentation est celui des divers points d'articulation, des lèvres au pharynx. Je prends mes exemples de préférence chez S.P. pour illustrer les grandes variations formantiques, et chez S.C. (sujet cultivé) pour montrer le jeu des compensations articulatoires qui empêchent la diphtongaison.

On remarquera, enfin, que je ne m'en suis pas tenu à ce qu'on appelle les cibles articulatoires et acoustiques des voyelles; ces dernières varient considérablement sous l'effet de la coarticulation consonantique; j'ai donc été amené à examiner les positions et les transitions des consonnes, aussi bien sur le plan articulatoire que sur le plan acoustique; les variations de F2 sont particulièrement éclairantes pour juger de ce qu'on a voulu définir comme les locus fixes des consonnes et les transitions de voyelles dans la parole synthétique.

Remarques générales:

1. Quand la consonne précède la voyelle (CV), la transition entre la consonne et la voyelle est explosive ou de détente; je trouve commode de parler de transition de départ des formants sur les sonagrammes, ou encore du point de départ de la transition; on pourrait la symboliser par TR 1 ex. ou TR 2 ex., selon qu'il s'agit du formant 1 (F1) ou du formant 2 (F2). Si la consonne est en position implosive (VC), je parle de transition d'arrivée; on peut la désigner par TR 1 im. ou TR 2 im.

2. Sur le sonagramme, toute transition, explosive ou implosive, est faite de gauche à droite; si elle monte en fréquence, elle sera dite ascendante et symbolisée par une flèche ascendante (C↗V ou V↗C); si, au contraire, elle descend en fréquence, elle sera dite descendante et représentée par une flèche descendante (C↘V ou V↘C); les transitions qui ne comportent pas de variation de fréquence évidente seront marquées par une flèche horizontale (C→V ou V→C). Je préfère cette manière de présenter les choses aux expressions de "transitions positives" ou "transitions négatives", qui peuvent prêter à confusion en raison des phénomènes inverses que sont les explosions et les implosions.

3. Je donnerai la fréquence de départ ou celle d'arrivée de la transition de F2 (TR 2), et non pas celle de TR 1; TR 2 est d'ordinaire plus facile à mesurer que TR 1.

4. Il s'agit d'ordinaire de syllabes sous l'accent.

5. Pour les constrictives, je parlerai de détente ou de relâchement plutôt que d'explosion.

6. J'examine les transitions 1 ou 2, d'abord en position de départ ou explosive (TR 1 ex. ou TR 2 ex.), puis en position d'arrivée ou implosive (TR 1 im. ou TR 2 im.); je tire les conclusions qui se dégagent séparément pour les transitions de F1 et pour les transitions de F2, quitte à reprendre ensuite les variations des deux formants pour un même point d'articulation.

LES LEVRES

Transitions de départ (position explosive), pour le formant 1

(TR 1 ex.)

P↗α (paille)

B↗œ (beurre)

m↗a/α (mal)

v↗a/ɛ (archevêque)

v↗ɛ (avec; verre)

Transitions d'arrivée (position implosive), pour le formant 1

(TR 1 im.)

ʊ→P (loupe)

α→m (l'âme)

α↘m (l'âme)

a→m (lame)

ɛ↘m (blasphème)

α→v (cadavre)

ɛ↘v (lèves)

ɔ↘v (rénove)

œ↘v (veuve)

Observations:

A l'examen, on remarque une différence entre les transitions d'arrivée et celles de départ; la position implosive, plus faible que la position explosive, entraîne des transitions moins nettes; il arrive très souvent que les consonnes finales de mot en franco-québécois, même sous l'accent, ne comportent que la phase implosive, l'explosion restant à peu près inaudible.

De plus, les sourdes ont des transitions de départ difficiles à observer, à cause du retard de sonorité qui les caractérise, surtout pour F1. Ce retard des cordes vocales a été maintes fois signalé; il semble le trait distinctif le plus stable entre les occlusives sourdes et les occlusives sonores, du moins en français et en danois (1). L'explosion de la forte [P] est plus rapide que celle de la faible [B] (2), surtout devant les voyelles très ouvertes. La raison pour laquelle la fréquence de départ pour F1 n'est pas très basse me semble la suivante: quand la résonance s'institue dans la bouche, les lèvres sont déjà trop ouvertes pour que la résonance du canal buccal soit encore abaissée par la fermeture labiale.

Durant la tenue de [B], la résonance, quand elle a lieu, peut descendre jusque vers 150 Hz (Fant, 1969, p. 13); à l'explosion, F1 monte vers la cible de la voyelle.

Au départ de [v], la transition 1 monte, et à l'arrivée sur la même consonne, elle descent. Mais la fermeture labiodentale est beaucoup moins prononcée que celle des occlusives bilabiales, de sorte que, si la constrictive n'est pas serrée, elle n'entraîne pas forcément l'abaissement de F1. La projection labiale a un effet abaisseur beaucoup plus marqué que la seule fermeture, comme on le verra plus loin.

L'occlusive labiale de [m] n'affecte à peu près pas F1, en position explosive comme en position implosive. Il faut croire que l'ouverture du résonateur nasal compense l'occlusion buccale et empêche la fréquence de F1 de s'abaisser. Les cas où F1 baisse devant [m] sont tous des diphtongues terminées par [ʊ] ou [i] qui ont des F1 très bas, et cela avant l'arrivée sur [m].

Conclusion:

Globalement, on peut dire, en termes de corrélation, que la fermeture labiale abaisse F1; la transition est bien marquée pour [B], elle l'est moins pour [P]; elle est aussi marquée pour [v]; quant à [m], il a peu d'influence sur F1.

---

1 Eli Fischer-Jørgensen, "Voicing, Tenseness and Aspiration in stop consonants, with special reference to French and Danish", in Annual Report III, 1969, Inst. Phon. Univ. Copenhague, pp. 63-114.

2 O. Fujimura, "Bilabial Stop and Nasal Consonants: a Motion Picture Study and its Acoustical Implications" in J. of Speech and Hearing Research 4, 1961, pp. 233-247.

TR 2 ex. Fréquences de départ :

P↗α	960 Hz	(Retourne la page)	
P→ɛ	1360 Hz	(La neige épaisse)	
B↗α	880 Hz	(Pas ce plombage-là)	
	1000 Hz	(Fais un plombage)	
B→α	1120 Hz	(Frappe à la base)	
B↗y	1400 Hz	(Une tête de buse)	
m→α	1080 Hz	(C'est un beau mâle)	
	1120 Hz	(C'est un beau mâle)	
m→α	1360 Hz	(Ca me fait mal)	
	1320 Hz	(Ca me fait mal)	
m↗e	920 Hz	(Faut pas chômer)	
	840 Hz	(Faut pas chômer)	
m↗i	2000 Hz	(La mise en scène)	sona*
v↗α	800 Hz	(Il est bavard)	
v→α	1080 Hz	(Il est bavard)	
v↗a	1200 Hz	(Elle est bavarde)	
	1160 Hz	(Elle est bavarde)	
v↗ɛ	960 Hz	(Un archevêque)	
	1200 Hz	(Un archevêque)	
	1320 Hz	(Tu viens avec)	
	1320 Hz	(Tu viens avec)	
	1360 Hz	(Faut que tu te réveilles)	
v↗i	1800 Hz	(Pendant sa vie)	
	1240 Hz	(Va pas trop vite)	

* Sonagramme seulement: l'abréviation "sona" indique que je n'ai pas fait le croquis des positions articulatoires, et que je m'en suis tenu au sonagramme.

TR 2 im. Fréquences d'arrivée :

a↘P	1280 Hz	(Appelle Thérèse)	sona
	1360 Hz	(Appelle Thérèse)	sona
	1200 Hz	(Apporte un câble)	sona
α↘P	1200 Hz	(Une râpe à fer)	sona
œ̃↘B	1280 Hz	(Un bon arôme)	sona
o→B	800 Hz	(Reviens à l'aube)	sona
a↘m	1240 Hz	(Aiguise la lame)	
	1320 Hz	(Aiguise la lame)	
α↘m	1120 Hz	(La mort dans l'âme)	
ɛ↘m	1360 Hz	(J'aime les fèves jaunes)	sona
	1360 Hz	(J'aime mieux la jaune)	sona
o↘m	720 Hz	(Un bon arôme)	sona
α↘v	1200 Hz	(C'est un cadavre)	
ɛi↘v	1840 Hz	(Y a des rêves fous)	
	1240 Hz	(Un rêve affreux)	
i↘v	1920 Hz	(Faut qu'on s'esquive)	

Observations:

L'explosion ou la détente des consonnes qui ont leur point d'articulation aux lèvres n'abaisse jamais la fréquence de F2, mais la laisse s'élever en libérant le résonateur buccal. A l'implosion, c'est le contraire qui se produit: la transition est descendante.

La langue n'étant pas directement engagée dans l'articulation des bilabiales et des labiodentales, les fréquences de résonance propres aux différentes voyelles ne se trouvent abaissées que par la fermeture antérieure; la différence de fréquences causée par cette fermeture, indépendamment de la projection labiale, paraît varier entre 0 Hz et 300 Hz; quand elle est supérieure à 300 Hz, c'est que la langue concourt à abaisser la résonance; quand elle est nulle, la raison s'en trouve ou bien dans la résonance nasale ou dans le retard de sonorité après les explosions sourdes, ou encore dans le fait que la résonance de F2 est trop basse pour s'abaisser davantage. J'ai pu constater que F2 ne pouvait pas descendre sous 650 Hz chez mes sujets.

Le retard de sonorité à l'explosion de [P] s'observe pour F2 comme pour F1; les transitions sont plus faciles à voir pour F2 que pour F1, parce que

les écarts y sont d'ordinaire plus prononcés.

[m] en position explosive influence peu le F2 des voyelles ouvertes; en position implosive, il l'abaisse davantage, sans doute à cause du retard de nasalité, quand il se produit.

Conclusions:

La fermeture labiale abaisse aussi bien F2 que F1, comme le laissait prévoir le résonateur théorique.

L'écart de fréquence imposé par les transitions n'est pas très grand, la langue restant libre de prendre et de conserver les positions articulatoires des voyelles, indépendamment des transitions.

Il ressort encore de ces observations que les consonnes bilabiales et labiodentales n'ont pas un locus acoustique qui serait leurs fréquences de résonance propres, mais qu'elles ne font qu'abaisser les résonances: on le voit clairement au fait que les fréquences de départ ou d'arrivée que j'ai relevées ici varient de 800 Hz à 2000 Hz selon la voyelle d'accompagnement.

On peut déjà penser que la parole synthétique serait de beaucoup améliorée si, au lieu de synthétiser les consonnes comme ayant des locus acoustiques fixes, on tenait davantage compte de la forme du résonateur imposée par la voyelle durant la constriction et au moment de l'explosion ou de l'implosion.

AXES 30° - 50°

[T], [d], [N], [s], [z]

TR 1 ex.

T→ɑ (Mort à la tâche)

(Il casse la tasse)

(Départ en retard)

(Monte à l'étage)

T→a (Monte à l'étage)

(De la potasse)

(Tu fais des taches)

D→ɑ (Une vraie tête d'âne)

(Une vraie tête d'âne)

S→ɑ (Il faut être sage)

S→a (Il faut être sage)

S↘ɛ (Une eau malsaine)

Z↘ɔ (Faut que tu les ôtes)

TR 1 im.

α→T (Aimes-tu les pâtes)

ɔ↘T (Ca sent l'azote) sona

(Manque de jugeotte) sona

a→T (Coupe-lui les pattes)

(Coupe-lui les pattes)

ɛ↘T (L'enfant qui tète)

ɛ→T (L'enfant qui tète)

ɛ↘D (Elle est trop laide)

ɛ→D (Elle est trop laide)

(Il veut de l'aide)

(J'ai peur qu'il cède)

(J'ai peur qu'il cède)

a→S (De la potasse)

ɛ↘S (D'une autre espèce)

Observations:

Les transitions sont montantes à l'explosion ou à la détente des consonnes à ces points d'articulation. On peut penser que, puisque le F1 des voyelles ouvertes remonte, une constriction ou une occlusion dans cette zone ferait baisser la fréquence de cette résonance.

Pourtant, il faut y regarder de plus près; en effet, les mêmes consonnes en position implosive ne font pas descendre TR 1. A l'examen, on constate qu'il y a beaucoup de différences dans les positions articulatoires entre l'explosion et l'implosion d'une consonne prononcée avec la même voyelle. Il suffit de comparer la forme et la hauteur de la langue dans la syllabe ... tète (Fig. 3); on voit qu'au moment de l'explosion du premier [T], la distance de la langue au palais dur sur l'axe de 90° n'est que de 4,5 mm à cause du [i] qui précède, tandis qu'elle est de 8 mm au moment de l'implosion du [T] final. Un autre fait décisif qui s'ajoute à la distance sur 90° est le passage pharyngal; le [T] explosif comporte une distance de 9 mm sur l'axe 180°, (mesure 14), tandis que le [T] implosif n'a plus que 6 mm. Si l'on se reporte au résonateur théorique, on constate qu'une déconstriction sur 90° et une constriction sur 180° font toutes deux monter F1. Ces deux actions conjuguées annulent et même renversent l'effet de la constriction au

point d'articulation consonantique.

Conclusion:

La constriction du canal buccal dans la zone située autour de l'axe de 40 abaisse F1 mais n'est pas le seul élément déterminant dans la direction des transitions; les positions articulatoires imposées par la voyelle conditionnent à tout moment les fréquences du résonateur.

TR 2 ex. Fréquences de départ :

T→α	1360 Hz	(Il casse la tasse)
	1360 Hz	(Il n'est pas tard)
	1280 Hz	(Départ en retard)
	1360 Hz	(Mort à la tâche)
	1360 Hz	(Mort à la tâche)
T→a	1560 Hz	(Monte à l'étage)
	1440 Hz	(Tu fais des taches)
	1640 Hz	(De la potasse)
D→α	1400 Hz	(C'est un cadavre)
	1360 Hz	(C'est un cadavre)
D↘α	1360 Hz	(Une vraie tête d'âne)
	1440 Hz	(Une vraie tête d'âne)
D↗i	1600 Hz	(Il faut le dire)
	1560 Hz	(Il faut le dire)
	1600 Hz	(C'est un prodige)
	1680 Hz	(C'est un prodige)
N→a	1400 Hz	(Va à la nage)
N↗e	1800 Hz	(Sur Côte-des-Neiges)
	1560 Hz	(Toute cette neige-là)
	1760 Hz	(Dans un banc de neige)
S→a	1320 Hz	(Il faut être sage)

Observations:

Pour les voyelles ouvertes, la transition de départ TR 2 est sensiblement à la fréquence du formant de la voyelle; pour les voyelles fermées, qui ont un F2 plus élevé, la transition est montante. Il ne me semble pas que le fait que l'apex fasse contact dans la zone de 40° affecte à lui seul la fréquence du résonateur; la forme et la hauteur de la langue sous le palais dur sont par contre primordiales.

TR 2 im. Fréquences d'arrivée:

α→T	1360 Hz	(Aimes-tu les pâtes)
	1360 Hz	(Prends du pâté)
a→T	1440 Hz	(Coupe-lui les pattes)
	1520 Hz	(Coupe-lui les pattes)
ai→T	2240 Hz	(Une mauvaise tête)
ɛ↘T	1760 Hz	(Une mauvaise tête)
ɛ↘T	1920 Hz	(L'enfant qui tète)
	1560 Hz	(L'enfant qui tète)
i↘T	1680 Hz	(Va pas trop vite)
	1480 Hz	(Va pas trop vite)
ai→D	2080 Hz	(Il veut de l'aide)
		(Son frère l'aide pas)
ɛ→D	1600 Hz	(Il veut de l'aide)
	1560 Hz	(Elle est trop laide)
ɛ↘D	1720 Hz	(Elle est trop laide)
a→S	1440 Hz	(De la potasse)
	1440 Hz	(Une carapace)
	1400 Hz	(Une carapace)
	1440 Hz	(La populace)
α→S	1200 Hz	(Beaucoup d'espace)
	1320 Hz	(Elle est très lasse)
a→Z	1360 Hz	(Frappe à la base)
	1360 Hz	(A base de gaz)

Observations:

L'implosion peut se faire aussi bien à 2240 Hz qu'à 1200 Hz, selon la fréquence du formant de voyelle au moment de l'arrivée sur la consonne; quand la transition est descendante à partir d'un F2 élevé, la raison s'en trouve dans l'abaissement de la langue sur l'axe de 90°. On peut comparer à ce point de vue Une mauvaise tête et L'enfant qui tète (Fig. 3 et 4).

Conclusions:

Les constrictions ou les occlusions du résonateur buccal dans la zone située autour de l'axe de 40° ne sont pas critiques pour les fréquences du formant 2, et elles ont un effet abaisseur pour les fréquences du formant 1.

Pour les deux régimes vibratoires, c'est la forme et la position de la langue sous le palais dur (axe 90°) au moment de l'explosion ou de l'implosion qui détermine les fréquences des transitions.

Les locus acoustiques des consonnes, fréquences fixes vers lesquelles se dirigeraient les formants de voyelles, n'existent pas dans la parole réelle.

Pour la parole synthétique, je ne vois pas d'inconvénient à conserver cette notion de locus, à condition de les faire varier en fréquences selon l'entourage vocalique; pour la commodité de l'écriture, je proposerais d'adopter pour TR 1 une fréquence légèrement inférieure à la fréquence du F1 soit de 300 Hz, et pour TR 2, deux fréquences proches de 1500 Hz du résonateur théorique: 1400 Hz pour les voyelles ouvertes, et 1600 Hz pour les voyelles fermées.

AXE 50°

[ʃ] et [ʒ]

Bien que le point d'articulation de [ʃ] et de [ʒ] soit assez voisin de celui des consonnes qu'on vient d'étudier, il m'a paru intéressant de relever séparément leurs transitions; on sait, en effet, que le spectre du bruit de constriction de [ʃ] est concentré en plus basse fréquence que celui de [s].

Mais il faut distinguer entre la fréquence du bruit de constriction et la fréquence de résonance du canal buccal; le bruit de constriction a sa source au point d'articulation, et a pour résonateur, la partie de la bouche antérieure à ce point; tandis que la résonance buccale a sa source dans les vibrations des cordes vocales et a pour résonateur, le canal buccal tout entier.

En vérité, [ʃ] et [ʒ] occupent une position légèrement plus reculée que [s] et [Z]; de plus, le point de constriction sur la langue est plus près de l'apex pour [s] que pour [ʃ]. Mais c'est la cavité antérieure qui, à cause de la projection labiale, contribue en partie à abaisser le spectre de bruit de [ʃ] et de [ʒ].

TR 1 ex.

ʃ→o (J'aime les pâtes chaudes)
ʃ→œ (Il est pêcheur)

(Il est pêcheur)

ʒ→ɑ (Envoie-moi Jacques)

(Envoie-moi Jacques)

ʒ→a (Prononce un "je")

(Prononce un "je")

ʒ→œ (Y est resté jeune)

(Y est resté jeune)

ʒ→ø (Pratique le jeûne)

(Pratique le jeûne)

Observations:

La transition remonte quand le F1 de la voyelle est supérieur à 300 Hz.

TR 1 im.

ɛ↘ʃ (Il aime la pêche)

ɛ↘ʒ (Un sacrilège)

ɔ↘ʒ (Remonte l'horloge)

Très souvent F1 s'affaiblit et disparaît au lieu de descendre en transition, mais il ne remonte jamais:

a→...ʒ (Retourne la page)

(Monte à l'étage)

(Un fou en cage)

(Il faut être sage)

Observations:

TR 1 a tendance à descendre à l'arrivée de l'apex ou de la partie post-apicale de la langue dans la région de 50°. On savait par le résonateur théorique que toute constriction du canal buccal dans la partie antérieure a pour effet d'abaisser la résonance de F1.

TR 2 ex. Fréquences de départ:

ʃ→α	1440 Hz	(Sac de couchage) sona
ʃ→a	1520 Hz	(Un siège de char) sona
ʃ→ɛ	1840 Hz	(Achète sept auges) sona
ʃ→e	1840 Hz	(Chez nos ancêtres) sona
ʒ→ɛ	1840 Hz	(J'aime mieux la jaune) sona
ʒ→e	1840 Hz	(Tout enneigé) sona
ʒ→i	1920 Hz	(Un gilet mauve) sona

Observations:

Les transitions se font dans l'ensemble à une fréquence de départ située autour de 1480 Hz quand la consonne précède un [α], autour de 1840 Hz devant [ɛ] et autour de 1900 Hz devant [i].

TR 2 im. Fréquences d'arrivée:

α→ʃ	1320 Hz	(Mort à la tâche)
α^{v}→ʃ	760 Hz	(Mort à la tâche)
a→ʃ	1440 Hz	(Tu fais des taches)
	1440 Hz	(Tu fais des taches)
ɛ→ʃ	1680 Hz	(Même les saints pêchent)
ε^{i}→ʃ	2080 Hz	(Même les saints pêchent)
a→ʒ	1440 Hz	(Retourne la page)
	1440 Hz	(Monte à l'étage)
	1440 Hz	(Un fou en cage)
	1400 Hz	(Va à la nage)
α^{v}→ʒ	880 Hz	(Pas ce plombage-là)
	720 Hz	(Fais un plombage)
ɛ→ʒ	1600 Hz	(Un sacrilège)
ε^{i}→ʒ	2160 Hz	(Un sacrilège)
e^{i}→ʒ	2080 Hz	(Toute cette neige-là)

ei→ʒ 2080 Hz (Siège d'autobus)

Observations:

Les transitions se font dans l'ensemble autour de 1440 Hz après [α], autour de 1600 Hz après [ɛ] et autour de 2000 Hz après [i].

Conclusions:

Le rétrécissement du résonateur buccal dans la région située autour de l'axe de 50° influence peu à lui seul la résonance de la transition; c'est la fréquence de F2 de la voyelle qui conditionne la fréquence de la transition dans cette région. On a vu, par le résonateur théorique, que les constrictions sur l'axe de 50° font baisser F1 et commencent à faire monter légèrement F2. Les observations sur la parole réelle nous permettent de constater qu'il en est de même dans le canal buccal.

AXES 80° - 110°

[K], [G], [ɲ]

TR 1 ex.

K→α (Gâteau moka) sona

K→a (Un fou en cage)

(Un fou en cage)

K→ɛ (Avec l'équerre)

(Enlève la caisse) sona

K→ɔ (On est d'accord)

G→α (A base de gaz)

(A base de gaz)

(Un train en gare) sona

(Fais tes bagages) sona

G→a (Gare à la faune) sona

G→œ (De quelle longueur)

(De quelle longueur)

ɲ→œ (Un monseigneur) sona

Observations:

La transition remonte toujours sur F1 quand il n'est pas trop bas

en fréquence.

TR 1 im.

a↘K (Finis ton bac) sona. Sujet témoin.

ɛ↘K (Un archevêque)

(Un archevêque) (Fig. 5)

o→K (Tu viens avec) (Fig. 6)

(Une grosse voix rauque)

a↘G (La nouvelle vague) sona. Sujet témoin.

ɸ→G (L'aide aux aveugles)

(L'aide aux aveugles)

a↘ɲ (Sur la montagne) sona. Sujet témoin.

ɛ↘ɲ (Passe-moi ton peigne) sona. Sujet témoin.

ɔ↘ɲ (Il est ivrogne) sona. Sujet témoin.

Observations:

La transition descend toujours sur F1 quand il n'est pas trop bas en fréquence.

TR 2 ex. Voyelles antérieures. Fréquences de départ:

K↘ɛ 2000 Hz (De quelle grandeur) sona

K→i 2240 Hz (Viens jouer aux quilles) sona

K↗i 2080 Hz (Faut qu'on s'esquive) sona

G↘a 2000 Hz (Regarde la feuille) sona

1920 Hz (Gare à la faune) sona

1920 Hz (Dans le garage) sona

G↘ɸ 2000 Hz (Une pauvre gueuse) sona

G↘y 2000 Hz (Une belle figure) sona

G↗i 2120 Hz (Fais à ta guise) sona

ɲ↘œ 2160 Hz (Un monseigneur) sona

Observations:

Toutes les transitions sont descendantes vers les F2 des voyelles antérieures qui ne sont pas très fermées.

TR 2 ex. Voyelles postérieures. Fréquences de départ:

K→α 1280 Hz (Un fou en cage) axe 105°

K↘α 1360 Hz (Gâteau moka) sona

K↘ɔ 1440 Hz (D'accord sur tout) sona

K↘o 1200 Hz (Dans cette côte-là) sona

1120 Hz (Dans cette côte-là) sona

K↘ɜ 1360 Hz (Qu'est-ce qu'on sort là) sona

K↘ʊ 1040 Hz (Une poule qui couve) sona

G→α 1400 Hz (A base de gaz) axe 108°

G↘o 1360 Hz (Elle est nigaude) sona

Observations:

Les transitions sont descendantes ou horizontales pour les voyelles postérieures dont le F2 ne dépasse pas 1440 Hz. Le point de départ varie entre 1440 Hz et 1040 Hz, selon la fermeture de la voyelle qui suit.

TR 2 im. Voyelles antérieures. Fréquences d'arrivée:

a↗K 1680 Hz (Sac de couchage) sona

ɛ↗K 1680 Hz (Faut que tu t'excuses) sona

ɛ^{i}→K 2360 Hz (Un archevêque) axe 85° (Fig. 5)

i↗K 2080 Hz (Pratique la ruse) sona

ɸ↗G 1600 Hz (Cet aveugle-là) sona

1600 Hz (L'aide aux aveugles) sona

Observations:

La transition est montante si le F2 de la voyelle qui précède n'est pas supérieur à 2080 Hz.

TR 2 im. Voyelles postérieures. Fréquences d'arrivée:

$\alpha^{u} \rightarrow K$ 640 Hz (Envoie-moi Jacques)

800 Hz (Reviens à Pâques) sona

$o \rightarrow K$ 800 Hz (Donne-moi un Coke) sona

$\tilde{\alpha} \rightarrow K$ 1120 Hz (Manque de jugeotte) sona

Observations:

L'implosion des consonnes après les voyelles postérieures est faible et difficile à mesurer. La transition ne semble pas faire remonter F2, même quand il est en très basse fréquence.

Conclusions:

Le rétrécissement du canal buccal dans la région de 80° - 110° abaisse la fréquence de F1.

Pour F2, il faut distinguer entre les voyelles antérieures et les voyelles postérieures. Pour les voyelles antérieures, l'occlusion, qui se fait autour de l'axe de 85°, fait monter la fréquence; pour les voyelles postérieures, l'occlusion, qui se fait autour de l'axe de 105°, fait aussi remonter la fréquence, mais moins sensiblement que pour les voyelles antérieures. On sait qu'à mesure qu'on recule vers le voile du palais, les constrictions sont beaucoup moins critiques pour les résonances de F2. Mais c'est surtout la projection et la fermeture labiales qui abaissent considérablement les fréquences du résonateur et obligent les implosions surtout à se faire en basse fréquence.

Il est à remarquer qu'une occlusion située entre deux voyelles dont l'une est antérieure et l'autre postérieure, peut voir son point d'articulation influencé par les deux voyelles, ou par une seule; ce sera par la voyelle qui suit, si la coupe syllabique est faite avant la consonne (v/cv), et par celle qui précède, si la coupe syllabique est faite après la consonne (vc/v); les vraies syllabes ouvertes à l'intérieur de la chaîne parlée me paraissent rares, du moins dans la prononciation de mes locuteurs.

Pour la synthèse de la parole artificielle, on pourrait proposer 2000 Hz à l'explosion devant les voyelles antérieures, et entre 1440 Hz et 1040 Hz devant les voyelles postérieures, selon le F2 de la voyelle; à l'implosion, 2000 Hz aussi après les voyelles antérieures, et entre 1200 Hz et 800 Hz après les voyelles postérieures.

AXES 130° - 140° [R]

On ne saurait établir la corrélation dans la région uvulaire sans isoler les mouvements de la langue de ceux des lèvres qui d'ordinaire les accompagnent pour les articulations dans cette zone. Il a donc fallu trouver des exemples où seul le déplacement de la langue est responsable des variations formantiques. C'est ce qui se produit dans Il vous aide trop.

On voit la langue passer de [ɹ] apical à la position de [o]; F2

s'abaisse de 1200 Hz à 800 Hz et F1 remonte légèrement, tandis que l'ouverture labiale et la distance des incisives ne varient pas.

Autre exemple: Une tête de loup. Pour à peu près la même distance de la langue au palais sur l'axe de 90° et la même embouchure, F2 passe de 1040 Hz à 800 Hz entre les images 42 et 45.

Ces exemples ne peuvent servir que d'indication; ils montrent que F2 descend quand la langue recule dans la bouche, mais ils ne prouvent pas explicitement qu'une constriction dans la zone uvulaire fait descendre F2.

L'articulation de [R] uvulaire peut nous fournir des exemples plus décisifs, quand elle est assez serrée, puisqu'elle ne comporte pas en elle-même de labialité.

TR 1 ex. Fréquences de départ:

R→ɑ 320 Hz (Donne un coup de râpe)

480 Hz (Donne un coup de râpe)

R→a 400 Hz (Le fédéral)

R→ɛ 400 Hz (Fais ça pour elle)

400 Hz (Fais ça pour elle)

R→o 400 Hz (Une grosse voix rauque)

Observations:

La transition de [R] n'est pas souvent facile à mesurer; elle semble commencer autour de 400 Hz pour les voyelles ouvertes, et elle est ascendante.

TR 1 im. Fréquences d'arrivée:

a→R 720 Hz (Elle est bavarde)

680 Hz (Elle est bavarde)

ɑ→R 600 Hz (Il n'est pas tard)

ɑ→R 600 Hz (Il est bavard)

720 Hz (Il est bavard)

640 Hz (Départ en retard)

ɔ→R 600 Hz (On est d'accord)

560 Hz (Il est très fort)

560 Hz (Elle est très forte)

	640 Hz	(Il n'a pas tort)
ɛ→R	520 Hz	(Avec l'équerre)
	560 Hz	(Viens prendre un verre)
ε^{i}↗R	400 Hz	(Viens prendre un verre)
	560 Hz	(Il faut se taire)
œ→R	600 Hz	(En ta faveur)
	600 Hz	(Avec sa soeur)
	600 Hz	(Il est pêcheur)
	600 Hz	(Veux-tu du beurre)
i→R	240 Hz	(Il faut le dire)
	200 Hz	(Il faut le dire)
y→R	240 Hz	(Soigne ta coupure)
ʊ→R	360 Hz	(Beaucoup trop lourd)
	320 Hz	(Beaucoup trop lourd)
	240 Hz	(C'est à ton tour)

Observations:

La transition de [R] final n'affecte pas la fréquence de F1 qui précède. L'articulation de [R] en position finale est d'ordinaire très faible chez mes informateurs; elle fait très peu varier les position articulatoires de la voyelle qui précède. Un abaissement du voile du palais, sans ouverture des fosses nasales, n'a pas le même effet sur le résonateur qu'une constriction faite par une remontée de la partie postérieure de la langue sous la luette. Un [R] réalisé par constriction dans la zone de 130° - 140° ferait remonter légèrement la transition implosive; c'est du moins ce qu'on peut supposer, puisque la transition explosive est montante. Delattre (1) propose, pour la synthèse de la parole artificielle, de faire remonter F1 au-dessus de 500 Hz pour distinguer le [R] des autres consonnes et des semi-consonnes.

---

1 P. Delattre, Studies in French and comparative Phonetics, Mouton, 1966, p. 284.

TR 2 ex. Fréquences de départ:

R↙α 1280 Hz (Une râpe à fer)

R↘a 1640 Hz (Le fédéral)

R→a 1440 Hz (Le fédéral)

R↙ɛ 1440 Hz (Un thérapeute) sona

1440 Hz (Tempête de grêle)

1520 Hz (Fais ça pour elle)

1360 Hz (Appelle Thérèse) sona

Observations:

Bien que j'aie relevé l'ensemble des réalisations chez trois informateurs, je n'ai pas trouvé de constante significative dans les transitions de [R] vers les voyelles. Plus que toutes les autres consonnes, le [R] semble coarticulé non seulement à travers la voyelle qui suit, mais en même temps à travers le phonème, consonne ou voyelle, qui précède.

Pour ne prendre qu'un exemple (VCV), on peut comparer Fais ça pour elle et Tempête de grêle:

ʊ = 840 Hz,	R = 1120 Hz,	ɛ = 1440 Hz
G[ɸ] =1600 Hz,	R = 1440 Hz,	ɛ = 1520 Hz

Dans le premier exemple, la fréquence de [R] sur F2 est à mi-chemin entre les deux voyelles; dans le second exemple, elle reste beaucoup plus haute sous l'influence de la position articulatoire du [G] qui précède. Öhman (1) a signalé un phénomène à peu près semblable pour les occlusives en position VCV.

Pour ces raisons de coarticulation, la dispersion des transitions du [R] explosif se situe entre 1840 Hz et 800 Hz, selon la fréquence des F2 qui précèdent et qui suivent. Quand la constriction uvulaire est assez ferme entre deux formants 2 de fréquences comparables, elle abaisse la résonance de quelque 250 Hz. Exemple: Appelle Thérèse (sona).

e = 1680 Hz	R = 1360 Hz	ɛ = 1520 Hz

Quand la langue ne s'élève pas vers la luette, mais que cette dernière s'abaisse sur la langue, la fréquence ne semble pas varier. Il y a très peu de bruit de constriction dans les [R] de mes sujets. J'ai d'ailleurs relevé quatre façons de faire ce phonème chez le même sujet populaire: [R], [ʁ], [ɾ] et [ɹ].

---

1 S.E.G. Öhman, "Coarticulation in VCV Utterances: Spectro-graphic Measurements" in J.A.S.A. 1965, 39, pp. 151-168.

TR 2 im. Fréquences d'arrivée:

α→R	1200 Hz	(Il est bavard)
α↘R	1080 Hz	(Il est bavard)
	800 Hz	(Départ en retard)
	1040 Hz	(Il n'est pas tard)
ε↘R	1440 Hz	(Sortons prendre l'air)
	1440 Hz	(Viens prendre un verre)
	1400 Hz	(Avec l'équerre)
ε^{i}↘R	2000 Hz	(Sortons prendre l'air)
	2000 Hz	(Avec l'équerre)
	2000 Hz	(Il faut se taire)
ɔ→R	1120 Hz	(Il n'a pas tort)
	1040 Hz	(Il est très fort)
	1080 Hz	(On est d'accord)
œ→R	1360 Hz	(Avec sa soeur)
	1400 Hz	(Avec sa soeur)
œ↘R	1200 Hz	(Il est pêcheur)
y↘R	1200 Hz	(Soigne ta coupure) (Fig. 8)

Observations:

La transition implosive est mieux marquée pour F2 que pour F1, surtout chez S.C.; chez S.P., F2 disparaît souvent avant la transition. Elle est descendante ou droite, mais jamais montante; la régularité de variations est plus facile à observer pour les transitions du [R] final, parce qu'elle n'est pas contrariée par l'action d'une articulation subséquente. Dans la majorité des cas, la transition est descendante.

Conclusions:

Pour nous en tenir à la corrélation, une constriction dans la zone de 140°, quand elle se produit réellement et indépendamment de la fermeture et de la projection labiales, abaisse la fréquence de F2 et élève légèrement celle de F1. C'est ce que laissait prévoir le résonateur théorique.

AXE 180°, Passage pharyngal

Le passage pharyngal le plus étroit se trouve dans les articulations [a], [ɑ] et [ɔ]; pour [o] et [ʊ], il y a remontée de la langue et dégagement du pharynx sur l'axe de 180°.

Il m'a été impossible d'établir la corrélation à ce point du résonateur buccal, parce que je n'ai pu isoler des articulations où le passage pharyngal était seul ou principalement responsable de la variation des formants. Le passage pharyngal ne semble pas le point déterminant pour la distinction des voyelles; voici quelques exemples tirés de comparaisons entre [a] et [ɑ]: Le principal; Il est bien pâle.

			ouverture labiale	45°	90°	130°	180°	F1	F2
S.P.	a	(image 45)	10,5 mm	12,5	11,5		5 mm	800	1360
	ɑ	(image 44)	6,	17	17		5 mm	640	1080
S.C.	a	(image 41)	8 mm	13	13	27,5	6,	680	1320
	ɑ	(image 39)	10,5 mm	18	15	31	5,5	640	1200

On voit que le passage pharyngal varie peu.

Dans d'autres exemples les différences y sont plus marquées. Coupe-lui les pattes; Aimes-tu les pâtes:

			ouverture labiale	45°	90°	130°	180°	F1	F2
S.P.	a	(image 53)	12	13	11,5		7,5	800	1440
	ɑ	(image 57)	10	20	15		5	640	1040
S.C.	a	(image 51)	8,5	11,5	13		8,5	720	1440
	ɑ	(image 38)	8	14	15		5	640	1200

Voici, avant de conclure sur ce point, un tableau qui met en regard les passages pharyngaux les plus étroits et les F2 correspondants, chez le même sujet:

Une carapace [a]:	5,5 mm	1400 Hz
Beaucoup d'espace [ɑ]:	5, mm	1200 Hz
La populace [a]:	7, mm	1480 Hz
Elle est très lasse [ɑ]:	6,5 mm	1280 Hz
Le fédéral [a]:	5, mm	1320 Hz
J'entends des râles [ɑ]:	5, mm	1280 Hz
Aiguise la lame [a]:	7, mm	1400 Hz
La mort dans l'âme [ɑ]:	5,5 mm	1120 Hz
Tu fais des taches [a]:	8, mm	1440 Hz
Mort à la tâche [ɑ]:	5,5 mm	1280 Hz
Ca me fait mal [a]:	7,5 mm	1360 Hz
C'est un beau mâle [ɑ]:	6, mm	1200 Hz
Elle est bavarde [a]:	5,5 mm	1360 Hz
Il est bavard [ɑ]:	5, mm	1160 Hz

Observations:

1. Il peut y avoir recul de la langue de [a] à [ɑ], mais non l'inverse.

2. Même quand le recul est assez prononcé, il n'est pas seul responsable de l'abaissement du formant.

La meilleure illustration de l'action du recul de la langue dans le pharynx se trouve dans Elle est trop laide. Pour un recul de 5 mm, F1 passe de 320 Hz à 580 Hz, et F2 varie de 1760 Hz à 1440 Hz; mais l'abaissement de la langue sur l'axe de 90° en est peut-être la principale cause. Pour ce qui est du dégagement du passage pharyngal dans l'articulation de [ʊ], il suffit de comparer les positions articulatoires dans la diphtongaison [ɑʊ] de C'est un cadavre. On voit que la langue dégage l'axe 180° de [ɑ] à [ʊ]. A vrai dire, il s'agit d'un dégagement peu prononcé et causé par la remontée du dos de la langue sur l'axe 140 .°

Conclusion:

La corrélation ne peut être établie isolément sur l'axe de 180° dans le résonateur buccal. Le résonateur théorique nous enseigne que, pour un rétrécissement à cet endroit, F2 descend et F1 remonte; je puis seulement dire qu'il ne semble pas que les données de la parole réelle contredisent la théorie.

[L]

La corrélation aux divers points d'articulation m'a amené à passer en revue l'ensemble des consonnes, moins [L]; bien que ce travail ne soit pas une étude consonantique, je m'arrêterai un instant sur cette dernière consonne, afin de compléter le tableau des transitions.

[L] est une consonne qui se vocalise facilement et dont le spectre est, comme celui de [R], fortement influencé par la coarticulation vocalique. Elle comporte nécessairement un contact de l'apex dans la région alvéolaire, en français, et accessoirement une constriction bilatérale dans les articulations assez durables.

Nous avons vu qu'une occlusion ou une constriction dans la région située autour de l'axe de 40° n'est pas critique pour la résonance de F2, mais abaisse F1. Quant à la constriction sonore qui a lieu entre les deux côtés de la langue et les joues, elle se fait indépendamment de l'ouverture des maxillaires mais elle est liée à la hauteur de la langue commandée par la coarticulation vocalique; cette constriction latérale ne se réalise pas dans la parole rapide.

TR 1 ex. Fréquences de départ:

L↗a 480 Hz (Aiguise la lame)

320 Hz (Aiguise la lame)

L↗ɑ 400 Hz (La mort dans l'âme)

320 Hz (La mort dans l'âme)

Observations:

La transition remonte vers F1 des voyelles ouvertes.

TR 1 im. Fréquences d'arrivée:

L'implosion de [L] se fait en deux temps: d'abord à l'arrivée de l'apex sur les alvéoles, ensuite au début de la constriction.

a↖L v ↘ 640 Hz ↘ 360 Hz (Ca me fait mal)

v ↘ 400 Hz ↘ 240 Hz (Ca me fait mal)

v ↘ 560 Hz ↘ 320 Hz (Le fédéral)

v ↘ 480 Hz ↘ 320 Hz (Le fédéral)

Observations:

La transition descend pour l'occlusion de l'apex et subit une autre déflexion durant la constriction bilatérale.

TR 2 ex. Fréquences de départ:

L↘α	1440 Hz	(La mort dans l'âme)
	1320 Hz	(La mort dans l'âme)
L↘α	1560 Hz	(Aiguise la lame)
	1840 Hz	(Aiguise la lame)
	1600 Hz	(Dans le village)
L↗ε	1760 Hz	(Bois un peu de lait)
	1440 Hz	(Bois un peu de lait)
	1760 Hz	(Son frère l'aide pas)
	1640 Hz	(Il faut que tu te lèves)
L→ε	1480 Hz	(Il veut de l'aide)
L↗y	1760 Hz	(Un vrai déluge)
L↘u	1440 Hz	(Avec la loupe)
	1640 Hz	(Beaucoup trop lourd)
	1440 Hz	(Beaucoup trop lourd)
	1760 Hz	(Une tête de loup)
	1520 Hz	(Une tête de loup)

Observations:

La fréquence de résonance de [L] au moment de la rupture de l'occlusion apicale se situe entre 1840 Hz et 1320 Hz, puis descend ou monte selon la fréquence du F2 qui suit. Si F2 est inférieur à 1320 Hz, la transition est toujours descendante; si F2 est supérieur à 1840, elle est toujours ascendante. Entre ces limites, la corrélation est difficile à faire. Il est possible que la transition sur F1, qui est constante soit beaucoup plus informative pour la perception de [L] que la transition sur F2.

TR 2 im. Fréquences d'arrivée:

a↗L	1440 Hz	(Le fédéral)
a→L	1360 Hz	(Ca me fait mal)
	1360 Hz	(Ca me fait mal)
α↗L	1320 Hz	(Elle est bien pâle)

1360 Hz (J'entends des râles)

1440 Hz (C'est un beau mâle)

ɛ→L 1600 Hz (Tempête de grêle)

1440 Hz (Fais ça pour elle)

o↗L 1400 Hz (Sur un toit de tôle)

i↘L 1800 Hz (Il le dévore) sona

1840 Hz (Y faut qu'il t'aide) sona

1840 Hz (Y faut qu'il t'aide) sona

2000 Hz (Faudrait qu'il ose) sona

i→L 1840 Hz (Faut qu'il l'épouse)

Observations:

La transition est d'ordinaire ascendante, si le F2 de la voyelle qui précède est inférieur à 1320 Hz; elle est descendante, si le F2 qui précède est supérieur à 1840 Hz.

Conclusions:

En termes de corrélation articulatoire et acoustique, on peut dire que le fait que l'apex touche les alvéoles oblige le résonateur buccal à des fréquences qui se situent entre 1320 Hz et 1840 Hz pour F2, et à des fréquences autour de 320 Hz pour F1.

Les mouvements de la langue considérés dans leur ensemble

Les mouvements de la langue ne sont pas souvent uniquement horizontaux ou verticaux. On peut tout de même dégager les grandes lignes de la corrélation articulatoire et formantique sur deux axes perpendiculaires, tels qu'on les trouve en abscisse et en ordonnée sur les trapèzes acoustiques.

1. Incidences des mouvements horizontaux sur F1:

Pour les voyelles très fermées, les déplacements de la langue d'avant vers l'arrière ou d'arrière vers l'avant n'entraînent pas de variation sensible de F1. Exemple: Une tête de loup (Fig. 7), Faut qu'il l'épouse, Il faut que ça bouge. De [L] à [ʊ] et de [ʊ] à [Z] et à [ʒ], F1 ne varie pas.

Pour les voyelles ouvertes, les mouvements horizontaux de la langue sont toujours accompagnés de mouvements verticaux et d'une variation de l'ouverture des incisives et des lèvres; pour cette raison, on ne peut les isoler et les analyser séparément; je peux seulement savoir indirectement que le recul de la langue dans le pharynx élève F1.

2. Incidences des mouvements verticaux sur F1:

Les mouvements de la langue en direction du palais dur font baisser F1; exemples: Avec l'équerre; Sortons prendre l'air; dans les deux cas, on voit F1 passer de 720 Hz à 320 Hz quand la langue remonte sur l'axe de 90°.

Seuil critique pour F1:

Il y a dans ces mouvements verticaux un seuil critique situé à peu près à 7,5 ou 8 mm du palais dur sur l'axe de 90°. On peut en voir une illustration dans la diphtongaison de [ɛ], Il faut qu'il s'aide. F1 varie peu jusqu'à ce que la langue arrive à la hauteur de 7,5 mm. Autre exemple: Raconte ton rêve; F1 passe de 720 Hz à 480 Hz quand la distance de la langue au palais dur sur 90° passe de 8mm à 4,5 mm.

3. Incidences des mouvements horizontaux sur F2:

Les déplacements de la langue de l'avant vers l'arrière dans les articulations fermées font baisser rapidement F2. Exemples:

Une tête de loup; de [L] à [ʊ], F2 passe de 1760 à 800 Hz,

Loup, F2 : 1520 ↘ 800 Hz.

D'accord sur tout; de [T] à [ʊ], F2: 1440 ↘ 720 Hz,

Beaucoup trop lourd; de [L] à [ʊ], F2 : 1640 ↘ 640 Hz.

Beaucoup trop lourd; de [L] à [ʊ], F2 : 1400 ↘ 640 Hz.

Les déplacements en sens inverse font remonter F2, s'ils se produisent sur une ligne horizontale assez proche du palais. Exemples: Faut qu'il l'épouse; de [ʊ] à [Z], F2 passe de 840 à 1320 Hz.

4. Incidences des mouvements verticaux sur F2:

Les mouvements verticaux de la langue vers le palais dur font remonter F2; des mouvements en sens inverse le font descendre. Exemples: Il faut qu'il s'aide; de [a] à [i], F2 passe de 1280 à 2120 Hz. J'ai peur qu'il cède; à l'intérieur du [ɛ], F2 passe de 1680 à 1280 Hz.

Seuil critique pour F2:

Quand la langue s'élève sur l'axe de 90°, F2 commence à monter en fréquence avant que F1 ne commence à baisser; le seuil critique est plus distant du palais dur pour F2 que pour F1; on observe entre les deux variations un délai de deux ou trois images, ou une différence de 2 ou 3 mm sur l'axe de 90°. Exemple: Il faut qu'il s'aide; F2 commence à monter à l'image 45, soit à 10 mm de distance de la langue au palais; tandis que F1 commence à baisser à l'image 47 seulement, soit à une distance de 7,5 mm. Autre exemple: Viens prendre un verre; F2 varie à l'image 56 (10 mm sur 90°), F1 varie à l'image 59 (7,5 mm sur 90°).

Le seuil critique pour F2 se trouve autour de 10 mm sous le palais dur, axe de 90°.

5. Corollaire:

Il découle du paragraphe no 3 que F2 ne peut être haut en fréquence quand la langue est haute dans la partie postérieure de la bouche, quelle que soit sa hauteur sur l'axe de 90°. D'ailleurs, la position haute sur les axes postérieurs s'accompagnent toujours, en français, excepté pour [R], de la projection et de la fermeture labiales qui abaissent considérablement les fréquences du résonateur.

Résumé et conclusions

Les grandes lignes de la corrélation entre les mouvements articulatoires et les variations formantiques peuvent se ramener à quelques principes qui, pris séparément, sont assez simples:

1. La fermeture et la projection labiales abaissent à la fois F1 et F2. Pour cette raison, les transitions de départ des consonnes bilabiales et labiodentales sont ascendantes, et celles d'arrivée sont descendantes, pourvu que F1 et F2 ne soient pas déjà très bas.

2. La constriction et l'occlusion de la langue dans la résion alvéodentale abaissent la résonance de F1, mais ne font pas varier à elles seules la fréquence de F2.

3. La remontée et le contact de la langue dans la région post-alvéolaire et palatale font baisser la fréquence de F1 et monter celle de F2.

4. La remontée et le contact de la langue dans la région postérieure du palais n'affecteraient pas sensiblement les résonances de F1 et de F2 s'ils n'étaient accompagnés d'une fermeture prononcée des incisives et des lèvres et de projection labiale qui abaissent toutes les fréquences du résonateur buccal.

5. Le recul de la langue dans la région de la luette (140°) fait à lui seul monter F1 et descendre F2.

6. Le recul de la langue dans le pharynx contribue à faire monter F1 et baisser F2, mais il n'est pas prédominant dans l'ensemble des mouvements articulatoires.

Remarques:

1. Il est rare que les mouvements articulatoires de la parole réelle ne mettent qu'un seul de ces principes en cause; les déplacements de la langue dans une direction occasionnent en même temps un rétrécissement du canal buccal dans plus d'une région à la fois, et des dégagements dans d'autres régions.

2. Dans certains cas, l'action de la contriction et celle du dégagement s'ajoutent pour faire varier les formants dans le même sens; dans d'autres cas, les deux actions se contrarient et s'annulent, ou renversent le sens des variations formantiques. C'est ainsi, par exemple, que la fermeture et la projection labiales contrarient l'effet sur F2 de la position haute de la langue dans les voyelles fermées antérieures et accentuent l'effet de la position haute de la langue dans les voyelles fermées postérieures.

Parole synthétique

Pour appliquer ces données à la programmation de la parole de synthèse, on pourrait faire remonter TR 1 ex. à partir de la consonne quelle qu'elle soit (excepté R postérieur), si le F1 n'a pas à rester bas à cause d'une voyelle haute; de même on peut faire descendre TR 1 im., à moins que F1 soit déjà assez bas. On pourrait programmer comme suit: TR 1 = F1-200 Hz, mais TR 1 ≥ 200 Hz. Pour R: TR 1 = F1+200 Hz, mais TR 1 ≤ 700 Hz.

Pour les labiales, TR 2 ex. remonte de 200 ou 300 Hz vers F2 de la voyelle; TR 2 im. descend d'autant à partir de F2.

Pour les dentales, TR 2 ex. reste droit si la langue reste à la même hauteur sous le palais dur après l'explosion ou la détente; elle monte ou descend selon que la langue monte sous le palais dur et dégage le passage pharyngal, ou descend sous le palais dur et engage le passage pharyngal, après l'explosion ou la déconstriction; pour TR 2 im., même raisonnement, inversé.

Pour les palatales, leur articulation oblige le résonateur buccal à résonner en haute fréquence, 2300 Hz environ, si le point d'articulation est antérieur, et en moyenne fréquence, 1500 Hz environ, si le point d'articulation est postérieur. De là, il y a mouvement ascendant ou descendant, ou droite ligne, selon la fréquence de F2 de la voyelle. Pour TR 2 im., même raisonnement, mais inversé.

Dans tous les cas, il faut tenir compte des phonèmes (consonnes ou voyelles) qui précèdent et qui suivent la syllabe en cause. Ainsi la TR 2 ex. sera différente entre /T/ et /a/ selon qu'il s'agira de /ita/ ou de /ata/; et TR 2 im. différera entre /a/ et /T/ selon qu'il s'agira de /ati/ ou de /ata/.

Cette corrélation des mouvements articulatoires, mesurés sur les deux seules dimensions du plan sagittal médian, avec les données acoustiques réduites à deux formants, est un compromis entre les simplifications de la parole synthétique et la complexité de la parole réelle.

La théorie des résonateurs et l'application que j'en ai faite pour dégager les grandes lignes d'une corrélation des données articulatoires et acoustiques n'avaient pour but que de servir de support aux considérations que j'ai voulu faire sur les transitions articulatoires et acoustiques.

Bibliographie

Bibal (M.) et Heiny (P.), Electronique appliquée, Delagrave, éd. de 1968.

Chiba (T.) et Kajiama (M.), The Vowel - its Nature and Structure, Tokyo, 1941.

Delattre (P.), Studies in French and Comparative Phonetics, Mouton, 1966.

Didier (A.), Reproduction des sons et des images, Tome 1, Masson, Paris, 1964.

Fant (G.), Acoustic Theory of Speech Production, Mouton, 1960.

"Analysis and Synthesis of Speech Processes", in Malmberg, Manual of Phonetics, North-Holland Publishing Co., Amsterdam, pp. 173-277, 1968.

"Stops in CV-Syllables" STL-QPSR 4/1969, Stockholm, 1969.

Fischer-Jørgensen (E.), "Voicing, Tenseness and Aspiration in Stop Consonants, with Special Reference to French and Danish", in Annual Report III, 1969, Inst. Phon. Univ. Copenhague, pp. 63-114.

Flanagan (J.L.), Speech Analysis, Synthesis and Perception, Springer-Verlag, Berlin, Heidelberg, New York, 1965.

Fujimura (O.), "Bilabial Stop and Nasal Consonants: a Motion Picture Study and its Acoustical Implications", in J. of Speech and Hearing Research 4, 1969, pp. 233-247.

Heinz (J.M.) and Stevens (K.N.), "On the Relations between Lateral Cineradiographs, Area Functions, and Acoustic Spectra of Speech", in Proc. 5th Int. Cong. Acoust., Liège, 1965.

Joos (M.), Acoustic Phonetics, Supp. to Language, vol. 24, no 2, 1948.

Lindblom (B.E.F.), "Numerical Models in the Study of Speech Production and Speech Perception: Some Phonological Implications.", in Actes du VIIe Cong. Int. des Sc. Phon. Montréal, 1971 (à paraître chez Mouton, 1972).

Ochiai (Y.), "Recherches sur les voyelles françaises au point de vue du TIMBRE PHONETIQUE ET VOCAL", in Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung, Band 18, Heft 3, 1965, pp. 243-79.

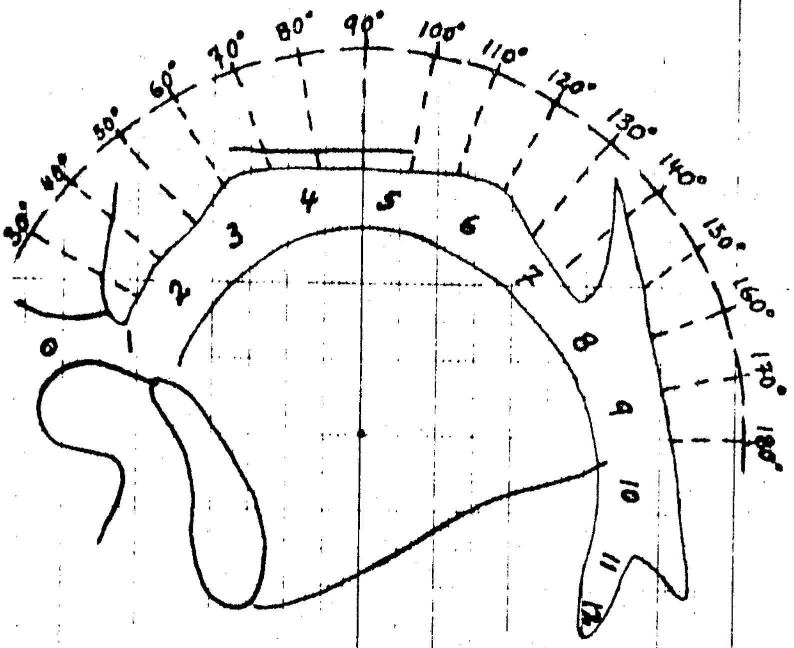
Öhman (S.E.G.), "Coarticulation in VCV Utterances: Spectrographic Measurements", in J.A.S.A. 1965, pp. 151-168.

Piraux (H.), Dictionnaire général d'acoustique et d'électroacoustique, Eyrolles, Paris, 1964.

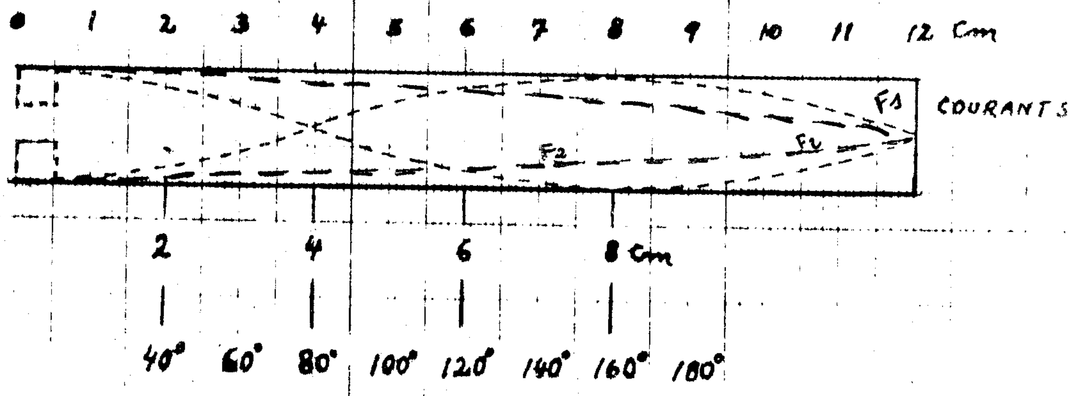
Santerre (L.), Les voyelles orales dans le français parlé à Montréal, Klincksieck, Paris, 1972 (sous presse).

Sovijärvi (A.), "On Transition in the Light of X-Ray Films", in Proc. 6th Int. Cong. Phon. Sc., Prague 1967, Hieber, Munchen 1970, pp. 851-57.

CORRELATION ARTICULATOIRE ET ACOUSTIQUE



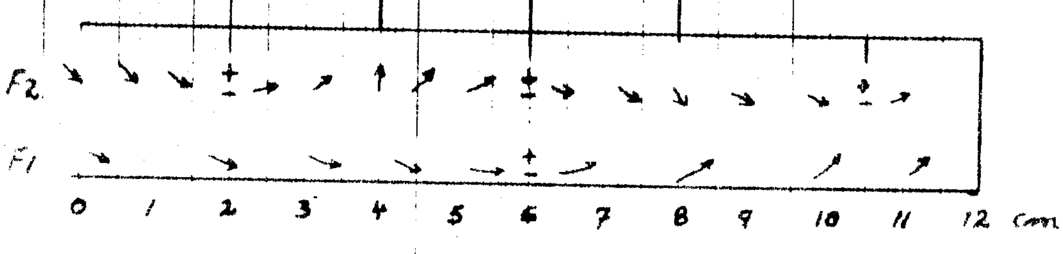
COUPE RADIOCINEMATOGRAPHIQUE SUR LE PLAN SAGITTAL MEDIAN
DIMENSIONS REDUITES AUX 2/3.
POSITIONS ARTICULATOIRES DU SCHWA.



— — — — — — : F2

- - - - - - - - : F1

LES CHIFFRES ET LES MESURES D'ANGLES SONT CORRESPONDANTS SUR LES TROIS GRAPHIQUES.



LES FLECHES [illegible] POUR F1 ET POUR F2 L'ABAISSEMENT OU LE RELEVEMENT DE LA FREQUENCE [illegible] DE 500 Hz ET DE 1500 Hz QUAND LE RESONATEUR EST RETRECI; EN CAS DE DEGAGEMENT AU LIEU DE RETRECISSEMENT AUX POINTS MARQUES, LE SENS DES FLECHES EST A INVERSER. LES POINTS MARQUES ± SONT NEUTRES, PARCE QUE LE COURANT [illegible] LA PRESSION [illegible] SONT EGAUX.

FIG. 1

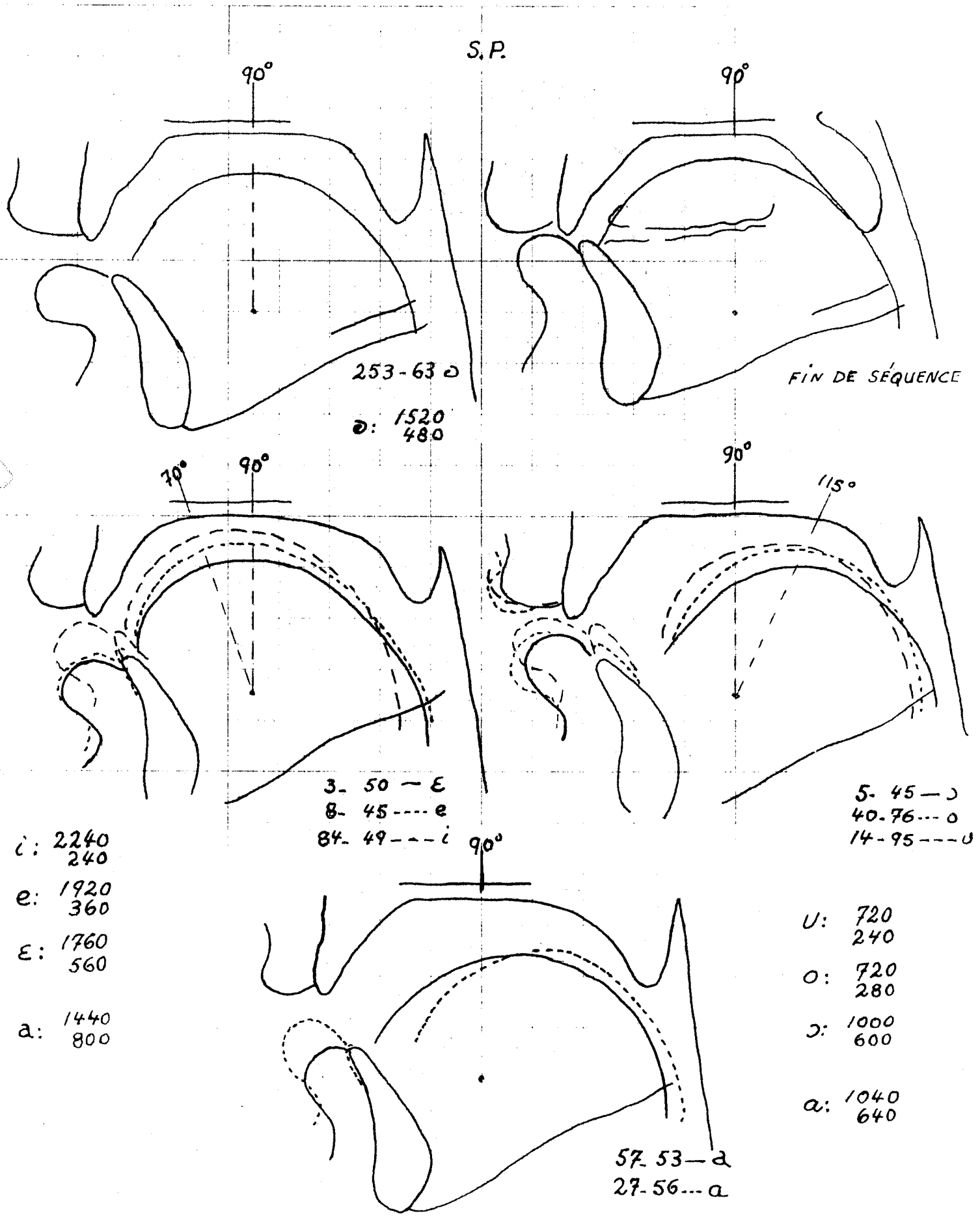
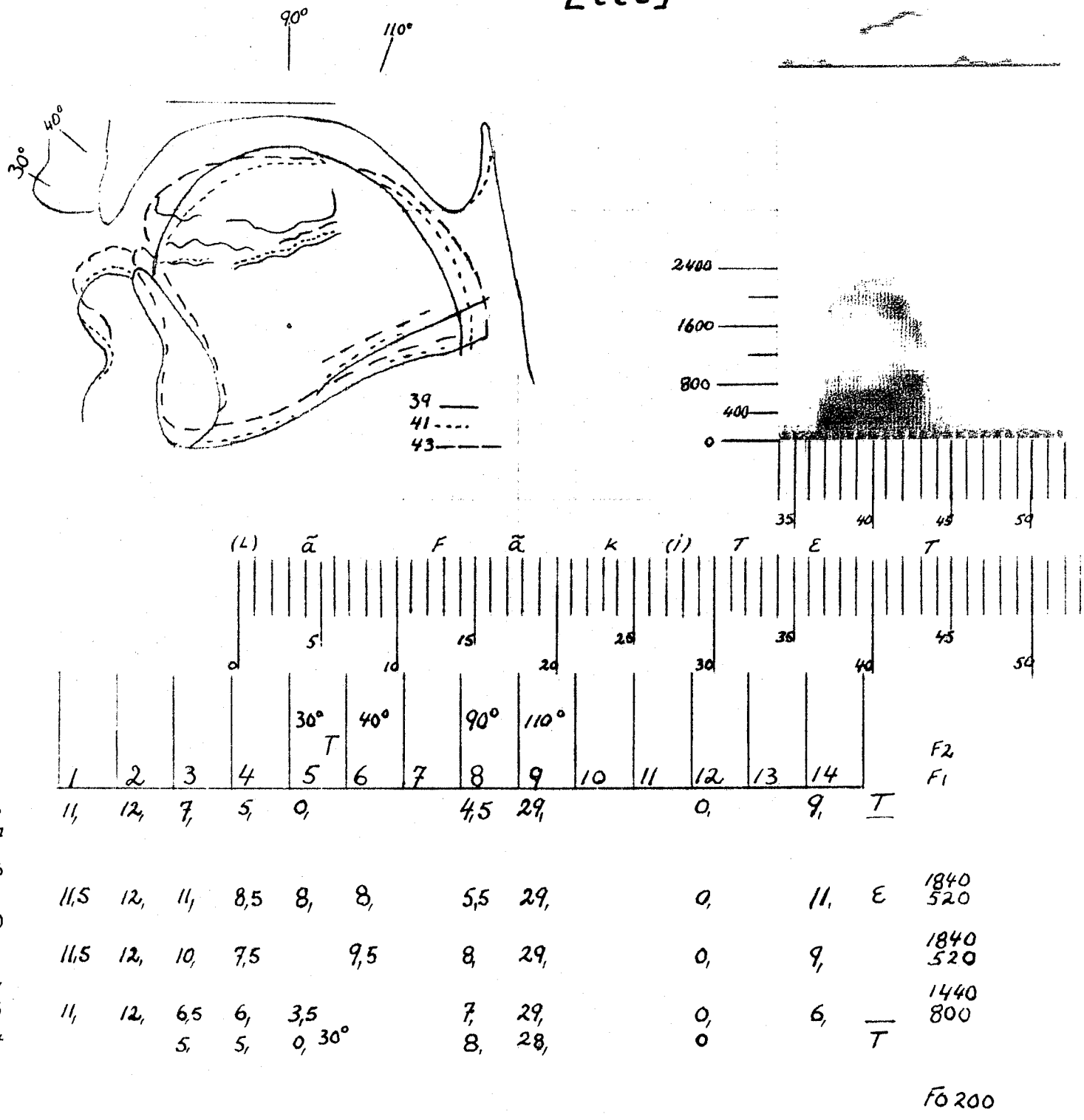

S. P.
90°
253-63 ɔ
ɔ: 1520
480
90°
FIN DE SÉQUENCE
70°
90°
3- 50 — ɛ
8- 45 ---- e
84- 49 - - - i
90°
115°
5- 45 — ɔ
40-76 --- o
14-95 - - - u
i: 2240
240
e: 1920
360
ɛ: 1760
560
a: 1440
800
90°
57- 53 — a
27-56 --- a
u: 720
240
o: 720
280
ɔ: 1000
600
a: 1040
640

Fig. 2

L'ENFANT QUI TÈTE
[tɛt]



36	11,	12,	7,	5,	0,			4,5	29,			0,		9,	T
37															
38															
39	11,5	12,	11,	8,5	8,	8,		5,5	29,			0,		11,	ɛ
40															
41	11,5	12,	10,	7,5		9,5		8,	29,			0,		9,	
42															
43	11,	12,	6,5	6,	3,5			7,	29,			0,		6,	—
44			5,	5,	0,	30°		8,	28,			0			T

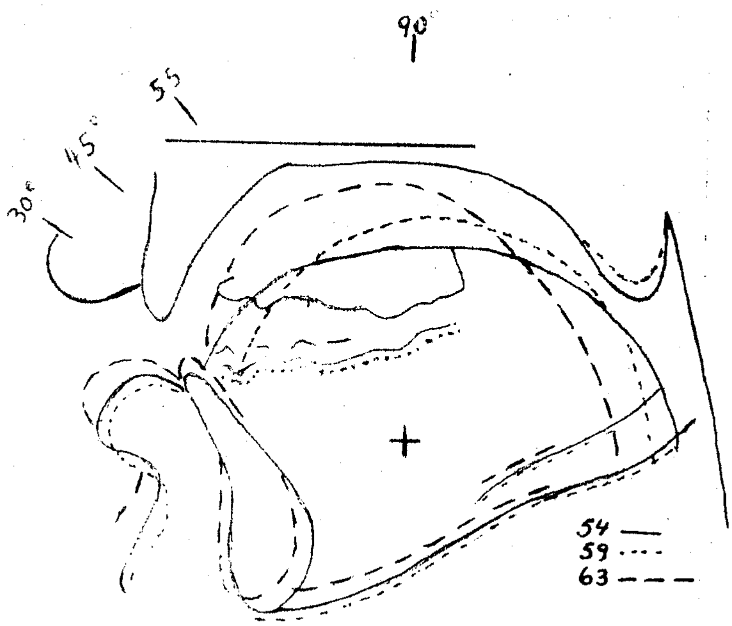
1840
520
1840
520
1440
800

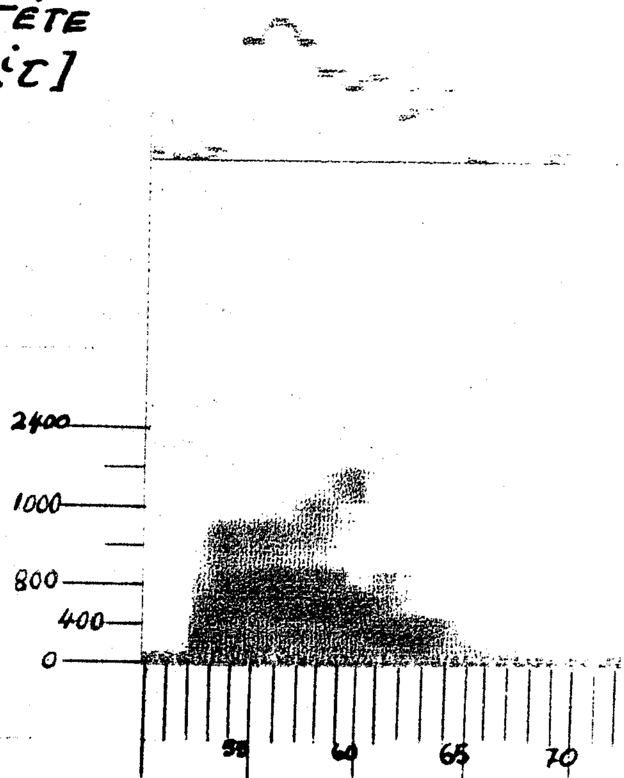
F0 200

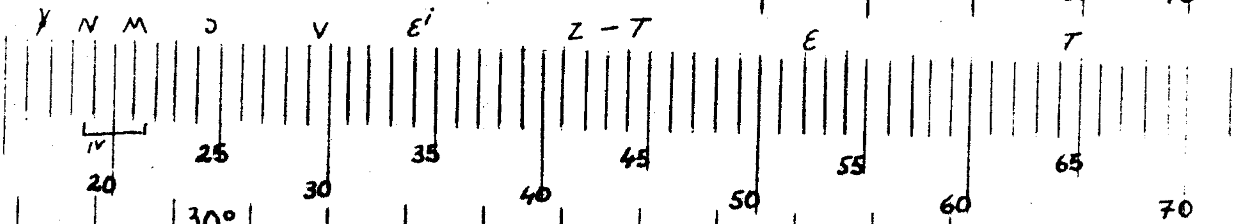
DURÉE TOTALE DE LA VOYELLE: 14 cs

Fig. 3

UNE MAUVAISE TÊTE
[ɛˑɛ]







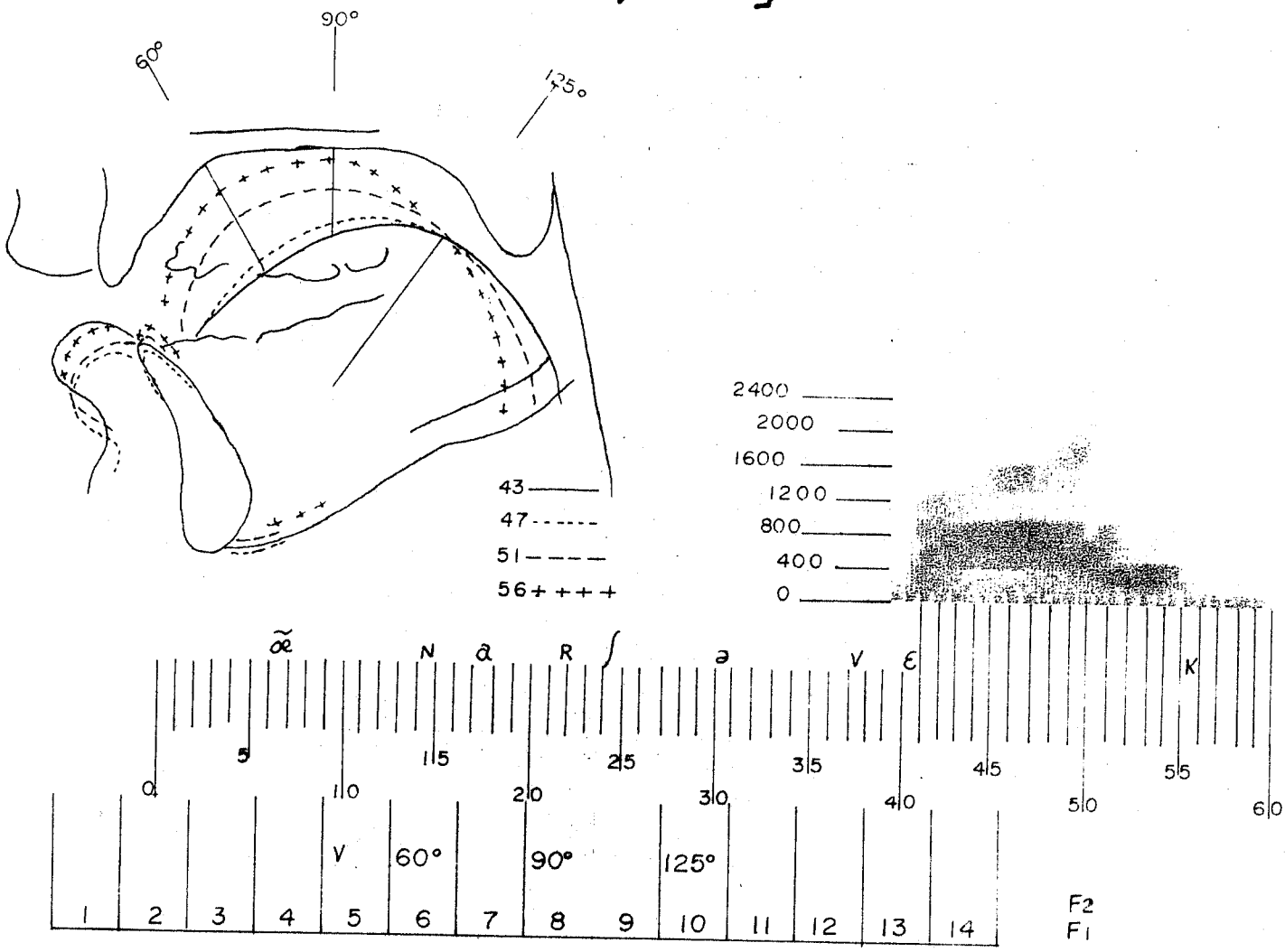
	1	2	3	4	5 (30° 45° T)	6 (55°)	7	8 (90°)	9 (135°)	10	11	12	13	14		F2 / F1
52			8	6	0			10,5						45	I	
53			8	6,5	5			11								
54	12	12	10,5	7,5	7,5	11		11			3	0		5,0	ə/a	1280 / 720
55			11,5	8	8	14		12								
56					(45°) ↓											
57			11,5	8	11,5	13		10								
58																
59	11,5	12	12	8,5	11	12,5		7,5			3,5	0		8,0		1520 / 720
60																
61			10	7,5	8	7,5		4						8,5		
62																
63	13,5	12	8	6	4,5	4		3			8	0		12,5	I	2240 / 320
64																
65			3,5	5	0	2,5		1,5						11,5	T	

Fo 220

DURÉE TOTALE DE LA VOYELLE : 26 cs

Fig. 4

UN ARCHEVÊQUE
[œ̃ naʀʃəvaik]

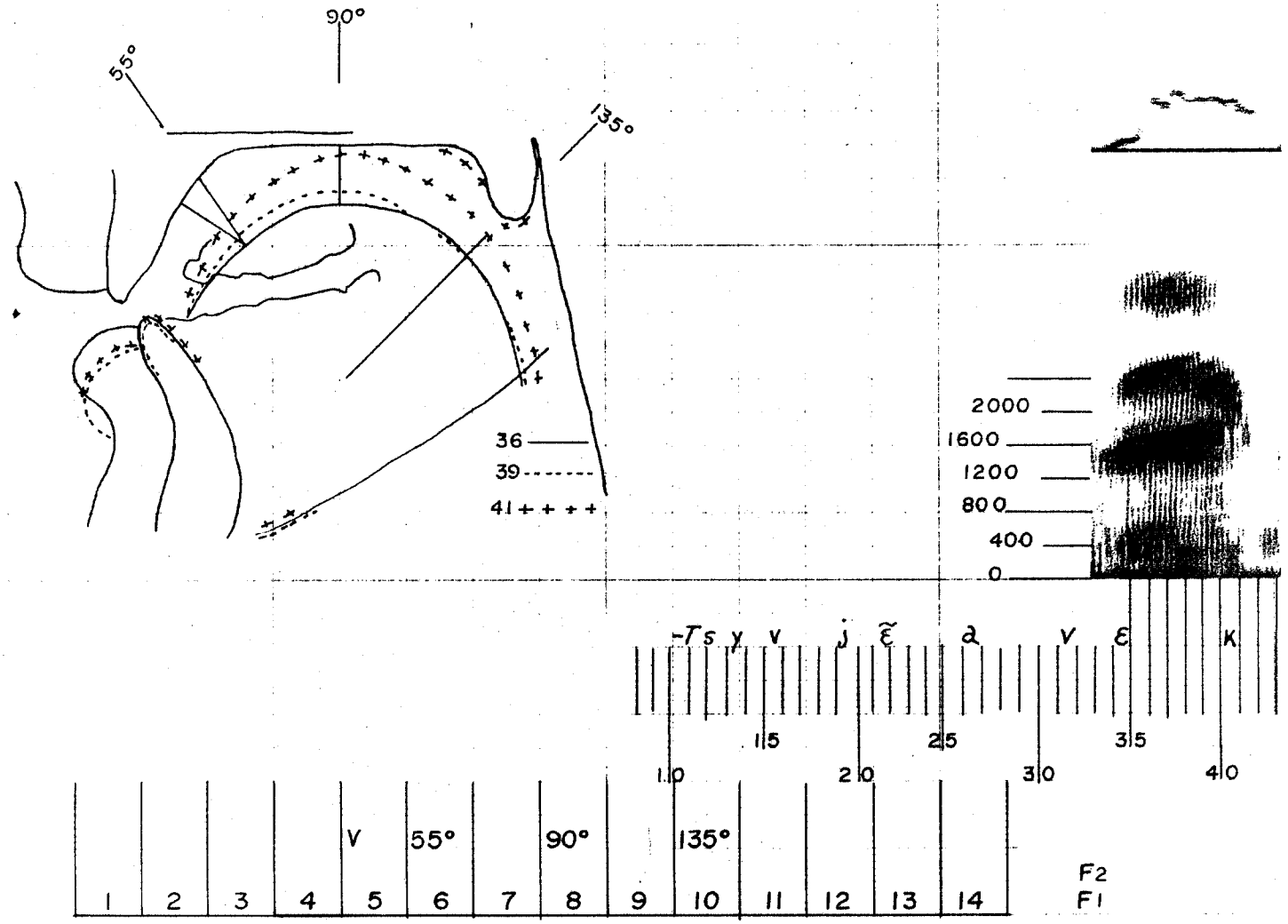


40				7,5	0	15	11	27,5	5	V	960
41				9	1		12				
42					1,5	16	13				
43	13	13,5	7,5	9	5,5	18	13	27,5	5	ɑ̃	1120
44											640
45			11	9		18	13	27,5			
46											
47	12	13,5	12	10,5		16	11	27,5	5	ɑ	1440
48											760
49				10		12,5	8,5	27,5	6,5		
50											
51	12	13,5	11	8,5		9	6	27,5	9	ɛ	1720
52											560
53			10	7,5		6,5	4	27,5	9,5	e	2040
54											
55			8	6,5		3	2,5		11,5	i	2360
56	12	13,5	8	7		4,5	1,5	27,5	13,5	K	320

DURÉE TOTALE DE LA VOYELLE: 31 cs

Fo 220

FIG. 5

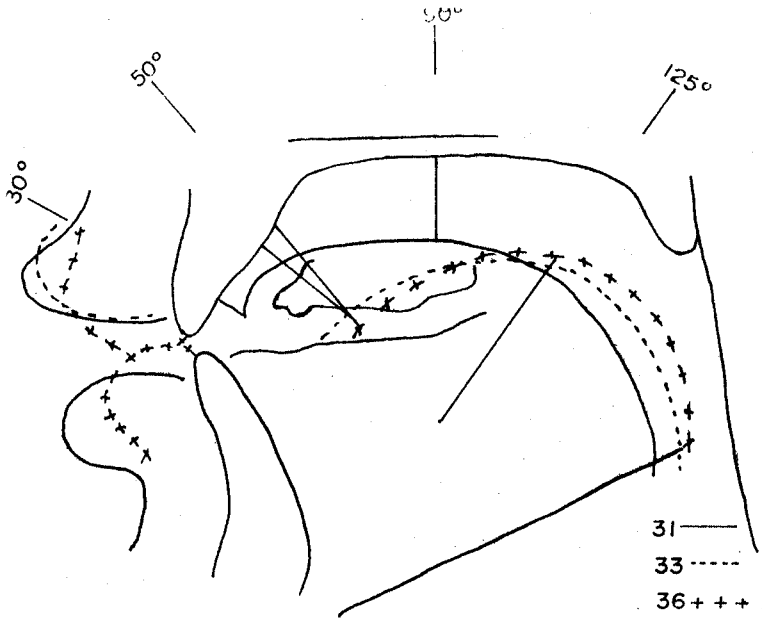


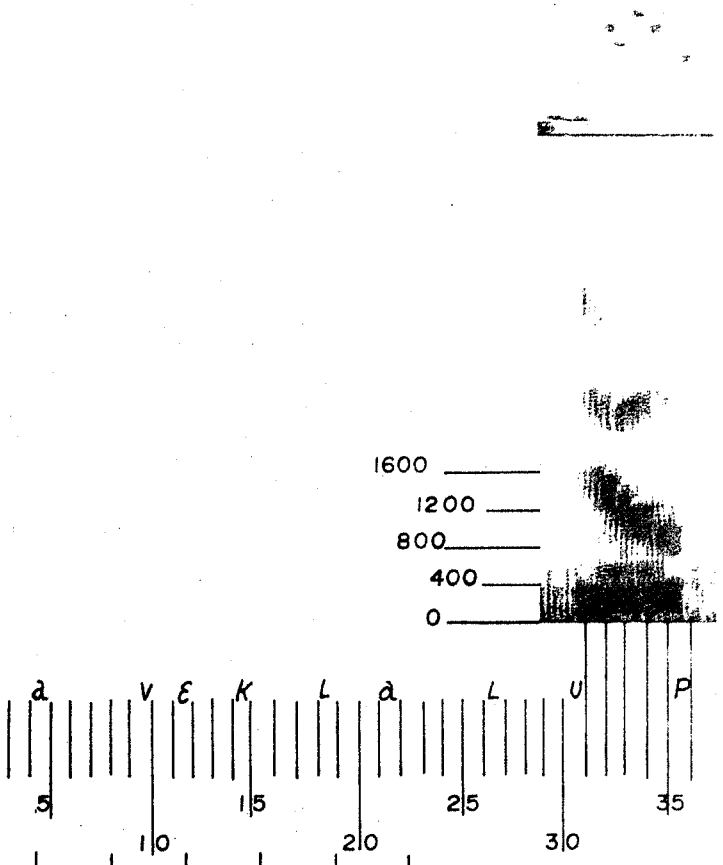
34				4	0	10	8,5	26,5	9	V	1360 320
35					2,5						1480
36	11	13	5,5	4	3,5	11	8,5	26,5	9		400
37				4		11	8,5			ɛ	
38						10,5	8,5				1600
39	11	13	10	5		9,5	6,5	26,5	8,5		480
40							3,5				1680
41	11	13	9	4		7	1	31	6,5	k	

DURÉE TOTALE DE LA VOYELLE: 12cs Fo 180

Fig. 6

AVEC LA LOUPE





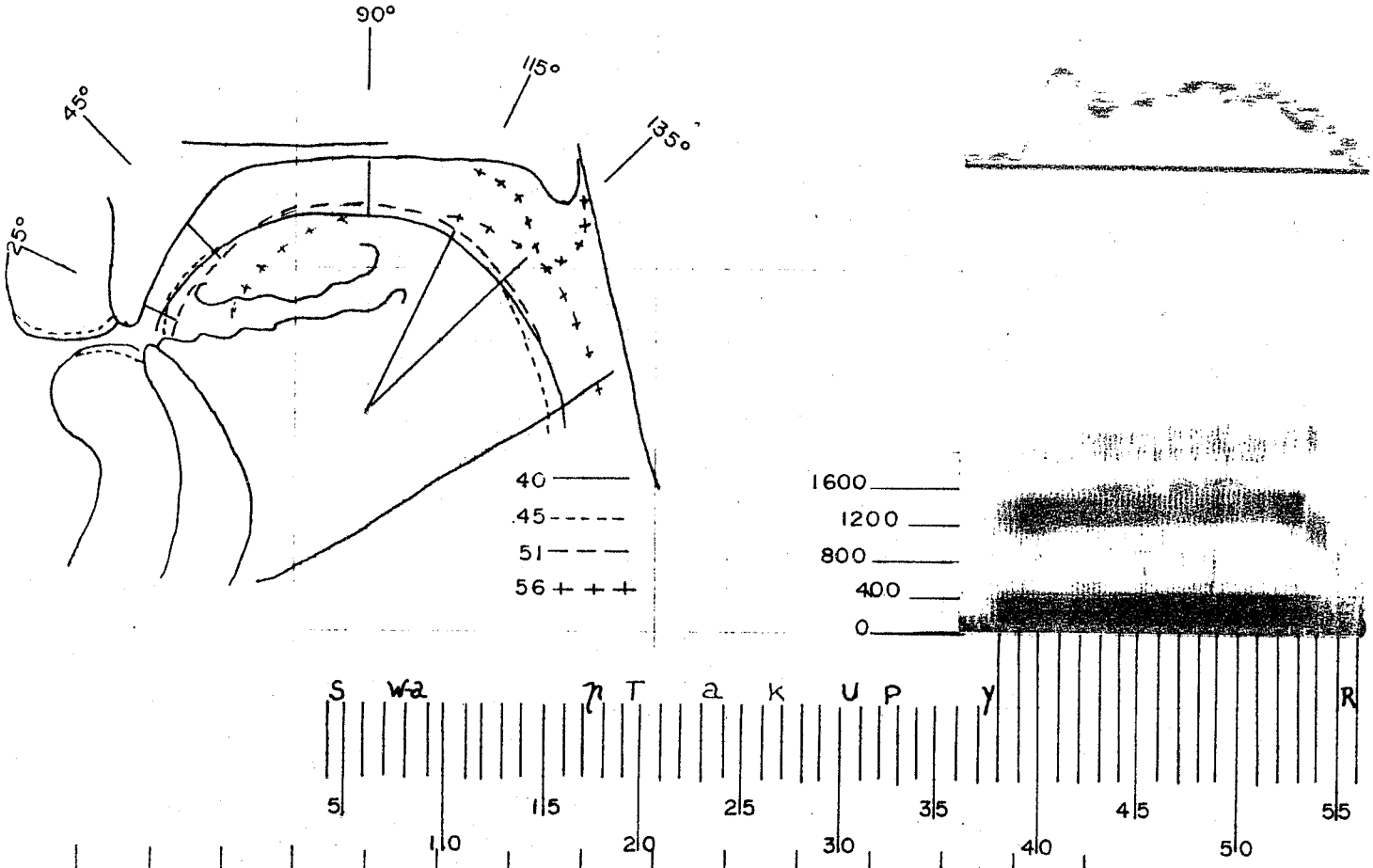
					30° L	50°		90°		125°					F2
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	F1
30	17,5	19		3	0			9		23,5				11	L 1440
31	18	20	6,5	3	3			11		24,5				10,5	240
32					11			13						9	
33	18	18	7	3	13	14		15		25,5				7	1240
34			7			14,5		16		27				5,5	320
35			2			16								5	U 800
36	12	14,5	0	3		16,5		16		27				5	P 240

DURÉE TOTALE DE LA VOYELLE: 10 cs

Fo 200

FIG. 7

SOIGNE TA COUPURE



38			1	3		5			26,5				10		
39															1280
40	13	14	1	3	2,5	5		8	26,5				10		240
41															
42			1	3	2,5				26,5				11		
43															
44															
45	13	14	2,5	3	3	4,5		8	26,5				12	y	1360
46															240
47															
48			2,5	3	4	5		7,5	26,5				11,5		
49															
50															
51	13	14	2,5	3	5	6,5		6,5	28	26			10		1500
52															240
53															
54										23,5	8		9		
55										29	3		8		1500
56			2	2,5	12					31	0,5		8	R	

DURÉE TOTALE DE LA VOYELLE : 35 cs

Fo 180

FIG. 8

Questions posées par R. CHOCHOLLE à M. SANTERRE

1°) Etant donné le très petit nombre de sujets examinés pensez-vous que votre exposé soit représentatif de l'ensemble d'une population ?

2°) N'y-a-t-il pas des différences interindividuelles importantes ? Quel est l'ordre de grandeur des variations interindividuelles et intraindividuelles ?

3°) N'y-a-t-il pas des différences suivant les langues parlées ? par exemple, Français, Russe, Arabe, Anglais, Suédois, Japonais, etc. ?

4°) Ceci est important si on veut synthétiser la parole sur ces bases ; ne serait-on pas amené à des synthétisations différentes selon des langues différentes, les dialectes, etc. ?

Réponses

1°) Les deux sujets que j'ai retenus pour les films ont été reconnus par plusieurs Montréalais auxquels je les ai présentés, comme bien représentatifs de leur classe sociale ; les autres locuteurs de leur propre milieu les ont aussi reconnus comme étant bien des leurs.

2°) Il y a toujours des différences chez le même locuteur selon les situations ; même dans des conditions absolument identiques, un même locuteur ne dit pas la même phrase exactement de la même manière ; sur 5000 images de films que j'ai tracées, je ne pense pas en avoir deux qui soient tout à fait semblables ; on peut dire la même chose sur le plan acoustique. Mais ces différences ne sont en rien significatives sur le plan phonologique. Sur le plan phonétique, il est très difficile de faire la part précise de ce qui est des caractéristiques individuelles, des caractéristiques sociales, et de ce qui est strictement pertinent pour la communication. C'est une difficulté dont les phonéticiens ont l'habitude et dont ils doivent s'accommoder, jusqu'à ce qu'on ait fait de très grands progrès dans l'étude des caractéristiques individuelles de la parole.

.../...

En attendant, nous n'avons pas l'impression de confondre ce qui est important pour la communication sémantique et les variations individuelles d'ordre phonostylistique ou sociologique. Les différences interphonémiques sont beaucoup plus grandes que les variations phonostylistiques.

3°) - 4°) Bien sûr, et il faut en tenir compte.

Question posée par M. ALINAT à M. SANTERRE

Pour les consonnes TDN vous avez montré que la notion de "Locus" n'existe pas sur des analyses de parole par sonographe.

Les expériences de DELATTRE se faisaient sur la synthèse, donc avec un appareillage différent. Quelle est, à votre avis, l'influence des caractéristiques (temps de retard et largeur de filtre par exemple) des appareils sur vos conclusions différentes ?

Réponse

Les moyens d'analyse acoustiques dont les constantes de temps seraient différentes de celles du sonographe (300 HZ) ne changeraient pas la direction des transitions que j'ai étudiées ; l'angle de montée ou de descente des transitions pourraient seulement varier un peu. Depuis assez longtemps, DELATTRE lui-même mettait des doutes sur la valeur des locus de consonnes dans la parole réelle ; et plusieurs l'avaient fait bien avant lui.

Question posée par J. S. LIENARD à M. SANTERRE

Je suis d'accord avec vous en ce qui concerne la théorie du locus, selon laquelle toute consonne est caractérisée par le point de convergence des formants vocaliques. Il est bien certain que la réalité n'est pas aussi simple : nous avons pu le constater dans nos expérimentations de synthèse avec l'Icophone, et DELATTRE, lui-même, dans ses derniers travaux a beaucoup atténué le caractère systématique de la théorie.

Réponse

Je vous remercie.

Questions posées par M. PAILLE à M. SANTERRE

1°) Connaissez-vous les travaux d'Öhman ?

2°) Avez-vous fait des mesures très précises ou ce que vous avancez est-il qualitatif et statistique?

Réponses

1°) Bien sûr, vous trouverez la référence à Öhman dans le texte ; le temps qui presse m'a obligé à passer outre à plusieurs parties de mon entretien.

2°) Mes mesures ont été scrupuleusement précises : le demi-millimètre dans l'articulation, et 40 Hz près dans les mesures formantiques. J'ai dû faire des erreurs inévitables, mais je pense que je les ai toujours faites dans le même sens, de sorte que les comparaisons ne s'en trouvent pas faussées pour autant.

Question posée par M. LINDBLOM à M. SANTERRE

I would like to reinforce your remarks concerning the importance of the position of the tongue body as a determinant of the formant pattern for apical consonants. If I understood you correctly you assign to the apex itself a subordinate role in influencing the formants. This is somewhat at variance with our own investigations of the acoustic correlates of apical movement. Could you clarify further how you arrive at this conclusion ?

.../...

Réponse

C'est l'ensemble des comparaisons que j'ai faites entre les mouvements articulatoires et les variations formantiques qui m'ont amené à penser que le seul fait pour l'apex de toucher les dents ou les alvéoles n'affectait pas la fréquence de F_2. Quand F_2 varie, c'est que par ailleurs la langue monte ou descend sous le palais dur, ou bien dégage ou rétrécit le passage pharyngal. J'en suis venu à cette conclusion avant même de connaître le modèle théorique que je propose. Mon entretien n'a pas fourni assez d'exemples à l'appui, mais l'ensemble de la documentation est beaucoup plus éclairant à ce sujet.

Question posée par M. PECKELS à M. SANTERRE

Influence du type d'analyseur (Sonograph) sur la précision des résultats obtenus. Est-ce que M. SANTERRE pense utiliser des méthodes numériques dans le futur ?

Réponse.

Je pourrais faire le même travail avec un analyseur en temps réel B et K dont nous disposons, mais il faudrait emmagasiner en mémoire d'ordinateur les données des 16 outputs et les reprogrammer pour les synchroniser avec l'articulation ; cela pose des problèmes assez difficiles. Je gagnerais sûrement en précision des résultats, mais il ne pourrait y avoir contradiction avec les conclusions auxquelles j'arrive maintenant.

SIMULATION DU CONDUIT VOCAL

SIMULATION DU CONDUIT VOCAL

J. GENIN

C.N.E.T. - LANNION

Pages

Le but de cet exposé n'est pas une étude exhaustive du problème de la synthèse de la parole par simulation du conduit vocal. Il s'agit au contraire d'une approche du problème dans laquelle on s'intéressera surtout à la validité des approximations nécessaires. On proposera enfin quelques méthodes de simulation sortant quelque peu des sentiers battus.

I. LA SYNTHESE DE LA PAROLE

I.1 Rappel historique

Il est inutile de souligner l'intérêt que peut présenter un système automatique de réponse vocale pour la gestion d'ensembles tels que magasins de stocks ou centraux téléphoniques. Plusieurs expériences de ce genre ont déjà été mises en oeuvre de par le monde et ont prouvé leur efficacité.

Le problème n'est pas nouveau. En effet il semble généralement admis d'attribuer à Van Kempelen (1773) la première expérience sérieuse dans ce domaine. On peut dire en quelques mots que sa machine parlante était un simulateur mécanique du conduit vocal fonctionnant à grand renfort de pistons, de valves et de soufflets. (1).

Les expériences de ce genre malgré les améliorations apportées par des gens comme Faber (1835) n'ont pas eu de suite, probablement pour la raison qu'elles n'avaient pas d'autre but que d'amuser les foules (ou les princes). D'autre part les technologies du moment ne leur permettaient absolument aucun développement.

C'est sans doute pour cela qu'il a fallu attendre au XXe Siècle le développement de l'électronique pour voir de nouvelles études sur la synthèse de la parole.

Les premières expériences intéressantes semblent être celles utilisant le Vocoder à canaux de Dudley (1939) et ses perfectionnements qui permettaient une représentation économique du signal de parole et présentaient l'avantage de séparer les informations portant sur le timbre la hauteur et le rythme de la parole.

Cet appareil ou des appareils s'en approchant de plus ou moins loin permettent de réaliser des synthèses de parole par juxtaposition de mots (3, 4) ou de syllabes (5, 6), la nature de l'appareil permettant des ajustements de la hauteur et du rythme de la parole.

Plus récemment est apparu le synthétiseur à formants (7, 8) qui tirait parti d'une meilleure connaissance du phénomène phonatoire. En plus des possibilités du Vocoder à canaux ce nouvel appareil a permis l'accès à la synthèse par règles (9).

Enfin pour serrer le plus près possible la réalité physique du phénomène, on est revenu au principe le plus ancien : la simulation du conduit vocal à l'aide de tout l'arsenal de l'électronique et de l'informatique modernes.

I.2 Quelques réalisations

Les premiers appareils construits étaient des réseaux statiques d'inductances et de capacités sélectionnées à l'aide de commutateurs et aptes à la seule production de voyelles soutenues ou de quelques consonnes fricatives (10, 11, 12).

Sont apparus ensuite des appareils plus intéressants, simulateurs dynamiques aptes à la production de syllabes ou de parole continue. Les méthodes utilisées sont des plus diverses : effet Miller et inductances à noyau saturé (13), circuits actifs (11), réalisation sur calculateur analogique (14), simulations sur calculateurs numériques...

II. THEORIE DU FONCTIONNEMENT DU CONDUIT VOCAL

II.1 Les équations

Avant de concevoir un appareil de simulation du conduit vocal il est bon de poser quelques éléments d'une théorie de son fonctionnement.

Quelques hypothèses préliminaires sont indispensables :

On supposera que le comportement du conduit vocal est identique à celui d'un tuyau rectiligne dans lequel ne se propagent que des ondes acoustiques planes normales à l'axe. Cette hypothèse qui consiste à ne pas tenir compte de la propagation des modes transversaux est valable en première approximation si les dimensions du tube sont inférieures aux longueurs d'onde étudiées, soit pour des fréquences inférieures à 4000 Hz (fig. 1)

.../...

Dans ces conditions le conduit vocal est caractérisé par sa longueur l et sa section $s(x,t)$, variable à la fois le long de l'axe et dans le temps, et son comportement est donné par les fonctions $p(x,t)$ et $v(x,t)$ pression acoustique du gaz et vitesse selon l'axe du tuyau supposées uniformes dans chaque section droite.

Enfin dans une première approche on ne traitera que la propagation des ondes sonores à l'intérieur du conduit supposé sans pertes, laissant de coté le problème des impédances de fermeture aux accès.

Quelques raisonnements simples permettent d'établir les équations du conduit vocal.

II.1.1 Equation de continuité (Fig. 2)

Considérons ce qui se passe dans une tranche d'épaisseur dx du tuyau. La masse gazeuse y contenue est égale à $\rho S dx$.

La masse de gaz qui y pénètre pendant le temps dt par la section droite d'abscisse x est $dt(\rho Sv)_x$, celle qui sort pendant le même temps par la section droite d'abscisse x+dx est : $dt(\rho Sv)_{x+dx} = dt\left[(\rho Sv)_x + \frac{\partial}{\partial x}(\rho Sv)dx\right]$ (1)

donc : $$\frac{\partial(\rho Sv)}{\partial x} = -\frac{\partial(\rho S)}{\partial t} \qquad (2)$$

$$\rho\frac{\partial Sv}{\partial x} + Sv\frac{\partial \rho}{\partial x} = -\rho\frac{\partial S}{\partial t} - S\frac{\partial \rho}{\partial t} \qquad (3)$$

$$\boxed{\frac{\partial Sv}{\partial x} = -\frac{S}{\rho}\frac{d\rho}{dt} - \frac{\partial S}{\partial t}} \qquad (4)$$

II.1.2 Equation thermodynamique

L'état du gaz est représenté par une équation telle que $\rho = \rho(P,T)$ T étant la température. Sous l'hypothèse de mouvements adiabatiques d'un gaz parfait cette équation devient $P/\gamma = cst$ (γ étant un nombre constant qui vaut 1.4 pour l'air). Introduisons la vitesse du son : $c = 1/\sqrt{\partial\rho/\partial P}$

L'équation (4) devient :

$$\boxed{\frac{\partial Sv}{\partial x} = -\frac{S}{\rho c^2}\frac{dP}{dt} - \frac{\partial S}{\partial t}} \qquad (5)$$

Remarquons que cette équation serait toujours valable sous l'hypothèse de mouvements isothermes ou $\frac{p}{\rho} = cst$ avec une valeur différente $c' = c/\gamma$ de la vitesse du son.

II.1.3 Equation fondamentale de la dynamique

Considérons la tranche de gaz située entre les sections droites d'abscisses x et x + dx.

La force appliquée à cette tranche, différence des forces appliquées sur chacune de ses faces est $dF = -S\frac{\partial p}{\partial x} dx$ (Figure 3).

L'équation fondamentale de la dynamique s'écrit :

$$-S\frac{\partial p}{\partial x} dx = \rho S dx . \frac{dv}{dt} \qquad (6)$$

$$\boxed{\frac{\partial p}{\partial x} = -\rho \frac{dv}{dt}} \qquad (7)$$

II.2 Approximations

Dans le comportement du conduit vocal les amplitudes des écarts de pression et donc de masse volumique restent très faibles (Au-dessous des cordes vocales l'excès de pression provoqué par l'action des poumons est de l'ordre de 10 cm H_2O soit 1 % de la pression atmosphérique).

On pourra donc dans les équations (5) et (7) dégagées ci-dessus remplacer ρ par ρ_0 masse volumique dans les conditions de repos du conduit vocal.

Ces mêmes équations contiennent des dérivées totales $\frac{dp}{dt}$ et $\frac{dv}{dt}$ qui rendent leur étude extrêmement délicate . Si on considère que la vitesse du gaz est très faible vis à vis de la vitesse du son* on pourra écrire :

$$\frac{dp}{dt} = \frac{\partial p}{\partial t} + v\frac{\partial p}{\partial x} \# \frac{\partial p}{\partial t}$$

*Pendant la production du son voisé, le débit maximum d'air fourni par les vocales est de l'ordre de 500 cm^3/s pour une section de CV supérieure à 0,5 cm^2 la vitesse correspondante est inférieure à 10 m/s.

(En régime cissoïdal par exemple cette relation s'écrit

$$p = p_0 + \Delta p\, e^{j\omega(t - x/c)} \qquad \frac{dp}{dt} = \Delta p\, e^{j\omega(t - x/c)} \cdot j\omega\left(1 - \frac{v}{c}\right) \quad)$$

Sous ces conditions les équations du mouvement deviennent :

$$\boxed{\frac{\partial S v}{\partial x} = -\frac{S}{\rho_0 c^2}\cdot\frac{\partial p}{\partial t} - \frac{\partial S}{\partial t}} \quad \text{et} \quad \frac{\partial p}{\partial x} = -\rho_0 \frac{\partial v}{\partial t}$$

Remarquons que :

$$\frac{\partial S v}{\partial t} = S\frac{\partial v}{\partial t} + v\frac{\partial S}{\partial t}$$

donc :

$$\boxed{\frac{\partial p}{\partial x} = -\frac{\rho_0}{S}\frac{\partial S v}{\partial t} + \frac{\rho_0}{S} v \frac{\partial S}{\partial t}}$$

II.3 Analogies

Parallèlement à l'étude précédente cherchons à déterminer les équations du comportement d'une ligne électrique sans perte, constituée d'une inductance linéique et d'une capacité linéique L(x,t) et C(x,t) (Figure 4).

Si $\varphi(x,t)$ représente le flux magnétique embrassé par la ligne jusqu'à l'abscisse x, et si q(x,t) représente la charge de la ligne jusqu'au même point.

On a $\frac{\partial \varphi}{\partial x} = L i \qquad \frac{\partial q}{\partial x} = C v$

i(x,t) et v(x,t) étant le courant et la tension présents sur la ligne.

En dérivant ces équations par rapport à t il vient

$$\frac{\partial v}{\partial x} = -\frac{\partial}{\partial t}\frac{\partial \varphi}{\partial x} = -\frac{\partial L i}{\partial t}$$

$$\frac{\partial i}{\partial x} = -\frac{\partial}{\partial t}\frac{\partial q}{\partial x} = -\frac{\partial C v}{\partial t}$$

ou

$$\boxed{\begin{aligned} \frac{\partial v}{\partial x} &= -L\frac{\partial i}{\partial t} - i\frac{\partial L}{\partial t} \\ \frac{\partial i}{\partial x} &= -C\frac{\partial v}{\partial t} - v\frac{\partial C}{\partial t} \end{aligned}}$$

.../...

En choississant des systèmes d'unités mécaniques et électriques tels que l'on puisse écrire $V = p$, $i = Sv$ les équations ci-dessus mettent en évidence une analogie de comportement du conduit vocal supposé invariant dans le temps et d'une ligne électrique définie par:

$$L(x) = \frac{\rho_0}{S(x)} \qquad C(x) = \frac{S(x)}{\rho_0 c^2}$$

La symétrie des équations montre une autre analogie définie par : $V = Sv$ $i = p$.

S'agissant du conduit vocal variant au cours du temps, on peut écrire, avec la première analogie :

$$i\frac{\partial L}{\partial t} = Sv\frac{\partial}{\partial t}\left(\frac{\rho_0}{S}\right) = v\frac{\rho_0}{S}\frac{\partial S}{\partial t}$$

et

$$V\frac{\partial C}{\partial t} = p\frac{\partial}{\partial t}\left(\frac{S}{\rho_0 c^2}\right) = \frac{p}{\rho_0 c^2}\frac{\partial S}{\partial t}$$

Les équations écrites jusqu'ici étaient valables dans les deux hypothèses de mouvement isotherme et de mouvement adiabatique . Il est indispensable ici de faire une distinction :

En régime isotherme, le gaz étant supposé parfait :

$$\frac{dp}{p_0} = \frac{d\rho}{\rho_0}$$

$$c^2 = \lambda/(d\rho/dp) = p_0/\rho_0$$

Le facteur $p/\rho_0 c^2 \# p_0/\rho_0 c^2$ vaut donc 1 et l'analogie électrique rend bien compte du comportement du conduit vocal.

En régime adiabatique où l'on a $\frac{dp}{p_0} = \gamma\frac{d\rho}{\rho_0}$

$$c^2 = \lambda/(d\rho/dp) = \gamma p_0/\rho_0$$

Le facteur $p/\rho_0 c^2 \# p_0/\rho_0 c^2$ vaut maintenant $\frac{1}{\gamma} = 0{,}7$ l'analogie électrique s'écarte du modèle acoustique quant aux variations du conduit vocal au cours du temps.

Sans pouvoir donner des valeurs numériques on constatera cependant que, le conduit vocal n'étant soumis qu'à des mouvements lents (constantes de temps supérieures à 50 ms) ce défaut perd de l'importance pour les fréquences du signal acoustique élevées.

II.4 Limites de validité des analogies

Rappelons que les analogies entre ligne et conduit vocal ne seront valables que pour de faibles signaux et pour de faibles vitesses de déplacement du gaz. Il faudra en particulier s'attendre à des écarts lors de la production de consonnes fricatives, pour lesquelles le débit d'air au niveau des cordres vocales peut atteindre 2000 cm^3/S et qui sont caractérisées par la présence d'étranglement du conduit vocal. Pour une section de 0,1 cm2, la vitesse atteindra 200 m/s.

Les mouvements de l'air étant supposés adiabatiques on a vu également que pour les fréquences basses lorsque le conduit vocal se déforme l'analogie perd de sa rigueur.

Enfin, utilisant cette théorie on réalisera un simulateur de conduit vocal en construisant une ligne électrique à constantes localisées constituée de cellules inductance-capacité correspondant chacune à une tranche du conduit vocal.

Il faut voir deux étapes dans cette quantification. En premier lieu, on approche la fonction $S(x,t)$ continue en x et en t par une fonction $S'(x,t)$ continue en t et "en escalier" par rapport à x. Cette approximation est valable si la longueur du pas Δx est petite devant la longueur d'onde des sons à transmettre qui vaut à 3000 Hz 10 cm environ ; on prend en général $\Delta x = 1$ cm.

La deuxième étape de la quantification consiste à approcher chaque tranche de tuyau cylindrique, analogue d'un tronçon de ligne électrique à constantes réparties, par un tronçon de ligne à constantes localisées. Afin de ne pas multiplier les inductances et les capacités à commander on utilise généralement une seule cellule en π ou en T ou même en Γ. L'impédance caractéristique d'une ligne à constantes localisées sans pertes d'inductance et capacité linéiques L et C vaut $Z_c = \sqrt{\frac{L}{C}}$

L'impédance caractéristique d'une cellule en π vaut

$$Z(\omega) = Z_c \frac{1}{\sqrt{1 - \frac{LC\omega^2\Delta x^2}{4}}} = Z_c \frac{1}{\sqrt{1 - \frac{\omega^2 \Delta x^2}{4c^2}}}$$

$$Z(\omega) \# Z_c\left(1 + \frac{\omega^2 \Delta x^2}{8 c^2}\right)$$

A 3 000 Hz, pour $\Delta x = 1$ cm, $Z(\omega)$ s'écarte de la valeur théorique Z_c de 5 %.

Cette considération peut apporter un certain intérêt à la méthode de simulation qui va être présentée ci-dessous dans laquelle les éléments inductances et capacités sont des éléments invariables permettant ainsi une quantification plus fine par rapport à x sans augmenter la nombre de circuits à commander.

Enfin si la commande des sections est échantillonnée dans le temps il faudra également s'attendre à voir apparaître un bruit de quantification aux fréquences harmoniques de la fréquence d'échantillonnage. Ce défaut, atténué si la mise à jour de la forme du circuit vocal est synchrone des vibrations des cordes vocales, ne disparaîtra complètement que pour une fréquence d'échantillonnage supérieure à 10 kHz ou pour une commande continue comme il est envisagé pour la méthode présentée plus loin.

II.5 Les pertes - les conditions aux limites

Dans une ligne électrique comportant des pertes, celles-ci sont représentées par une résistance et une conductance de fuite linéiques R et G. Il était naturel de représenter les pertes acoustiques du conduit vocal sous la même forme. On trouvera dans (2) une évaluation de ces termes.

Dans ces conditions les équations deviennent :

$$\frac{\partial S_v}{\partial x} = -\frac{S}{\rho_0 c^2}\frac{\partial p}{\partial t} - G p - \frac{\partial S}{\partial t} \qquad \frac{\partial p}{\partial x} = -\frac{\rho_0}{S}\frac{\partial S_v}{\partial t} - R S_v + \frac{\rho_0}{S}\frac{\partial S}{\partial t}$$

De la même façon, il est utile d'évaluer les impédances de fermeture de la ligne électrique analogue aux points correspondant à la bouche et aux cordes vocales (2).

.../...

III. UNE METHODE DE SIMULATION ANALOGIQUE

Rappelons les équations qui régissent le comportement du conduit vocal d'une part, et d'une ligne électrique d'autre part, tous deux supposés sans pertes et indépendants du temps pour alléger les équations.

Conduit vocal

$$\begin{cases} \dfrac{\partial p}{\partial x} = -\dfrac{\rho_0}{S}\dfrac{\partial S_v}{\partial t} \\ \dfrac{\partial S_v}{\partial x} = -\dfrac{S}{\rho_0 c^2}\dfrac{\partial p}{\partial t} \end{cases}$$

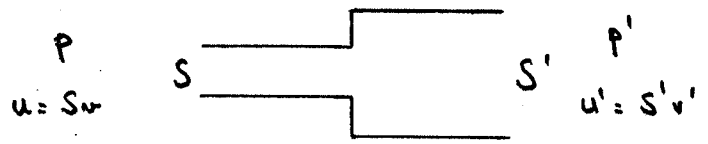
Ligne électrique

$$\begin{cases} \dfrac{\partial V}{\partial x} = -L\dfrac{\partial i}{\partial t} \\ \dfrac{\partial i}{\partial x} = -C\dfrac{\partial V}{\partial t} \end{cases}$$

L et C : inductance et capacité linéiques

La première méthode de simulation qui vient à l'esprit consiste à construire une ligne électrique composée de capacités ayant un point à la masse et d'inductances "en l'air", les valeurs de tous ces éléments pouvant être commandées par des tensions. La réalisation des inductances posant quelques problèmes nous allons nous intéresser à une méthode différente utilisant des inductances et des capacités fixes.

Approchant le tuyau acoustique par une succession de tronçons cylindriques, considérons la jonction entre deux d'entre eux :



Utilisons à gauche l'analogie définie par p = aV, u = bi nous devrons simuler la partie gauche par une ligne de caractéristiques linéiques :

$$L = \frac{b\rho_0}{aS} \qquad C = \frac{aS}{b\rho_0 c^2}$$

Utilisons à droite l'analogie définie par p' = av u' = βi

alors :

$$L' = \frac{\beta\rho_0}{\alpha S'} \qquad C' = \frac{\alpha S'}{\beta\rho_0 c^2}$$

En insérant entre ces deux lignes un quadripôle tel que $av = \alpha v'$, $bi = \beta i'$ on pourra utiliser des deux cotés de la jonction les mêmes valeurs de L et C.

Avec la condition $a/\alpha = \beta/b$, ce quadripôle est un transformateur de rapport $n = \frac{\alpha}{a}$.

En écrivant $L = L'$ et $C = C'$

On a $\frac{S'}{S} = \frac{a\beta}{\alpha b} = 1/n^2$

Cette méthode ramène le problème de la construction d'un simulateur du conduit vocal à celui de la construction de transformateurs dont on peut commander le rapport (fig. 5).

Ce problème semble d'une résolution tout aussi délicate que celui des inductances "en l'air".

III.1. Emploi des gyrateurs

D'une autre façon, gardant à gauche de la jonction la même analogie $p = aV$ $u = bi$, on peut utiliser à droite une analogie différente $p = \alpha i'$ $u = \beta v'$. Les caractéristiques linéiques de cette ligne sont :

$$L' = \frac{\gamma S'}{\rho_0 \beta c^2} \qquad C' = \frac{\rho_0 \beta}{\gamma S'}$$

Le quadripôle à insérer est alors défini par : $aV = \alpha i'$ $bi = \beta V'$

Il s'agit d'un gyrateur de résistance $Rgy = \alpha/a = b/\beta$

En introduisant $Rc^2 = \frac{L}{c} = \frac{L'}{c'}$ et en écrivant $Rc^2 = L/C'$

On a $\frac{S'}{S} = \frac{Rc^2}{Rgy^2}$

Un gyrateur dont on peut commander la résistance de gyration Rgy est réalisable à l'aide des circuits actifs modernes.

Remarquons que dans le cas de la simulation par inductance et capacité, la valeur de chacun de ces éléments devrait être commandée dans une dynamique correspondant à celle de la fonction d'aire S(x) soit de 1 à 100 selon (10).

Par contre ici c'est le carré de la résistance de gyration qui doit être commandé dans une dynamique qui n'est que le rapport maximum des aires de deux sections successives.

Il suffira donc de faire varier Rgy dans une gamme de l'ordre de 1 à 5 ou 10 au maximum.

Enfin pour affiner l'analogie entre une section de longueur 1 cm du conduit vocal et son équivalent électrique on peut envisager avec cette méthode et cela sans ajouter de circuits actifs d'utiliser deux cellules LC pour chaque section (Figure 6).

REFERENCES

(1) J.S. LIENARD. La Synthèse de la Parole : Historique et réalisations actuelles - Revue d'Acoustique n° 11 1970 pp. 204 à 213

(2) J.L. FLANAGAN. Speech Analysis, Synthesis and Perception - Springer Verlag Berlin - Heidelberg 1965

(3) R. BURON. Generation of 1000 words Vocabulary for a pulse excited Vocoder operating as an audio response unit.

I.E.E.E. Trans, AU, U.S.A. (Mars 1968), 46, n° 1 pp. 21 - 25

(3 bis) J. LOIZILLON, G. ROGER. Unités de réponse vocale - Congrès d'Informatique de l'A.F.C.E.T. 1970 - S4.3.135-158

(4) J. PONCIN - Etude d'un système de synthèse de messages vocaux
Annales des Télécommunications 1970, t. 25 n° 11-12 pp. 405-418
M. CARTIER, J. GENIN, P. LORAND - Synthèse de la parole : Une unité de réponse vocale.
L'Echo des Recherches Juillet 1971 n° 65 pp. 43-51

(5) A. NEMETH, R. BURON - Expérience de synthèse automatique de la voix à 200 bits par seconde de parole.
Colloque international de téléinformatique
PARIS (1969) t. II CHIRON, pp. 817-826

(6) J. QUINIO, G. RENARD, D. TEIL - Installation et mise en route d'une unité à réponse vocale couplée à un petit ordinateur.
Congrès d'Informatique de l'A.F.C.E.T. - 1970 S4.2.123-134.

(7) G. FANT, J. MARTONY, U. RENGMAN, A. RISBERG

OVE II - Synthesis Strategy.
Speech communication seminar - Stockholm - Août 1862 - F.5

(8) R. CARRE, J.P. BEAUVIALA, J. PAILLE : Synthèse de la Parole : Description et utilisation d'un synthétiseur du type à formants.
Revue de Physique Appliquée Vol. 5, 785 - 793 (1970)

(9) L.R. RABINER - A model for synthesizing speech by rule.
I.E.E. Trans on Audio - AU 17 pp. 7 - 13 (1969)

(10) K.N. STEVENS, S. KASOWSKI, E. FANT. Electrical analog of the vocal tract.
J.A.S.A. 25 (1953) pp. 734-742

(11) Simulateur vocal - Projet d'élèves.
E.N.S.E.R.G. (Grenoble).

(12) M. LUGAN - Simulateur Electrique du conduit vocal - Rapport de stage TMA/ETA - Publication interne C.N.E.T.

(13) G. ROSEN - Dynamic analog speech synthesizer.
J.A.S.A. n° 30 (1958) pp. 201 - 209

(14) TAKASHI HONDA, SEIICHI INOUE, Yasuo OGAWA
A hibrid control system of a human vocal tract simulator.

6e C.I.A. TOKYO - Août 1968 B5.7 pp. B 175-178

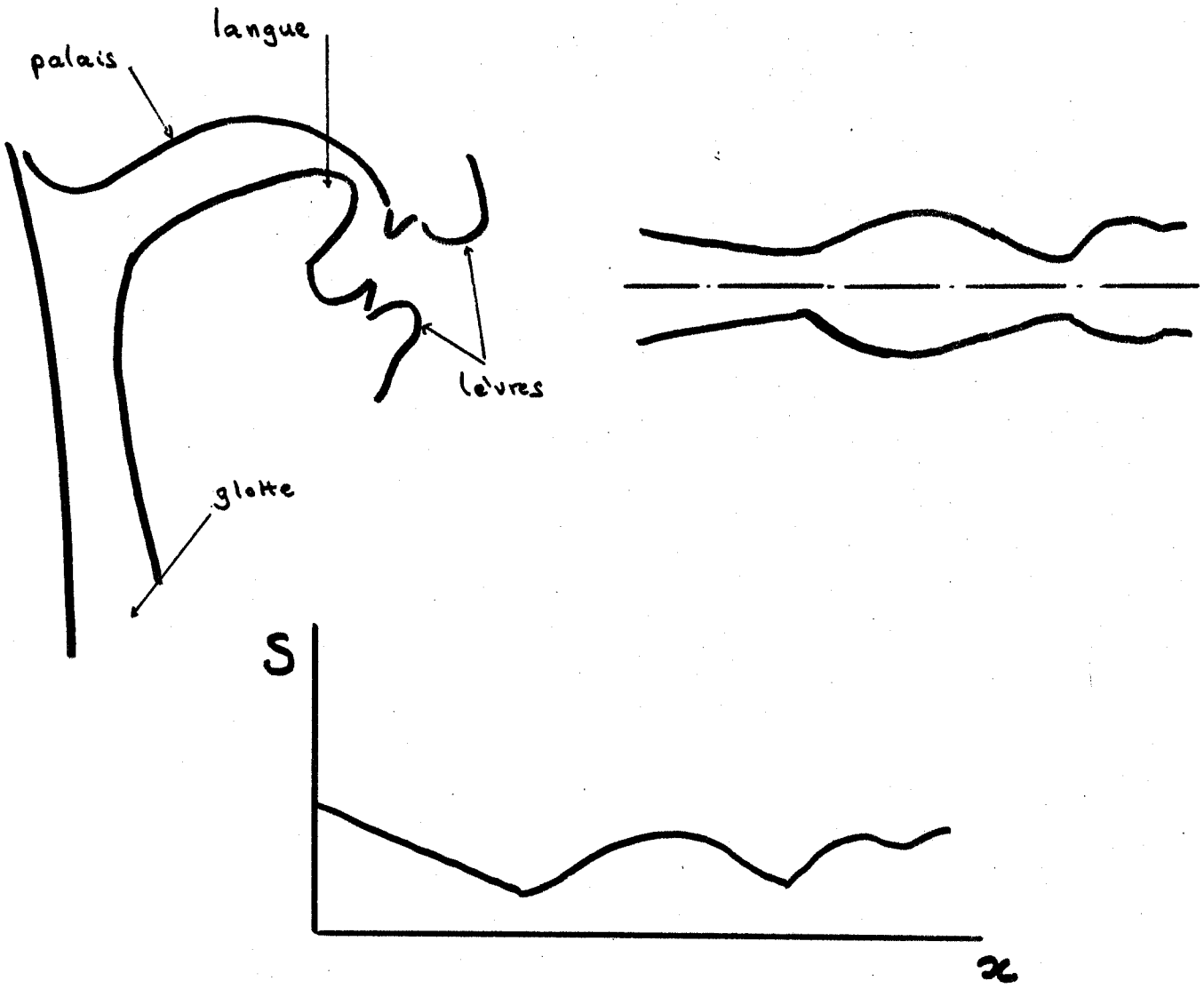
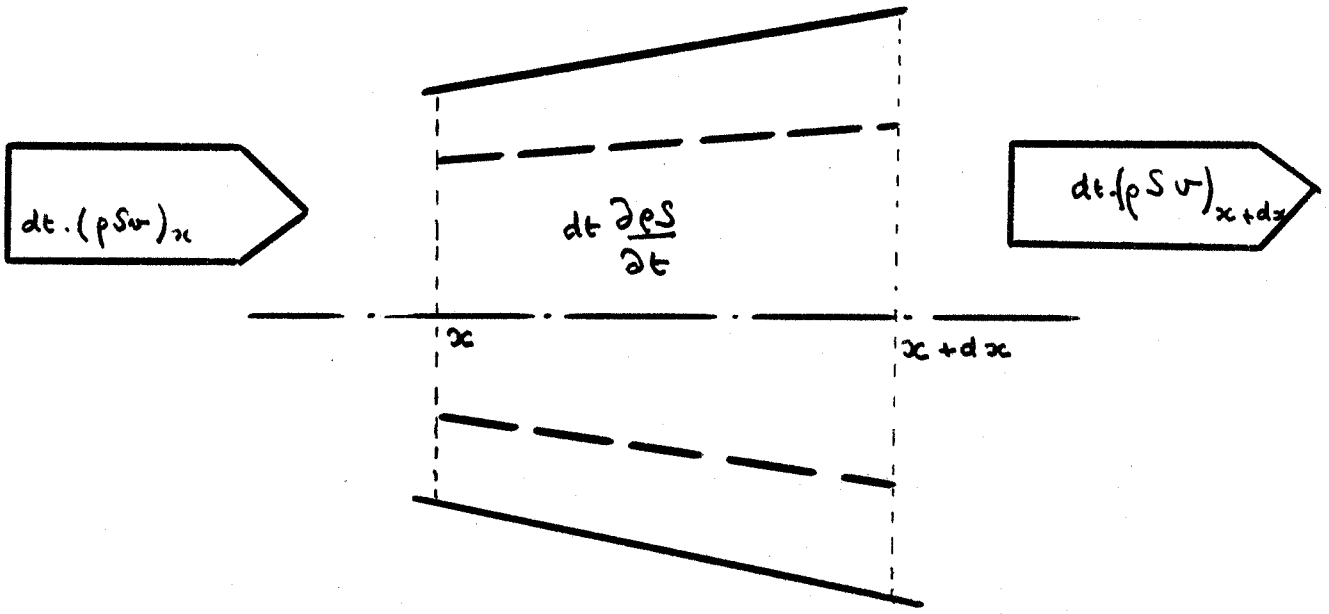


Figure 1

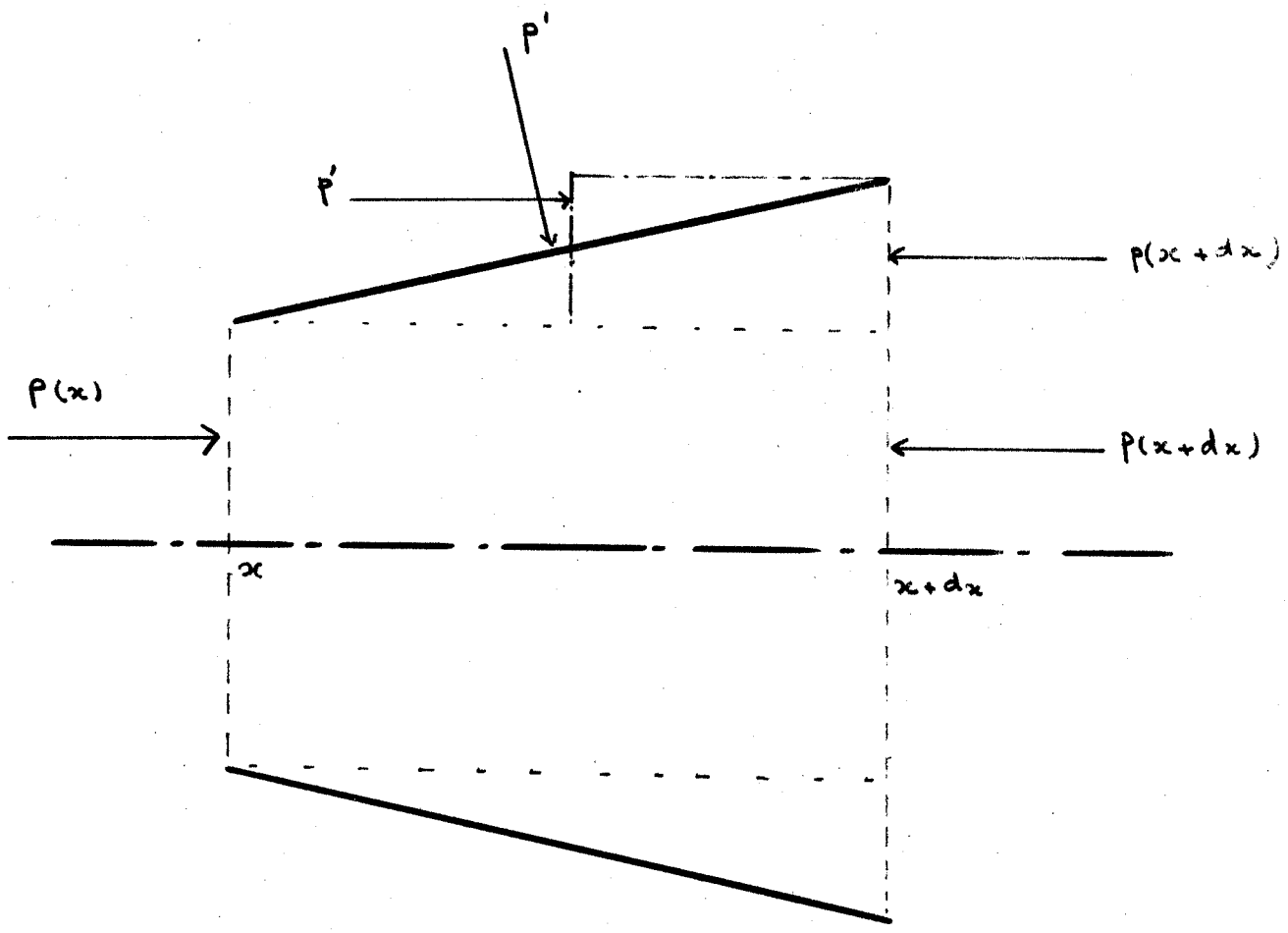


$$\frac{\partial(\rho S v)}{\partial x} = -\frac{\partial(\rho S)}{\partial t}$$

——— Position du conduit vocal à l'instant t

— — — Position du conduit vocal à l'instant $t + dt$

Figure 2



p' est comprise entre $p(x)$ et $p(x+dx)$

Si $p(x)$ est continue $dF = -S \dfrac{\partial p}{\partial x}$

Figure 3

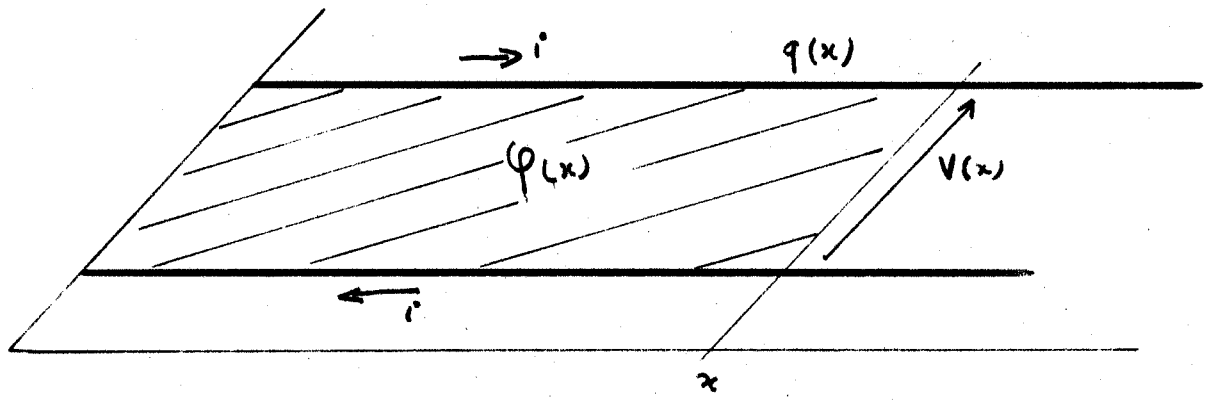


Figure 4

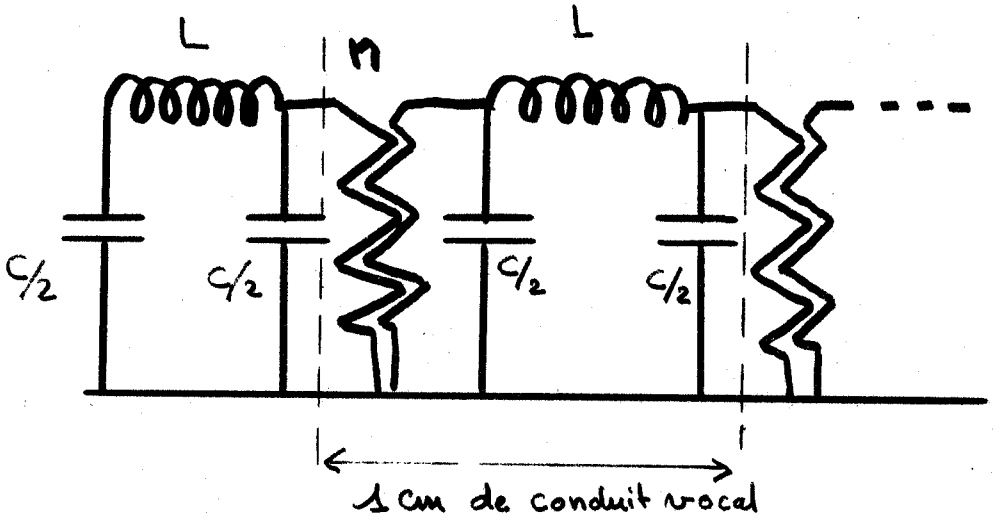


Figure 5

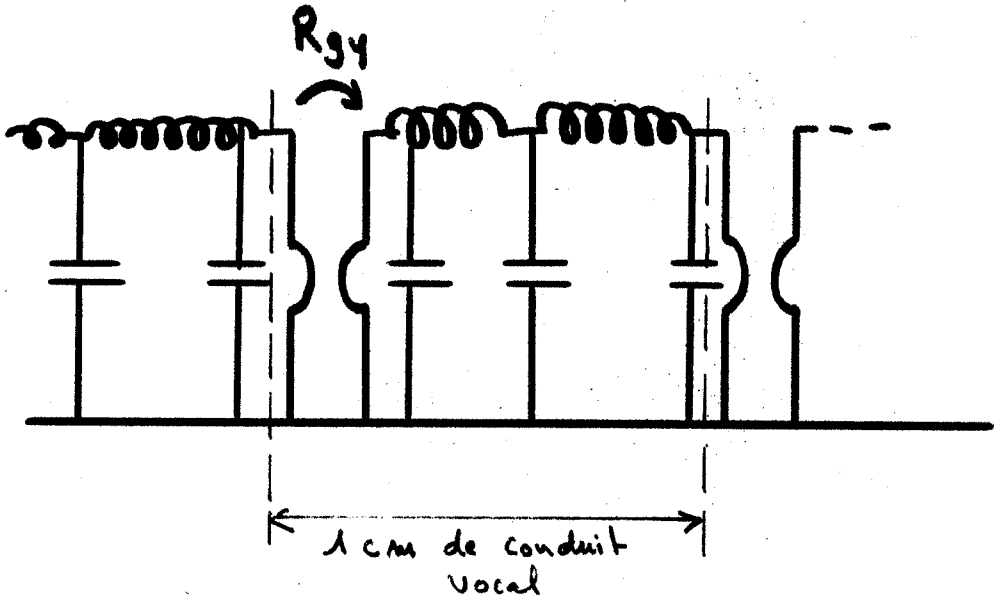


Figure 6

M. PAILLE :

Vous avez dit n'avoir pas trouvé de littérature sur les conditions de validité des hypothèses relatives à la théorie de la production de la parole. Pensez-vous qu'il n'en existe réellement pas ? En corollaire ne pensez-vous pas que les résultats que vous avancez en ce qui concerne le cas isotherme et le cas adiabatique montrent bien qu'il n'est pas nécessaire d'insister lourdement sur cette question ?

REPONSE :

Pour la première partie de la question je n'ai trouvé nulle part de développement faisant intervenir le fait que le conduit vocal se modifie au cours du temps. Tous les gens qui se sont penchés sur cette question ont dit : "le conduit vocal se modifie avec des constantes de temps de l'ordre de 50 ms et les signaux qu'on y fait passer ont des fréquences supérieures". Je n'ai pas connaissance d'étude plus approfondie sur ce point.

M. PAILLE :

Je ne suis pas tout à fait d'accord parce que si on regarde les techniques de codage prédictif, il est évident que ça ne marche que si l'évolution dynamique peut être considérée comme une succession d'états statiques, ce qui semble prouver que des gens se sont penchés sur cette question.

REPONSE :

Peut-on actuellement dire que le codage prédictif, c'est du conduit vocal ? Le but est le même : la parole, mais les approches sont peut-être différentes.

M. J.S. LIENARD :

Si j'ai bien compris vous représentez chaque section de 1 cm de conduit vocal par un circuit à paramètres fixes (self et capacité) et un gyrateur. Il faut donc 17 gyrateurs et ce sont les seules parties variables de votre système.
Ma question est la suivante : est-ce-que ça a été fait et est-ce que ça marche ?

REPONSE :

Si c'était fait et si ça marchait je vous aurais fait entendre une bande magnétique. Actuellement une cellule a été réalisée et ce qu'il faut faire maintenant c'est en mettre plusieurs bout à bout. Je m'attends à ce que de nouvelles difficultés apparaissent alors.

QUESTION :

La réalisation technique des gyrateurs est-elle une chose difficile ?

REPONSE :

Ce n'est pas très difficile si les conditions d'emploi sont bien définies et dans des plages restreintes. Dans mon modèle la dynamique de variation de la résistance de gyration est très faible mais il n'en est peut être pas de même pour les impédances auxquelles sont reliés ces gyrateurs.

QUESTION :

Serait-il plus simple de multiplier les gyrateurs pour avoir de plus petites variations ?

REPONSE :

Je considère qu'on n'a aucun intérêt à multiplier le nombre de circuits actifs dans un montage électronique. La première approximation faite oblige à se restreindre aux fréquences inférieures à 4000 Hz et ce pas de 1 cm est suffisamment fin dans ces conditions.

M. J.S. LIENARD :

Je voudrais apporter deux précisions, d'une part sur la machine de Kempelen, qui était effectivement le premier synthétiseur de parole que l'on ait connu, que j'ai reconstituée et que j'ai vue également à Munich où il en reste un petit morceau.

Je voudrais signaler que ce synthétiseur était intéressant à plus d'un titre, en particulier parce que Kempelen avait compris que la parole était mouvement et s'était arrangé pour faire une machine qui reproduisait les transitions de la parole.

Cette machine a été reprise et améliorée par Faber quelques temps plus tard. La machine de Faber a duré très longtemps et elle doit pourrir dans les caves de l'Ecole de Médecine de Paris.

Deuxième point, je crois qu'il n'est pas tout à fait faux de comparer l'Icophone à un synthétiseur de Vocoder parce qu'effectivemement, sur le plan technique, on a, comme dans le Play Back de Delattre, un certain nombre de lignes qui fonctionnent en parallèle, serait-ce en tout ou rien, mais sur le plan de l'utilisation de ce genre de synthétiseur ce n'est pas du tout comparable à un vocoder ; l'Icophone, pour nous est essentiellement une machine à reproduire des formes au sens de la Gestalt Theorie.

/

REPONSE :

Je pense que le Vocoder peut aussi être utilisé de cette façon.

M. J.S. LIENARD

Quand on dessine un lapin sur la feuille de mylar il sort un son spécifique du lapin.

REPONSE :

Dudley a conçu le Vocoder comme un transmetteur de parole et non comme un transmetteur de lapin.

M. PAILLE :

Vous avez dit que l'on ne savait pas faire de transformateur de rapport variable. Ne peut-on pas utiliser des noyaux plongeurs dans des bobines et réaliser ainsi un synthétiseur à arbre à cames en tête ?

REPONSE :

Que dire de la dynamique de variation et de la qualité ? Pour ce qui est des constantes de temps le conduit vocal est lui-même un système mécanique donc ça doit marcher.

QUESTION :

Comment pensez-vous commander cet appareil ?

REPONSE :

Dans le premier temps cet appareil sera alimenté par un calculateur, qui devra élaborer et envoyer toutes les 5 ou 10 ms la valeur de l'aire de chacune des 17 sections.

Par ailleurs je pense que la simulation du conduit vocal sera intéressante si on peut définir un modèle articulatoire et réaliser la transition des paramètres articulatoires aux aires des sections à l'aide d'un système cablé simple.

Si on est toujours obligé de sortir les 17 valeurs de sections d'un calculateur toutes les 5 ou 10 ms ça marchera mais ça sera toujours un gadget sans application pratique.

QUESTION :

Avez-vous étudié la simulation numérique de ce modèle analogique ?

REPONSE :

Non, je recule un peu devant les difficultés de programmation Je considère que si on veut parler de simulation numérique, il vaut mieux prendre une méthode totalement différente, parler non plus de selfs ni de capacités mais peut être de transformation de Fourier ou de constantes de temps de propagation avec coefficients de transmission et de réflexion. Il y a là des méthodes intéressantes qui peuvent conduire à des simulations numériques cablées peut être meilleures.

Je ne vois pas d'intérêt à une simulation numérique d'un simulateur analogique. Pour cela il est surement préférable de prendre un calculateur analogique.

M. VOMSCHEID :

Vos gyrateurs sont-ils des gyrateurs intégrés du commerce ?

REPONSE :

Non, car je crois que ces éléments présentent trop peu de marge dans la commande de la résistance de gyration.

SIMULATION DU CONDUIT VOCAL*

Bernard GUERIN

Laboratoire de la Communication Parlée & de l'Instrumentation de Mesures

E.N.S. d'ELECTRONIQUE et de RADIOELECTRICITE

23, rue des Martyrs - 38 . G R E N O B L E

* Etude entreprise avec l'appui du Comité de Recherches en Informatique.

Pages

1. DESCRIPTION D'UNE CELLULE ELEMENTAIRE POUR UN ANALOGUE DU CONDUIT VOCAL DYNAMIQUE

Si on utilise l'analogie pression-tension, vitesse volumique-intensité pour simuler un conduit vocal, on peut alors isoler, dans le réseau électrique analogue, une cellule élémentaire du type Γ composée d'une inductance variable $L = k_1/A(x)$ et d'une capacité variable $C = k_2 \cdot A(x)$;
$A(x)$ est l'aire de la section du conduit vocal à une distance x de la glotte ;
x ne prendra ici que des valeurs discrètes car l'ensemble LC simule un cylindre élémentaire du conduit vocal de longueur ℓ_t/N ;
ℓ_t est la longueur totale de la glotte aux lèvres ;
N est le nombre de sections.

Dans notre réalisation, l'inductance variable L est synthétisée à l'aide d'éléments actifs, de résistances, de capacités, en utilisant le principe du mutateur. La capacité variable C sera réalisée suivant le principe de la multiplication d'impédance, souvent appelé dans ce cas : *effet MILLER*.

1.1. Capacité variable

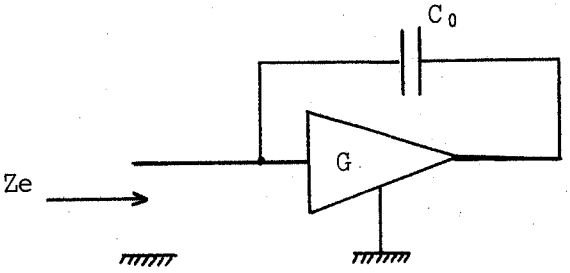
Les caractéristiques à obtenir sont les suivantes :
. 200 pF < C < 20 nF ;
. fréquence de travail utile : 50 Hz - 8 kHz ;
. facteur de qualité : > 50 pour f < 3 kHz , > 10 pour f > 3 kHz ;
. dynamique de commande : > 100.

Le schéma de principe est donné figure 1. Les contraintes sur le facteur de qualité nous conduisent, d'après un calcul rapide, à utiliser un gain faible pour la chaîne amplificatrice. En faisant varier G de 1 à -1 linéairement, on obtient une variation linéaire de la capacité de 0 à $2C_0$. L'utilisation d'un gain inférieur à 1 en valeur absolue pour la chaîne d'amplification n'apporte pas de contrainte sur la dynamique du signal d'entrée, seule la tension d'alimentation des éléments actifs fixera la limite supérieure.

FIGURE 1 - Principe de la capacité variable

$$Z_e = \frac{1}{j\, C_0(1-G)\, \omega}$$

G : gain d'un amplificateur à haute impédance d'entrée et faible impédance de sortie



La commande du gain est digitale et s'effectue avec 8 bits, ce qui permet d'obtenir facilement la dynamique désirée. En résumé, on peut obtenir, grâce à ce montage, les caractéristiques du cahier des charges.

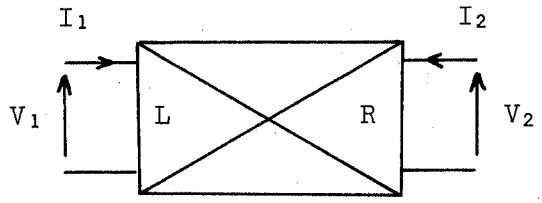
1.2. Inductance variable

Pour des fréquences relativement basses, f < 200 Hz, il est difficile d'obtenir des inductances ayant un bon facteur de qualité. De plus, la commande de sa valeur de self-inductance avec une bonne dynamique est également délicate.

Pour atteindre de bonnes performances, nous avons essayé de réaliser une inductance par synthèse à partir d'éléments actifs, de résistance, de capacités. On utilise pour cela le principe du mutateur [1]. Le mutateur est un quadripôle linéaire actif, à deux paires de bornes, qui transforme un type d'éléments en un autre type.

On distingue trois classes de mutateur : le *mutateur* R-L , le *mutateur* R-C , le *mutateur* L-C . Dans notre cas, nous utiliserons le mutateur R-L qui a la propriété de "transformer" une résistance en une inductance. Exprimons mathématiquement ses caractéristiques (figure 2).

FIGURE 2 - Mutateur R-L



$V_1 = K_1 \frac{dV_2}{dt}$; $I_1 = - K_2 I_2$;

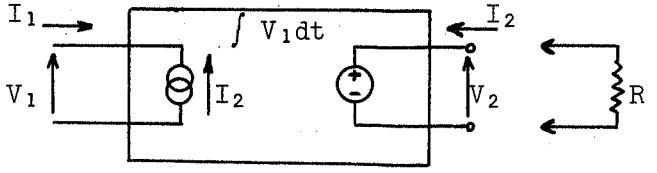
or : $V_2 = - R I_2$, donc :

$V_1 = R \frac{K_1}{K_2} \frac{dI_1}{dt}$.

On obtient ainsi l'équation caractéristique d'une inductance. On remarque que si on fait varier linéairement K_2, on réalise la relation $L = k/A$, A étant l'aire de la section variable du conduit vocal. Signalons que l'on peut également définir ce mutateur par les relations : $V_1 = - K_1'(dI_2/dt)$ et $I_2 = - K_2' V_2$.

De la première série de relations, on peut déduire, pour le mutateur, la structure donnée figure 3.

FIGURE 3 - Structure de base du mutateur



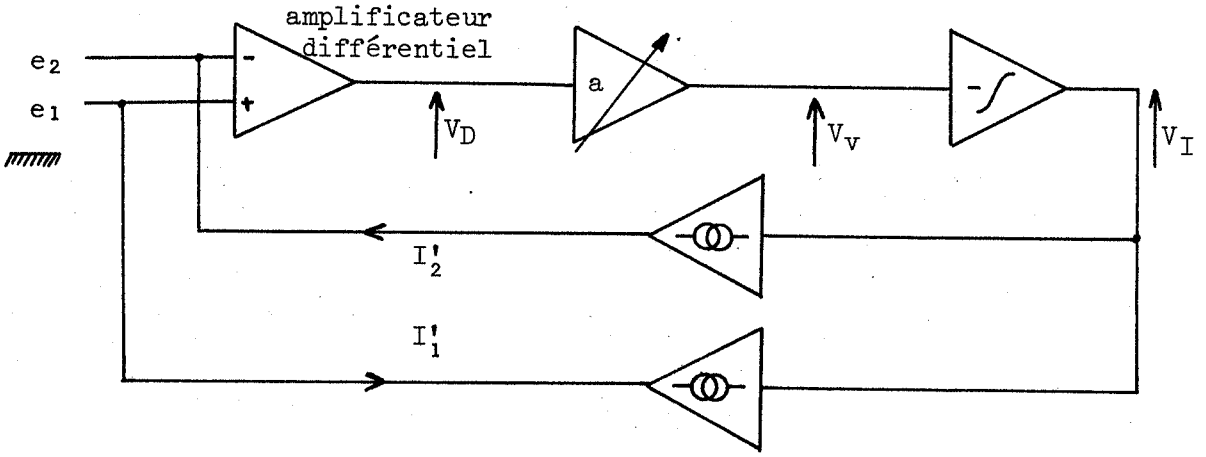
Une réalisation suivant cette structure donne une inductance avec un point à la masse. Pour synthétiser une inductance ayant les deux extrémités isolées de la masse, on a été conduit à utiliser le schéma suivant (figure 4).

FIGURE 4 - Synthèse d'une inductance isolée de la masse ; $V_D = e_1 - e_2$; $V_V = a V_D$;
a : commande linéaire

$$V_I = - \int K_1 a V_D \, dt$$

$$- I_1' = - I_2' = - K_1 K_2 a \int V_D \, dt$$

soit : $$e_1 - e_2 = \frac{1}{K_1 K_2 a} \frac{dI_1'}{dt} .$$

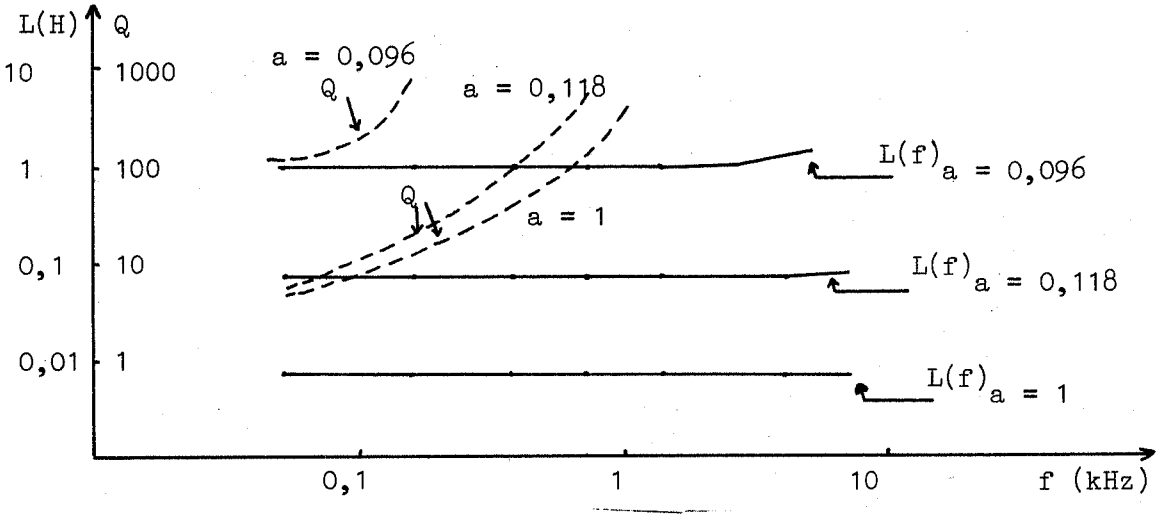


Dans ce schéma, on suppose que l'on a une résistance R fixe, ce qui implique que le courant I_1' (et I_2') est directement proportionnel à V_I.

Le cahier des charges de l'inductance à réaliser est le suivant :
. 20 mH < L < 2 H ;
. fréquence de travail utile : 50 Hz - 8 kHz ;
. facteur de qualité : Q > 10 à 50 Hz.

Les mesures montrent qu'une bonne linéarité est obtenue dans une dynamique de 100 quelle que soit la fréquence de travail, sauf pour a = 0,01 où on note une erreur, quand f > 5 kHz, atteignant 13 % à 10 kHz. Le facteur de qualité est supérieur à 10 dès que f > 100 Hz, mais il atteint très rapidement des valeurs importantes (figure 5).

FIGURE 5 - Courbes caractéristiques de l'inductance variable



1.3. Conclusion

On a cherché à réaliser des éléments variant linéairement avec une tension de commande. Le facteur de qualité doit également être élevé. Cette dernière caractéristique est difficile à contrôler directement sur l'élément considéré (L ou C), mais on pourra, grâce à des résistances additionnelles judicieusement placées, obtenir le facteur de qualité désiré, pourvu qu'il soit suffisamment important au départ.

2. SIMULATION DIGITALE DU CONDUIT VOCAL

Nous venons de voir que la réalisation d'un analogue électrique du conduit vocal soulève des problèmes d'ordre technologique. La construction d'une cellule élémentaire représente déjà un circuit électronique assez complexe. On peut alors penser que l'utilisation des techniques digitales serait plus satisfaisante du point de vue de la stabilité et de la précision. On peut également penser ne simuler qu'une cellule du conduit vocal et effectuer le calcul de tout le conduit par recirculation. On verra, dans les exemples qui vont suivre, qu'il est difficile de trouver un bon principe de simulation.

Nous allons donc décrire rapidement deux exemples de simulation digitale. Le premier, d'après EIICHI MATSUI, utilise l'algorithme de la transformée de FOURIER rapide, et le second, d'après ATAL repris par MERMELSTEIN, simule la fonction de transfert du conduit vocal par un filtre digital.

2.1. *Simulation utilisant la transformée de FOURIER rapide* [2]

Le conduit vocal est divisé en m sections cylindriques d'égale longueur ℓ, d'aire variable A_j ($j = 1, \ldots m$). Chacun des tubes est équivalent à une inductance série $L_j = \rho\ell/A_j$ et deux capacités parallèles $C_j/2 = A_j\ell/2\rho c^2$, où ρ est la densité de l'air et c la vitesse du son.

On suppose que le facteur de qualité Q de ces éléments augmente avec f jusqu'à une certaine fréquence f_c et décroît au-dessus. Cette hypothèse semble être confirmée par des données expérimentales. On introduit alors le paramètre :

$$\lambda = \varepsilon + \frac{j\omega\tau}{1 + j\,\omega\,\tau_H}$$

où $\tau = \ell/c$ et où la fréquence critique est donnée par $\omega_c^2\,\tau\,\tau_H = \varepsilon$.

Le facteur de qualité à 1 kHz est donc égal à : $Q = 2\pi.10^3\tau/\varepsilon$.

On calcule la fonction de transfert $H(\omega)$, définissant le rapport du débit au niveau des lèvres à celui au niveau de la glotte. On obtient :

$$H(\omega) = \frac{2\lambda}{i\,\omega\,\tau}\left(\frac{\ell}{\ell_{eff}} + i\,\omega\,\tau\right)\frac{2}{\Delta}\prod_{j=2}^{m}\frac{2A_j}{A_{j-1} + A_j}$$

où $\ell_{eff} = \frac{8}{3\pi}\sqrt{\frac{A_m}{\pi}}$ (correspondant à la correction d'extrémité du conduit vocal), et où Δ est le déterminant d'une matrice dont les termes sont fonctions de A_j et λ .

L'expression du signal de sortie du système peut être obtenue en tenant compte de l'excitation et du rayonnement, soit : $S(\omega) = E(\omega)\,H(\omega)\,R(\omega)$.

On calcule la transformée de FOURIER inverse de $S(\omega)$ et on en déduit la réponse impulsionnelle. Le calcul est effectué à l'aide de l'algorithme de la transformée de FOURIER rapide.

On considère ici la parole représentée par une suite de réponses impulsionnelles du conduit vocal, chacune d'elles étant appelée *élément de son* . La parole continue est obtenue en assemblant une série d'éléments de son synchronisés par la période de mélodie.

Si on accepte une bande passante limitée à 5 kHz pour la parole synthétisée, la simulation est faite dans une échelle de temps égale à 5 fois le temps réel sur un IBM.360/75.

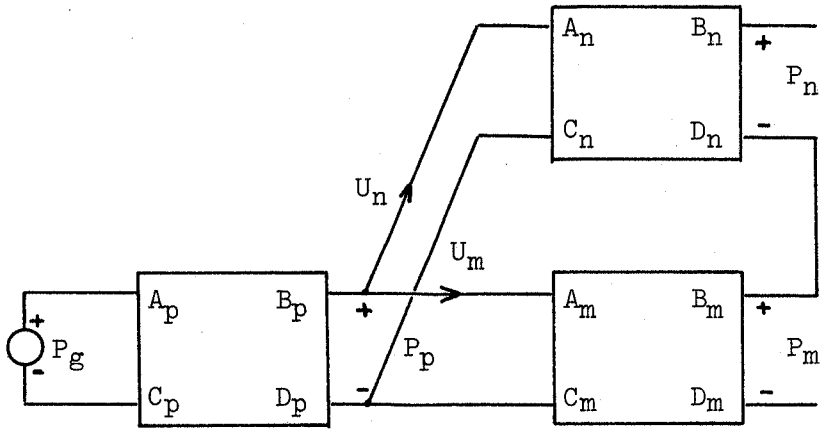
2.2. *Simulation du conduit vocal par un filtre digital* [3,4]

On peut exprimer la fonction de transfert du conduit vocal à partir de sa fonction d'aire et l'exprimer comme une fraction rationnelle dans le plan z. Cette expression se déduit directement du signal de parole mais également de la forme du conduit vocal. La ligne artificielle servant de modèle comprendra des éléments réactifs et dissipatifs représentatifs du conduit vocal et du conduit nasal. Le signal de parole de synthèse est alors obtenu à la sortie d'un filtre digital récursif déduit de la fraction rationnelle en z.

2.2.1. *Fonction de transfert du conduit vocal*

Le modèle complet du conduit vocal est donné figure 6.

FIGURE 6 - Représentation du circuit équivalent
pour le calcul de la fonction de transfert



Le conduit vocal est divisé en tubes cylindriques uniformes, de longueur 0,875 cm. Les sections du conduit vocal sont variables ainsi que les deux premières sections du conduit nasal. Si les impédances de source et de rayonnement sont incluses dans les coefficients A, B, C et D, on peut trouver la relation qui relie les pressions à la sortie du conduit vocal (p_m) et nasal (p_n) et U_g, débit issu de la glotte :

$$\frac{p_m+p_n}{U_g} = \frac{A_n + A_m}{A_n(C_p A_m + C_m D_p) + A_m D_p C_n}$$, $C_n \neq 0$ dans le cas d'un couplage du conduit nasal,

sinon : $$\frac{p_m}{U_g} = \frac{1}{C_p A_m + C_m D_p}$$, $C_n = 0$.

Considérons maintenant Y_N, admittance caractéristique de la $N^{ième}$ section, A_N : aire de sa section.

Soit $r_N = (A_N - A_{N-1})/(A_N + A_{N-1})$ le coefficient de réflexion de la jonction des sections N et N-1, α_N est l'atténuation apportée par une section. Si $\Gamma_N(z^{-1})$ est le coefficient de réflexion généralisé, vu de la $N^{ième}$ section vers la glotte, et si $\Gamma_N(z^{-1})$ est de la forme $P_N(z^{-1})/Q_N(z^{-1})$, l'application itérative des équations suivantes (le paramètre de retard étant z^{-1}) :

$P_N(z^{-1}) = \alpha_{N-1} z^{-1} P_{N-1}(z^{-1}) + r_N Q_{N-1}(z^{-1})$ et : $Q_N(z^{-1}) = Q_{N-1}(z^{-1}) + \alpha_{N-1} z^{-1} r_N P_{N-1}(z^{-1})$

pour N = 2, ... k (k étant environ le nombre de sections), permet de calculer Γ_k.

Pour le conduit nasal, on obtient les relations :

$A_n = A_k(z^{-1}) = Q_k(z^{-1}) + P_k(z^{-1})$; $C_n = C_k(z^{-1}) = y_k\left(Q_k(z^{-1}) - P_k(z^{-1})\right)$

y_k : impédance de charge du conduit nasal.

On détermine de la même manière les autres coefficients. La configuration du conduit vocal est supposée stationnaire pendant une période de mélodie.

La fonction de transfert alors déterminée a comme dénominateur un polynôme du $34^{ème}$ ordre (approximativement) pour un son nasalisé. Un si grand nombre de pôles demanderait un temps de calcul assez long pour un signal échantillonné à 10 kHz. Aussi, il est intéressant d'exprimer la fonction de transfert avec le paramètre $z^1 = z^2$ (transformation valable pour f < 5 kHz).

De plus, on montre expérimentalement qu'un prédicteur d'ordre 12 donne une parole de synthèse de qualité sensiblement comparable à celle d'un prédicteur d'ordre supérieur. On pourra donc utiliser un filtre digital d'ordre 10 ou 12 (5 ou 6 pôles complexes).

2.2.2. Résultats

Pour une simulation complète, les temps de calcul sont assez longs : 25 fois le temps réel avec un processeur HONEYWELL 635 pour 5 kHz de bande passante, les données étant la fonction d'aire. MERMELSTEIN a pu faire des études comparées, grâce à cette méthode, sur des voyelles anglaises, nasalisées ou non. Il remarque que la synthèse avec un filtre digital ne contenant que des pôles donne des sons moins naturels que celle tenant compte des pôles et des zéros.

En résumé, on peut dire que l'inconvénient majeur de cette méthode de simulation est qu'elle met en oeuvre des moyens de calcul très importants.

3. SIMULATION DE LA SOURCE DE BRUIT [5]

3.1. Introduction

Pour concevoir un analogue du conduit vocal, il est nécessaire d'étudier la simulation de tubes cylindriques élémentaires de section variable, de la source vocale, de la source de bruit. Ensuite, se pose le problème de la commande générale de cet analogue. Mais on peut penser que le nombre de paramètres sera réduit avec ce type d'appareil, en particulier si on simule certaines contraintes articulatoires. Par exemple, la simulation correcte d'une source de bruit devrait être effectuée sans paramètres particuliers. En effet, dans le fonctionnement réel, cette source de bruit prend naissance automatiquement lorsque certaines conditions de constriction et de pression sont réalisées.

Nous allons d'abord étudier théoriquement le phénomène de production de bruit, puis la simulation d'une source de bruit et enfin son introduction automatique en un endroit déterminé d'un analogue électrique du conduit vocal.

3.2. Généralités

Un certain nombre de sons de la parole sont produits à partir d'un flux d'air turbulent provoqué par une constriction très étroite du conduit vocal. L'instabilité de cet écoulement turbulent donne naissance à une source de bruit dans le conduit vocal. Il y a deux cas principaux de production de ce bruit : lors d'une constriction très petite, quasi stationnaire, et lors d'un changement brutal de configuration articulatoire. Dans certains cas, on peut rencontrer ce phénomène au niveau des cordes vocales qui ne modulent plus le flux commun dans le cas des voyelles, mais donne lieu à une source de bruit ; exemple : [h].

3.3. Considérations théoriques

Une constriction peut avoir quelques millimètres de long (telle celle à la glotte), ou plusieurs centimètres (telle celle formée par le palais et la langue). Le conduit vocal peut se contracter ou s'étendre brutalement à la constriction, ou bien cette variation peut être graduelle. On remarque ainsi immédiatement que les variables commandant la source de bruit sont nombreuses.

Quand l'air s'écoule à travers le conduit vocal, dans une constriction, la chute de pression est due aux pertes énergétiques du flux d'air tourbillonnant lors de la

contraction ou de l'expansion plutôt qu'à la constriction elle-même, sauf dans le cas où les dimensions de celle-ci sont extrêmement faibles. Les tourbillons peuvent être vus comme des éléments volumiques d'air qui effectuent des rotations. A cause de la viscosité de l'air, les pertes d'énergie se manifestent comme une chute de pression additionnelle ou une résistance à l'écoulement. On trouve expérimentalement que cette chute de pression vaut :

$$\Delta P_L = k_L \frac{\rho U^2}{2 A^2}$$

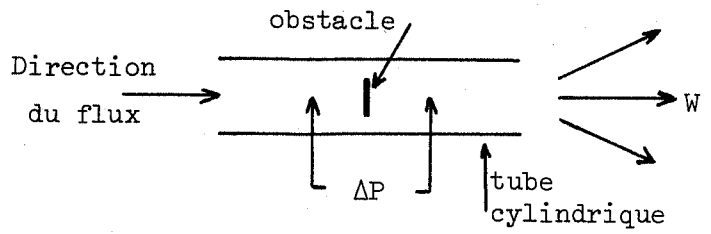
où A est l'aire de la constriction ; U : vitesse volumique ; ρ : densité de l'air ; k_L : constante.

La valeur de k_L dépend des conditions de formation de la constriction, de ses dimensions. Pour des variations brutales des dimensions de la constriction, k_L varie de 1,2 à 1,7 et, pour des transitions graduelles, $k_L = 0,87$; $k_L = 0,9$ est probablement une valeur raisonnable quelle que soit la place de la constriction. Il existe un autre terme de perte, dû à la résistance laminaire de la constriction, mais il est en général négligeable. En définitive, on peut écrire : $\Delta P_L = 0,9\ \rho U^2/2A^2$.

Etudions maintenant les caractéristiques du bruit alors produit. L'écoulement turbulent à travers une constriction conduit à une distribution aléatoire des vitesses des particules et donc à la création d'un son de bruit. S'il est créé près de la glotte, on l'appelle quelquefois *bruit d'aspiration*, s'il est créé près d'une constriction dans le conduit vocal, on l'appelle alors *bruit de friction*.

Les caractéristiques de la source de bruit créé par un obstacle dans un tuyau cylindrique ont été étudiées par GORDON et par HELLER et WIDNALL (figure 7).

FIGURE 7 - Configuration du montage utilisé par GORDON et par HELLER et WIDNALL pour l'étude de la source de bruit



Ils en ont déduit que :

- la puissance rayonnée du son résultant est proportionnelle à V^4 ($V = U/A$) ;
- la fluctuation de la force agissant sur l'obstacle a une distribution de fréquence centrée sur une fréquence proportionnelle à V/D où D est une dimension caractéristique de l'obstacle (figure 8.a, page suivante) ;
- la puissance du son rayonné dépend de la fréquence et on obtient alors un spectre représenté à la figure 8.b de la page suivante ;
- on peut exprimer la puissance rayonnée par la relation : $W = k\Delta P^3 d^2/\rho^2 c^2$ où d est le diamètre du tube, c la vitesse du son et k une constante.

Ces résultats vont pouvoir être appliqués au cas du conduit vocal.

3.4. *Application à la génération du bruit dans le conduit vocal*

La configuration du conduit vocal pendant la production de bruit est donné par la figure 9.a, page suivante). Les résultats énoncés ci-dessus sont applicables et on peut alors en déduire le circuit équivalent (figure 9.b).

Le spectre de la source de bruit est équivalent à celui donné figure 8.a. La fréquence centrale vaut $0,2\ U/A^{3/2}$. Usuellement, on trouve de 500 à 3000 Hz, la plus

basse étant obtenue pour un bruit d'aspiration, ou un bruit produit par une constriction large et un écoulement lent ; la plus haute étant obtenue pour des fricatives.

FIGURE 8

Figure 8.a
Spectre de la force aléatoire sur l'obstacle

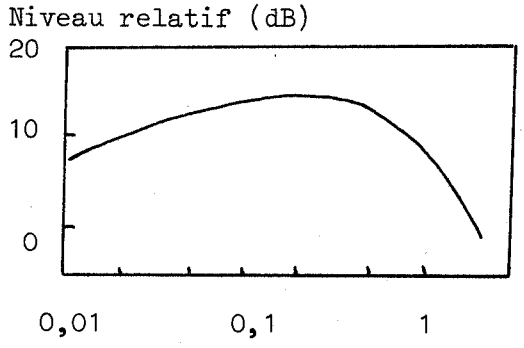


Figure 8.b
Spectre du son rayonné par un tube semi-infini

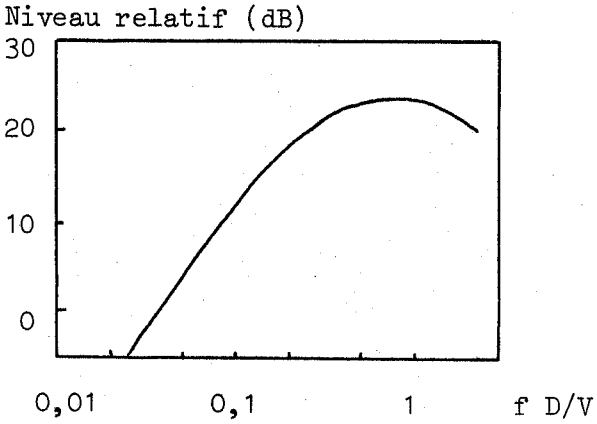


FIGURE 9

Figure 9.a
Figuration d'une constriction du conduit vocal

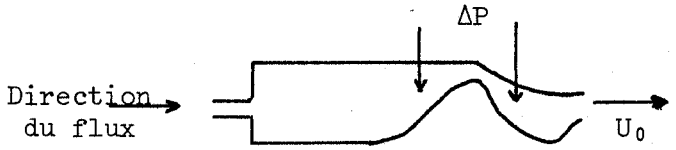
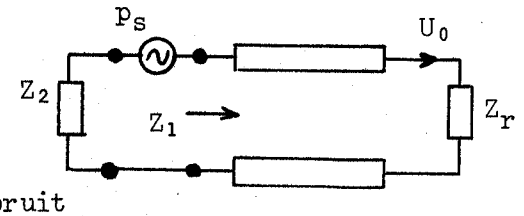


Figure 9.b
Circuit équivalent



p_s : pression du son de bruit

p_s est proportionnel à ΔP et, pour une chute de pression donnée, elle est sensiblement indépendante des dimensions de la constriction et de sa position. Elle peut cependant être légèrement réduite dans le cas où la source est près de la bouche (comme pour le [f]).

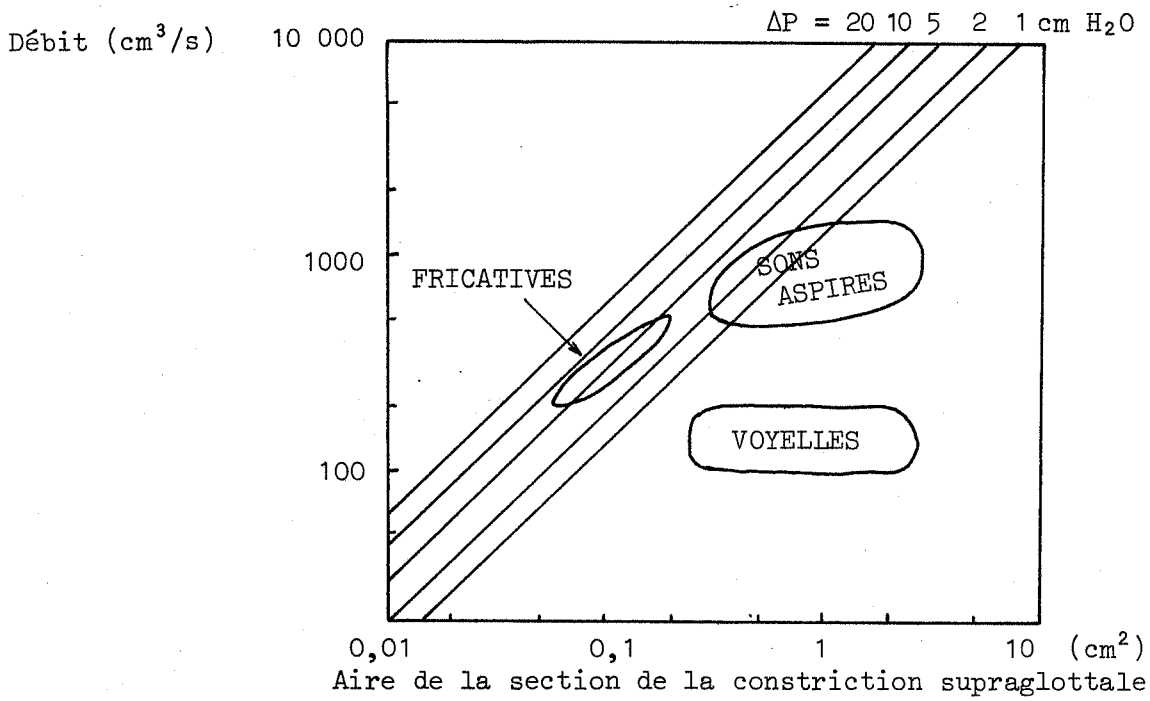
Le spectre du signal rayonné à la bouche sera modulé par la fonction de transfert du conduit vocal. Le premier pôle peut varier de 500 Hz pour le [h] jusqu'à 4000 Hz pour le [s]. Le sommet du spectre du bruit produit à la bouche par la plupart des constrictions se situe dans la bande de fréquence 2500 - 15 000 Hz.

Les mesures faites par HIXON, MINIFIE et TAIT [6] pour le [s] et [$\overset{\vee}{s}$] viennent confirmer les résultats déduits du modèle étudié ci-dessus. MEYER-EPPLER [7] a montré que la pression du son rayonné est proportionnelle à la différence entre la pression intra-orale et une pression seuil intra-orale p_t au-dessous de laquelle aucun bruit n'est produit. On vérifie également que la forme de la constriction influence peu le niveau du bruit.

3.5. *Conclusion - Mise en place automatique de la source de bruit*

En fonction des différentes caractéristiques d'une constriction, on a pu déterminer dans un plan U(A) avec ΔP comme paramètre, les domaines de formation d'une source de bruit dans le conduit vocal (figure 10).

FIGURE 10 - Lieu des différentes classes de son produit selon la dimension de la constriction, de la chute de pression à travers celle-ci et la vitesse volumique du flux



Essayons d'étudier comment on pourrait appliquer ces résultats à mise en place du bruit dans un analogue du conduit vocal. Dans ce type de synthétiseur, il est intéressant de limiter au minimum le nombre des paramètres de commande. Or, dans le conduit vocal réel, la naissance d'une source de bruit dépend de la configuration à un instant donné et des conditions de pression et de vitesse de l'air en ce point. On doit donc pouvoir imaginer un circuit qui produira automatiquement une source de bruit quand les conditions adéquates seront remplies. En chacun des points du conduit vocal, on disposera de points de mesures et de renseignements sur l'aire de la section (figure 11.a). On pourra en déduire les valeurs équivalentes de chute de pression et on comparera tous ces renseignements aux caractéristiques de la naissance d'une source de bruit. Si la comparaison est positive, on insèrera une source de bruit à l'endroit considéré. Cette insertion peut se concevoir de deux manières :

. la première, très séduisante, utiliserait un élément non linéaire en série dans la self-inductance de la cellule élémentaire, qui, sous certaines conditions, deviendrait générateur de bruit (figure 11.b) ;
. la seconde comprendrait un circuit de décision insérant à l'endroit voulu une source de bruit externe (figure 11.a).

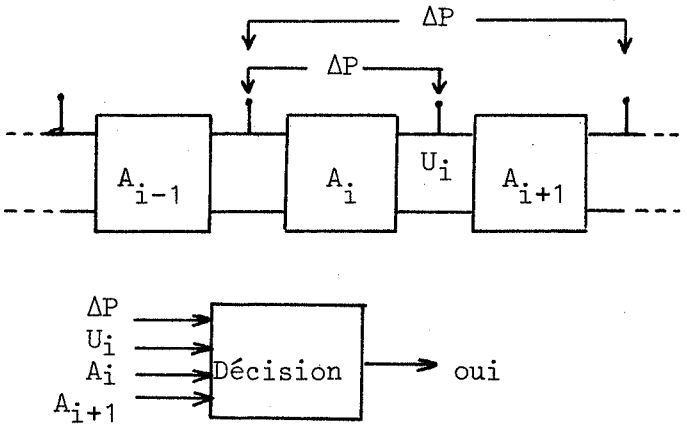
Les domaines définis dans le plan U(A) donnent une meilleure indication des conditions déterminant la création d'une source de bruit que le seuil défini par le nombre de REYNOLD (cf. MEYER-EPPLER). Le seuil déterminé par ce nombre est imprécis car il dépend de la forme de la constriction et de l'état de surface des parois.

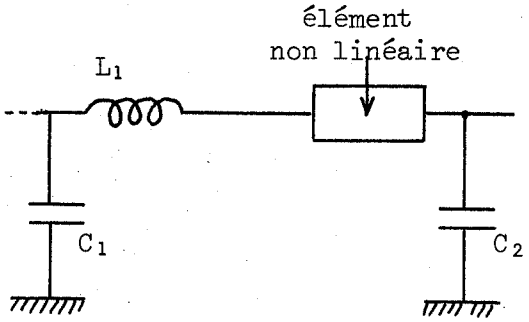
Toutes ces études ont été faites sur des configurations statiques ; or, il serait intéressant d'étudier de près le comportement de la pression avec le temps et les conditions d'écoulements lors d'une variation brusque de la configuration. Nous n'avons donc pu étudier théoriquement que le cas des fricatives et des sons aspirés, mais pas celui des consonnes où il y a une ouverture rapide du conduit vocal en l'un de ses points.

Figure 11.a

FIGURE 11
Insertion automatique
du bruit dans un analogue
du conduit vocal

Figure 11.b





BIBLIOGRAPHIE

[1] L.O. CHUA
Synthesis of new nonlinear network elements ;
Proc. I.E.E.E., 56 n° 8, 1325-1340 (1968).

[2] EIICHI MATSUI
Computer-simulated vocal organs ;
The 6th I.C.A., B.5.1, B151-B154 (1968).

[3] B.S. ATAL
Sound transmission in the vocal tract with applications to speech analysis and synthesis ;
The 7th I.C.A., 23C12, 169-172 (1971).

[4] P. MERMELSTEIN
Calculation of the vocal-tract transfert function for speech synthesis applications ;
The 7th I.C.A., 23C13, 173-176 (1971).

[5] K.N. STEVENS
Airflow and turbulence noise for fricative and stop consonants : static considerations ;
J.A.S.A., 50 n° 4, 1180-1192 (1971).

[6] T.J. HIXON, F.D. MINIFIE, C.A. TAIT
Correlates of turbulence noise production for speech ;
J. Speech and Hearing Res., 10, 133-140 (1967).

[7] W. MEYER-EPPLER
Zum Erzengungsmechanismus der Geräuschlante ;
Z. Phonetik, 7, 196-212 (1953).

DISCUSSIONS ET INTERVENTIONS

M. EL MALLAWANY
Dans le cas de la simulation numérique du conduit vocal, vous avez dit que la synthèse effectuée en maintenant les zéros du conduit nasal donnait de meilleurs résultats que celle les remplaçant par des pôles équivalents.

M. B. GUERIN
Dans ses travaux, MERMELSTEIN a constaté effectivement que la qualité de la parole obtenue en simulant les pôles du conduit nasal était meilleure que celle obtenue en considérant les pôles équivalents. Cela revient effectivement à augmenter la complexité de la simulation.

M. GIMONET
Votre simulation du conduit vocal conduisant à une représentation en 10 pôles réalise-t-elle une réduction d'information satisfaisante ? Autrement dit, ces "paramètres" évoluent-ils suffisamment lentement ?

M. B. GUERIN
On peut comparer les différents systèmes de synthèse de la parole et leur quantité d'information nécessaire à leur commande. Le téléphone, admettant une bande passante 300 Hz - 3400 Hz, demande environ de 25 000 à 50 000 bits/s pour la transmission de la parole. Le vocoder à canaux permet déjà de diminuer ce nombre et on considère qu'un vocoder à formants ne nécessite plus que 1000 à 2000 bits/s environ. La transmission des informations de commande d'un analogue du conduit vocal sera encore plus économique car on se rapproche de la commande du conduit réel qui doit être optimale. Les paramètres de commande du filtre digital sont directement fonctions des paramètres articulatoires ; on doit donc réaliser ainsi une réduction de la quantité d'information.

M. J.P. PECKELS
Certains auteurs semblent avoir montré que, dans la parole naturelle, il existe un bruit superposé, même dans les voyelles à l'état stable. C'est cette superposition qui contribuerait, pour une part importante, à l'aspect "naturel" des sons, même des sons voisés.
Est-ce que vous avez vu quelque chose dans ce domaine ? Est-ce que ce bruit est le résultat d'une irrégularité naturelle du pitch (dans le sens de LIBERMAN) ou d'une source de bruit naturelle fonctionnant simultanément avec les cordes vocales ?

M. J. PAILLÈ
Les hypothèses relatives à la production de la parole sont des hypothèses simplificatrices. En particulier, on suppose que le conduit vocal n'a pas un profil trop heurté. Dans le cadre de la production de la parole naturelle, il semble qu'il existe des "accidents" de profil. Au niveau de ces accidents, on peut voir apparaître des sources de bruit qui interviendront même dans les portions les plus stationnaires du signal.

T A B L E R O N D E

SIMULATION DE LA SOURCE VOCALE

Animateurs : J. PAILLE (E.N.S.E.R.G. Grenoble)
J. GENIN (C.N.E.T. - Lannion)
R. DESCOUT (C.N.E.T. - Lannion)

T A B L E R O N D E

NORMALISATION DES TESTS D'INTELLIGIBILITE POUR LA PAROLE SYNTHETIQUE

Animateurs : MM. ROSSI (Aix-en-Provence)
CARTIER (C.N.E.T. Lannion)
PECKELS (I.B.M. - La Gaude)

LINGUISTIQUE et RECONNAISSANCE

de la PAROLE

LA THEORIE CHOMSKIENNE ET LE LANGAGE DES MACHINES

Melle Mitsou RONAT
Un. de Paris VIII - VINCENNES

La théorie chomskienne et le langage des machines

"Pour ceux qui espèrent appliquer les résultats d'une discipline aux problèmes d'une autre, il est fondamental de rendre très claire la nature exacte, non seulement de ce qui a été atteint, mais aussi des limites de ce qui a été atteint".

Noam Chomsky,

("La forme et le sens dans le Langage naturel", traduit dans Hypothèses, coll. Change, Seghers/Laffont)

Si j'ai commencé par cette phrase de Chomsky, adressée en fait à des psychologues et à des éducateurs spécialisés dans l'organisation des programmes scolaires, c'est qu'elle peut résumer le but que je me suis fixée dans cette intervention, pour les journées de Lannion.

La majorité des interventions de cette rencontre, d'après le programme, fait état des derniers résultats des travaux en phonétique acoustique et articulatoire, c'est-à-dire des études faites sur la face matérielle et physique du langage, et "sensible" recherche sur la parole. De ce point de vue, ils peuvent se rattacher aux sciences physiques. Mon point de vue sera différent, quoiqu'évidemment relié : il sera linguistique, au sens où la linguistique est la première des sciences humaines à avoir reçu une formalisation mathématique, axiomatisée, due à Chomsky. Son objet, c'est la langue ou plutôt la compétence, mécanisme abstrait qui permet de comprendre et de traduire les phrases nouvelles.

Dans la mesure où les perspectives d'avenir des recherches entreprises sur la reconnaissance de la parole se présentent comme la possibilité d'ouvrir un "dialogue" avec les machines, il me paraît intéressant de rappeler ici, très brièvement, puisque cela est connu de certains d'entre vous déjà, des résultats, positifs et négatifs, des recherches linguistiques dites "avancées", en ce qui concerne le langage humain.

Bien que les techniques de segmentation phonologique ou morphématique soient relativement solidement fondées et pratiquées, Chomsky a montré qu'elles n'atteignent que la structure superficielle des phrases, sans donner d'indications sur leur structure profonde :

.../...

Cette dichotomie, structure superficielle / structure profonde, si importante pour l'interprétation sémantique, Chomsky l'a démontrée en plusieurs étapes. Il a commencé par prouver que des modèles établis d'une part par des théoriciens de la communication, comme Shannon et Weaver, et d'autre part par des linguistes structuralistes, comme Hockett, Wells, étaient inadéquats pour la description des langues naturelles : il a montré qu'une langue naturelle comme l'anglais comporte des phrases tout à fait courantes que ces modèles sont incapables de décrire.

Par exemple, en ce qui concerne le "processus de Markov à nombre fini d'états", relié aux probabilités et à certaines analyses fonctionnalistes, il montre qu'une phrase comme :

(1) Si, soit l'homme qui a dit que Marie est belle arrive aujourd'hui, soit la femme de Paul me téléphone, alors je pars

c'est-à-dire la structure :

(2) Si, [soit (S_1 (S_2 (S_3))), / soit (S_4)] / alors (S_5)

qui comporte d'une part des phrases enchassées (l'homme qui a dit que Marie est belle arrive aujourd'hui) et d'autre part les dépendances de part et d'autre de la barre "/", - à savoir que "si" implique "alors" et "soit" implique "soit". Ces phrases peuvent correspondre à des langages formels du type $\{x\ x\}$ ou $\{x\ \tilde{x}\}$, donnant, si on limite l'alphabet à $\{a, b\}$, respectivement des phrases comme aaaa, abbabb ou comme abba, ababababbababa, etc. dont il est démontré qu'ils ne sont pas à état fini (cf. Chomsky, "Trois modèles de description du langage", traduit dans la revue Langages, n° 9) (Dans les langages de type $\{x\ x\}$, il existe une dépendance du type si/alors, dans les langages dits à 'image miroir', on retrouve le problème de l'enchassement).

Deuxième exemple : les ambiguïtés du langage et les relations entre les phrases. Les modèles établis à l'aide de la segmentation par les linguistes structuralistes aboutissent à une symbolisation de classification. Pour reprendre l'exemple classique, ils obtenaient, grâce à l'opération supplémentaire de substitution, les descriptions :

(3)

le	garçon	mange	ait	des	pomme	s
		racine	suf.	article	nom	pluriel
article	nom	verbe		syntagme nominal		
syntagme nominal		syntagme verbal				
phrase						

.../...

Avec une telle méthode de description, on obtient des descriptions semblables pour des phrases superficiellement semblables, mais en réalité très différentes :

(4) ce garçon est facile à convaincre
(5) ce garçon est content de partir
(6) j'ai cassé le pied du tabouret
(7) j'ai cassé la figure à Pierre
(8) je dors le jour
(9) je mange le biscuit, etc.

on ne peut expliquer l'ambiguïté des phrases :

(10) j'ai trouvé ce livre intéressant
(11) Flying planes can be dangerous
(12) /lab nf rmlap rt/

ni les relations entre :

(13) l'acteur chante une chanson
la chanson est chantée par l'acteur
(14) il est facile de convaincre ce garçon (cf. (4))
(15) j'ai trouvé que ce livre était intéressant (cf. (10))
(16) Pierre me semble bizarre
il me semble que Pierre est bizarre, etc.

Ces relations sont décrites par des transformations ; c'est pourquoi la théorie s'appelle théorie des grammaires génératives et transformationnelles. Le modèle construit par Chomsky et son école doit rendre compte de toutes les phrases grammaticales d'une langue et rien qu'elles, en leur donnant une description structurale rendant compte aussi bien des ambiguïtés que des différences. La structure profonde montre les "relations grammaticales significatives", à savoir ce qui devrait être l'input d'une interprétation sémantique. Pour (4), nous aurions une structure profonde :

(17) ((convaincre ce garçon) est facile)

qui montre que le sujet de "est facile" est toute une proposition, ce qui exclut l'interprétation "ce garçon est facile", dans l'autre sens de facile, tandis que pour (5), on aurait :

(18) (ce garçon$_i$ est content de (ce qu'il$_i$ parte))

où la phrase enchassée est <u>complément</u> de l'adjectif "content", et où le sujet de "est content" est le même que le sujet de "partir". La similitude en structure superficielle de (4) et de (5) ne permet pas de décrire ces différences perceptibles immédiatement pour tout locuteur natif français.

.../...

Seule la structure profonde, qui fait partie du savoir du locuteur sur sa propre langue, au même titre que la connaissance des phonèmes, peut expliquer ces interprétations. L'accès à cette structure profonde n'est pas déductible de la structure superficielle qui, elle, reçoit l'interprétation phonologique et phonétique. C'est plutôt l'inverse.

Ainsi, si dans le domaine de la recherche électronique on veut dépasser le stade de la traduction biunivoque "un son = un signe graphique" et, si l'on veut parvenir à une "communication" avec la machine, il faudra lui enseigner non seulement la segmentation des unités minimales ou phonèmes, ou à un niveau supérieur à distinguer les mots, mais encore à distinguer structure profonde et structure superficielle, et connaître le système de transformations qui unit ces deux structures.

Je suppose aussi que ces recherches sont entreprises dans un but d'efficacité, et que ce serait un très mauvais résultat si la machine se méprenait sur la signification de l'ordre ou de la question posée, comme cela pourrait être le cas avec les phrases ambiguës. A moins d'essayer au maximum d'éviter ces phrases, pourtant essentielles aux langues naturelles, il faudra lui enseigner à <u>choisir</u> selon le contexte. Sans oublier la polysémie (cf. une maison particulière, bizarre, individuelle, les différents "sens" du mot "sens", etc. Ces propos sont sans doute décevants, mais ils posent le problème de la différence "<u>qualitative</u>" du langage humain et du langage des animaux-machines, et des limites qu'il faut se fixer dans ce genre de démarche.

La conception selon laquelle l'homme est une machine seulement un peu plus complexe que les autres, seul le rapport <u>quantitatif</u> étant en jeu, est liée à une tendance particulière de la psychologie du 20e siècle, la psychologie du comportement, elle-même liée au behaviorisme et aux philosophies empiristes anglo saxonnes. Très schématiquement, cette dernière prétend que l'apprentissage des langues, entre autres, se fait par des renforcements de stimuli, et que les généralisations sont conduites par induction, à partir d'un "cerveau en friche", d'une <u>tabula rase</u>.

La complexité des faits de langage, les contraintes imposées à la forme des grammaires, les processus inconscients qui conditionnent l'intuition grammaticale, tout cela a amené Chomsky à rejeter cette conception au profit d'une conception voisine, mais repensée dans les termes de la science moderne, de la conception cartésienne :

.../...

l'homme possède une faculté de langage innée, qui lui est propre, faculté qui lui permet d'acquérir n'importe quelle langue particulière, et de produire et de comprendre une infinité de phrases nouvelles. C'est ce qu'on nomme la "créativité".

La théorie des grammaires génératives commence à donner des hypothèses empiriquement fondées sur la nature de ce langage humain. Mais au stade actuel des connaissances, nous sommes loin de connaître tous les processus à l'oeuvre dans les langues naturelles, et cette connaissance est fondamentale si l'on veut construire une machine parlante, dans l'éventualité la plus ambitieuse.

C'est en partant d'une idée assez simpliste du langage que certains chercheurs ont tenté de constituer, par exemple, des techniques de traduction automatique. A moins de s'en tenir à un vocabulaire extrêmement réduit et à des constructions très simples, elles ont échoué avec celles qui sont utilisées dans le langage quotidien ou poétique (et encore !).

Ainsi, puisque les recherches sur la segmentation phonétique semblent avoir obtenu des résultats très positifs, et puisqu'il semble aussi que la prochaine étape soit la reconnaissance du mot et de la phrase dans son interprétation sémantique, les questions soulevées et les réponses données par la grammaire générative intéresseront les ingénieurs des télécommunications.

Mais il faut rappeler qu'avant tout la théorie chomskienne est une théorie de la syntaxe, qui cherche à poser des lois sur la succession des mots, à décrire les suites possibles d'une langue particulière. Nous pourrons discuter, si vous le voulez, des derniers développements de la théorie (sémantique générative, ou théorie interprétative).

Le problème est de décider s'il faut réduire ce que nous voulons savoir du langage humain aux possibilités de la calculatrice, ou si l'on doit étudier le langage dans toute sa complexité - avant de penser à le maîtriser mécaniquement. Ce qui est en question, pour elle plus loin, c'est la définition de la langue comme "instrument de communication", admise généralement comme allant de soi, ce qui élimine l'action du langage dans la différenciation de classes sociales, ou sur un autre plan, les malentendus, quiproquos, les lapsus, les mensonges ou les calembours, qui jouent sur l'ambiguité.

.../...

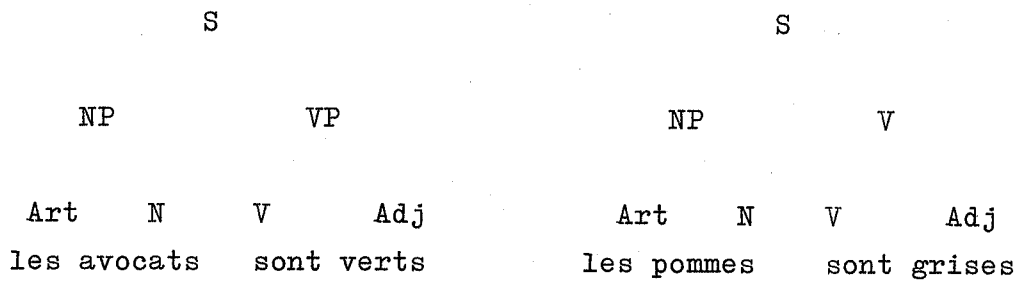
M. ROCHE à Melle RONAT

1. - Comment se font les accords de genre avec les règles de réécriture ?
2. - Le cas des phrases où intervient "respectivement".

Réponse

a) Généralement, on utilise des règles transformationnelles, ou de copie (voir la thèse de Gilles Fauconnier). Mais ce phénomène a été peu étudié jusqu'à présent.

b) On peut proposer une solution identique à la solution concernant la coordination (mais il peut y en avoir d'autres) par exemple, pour expliquer l'accord de : "les avocats et les pommes sont respectivement verts et grises", on pourrait supposer deux arbres initiaux :



La transformation d'accord a été faite à ce niveau ; puis il y aurait des transformations de coordination reliant les deux NP et les deux Adj. Cela ne présente pas de problèmes théoriques particuliers ; sur le plan pratique (automatique) je ne peux rien dire de ces propositions.

.../...

M. J.S. LIENARD à Mlle RONAT

Dans les deux phrases que vous avez citées ("le garçon est facile à convaincre" et "le garçon est content de partir"), les structure profondes ne se différencient des structures superficielles que par une analyse sémantique. Les structures syntaxiques ne vous semblent-elles pas secondaires par rapport aux structures sémantiques, sur cet exemple ?

Réponse

Effectivement, il y a une correspondance ici entre l'interprétation sémantique et la structure profonde, mais ce n'est pas suffisant. Ici on peut opposer :

*(1) le garçon est facile à se regarder souvent dans la glace

(2) le garçon est content de se regarder souvent dans la glace

Pour "ce garçon est facile à convaincre, on pose la structure profonde :

((convaincre ce garçon) est facile)

Or, on n'a pas

* ((se regarder ce garçon) est facile)

ni

*? il est facile que ce garçon se regarde souvent dans la glace

Cela prédit l'agrammaticalité de (1).

Tandis que pour (2) on a bien

(S_1 (ce garçon est content S_2 ce garçon se regarde souvent dans la glace))

Ce qui prédit également la correction de cette phrase.

De même, en anglais, on a :

John appeared to Mary to like himself
* him
* herself
her

uniquement parce qu'on propose la structure profonde :

it appeared to Mary John to like himself
* him
* herself
her

puisqu'il y a une règle générale de l'anglais impliquant que l'antécédent du pronom réfléchi soit dans la même phrase simple (avec quelques exceptions, elles-mêmes explicables, voir

Cela explique aussi :

John appeared to Mary to like * himself
him
herself
* her

Car la structure profonde est :

John appeared to Mary Mary like * himself
                                  him
                                  herself
                               * her

Ainsi, la structure profonde est fondée aussi sur des arguments syntaxiques (agrammaticalité ou grammaticalité) et non plus seulement sur le sens de la phrase.

Question

Pensez-vous que certaines ambiguités soient levées par la prosodie ?

Effectivement, pour /la bɔn fɛrm la port/ l'intonation peut aider à la reconnaissance = différence entre

la bonne ferme la porte = la domestique ferme la porte

ou

la bonne ferme l'apporte = l'exploitation agricole excellente apporte quelque chose

Cela dit, il existe bien des cas où une telle différenciation est impossible :

j'ai trouvé ce livre intéressant = 1 quelque chose d'intéressant
2 que ce livre était intéressant

j'ai acheté un jouet à Justine
1 Justine vend des jouets
2 un jouet destiné à Justine
3 jouet à Justine brosse à dents

.../...

REMARQUES ET PREMIERS RESULTATS SUR L'ANALYSE SEMANTIQUE DISCURSIVE

M. HERAULT D.

Centre de Linguistique Quantitative - Université PARIS VI -
St SULPICE DE FAVIERES

REMARQUES ET PREMIERS RESULTATS SUR L'ANALYSE SEMANTIQUE DISCURSIVE

SOMMAIRE DE L'EXPOSE

I. - Définition, intérêt et limites de l'analyse discursive.

- Prise en compte de discours réels.
- Dépassement du niveau traditionnel de la phrase.
- Limitation aux textes scientifiques.
- Possibilité de mise en évidence de la structure interne des discours cohérents.

2. - Place et rôle de la linguistique discursive dans la linguistique actuelle.

- Méthode inséparable de la "linguistique automatique".
- Place importante de la linguistique discursive dans les pays de l'Est.

3. - Résultats syntaxiques.

- Régularités syntaxiques au sens probabiliste.
- Absence de "mémoire syntaxique" ; conséquences.

4. - Résultats sémantiques.

- Organisation sémantique du discours scientifique indo-européen. Résultats détaillés sur les langues slaves.
- Extension à d'autres systèmes linguistiques (hongrois, japonais).
- Possibilités de construire une métalangue sémantique "naturelle".
- Obtention de types de "signifiés" en ne manipulant que des signifiants ; notion de sémantique "interne".
- Signification de telles analyses par rapport aux analyses syntaxico-sémantiques existantes (Moscou, Grenoble).

* Service de Linguistique Quantitative de l'Université de PARIS VI.
Adresse du Secrétariat : 91910 - ST SULPICE FAVIERES
Une documentation sur les enseignements et publications est disponible et sera envoyée à tout demandeur.

.../...

5. - Quelques conclusions (1).

- Structure et contenu des langages de programmation (1).
- Problèmes de documentation automatique.
- Possibilités nouvelles du stockage des textes en calculatrice.

Résumé en anglais.

REMARKS AND FIRST RESULTS ON SEMANTICAL DISCURSIVE ANALYSIS.

After having defined what it is meant by a "discursive analysis", having described the aims and the actual role of such a method, various results are presented, dealing partly with syntax and chiefly with semantics. A semantical organization of scientific indo-european discourses is proposed and developped in details for slavic languages. Various consequences are expounded, principally in the field of "Computational Linguistics".

On trouvera un assez long développement de cet exposé dans la Préface du livre (à paraître fin 1972) Le Spectre sémantique du discours scientifique en indo-européen, D. HERAULT et A. LJUDSKANOV (Sofia), Document de Linguistique Quantitative - 18, (diffusion DUNOD éditeur, 92 rue Bonaparte, 75006 PARIS). On indique à ce propos qu'une grande partie des résultats a été obtenue en collaboration avec l'Académie des Sciences de Bulgarie. Des éléments de cet exposé ont déjà paru dans D. HERAULT.
Remarques sur le discours scientifique, Mathématiques et Sciences Humaines (MSH), n° 34 (1972) et D. HERAULT, Préface au livre de L. TESNIERE, Table étymologique : les mots russes classés par racine, Document de Linguistique Quantitative - 8, 1970 (diffusion DUNOD).

.../...

A Monsieur HERAULT, de Mlle RONAT

Au sujet de "verbes" comme sommets de phrases, je rappelle que la théorie lexicaliste de Chomsky (1967), exposée dans l'article "Remarks on Normalization" met l'accent précisément sur le rôle verbal de noms et sur la structure phrastique du syntagme nominal :
"le développement de la recherche" = la recherche se développe
"la construction de la ville par l'architecte" = l'architecte construit la ville.

REPONSE.

En effet, depuis quelques années, N. Chomsky et ses disciples ont largement assoupli leur théorie de façon à pouvoir y faire rentrer naturellement des modifications importantes, telles que celle signalée ici. Cependant, l'optique fondamentale reste la théorie transformationaliste de la phrase, qui ne me semble pas être en mesure actuellement de prendre en compte la structure de discours réels (paragraphes, succession de paragraphes, etc...).

En tout état de cause, ceci permet de comprendre pourquoi, sémantiquement, il n'est pas utile de tenir compte, devant un mot, de sa partie suffixale (en indo-européen), laquelle joue presque uniquement un rôle syntaxique.

.../...

A Monsieur HERAULT de Mario ROSSI

1ère question.

Quelle procédure utilisez-vous pour identifier les racines ?

2ème question.

Dans un exemple que vous avez donné, vous semblez ne pas être d'accord avec l'analyse de TESNIERE selon laquelle le verbe constitue le noeud de la phrase. Quels critères utilisez-vous pour déterminer le noeud syntaxico-sémantique de la phrase ?

REPONSES.

1°) Les procédures les plus traditionnelles, mais utilisées avec le plus grand soin, sont utilisées pour une telle mise en évidence. Dans les domaines indo-européen et sémitique, de nombreux documents (dictionnaires, dictionnaires étymologiques, dictionnaires de racines, etc...) sont disponibles. Dans les autres domaines, il devient nécessaire de créer l'outil et ou de le perfectionner considérablement. C'est ce qui est fait en ce moment pour le domaine finno-ougrien.

2°) Dans l'ensemble, on peut considérer que les travaux de Lucien TESNIERE sont absolument remarquables, compte tenu du fait que l'essentiel de ses conceptions a été fixé vers 1945. Les nombreuses discussions que j'ai eu la chance d'échanger avec Madame Lucien TESNIERE, (ainsi que la consultation de nombreux documents inédits), m'ont convaincu que, dans l'optique "calculatrice", Lucien TESNIERE aurait infléchi sa théorie dans le sens que j'ai proposé.

Quant aux critères de détermination des "noeuds syntaxico-sémantiques", il relève essentiellement de l'analyse de la "syntaxe du discours", et, en particulier, des "schémas d'énonciation" tels qu'ils ont été définis dans l'article précédemment cité dans la revue MSH.
Il va sans dire qu'une réponse définitive à cette question ne pourra être donnée qu'après l'analyse automatique d'au moins 100000 à 200000 phrases. Une telle entreprise est actuellement en cours dans cet esprit au GETA (Groupe d'Etudes pour la Traduction Automatique - Grenoble), avec le soutien financier et moral de la DRME qu'il est particulièrement agréable de pouvoir ici remercier.

.../...

Remarques additionnelles.

(1) Cet exposé a précédé celui de C. ROCHE qui a donné une remarquable présentation de la thèse de WINOGRAD. Les travaux ne m'étaient connus que de seconde main, mais à leur propos, on peut faire la remarque suivante : dans le travail de WINOGRAD, un rôle important est joué par le "Planner" de Carl HEWITT, qui ne m'était pas connu auparavant. Mais, il est très intéressant de constater que ce langage existe dans la langue naturelle et se situe au niveau de la structure du paragraphe, dans le cours d'un raisonnement déductif. C'est ce qui a été souligné dès 1968-1970 par le groupe de recherches de l'Université de PARIS VI (ex- Faculté des Sciences de PARIS) dont une des activités essentielles est d'étudier les "moyens linguistiques propres" que possèdent les systèmes linguistiques pour transmettre un raisonnement déductif. Voir l'article déjà cité de D. HERAULT dans MSH.

(2) Parmi les conclusions qui n'ont pas été présentées et qui pourraient intéresser tout particulièrement les spécialistes de la reconnaissance de la parole, on peut indiquer l'idée suivante : aucun phénomène linguistique n'a une existence propre. En conséquence, la reconnaissance de la parole ne pourra demeurer strictement au niveau phonétique si elle veut dépasser une certaine qualité qui semble se situer actuellement aux environs de 70 %. Par ailleurs, des considération syntaxiques, comme en ce qui concerne la Traduction Automatique, seront largement insuffisantes. Donc, là encore, on voit apparaître la nécessité d'utiliser d'une certaine façon le niveau sémantique. Par ailleurs, la connaissance de la structure interne du discours semble devoir permettre d'envisager une correction des erreurs, ce qui n'est pas réalisable en ne manipulant que des phrases ou des éléments de phrase.

ANALYSE LINGUISTIQUE SELON WINOGRAD

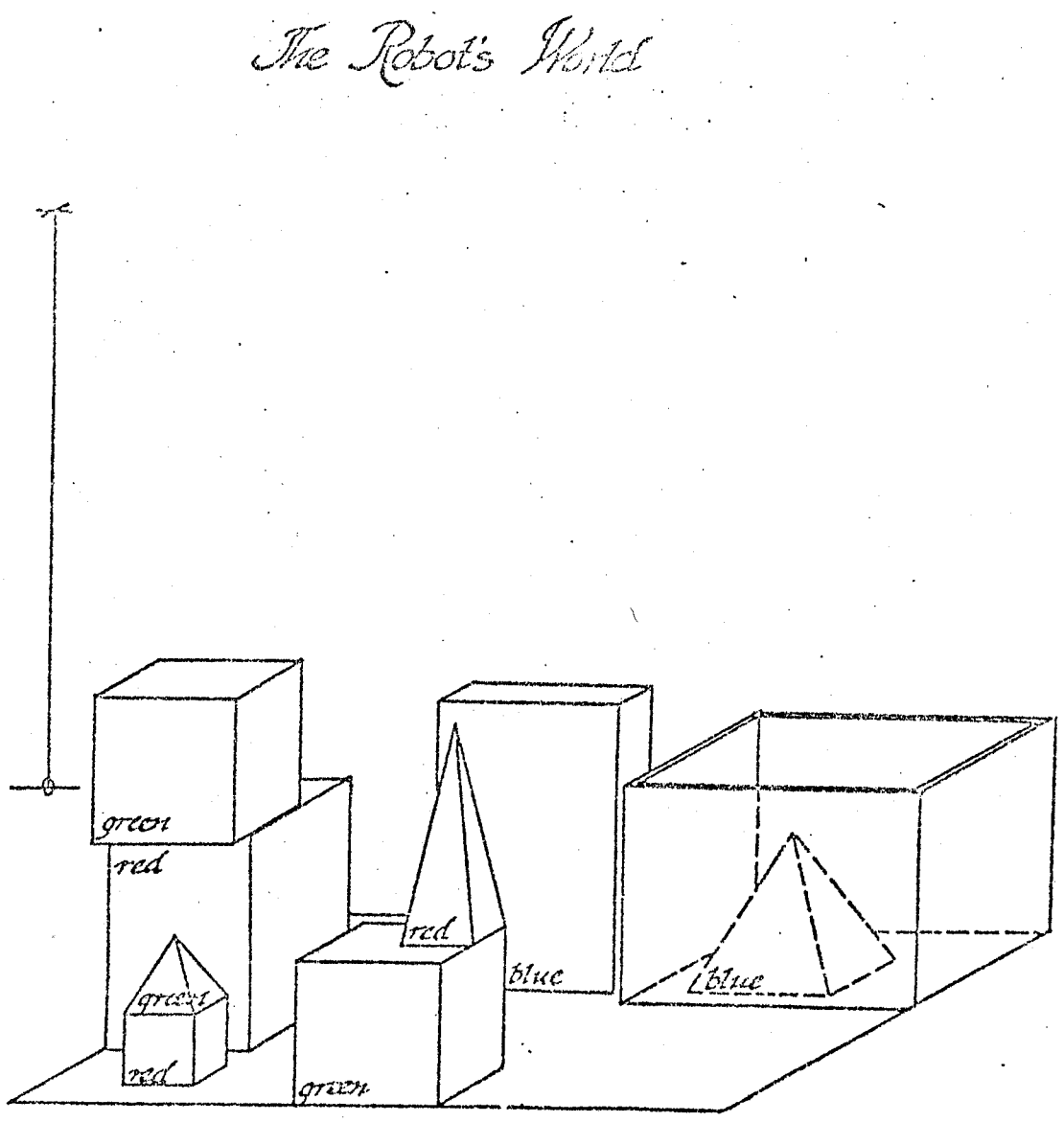
C. ROCHE

Institut de Programmation
Quai Saint Bernard
75 - PARIS 5e

Résumé de la thèse de Terry WINOGRAD : la représentation de données sous forme de procédures, dans un programme d'ordinateur pour la compréhension du langage naturel (août 1970).

Cette thèse "décrit un système sur ordinateur pour la compréhension de l'anglais. Le système répond aux questions, exécute les commandes et accepte les informations en anglais courant. Il utilise l'information sémantique et le contexte pour comprendre la phrase et lever les ambiguités.
Il fait une analyse syntactique complète de chaque phrase avec un programme heuristique de compréhension qui utilise différentes informations sur la phrase, sur le reste de ce qui a été dit et l'information générale sur son univers, pour décider de la signification de la phrase".

UNIVERS DU SYSTEME :



EXEMPLE DE DIALOGUE :

M - Pick up a big red block.
R - O. K.
M - Grasp the pyramid.
R - I don't understand which pyramid you mean.
M - Find a block which is taller than the one you are holding and put it into the box.
R - By "it", I assume you mean the block which is taller than the one I am holding. O. K.
M - What does the box contain ?
R - The blue pyramid and the blue block.
M - What is the pyramid supported by ?
R - The box.
M - How many blocks are not in the box ?
R - Four of them.
M - Is at least one of them narrower than the one which I told you to pick up ?
R - Yes, the red cube.

I. INTRODUCTION

- Le monde du robot : une table, une boite non déplaçable, objets géométriques, couleur des objets, main du robot. Notions de déplacement, de volume, de pesanteur, de temps.

- Problème posé : donner des ordres en anglais courant au robot. D'où trois difficultés :
  - reconnaissance de formes : simulé
  - résolution de problèmes : limité
  - linguistique : très sophistiqué.

- Exemples des difficultés rencontrées et résolues : analysons la "conversation" que nous venons de présenter :

Pick up... : il existe 2 blocs rouges, big veut dire donc le plus grand. Il faut (voir la figure) donc déplacer le cube vert avant de prendre

le bloc rouge. Il faut qu'il y ait de la place ailleurs. Sinon, on déplace d'autres blocs etc...

Grasp the pyr.... : il y a ambiguité détectée.

Find a block... : compréhension de the one et it. Choix heuristique d'après ce qui est le moins stupide.

What does... : nécessité de garder en mémoire tout ce qui s'est passé.

What is... : l'ambiguité est levée heuristiquement pour décider de quelle pyramide il s'agit : celle dont on parle.

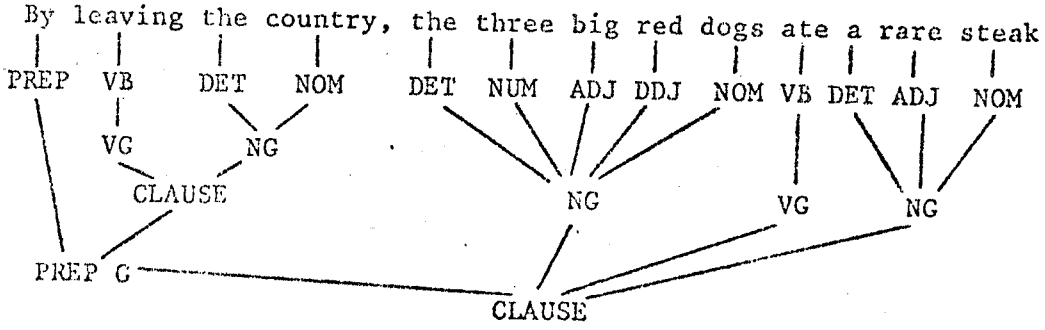
Is at least... : garder en mémoire ce qui s'est passé avec les successions des différentes actions dans le temps.

- Le programme : il comporte trois parties intéragissant et traitant les différents aspects :
  - sémantique
  - syntactique
  - résolution de problème.

II. ASPECT SYNTACTIQUE

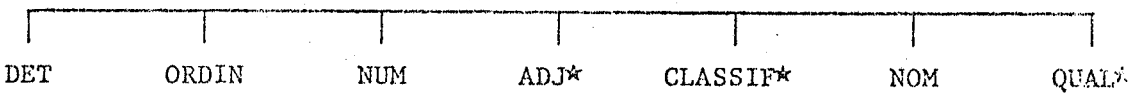
1°) Les notions de grammaire utilisées sont celles des grammaires systémiques (Halliday). Etant donnée une phrase, le programme PARSE engendre une arborescence à plusieurs niveaux correspondant à 3 notions syntactiques :
   - les clauses
   - les groupes (nominal, verbal, prepositionnel, adjectival)
   - les mots (nom, adjectif, déterminateur...)

Exemple :



2°) A chaque notion syntactique correspond un programme qui la détecte dans la phrase (programme PARSE)

Exemple de la structure du groupe nominal :

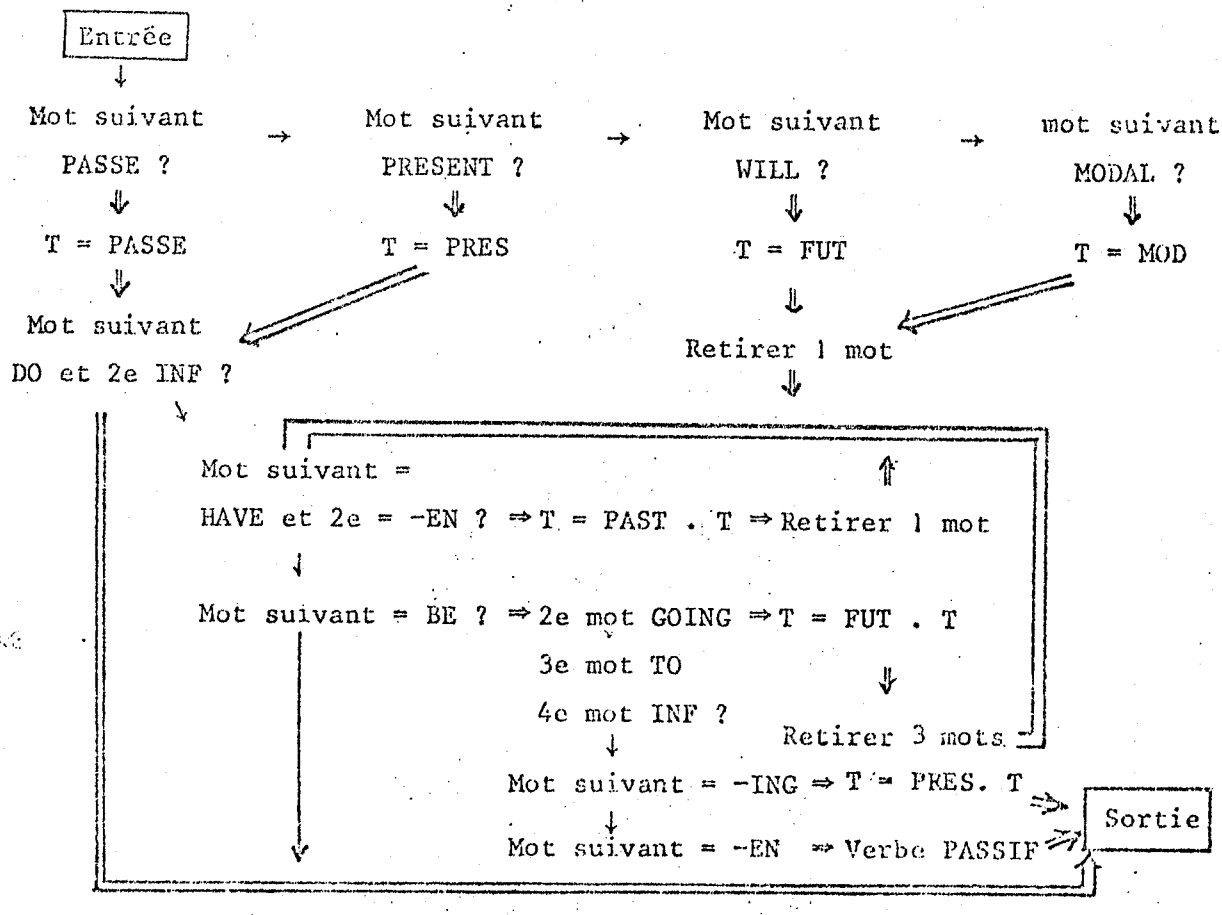


(L'un quelconque peut manquer, même le nom - * veut dire qu'on peut en trouver plusieurs).

Chaque mot rencontré est dans un dictionnaire qui indique en particulier si on doit le considérer comme NOM, DET, ..., QUAL etc...

3°) A chaque noeud de l'arborescence sont indiquées les caractéristiques correspondantes. Pour un verbe 4 caractéristiques : son temps, s'il est singulier ou pluriel, s'il est négatif, s'il est modal (avec can...).

Chaque caratéristique est déterminée par un programme.
Exemple de détermination du temps d'un verbe.



4°) Tous les programmes travaillent de gauche à droite, en gardant en mémoire toutes les ambiguités et en les levant successivement au fur et à mesure de la lecture.

5°) L'analyse syntactique se fait en même temps que l'analyse sémantique et les ambiguités syntactiques sont levées par la sémantique et réciproquement.

III. ASPECT DEDUIT

Le langage utilisé est le Planner de Carl HEWITT. C'est un langage qui a été créé en vue de la formalisation des problèmes pour leur résolution automatique.

Une "base de donnée" peut être modifiée par des ordres codés en Planner. Par exemple :

Th assert (human Turing) : ajoute dans la base de données le fait que Turing est humain.

Th erase (human Turing) : retire le même fait de la base.

Th goal (X) (Th use y_1 y_2 y_3) : recherche si le fait X est dans la base sinon on utilise y_1, y_2 et y_3 pour démontrer qu'il existe à partir des données de la base (procédé de démonstration automatique de théorème, avec parcours d'arbre exhaustif). Dans les deux cas on exécute l'ordre suivant si X est vrai sinon on passe à l'éventualité suivante.

Etc...

Exemple de programme écrit en planner et traduisant l'action de "prendre" (un cube, ...).

```
Th goal (Grasping B1) (Thuse TC Group)
Def Theorem TC-Group(X)
    Th goal (Manip X)
    Th cond Th Goal (Grasping X)
            Th card Th goal (Grasping Y)
                    Th goal (Get rid of Y)
                            (The use TC Get rid of)
    Th goal (Cleartop X) (Thuse TC cleartop)
    Th set Y (Top center X)
    Th goal (Movehand Y) (Thuse TC Movehand)
    Th assert (Grasping X).
```

IV. ASPECT SEMANTIQUE

Le problème est donc de traduire la phrase anglaise en "Planner". Le sens de la phrase comprend :

- les objets ou notions dont il est question dans la phrase
- leurs relations que la phrase exprime.

Objet ou relations sont représentés par des programmes planner :

Exemple de la traduction en programme planner de :

A red cube which supports three pyramids, but is not contained in a box.

Th prog X_1	
Th goal is X_1 block	Y-a-t'il un block X_1
Eg dim X_1	à côtés égaux
Th goal color X_1 red	rouge
Th find 3 X_2	tel que :
	si on cherche trois objets X_2
Th goal is X_2 pyramid	qui sont des pyramides
Th goal support X_1 X_2	qui sont supportées par X_1
Th not Th prog X_3	et tel qu'il ne faut pas qu'il existe X_3
Th goal is X_3 box	qui est une boîte
Th goal contain X_3 X_1	qui contient X_1

On considère aussi des marqueurs sémantiques (analogues aux caractéristiques syntactiques) qui indiquent pour chaque notion : forme, taille, couleur, si l'objet est animé ou non ...

Par exemple, la notion "red" ne peut s'appliquer qu'à des objets physiques. C'est l'accord sémantique à mettre en correspondance avec l'accord grammatical.

Les marqueurs des clauses indiquent si on a une question, une phrase impérative ou déclarative.

V. ASPECT REPONSE

Le robot doit répondre aux questions, et être capable de décrire les évènements. Cet aspect a été moins bien traité par WINOGRAD que les autres. Le problème essentiel est de "deviner" si le questionneur veut des détails ou non : par exemple, à la question "avez-vous l'heure?" il s'agit de donner effectivement l'heure qu'il est ...

VI. CONCLUSION

WINOGRAD a passé 2 années de recherche et 2 années de programmation pour terminer ce travail remarquable. Ce programme existe sur PDP 10 au Massachusetts Institute of Technology. On l'utilise en tapant sur un Télétype la phrase anglaise, la réponse est donnée sur un écran cathodique. L'univers du robot est représenté sur ce même écran et modifié au fur et à mesure de la "conversation".

.../...

Le programme est écrit en Lisp (sous-programmes heuristiques), en Planner (résolution de problème et définition des notions sémantiques) et en Programmar (langage conçu spécialement par WINOGRAD pour l'aspect syntactique).

Si un chercheur français voulait s'inspirer de ce travail, il lui serait à notre avis nécessaire dans un premier temps de définir l'équivalent d'une grammaire systémique pour le français. Ensuite, ainsi que l'a fait WINOGRAD, travailler sur un univers limité qui ne soit pas nécessairement celui d'un "robot" du type du MIT, mais tel que les définitions sémantiques des différentes notions soient en nombre raisonnable (stade actuel de la technologie des ordinateurs).

On remarque que ce travail a été fait indépendamment de toutes les notions théoriques de monoïde libre et de grammaires formelles dépendant du contexte. Il a été volontairement fait dans le but de montrer les capacités des ordinateurs considérés non plus comme des machines à manipuler des fichiers mais comme des supports d'algorithmes. Ce sont ces algorithmes qui représentent toutes les notions syntactiques et sémantiques analysées.

En clair, le précédent programme signifie que pour prendre X, il faut vérifier que X est manipulable, rechercher si on n'est pas déjà en train de le saisir, sinon regarder si on n'est pas en train de saisir un autre objet Y, dans ce cas se débarrasser de Y par l'action correspondante (TC Get rid of Y) ; ensuite s'arranger pour qu'il n'y ait rien sur X par l'action TC Cleartop, bouger la main pour la mettre sur le dessus de X, mettre dans sa mémoire qu'on est en train de saisir X.

Chaque programme peut appeler les autres de manière entièrement récursive. En particulier ici, TC Get rid of Y utilisera l'action Grasp Y.

.../...

Question posée par J. P. HATON à C. ROCHE

Savez-vous comment l'estimation de plausibilité est effectuée par le programme de WINOGRAD, dans le cas où une ambiguité se présente ?

Réponse

Cette estimation est faite de manière extrêmement rapide et simplifiée sans utiliser aucune notion de probabilité et utilise des nombres affectés à chaque notion sémantique.

Une structure sémantique a comme "plausibilité" la somme de celle de ses constituants. Pour répondre à une question ambigue, on répond d'abord à la plus plausible, puis on recommence en utilisant simplement les renseignements donnés par la précédente question et sa réponse, si on peut alors répondre en trouvant la même réponse on déduit 500 de la plausibilité. Si l'information de la phrase précédente est insuffisante, le système cherche la réponse en utilisant ce qui a été mentionné dans tout le discours précédent (sans donc tenir compte des données de base et de ce qui peut en être déduit logiquement), s'il y a succès on déduit 200 de la plausibilité. Si la plausibilité de la question reste supérieure à celle de la question suivante on donne la réponse. Sinon on examine les questions suivantes dans l'ordre de plausibilité. Lorsqu'à la fin l'ambiguité subsiste avec des plausibilités voisines, le système répond par exemple :

I'm not sure what you mean by "on top of" in the phrase "on top of green cubes".

Do you mean :

1. Directly on the surface
2. Anywhere on top of ?

Question posée par M. ROSSI à C. ROCHE

Comment est défini l'alphabet terminal et comment est fait l'analyse grammaticale chez WINOGRAD ?

.../...

Réponse

Il serait intéressant d'étudier ce qui pourrait s'appeler alphabet terminal et non terminal dans le travail de WINOGRAD. WINOGRAD lui-même ne s'est pas apparemment occupé de ce problème. L'analyse grammaticale est faite non en simulant des règles de réécriture ou transformationnelles par des programmes en ordinateur, mais en construisant directement des programmes (récursifs, avec programme de "contrôle" ...) qui ne se ramènent pas nécessairement à des règles de type grammatical classique.

.../...

LANGAGE ET COMMUNICATION

M. Jean GAGNEPAIN

Directeur de l'U.E.R. du Langage à l'Université de Haute Bretagne
RENNES

LANGAGE ET COMMUNICATION

Une certaine confusion règne actuellement chez les linguistes en quête de scientificité. Les uns, parce qu'ils veulent apprendre aux machines à parler, ont moins pour but de comprendre le fonctionnement intrinsèque du langage que d'en mimer artificiellement, voire d'en améliorer un jour, la performance. Ce sont les informaticiens. Les autres, parce qu'ils ont cru trouver dans l'expérimentation pathologique le moyen de fonder une théorie à visée finalement thérapeutique, tendent à mettre l'accent moins sur l'information produite que sur l'instance qui en permet à l'homme spontanément la production. Ce sont les cliniciens.

Or, il se trouve qu'en dépit d'un vocabulaire commun la démarche est inverse - de ceux qui, pour la mettre à la portée d'ordinateurs dont les capacités, qualitativement, ne diffèrent guère de celles du lave-vaisselle ou de l'aspirateur, entendent traiter mathématiquement la parole en faisant de la linguistique seulement un nouveau champ d'application - et de ceux qui, se gardant à la fois du psittacisme et de la mise en boîte et refusant de chercher le modèle de ce que nous sommes tant dans l'animal que dans les robots par nous-mêmes fabriqués, essaient d'élaborer les sciences de la culture en préservant d'abord la spécificité de leur objet.

*

* *

Parce que, précisément, la parole en est un, sa réalité n'est point toute, il s'en faut, du domaine de l'observable. Et cela non seulement parce que le phonème ne se confond pas avec la réalité naturellement continue du son, n'en déplaise aux acousticiens qui, prenant pour argent comptant ses frontières et comme si le tracé permettait de distinguer, par exemple, l'éternuement de l'éjective, croient pouvoir le définir en formants ;

.../...

mais parce qu'également le signifié n'est pas le sens auquel, symboliquement, le signe vient à s'attacher. La relation son-sens survit chez l'aphasique qui souvent obéit aux consignes et peut, en situation, s'exprimer de manière opportune. Elle est, par dressage, accessible au moins à certains animaux qui, s'ils mémorisent et, dans le cas du perroquet, imitent, pourtant ne parlent pas.

Il en va tout autrement du mot qui exige, pour être reconnu comme tel, d'être non seulement appréhendé sous son double aspect, mais compris. A l'instar de l'outil qui, hors de l'univers technique dans lequel il s'inscrit, ne diffère en rien d'un objet, le signe, en fait, est analyse et nul ne le déchiffre qui d'abord ne maîtrise sa structure logique implicite.

Ce qui nous est livré, autrement dit, dans un message, c'est moins de l'information que la possibilité d'en créer. La langue n'est que forme et phonétiquement aussi bien que sémantiquement ne s'achève que par la nécessaire intervention du locuteur qui lui donne son contenu. Tout phonologue sait que les identités dont il s'occupe et qu'il nomme traits pertinents ne sont pas des substances mais des cadres de variation ; que les unités dénombrées par lui dans la chaîne ne sont point articulations naturelles mais n'ont d'autre critère que leur irréductibilité. De même le sémiologue aurait-il peu de peine à montrer qu'un vocabulaire n'est pas un lexique, ni une phrase un texte ; qu'en chaque cas le sens se déduit et s'engendre à partir d'une signification qui le classe et le distribue à telle enseigne qu'un mot, par essence ambigu, n'est, d'une part, épuisé par aucune organisation conceptuelle, et que, le tout d'une séquence n'étant jamais égal, d'autre part, à la somme de ses parties, il est vain d'espérer fonder sur l'inventaire d'un corpus le système des règles qui grammaticalement le sous-tend.

C'est à tort, selon nous, qu'on oppose couramment aujourd'hui linguistique taxinomique et linguistique générative. Loin d'être propriétés d'une éventuelle science du langage, ces deux capacités, sélectivement atteintes dans les aphasies, nous semblent, en définitive, comme abscisse et comme ordonnée de toute grammaire, être inhérentes à son objet. L'ambiguité, en bref, nous apparaît fondamentale d'un univers logique médiatisant spontanément et toujours à priori l'univers perçu dans lequel le sujet parlant désespérément l'investit. C'est la condition même de ce que certains peut-être nommeraient la "modularité", nous dirons, nous, de la pensée.

*

* *

Il y a plus ; car ce qui est vrai du langage est vrai de tout fait de culture et l'analyse aussi bien que la dialectique grâce auxquelles l'homme élabore des structures qu'explicitement il conteste sont sous-jacentes, en même temps qu'à son savoir, à son pouvoir, à son être et à son vouloir. Une même rationalité se révèle dans l'impropriété logique du signe, la disponibilité technique de l'outil, l'arbitraire ethnique de la personne, l'ascèse éthique de la norme.

On ne saurait du même coup isoler la parole de l'ensemble des phénomènes qui lui sont analogues et constituent, au sens pascalien du terme, un ordre dont l'étude ressortit aux sciences humaines. De même, en effet, que la signification n'est qu'un aspect d'un processus général englobant avec elle fabrication, institution et règlementation, de même, à ce niveau, ne saurait-on rien affirmer du langage qui ne soit également vérifiable au plan de l'art, de la société ou du droit. L'originalité commune des disciplines qui en traitent tient à ce que la formalisation n'est point ici, comme dans les sciences de la nature, le fait du descripteur mais de l'objet décrit et que les modèles y seront plus exactement révélés que construits.

Et qu'on ne nous dise pas qu'il s'agit là de considérations accessoires pour ceux qui s'intéressent au seul problème du langage. Car toutes les analyses, en raison de l'interférence des médiations, se recoupent et s'il est vrai que, du point de vue de la conscience, ce dernier s'avère être spécifiquement grammaire, il ne laisse pas d'intéresser pour autant notre conduite, notre condition, notre comportement et d'être simultanément sur ces plans écriture, idiome et réticence. Or ce qu'on nomme la parole est, très évidemment, tout cela à la fois. A la polysémie qu'il tient déjà de son statut de signe et qui force à l'actualiser, le mot écrit joint le silence qu'avec tous les aléas qu'elle comporte vient interpréter la lecture. Et parce que les universaux sont un mythe, que, si le rossignol chante partout en rossignol, l'homme, lui, ne parle humain qu'au prix d'un effort de communication par lequel les divergences qui l'identifient se trouvent momentanément transcendées, on comprend que les langues ne soient jamais que des interférences d'idiolectes, qu'avec le polyglottisme la traduction, même discrète, soit la loi et qu'au terme le malentendu, de façon non fortuite, mais constitutive, soit roi. L'herméneutique, enfin, vient au secours de quiconque veut, au-delà du sens du message, décrypter le sens du discours et déceler le non-dit latent sous l'intention manifeste qui n'est jamais qu'un compromis. "Omnis homo mendax" dit l'Ecriture.

C'est somme toute, une autre façon de dire qu'il est raisonnable ... mais aussi un obstacle de plus à la reconnaissance ainsi qu'à la communication.

*

* *

L'ampleur des difficultés rencontrées dans un domaine qui n'est pas le leur et que les spécialistes n'ont encore que très superficiellement exploré fait apprécier d'autant mieux l'ingéniosité des constructeurs qui, d'ores et déjà, si modestement que ce soit, conversent avec l'ordinateur. Il faut avouer, pourtant - et la naïveté, linguistiquement parlant, en même temps que l'extrême sélectivité des programmes en fait foi - qu'ils n'y trouvent que ce qu'ils y ont mis, la machine restant stupide et le dialogue, tout compte fait, narcissique. Certains, je le sais, s'en consolent en pensant qu'il s'agit d'une étape dans la voie toujours ouverte de la complexité ; d'autres, dont je suis, tiennent que le passage, au contraire, du cerveau à l'ordinateur ne saurait être continu et que, la différence étant de qualité, le linguiste n'a guère plus d'enseignement à tirer de la cybernétique que l'ornithologue des progrès de la balistique ou de l'aviation.

Ce que l'ingénieur appelle grammaire, en effet, n'a que peu de rapports avec celle qui, à son insu, structure son propre langage. Loin de prétendre en rendre compte, elle vise à la remanier au contraire pour qu'émergent algébriquement à la forme des relations que rien linguistiquement ne suggère mais qu'il croit saisir dans les choses. C'est assez dire qu'en prenant le parti du consommateur et en créant des mécanismes capables, par distribution ou transformation, de réduire la polysémie et de conjurer l'ambiguité, il renonce à nous éclairer sur le principe d'une activité que sa démarche même contredit. S'il est normal, dans ces conditions, qu'il s'annexe Chomsky, il l'est moins de voir de nos jours tant de linguistes succomber, eux, à la facilité et lui emboîter aussi gaillardement le pas. Appeler un chat un chat, ce n'est plus du langage, mais de l'épicerie ou de la pharmacie.

.../...

C'est dans une impasse analogue que nous semble s'être d'emblée engagée la traduction automatique. L'idée d'améliorer la communication par l'identification des codes et la perméabilité des circuits, si partagée qu'elle soit par l'usager, n'est qu'un espoir régulièrement déçu. N'en sont surpris que les tenants d'un platonisme simpliste qui croient que partout l'homme est l'homme et qu'ayant la même chose à dire, il se contente de l'habiller diversement. Si l'on constate, en revanche, qu'ici l'habit fait le moine et que chacun, loin de représenter la combinaison singulière de facteurs qui lui préexistent dans quelque monde intelligible, n'est autre que l'ensemble contingent de ses différences, on conçoit la vanité d'un effort visant - tout en en sauvegardant le caractère - à rendre transparents nos échanges. Cela revient à nous détruire, en somme, puisque les bruits ou parasites, en l'occurence, c'est nous.

*

* *

Un seul domaine jusqu'ici échappe à cette contagion, celui que nous avons nommé plus haut le discours. Le sens, on l'a dit, y résulte d'une dialectique des intentions qui, ressortissant en gros à la finalité, se trouvait marginalisée et du même coup protégée des excès d'une science fondée sur la seule efficience. De là vient le succès - sans doute provisoire, eu égard à celui de la théorie des décisions, mais d'autant plus caporaliste - de la psychanalyse auprès d'un nombre sans cesse croissant de littéraires qui la marquent de plus en plus de leurs défauts.

Tout se passe, autrement dit, comme si les sciences humaines n'avaient actuellement le choix qu'entre le verbiage et l'algorithme, l'illusionisme et le comput. Les uns, dans le langage, auscultant le désir sont, au fond, nos modernes poètes ; les autres éprouvent pour un type de calcul qu'ils sont eux-mêmes les premiers surpris de comprendre la confiante gratitude de ceux qui autrefois n'ont point su résoudre leurs problèmes de robinets ou de trains. Il fallait être ministre pour se figurer qu'il suffît de parler de pluridiscipline pour reconcilier l'inconciliable et faire une vertu de deux vices.

.../...

Sans doute la mathématique reste-t-elle l'irremplaçable langage de la science ; mais son extension à ce nouveau monde dont Janus est l'emblême et que Freud, Marx ou Saussure ont les premiers cerné ne saurait lui faire oublier qu'elle devra rester ici l'auxiliaire et, pour être efficace, accepter de se transformer. Est-il si paradoxal, après tout, d'espérer, au seuil d'une autre Renaissance, avec l'avènement d'un nouveau mode de pensée, celui d'une mathématique de la qualité ?

ROLE

DES CONTRAINTES LINGUISTIQUES EN RECONNAISSANCE DE LA PAROLE

par

Jean-Paul HATON

Laboratoire d'Electricité et d'Automatique
Université de Nancy I

* * * *

Journées d'Etude sur la Parole, Lannion, 31 mai-1, 2 juin 1972

Pages

I - INTRODUCTION.

Les résultats obtenus en reconnaissance de la parole depuis 1950, époque des premiers travaux, restent limités et ne concernent, pour la plupart, que la reconnaissance de mots isolés d'un vocabulaire restreint [1]. Devant cet échec relatif, on a pu constater un élargissement de ce domaine de recherche, et il est maintenant certain que les paramètres purement acoustiques sont insuffisants pour mener à bien la reconnaissance d'une conversation continue ("connected speech"). Seul un système travaillant aux différents niveaux : acoustique, linguistique, syntaxique et sémantique pourra réaliser cette entreprise. Ces niveaux, hiérarchiquement classés, sont de plus interconnectés, le système de reconnaissance devant pouvoir à tout instant et dans les deux sens, passer d'un niveau à un autre.

Ce schéma théorique est d'ailleurs en accord avec ce que l'on connaît du processus humain correspondant [2], encore que la complexité des analyses effectuées à chaque niveau, ainsi que les interactions entre différents niveaux, soient sans commune mesure entre l'homme et la machine [3].

Ces différents indices tenant compte du "contexte" (linguistique, syntaxique, sémantique) sont en un certain sens indépendants (cf. par exemple Chomsky [4]) et la présente étude ne s'intéresse ainsi qu'à l'amélioration de la reconnaissance de mots par l'utilisation de contraintes linguistiques. Ces contraintes sont constituées de résultats statistiques concernant la langue, en particulier les probabilités conditionnelles d'apparition des diphonèmes et des triphonèmes. Elles sont utilisées ici dans le cadre d'un système général de reconnaissance qui a déjà été décrit [5].

Des méthodes analogues ont été utilisées avec succès en reconnaissance de caractères écrits [6] → [8]. Dès 1958, Denes avait montré dans un cas très simplifié que les probabilités d'occurrence des diphonèmes pouvaient améliorer sensiblement le taux de reconnaissance de sons [9]. Plus récemment, Green utilise ces données statistiques pour améliorer la reconnaissance acoustique d'un système, en particulier pour déterminer les frontières de mots [10].

Dans tous les cas la méthode nécessite le stockage de quantités importantes de données, cette difficulté a été en partie tournée ici par l'utilisation de triphonèmes binaires (cf. par. II).

Après quelques études préliminaires, les résultats obtenus dans diverses expériences montrent l'avantage que l'on a à utiliser les données linguistiques en reconnaissance de la parole.

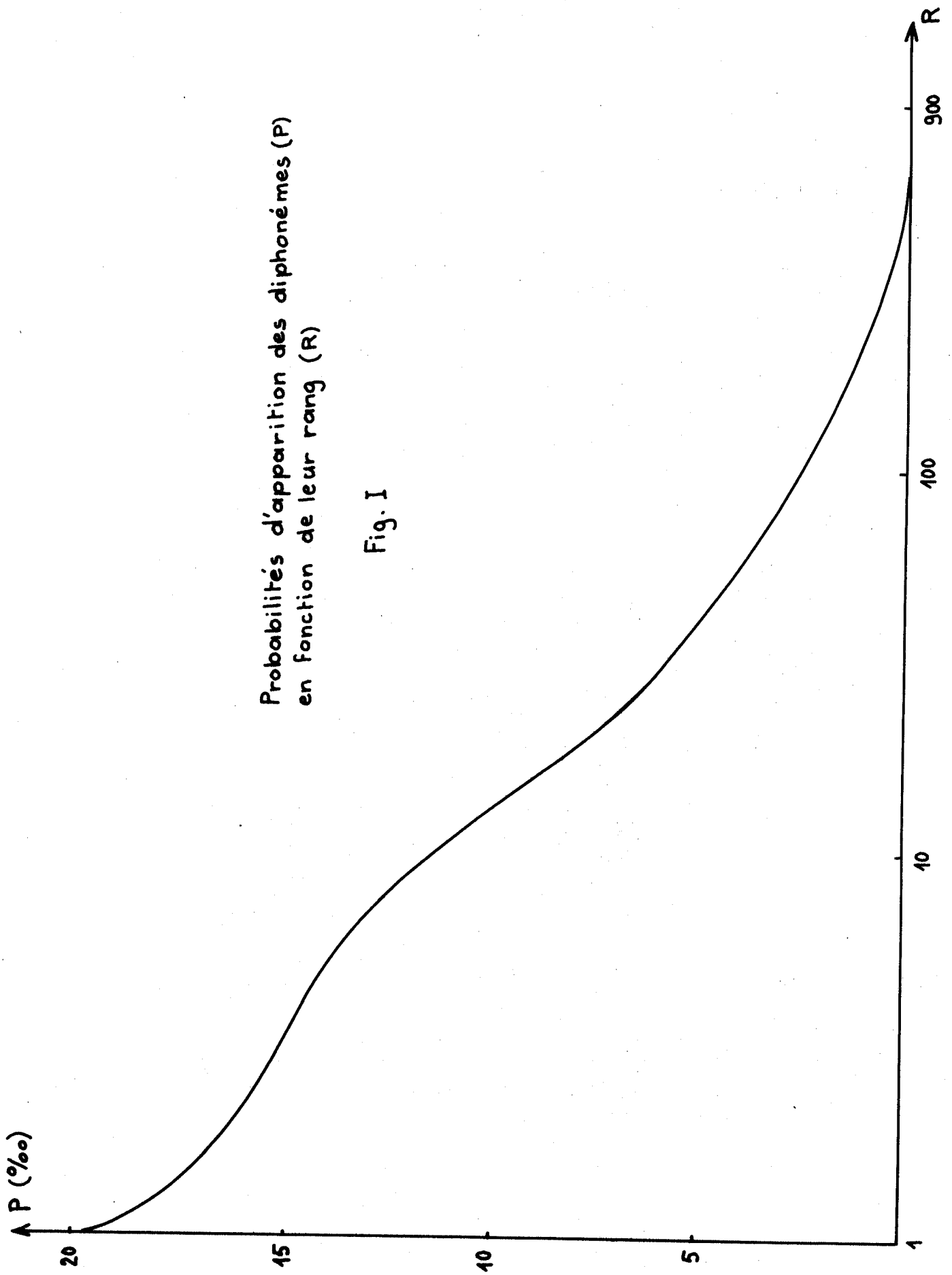
II - LES DONNEES STATISTIQUES UTILISEES.

A l'aide d'un programme spécial, diverses distributions statistiques concernant la langue française ont été déterminées [11] . Nous n'utiliserons ici que les probabilités d'apparition des diphonèmes et des triphonèmes.

L'étude statistique met en évidence tout le parti que l'on peut tirer de la redondance existant au niveau des associations phonétiques. En choisissant, comme nous l'avons fait, trente phonèmes pour décrire la langue française (ce qui n'introduit quasiment pas d'ambiguité), on constate que 600 seulement des 900 diphonèmes possibles sont effectivement utilisés. La disparité est encore plus importante au niveau des triphonèmes puisque, sur les 27 000 (30 x 30 x 30) possibles, environ 2520 sont seuls permis par les contraintes phonétiques.

On montre ainsi grossièrement l'existence d'une norme linguistique de la langue française. Cela se retrouve dans le fait qu'aucune relation simple n'existe entre les distributions de probabilités des phonèmes, des diphonèmes et des triphonèmes, ainsi que dans l'allure de la courbe donnant, par exemple, la probabilité d'apparition des diphonèmes en fonction de leur rang (figure I en coordonnées semi-logarithmiques).

Il est donc légitime d'attendre une amélioration du taux de reconnaissance par l'utilisation de ces données statistiques, puisque cela revient à tenir compte des contraintes phonétiques et linguistiques imposées par la langue. Il semble d'ailleurs que l'homme possède en mémoire, plus ou moins consciemment, ces statistiques, acquises par "apprentissage" [12]. Cependant, cette amélioration ne peut prétendre à autre chose qu'à une



Probabilités d'apparition des diphonémes (P)
en fonction de leur rang (R)

Fig. I

valeur statistique, aussi ne faut-il pas lui accorder trop d'importance.

Pour diminuer les données à stocker en mémoire, les triphonèmes ont été codés sous forme binaire :

$$\begin{cases} 1 & \text{si le triphonème considéré existe effectivement} \\ 0 & \text{sinon} \end{cases}$$

Ceci permet de coder un triphonème sur un élément binaire et non sur un mot de mémoire. Il en résulte un gain de place dans un rapport d'environ 30 à 1.

Le principe d'utilisation des contraintes statistiques qui va être utilisé peut alors se schématiser ainsi :

- élimination de certains "candidats" à la reconnaissance en considérant les triphonèmes binaires,

- classement des candidats restant en tenant compte des probabilités d'apparition des diphonèmes.

Les expériences reportées ici ont été effectuées avec un vocabulaire de 54 mots de longueur moyenne (au total 320 phonèmes) prononcés sans précautions particulières et enregistrés sur magnétophone. Pour mettre en évidence l'influence de la taille du vocabulaire, deux séries de résultats statistiques seront successivement utilisés :

- une statistique portant sur la langue française de façon générale (voir plus haut), que l'on appellera "statistique générale" ;

- une statistique portant exclusivement sur le vocabulaire de 54 mots utilisé, que l'on appellera "statistique réduite".

III - DESCRIPTION DU SYSTEME DE RECONNAISSANCE ET PRINCIPE GENERAL DE FONCTIONNEMENT.

La figure II donne un schéma-bloc du système de reconnaissance. Ici seuls les deux premiers niveaux (acoustique et surtout linguistique) nous intéressent.

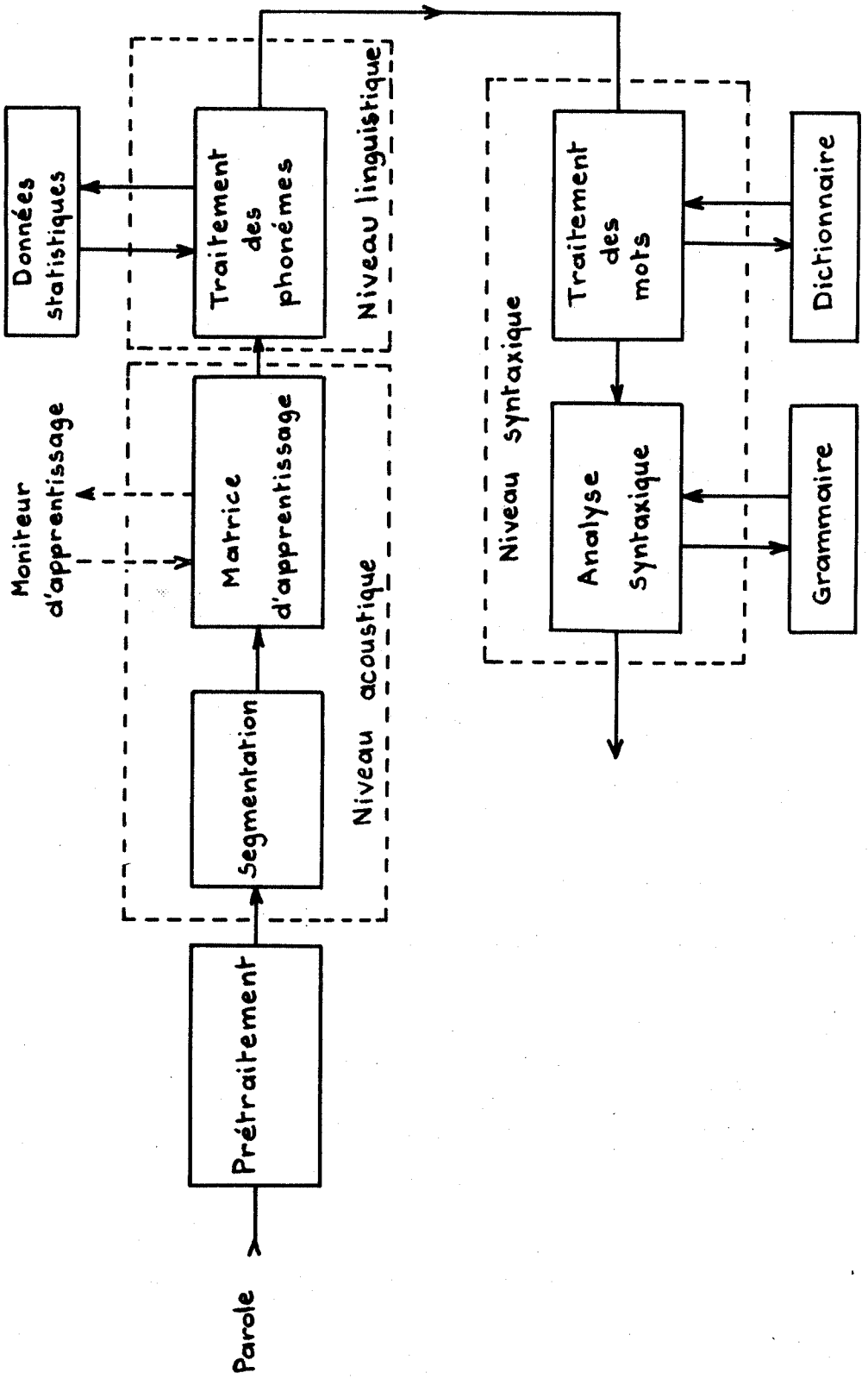


Schéma-bloc du système de reconnaissance

Fig II

Le premier niveau de reconnaissance phonémique, correspondant à l'utilisation de paramètres purement acoustiques, comporte un segmenteur et une matrice d'apprentissage à poids variables dont chacune des lignes est affectée à la reconnaissance d'une classe de phonèmes. A chaque fin de phonème, détectée par le segmenteur, la matrice fournit pour chacune des 30 classes de phonèmes un "score acoustique" et, à ce niveau, le système donne comme réponse la classe de phonème dont le score est maximal.

Actuellement ce niveau acoustique du système de reconnaissance est sur le point d'être entièrement réalisé en hardware.

En fonctionnement normal, les trente scores acoustiques fournis par la matrice d'apprentissage sont fournis au deuxième niveau, le niveau linguistique. Compte tenu des deux phonèmes précédemment reconnus par le système, les "candidats" correspondant à des triphonèmes non permis sont éliminés. Pour les autres, un score final est obtenu en ajoutant au score acoustique un score linguistique proportionnel à la probabilité d'apparition du diphonème correspondant. Ici encore, la réponse du système est le phonème dont le score final est maximal.

Nous nous limitons ici à l'étude de la succession de phonèmes ainsi obtenus. Un traitement ultérieur aux niveaux suivants permettra le décodage en mots puis en phrases par l'utilisation d'un dictionnaire et de règles de syntaxe.

IV - EXPERIENCES PRELIMINAIRES.

Le premier problème qui se pose dans l'utilisation du système qui vient d'être décrit est de savoir quelle importance donner aux informations linguistiques par rapport aux informations acoustiques. Il est impossible de répondre a priori à cette question, c'est pourquoi une étude systématique a été menée en testant le système avec différentes valeurs accordées aux informations linguistiques. D'autre part, il apparaît clairement qu'une erreur de reconnaissance au niveau d'un phonème peut être extrêmement préjudiciable pour la reconnaissance de la suite du message étudié.

Pour réduire au maximum ce risque d'erreur, il est intéressant de réduire le nombre de candidats possibles avant d'utiliser le traitement linguistique, et de ne garder ainsi que les quelques phonèmes dont le score acoustique est suffisamment important. On évite ainsi de reconnaître à tort un phonème dont le score linguistique serait important et le score acoustique faible. Ici encore, seule une étude systématique pouvait donner une solution.

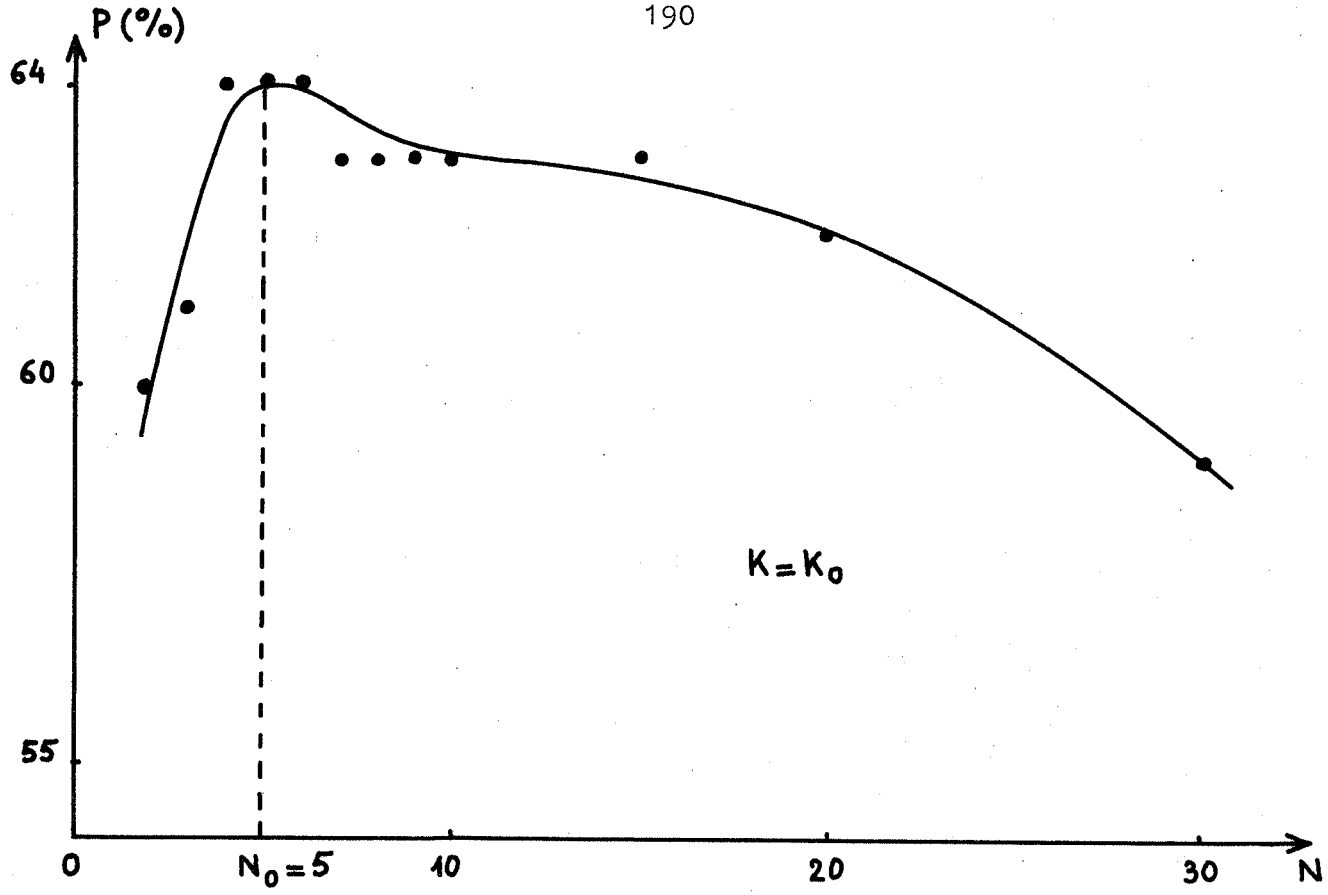
Le tableau I donne le pourcentage de reconnaissance obtenu en fonction du nombre de candidats et du coefficient accordé aux informations linguistiques (celles-ci étant prises au départ de l'ordre de grandeur des scores acoustiques).

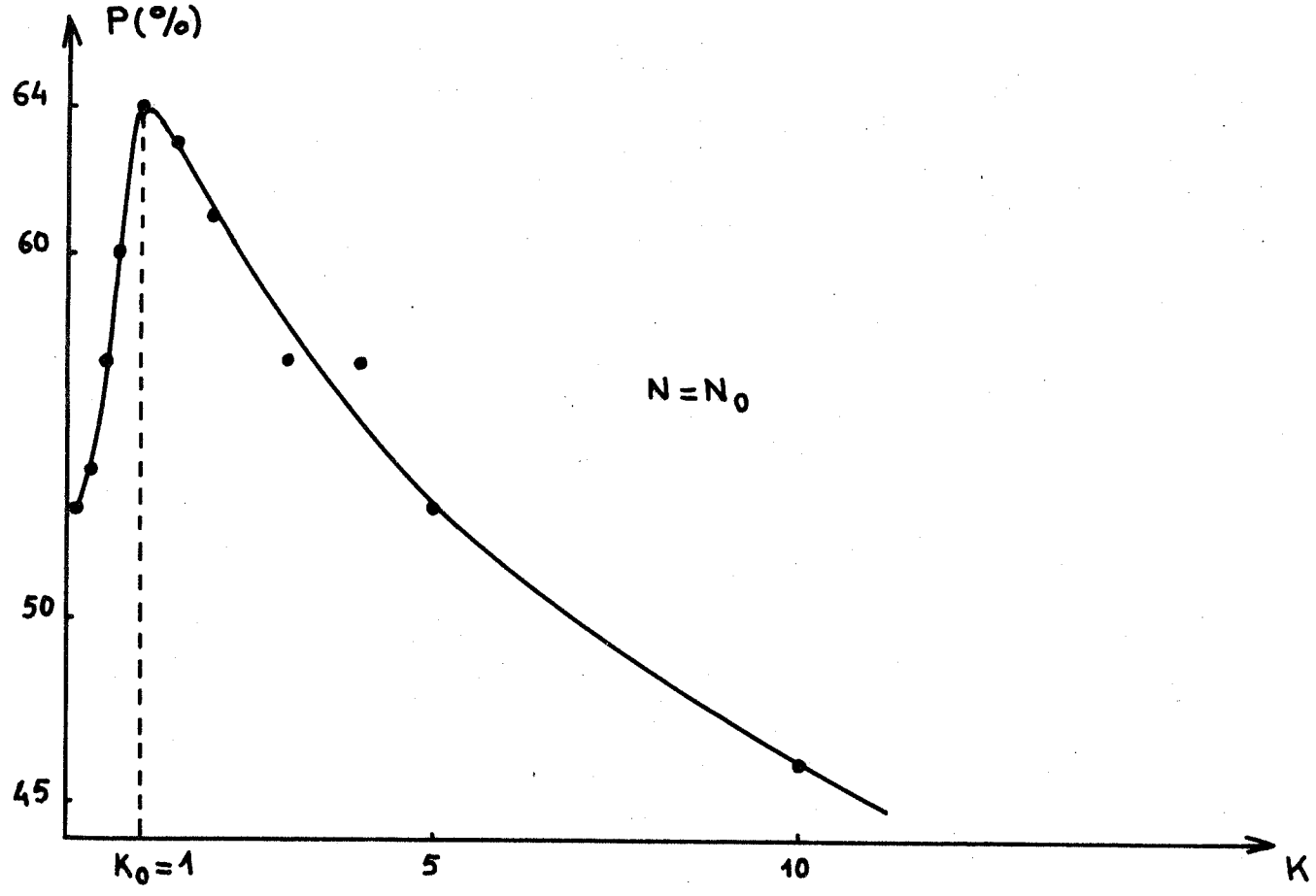
K \ N	2	3	4	5	6	7	8	9	10	15	20	30
0,1	53	53	53	53	53	53	53	53	53	53	53	53
0,3	53	54	54	54	54	54	54	54	54	54	54	54
0,5	57	57	57	57	57	56	56	56	56	56	56	53
0,7	58	58	60	60	59	59	59	59	59	59	58	56
1	60	61	64	64	64	63	63	63	63	63	62	59
1,5	58	60	63	63	62	61	61	61	61	61	61	56
2	56	59	61	61	59	59	59	59	59	59	58	54
3	53	56	57	57	56	56	56	56	56	56	54	51
4	53	55	56	57	55	55	56	55	55	55	53	49
5	51	53	53	53	52	52	52	52	52	52	51	48
10	46	46	47	46	45	45	44	44	44	43	43	43
0	53	53	53	53	53	53	53	53	53	53	53	53

N : nombre de candidats
K : coefficient des informations linguistiques

Tableau I.

Cette expérience a été effectuée sur les 54 mots du vocabulaire prononcés par un locuteur et n'avait d'autre but que la résolution des deux problèmes posés (en particulier les triphonèmes n'étaient pas utilisés).





Variation du pourcentage de reconnaissance

Fig III

D'autre part, la figure III montre les courbes de variation du pourcentage en fonction du nombre de candidats (pour la valeur optimale du pourcentage) d'une part, et en fonction du pourcentage (pour la valeur optimale du nombre de candidats) d'autre part.

On voit qu'un nombre de candidats égal à 5 et un coefficient multiplicatif des informations linguistiques égal à 1 fournissent le taux de reconnaissance le plus élevé. Ce sont ces valeurs qui ont été adoptées pour la suite. Il est remarquable que le choix du coefficient multiplicatif soit plus critique que celui du nombre de candidats à retenir (cf. figure III).

V - RESULTATS EXPERIMENTAUX DU SYSTEME COMPLET.

Compte tenu de ce qui précède, le fonctionnement du système de reconnaissance peut être schématisé comme l'indique la figure IV.

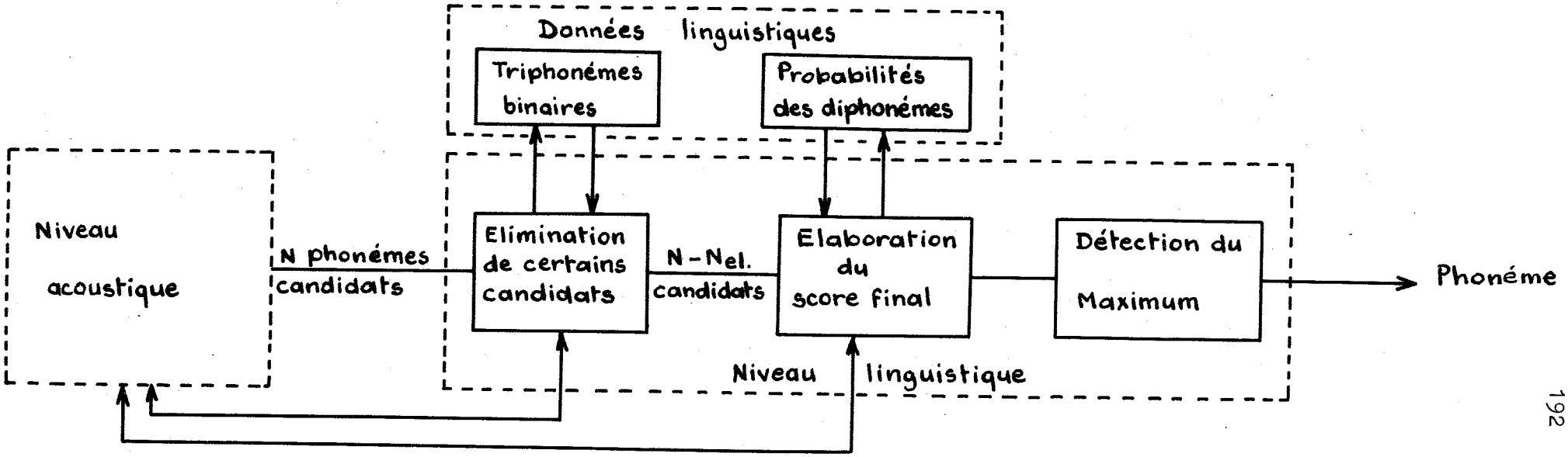
La "réponse" du niveau acoustique consiste en les N plus probables candidats, classés par score décroissant. Compte tenu des deux phonèmes qui viennent d'être reconnus et qui sont conservés en mémoire, N_o candidats sont immédiatement éliminés, correspondant à des triphonèmes non permis. Cette opération est très rapide puisqu'elle consiste simplement à tester l'état d'un bit d'un mot mémoire (un triphonème étant codé sur un élément binaire).

Ensuite le score final des $N - N_o$ candidats restant est obtenu en ajoutant au score acoustique un score linguistique obtenu à partir des probabilités d'apparition des diphonèmes, compte tenu du coefficient qui a été déterminé auparavant.

Les candidats sont à nouveau rangés par score décroissant et le système donne comme réponse le phonème dont le score final est maximal.

Le tableau II donne un exemple typique de reconnaissance.

La réponse du système au niveau acoustique (phonème /e/) a été corrigée au niveau linguistique pour donner finalement le phonème /i/ qui est la réponse exacte.



Fonctionnement du niveau linguistique

Fig IV

REPONSES PRECEDENTES : RPAS1= AI
RPAS2= T

SCORES ACOUSTIQUES

NUMERO	1	2	3	4	5
SCORE	35463	32902	29424	22013	19786
PHONEME	AI	S	I	A	L

REPONSE ACOUSTIQUE: AI

CANDIDATS ELIMINES PAR TRIPHONEMES

PHONEME	S	L

SCORES FINAUX

NUMERO	1	2	3
SCORE	63514	52391	39138
PHONEME	I	AI	A

REPONSE FINALE: I

Reconnaissance du phonème /i/

(dans le mot "étiquette")

TABLEAU II

En utilisant les résultats de la statistique générale (cf. par. II), le taux de reconnaissance de phonèmes obtenu pour un locuteur avec le vocabulaire de 54 mots est de 62 %.

Avec la statistique réduite, ce taux s'établit à 67 %. Cette amélioration était, bien entendu, prévisible puisque cela revient à augmenter les contraintes linguistiques ; malheureusement on diminue ainsi la généralité du système.

Une partie des erreurs commises provient du fait qu'une mauvaise reconnaissance à un moment donné se répercute sur la reconnaissance des phonèmes suivants et peut ainsi conduire à un mot tout-à-fait différent de celui qui est prononcé. Nous allons maintenant exposer les méthodes qui nous ont permis de limiter de telles catastrophes. Elles consistent à établir des interactions entre les niveaux acoustique et linguistique et, plus précisément, une sorte de contre-réaction du niveau linguistique sur le niveau acoustique.

VI - INTERACTION NIVEAU ACOUSTIQUE - NIVEAU LINGUISTIQUE.

Le processus de retour en arrière intervient dans deux cas (cf. figure IV):

1) Lorsque tous les phonèmes fournis par la reconnaissance acoustique ont été éliminés par les triphonèmes binaires. Le phonème précédemment reconnu est sans doute faux car il donne des associations incompatibles avec la norme de la langue.

2) Lorsque le phonème effectivement reconnu au niveau linguistique possède un score final inférieur à un seuil déterminé expérimentalement. Là encore, le phonème précédent doit être erroné car le score linguistique obtenu à partir des diphonèmes est insuffisant.

Dans les deux cas, le phonème candidat qui avait été classé en deuxième position est choisi comme réponse précédente, à condition qu'il forme lui-même avec les deux phonèmes précédents un triphonème effectivement permis.

Ensuite la reconnaissance du phonème actuel est reprise depuis le début.

Le nombre de retours en arrière, pour un phonème, a été limité à trois. Au-delà, la présence d'une anomalie est signalée par le système. L'expérience montre d'ailleurs qu'un nombre plus élevé de retours n'amène aucune amélioration à la reconnaissance.

Avec la statistique réduite, le taux de reconnaissance de phonèmes passe de 67 % à 70 % pour un locuteur en utilisant ces contre-réactions.

D'autres tests doivent encore être ajoutés au système pour permettre d'autres possibilités de retours en arrière.

Toutes ces expériences ont été effectuées en temps différé, mais elles peuvent facilement être menées en temps réel, permettant même un traitement ultérieur des séquences de phonèmes à des niveaux supérieurs.

VII - CONCLUSION.

La nécessité d'avoir un système hiérarchisé pour reconnaître la parole est maintenant chose admise. Un tel système doit pouvoir travailler simultanément à différents niveaux - acoustique, linguistique, sémantique - de façon à affiner peu à peu la reconnaissance.

Le but de cet article n'était pas de présenter un tel système complet mais simplement d'étudier les interactions entre le niveau acoustique et le niveau linguistique. On a montré que l'utilisation de données linguistiques simples sur la langue (probabilités d'occurrence des diphonèmes, triphonèmes binaires) permet d'améliorer sensiblement la reconnaissance au niveau du phonème.

Bien entendu, ceci ne constitue qu'une première étape, les séquences de phonèmes ainsi obtenues devant être décodées en mots, et les mots assemblés en phrases, aux niveaux supérieurs du système de reconnaissance.

Bibliographie.

1 . J.P. HATON, "Reconnaissance de la parole : bilan de vingt années de recherches et tendances actuelles", à paraître dans les Annales des Télécommunications.

2 . G.A. MILLER, "Decision units in the perception of speech", IRE Tr. on Inform. Theory, 8, n°2, pp. 81-83, 1962.

3 . J.C. RISSET, "Sur certains aspects fonctionnels de l'audition", Ann. Télécom., 23, n°3-4, pp. 91-120, 1968.

4 . N. CHOMSKY, "Syntactic structures", Mouton, The Hague, 1957.

5 . J.P. HATON, M. LAMOTTE, "Prétraitement et reconnaissance de la parole : simulation et réalisations pratiques", Journées d'Etude sur la Parole, Aix en Provence, 1-2 avril 1971.

6 . E.M. RISEMAN, R.W. EHRICH, "Contextual word recognition using binary digrams", IEEE Tr. Computers, C 20, n°4, pp. 397-403, 1971.

7 . R.W. DONALDSON, G.T. TOUSSAINT, "Use of contextual constraints in recognition of contour-traced handprinted characters", IEEE Tr. Computers, C 19, n°11, pp. 1096-99, 1970

8 . R. NARASIMHAN, V.S.N. REDDY, "A syntax-aided recognition scheme for handprinted English letters", Pattern Recognition, 3, n°4, pp. 345-361, 1971.

9 . P. DENES, "The design and operation of the mechanical speech recognizer at University College, London", J. Brit. IRE, 10, pp. 219-229, 1959.

10 . W.A. AINSWORTH, P.D. GREEN, "Toward the automatic recognition of spoken Basic English", Conf. on Mach. Perc. of patterns and pictures, Teddington, G-B., 12-14 avril 1972.

11 . J.P. HATON, M. LAMOTTE, "Etude statistique des phonèmes et des diphonèmes dans le français parlé", Revue d'Acoustique, 4, n°16, pp. 258-262, 1971.

12 . H.B. SAVIN, "Word frequency effect and errors in the perception of speech", J.A.S.A., 35, pp. 202-206, 1963.

****

Question posée par M. MAISSIS à J. P. HATON

L'utilisation des données linguistiques étant à deux niveaux (élimination de séquences et pondération de celles-ci), quelle est l'amélioration du taux de reconnaissance due à l'utilisation des probabilités d'apparition de diphonèmes ...? Je pense que cette amélioration ne doit pas être importante.

Réponse

Nous n'avons pas encore eu le temps de déterminer exactement la part de chacun des deux niveaux dans l'amélioration globale du taux de reconnaissance. Nous comptons cependant le faire car c'est un point intéressant à connaître. Il est déjà certain que l'amélioration due à l'utilisation des probabilités d'apparition des diphonèmes n'est pas négligeable, surtout dans le cas de la statistique réduite.

INTERACTION ENTRE FACTEURS SEGMENTAUX ET NON SEGMENTAUX DANS LA RECONNAISSANCE DE LA PAROLE *

Björn LINDBLOM et Stig-Göran SOENSSON
Département de phonétique
Université de Stockolm
Fack, S - 104 - 09 Stockolm 50, Suède
et
Département de communication acoustique
Institut Royal de Technologie (KTH)
S-100-44 Stockolm 70, Suède

* Un prolongement de cette recherche a été présenté pour être publié dans les IEEE Transactions - Audio and Electroacoustics.

Pages

RESUME :

Le but de la présente étude est de montrer qu'on peut lire les spectrogrammes de phrases du Suédois de manière précise sous des conditions non-triviales. La grande partie de cet article est consacrée à la description des expériences relatives à cette étude. A partir des résultats obtenus on développe une stratégie formalisée afin d'aider les lecteurs de spectrogrammes à retrouver la structure grammaticale de la phrase à partir de l'information obtenue sur les traits prosodiques, tels que l'accent tonique et l'accent d'intensité. On démontrera en particulier que les résultats de lecture sont considérablement améliorés lorsqu'est fournie au lecteur l'information prosodique et grammaticale simultanément avec la présentation des caractéristiques des segments phonétiques sur l'écran de visualisation.

En conclusion, nous attirerons l'attention sur le rôle important joué par la grammaire et la prosodiedans les expériences actuelles et nous discuterons les implications de ces résultats quant aux recherches futures sur la reconnaissance automatique de la parole et sur la perception.

Le raisonnement de base :

Il est bien connu que les langues naturelles sont très redondantes; par conséquent que les signaux de parole puissent subir une distorsion considérable sans devenir inintelligibles est un fait d'observation courante. On interprète généralement ces faits pour remarquer que les auditeurs humains sont capables de comprendre ce qui est dit parce qu'ils font usage d'une manière ou d'une autre de la redondance inhérente à la structure linguistique.

Cette interprétation étant admise, il semble naturel pour les recherches sur la perception de la parole de développer et d'évaluer des modèles possédant cette redondance et cherchant à l'exploiter.

Dans de tels systèmes on utilisera en entrée une représentation phonétique partielle, le caractère partiel de cette représentation étant dû au bruit et d'autres facteurs de dégradation de la parole.

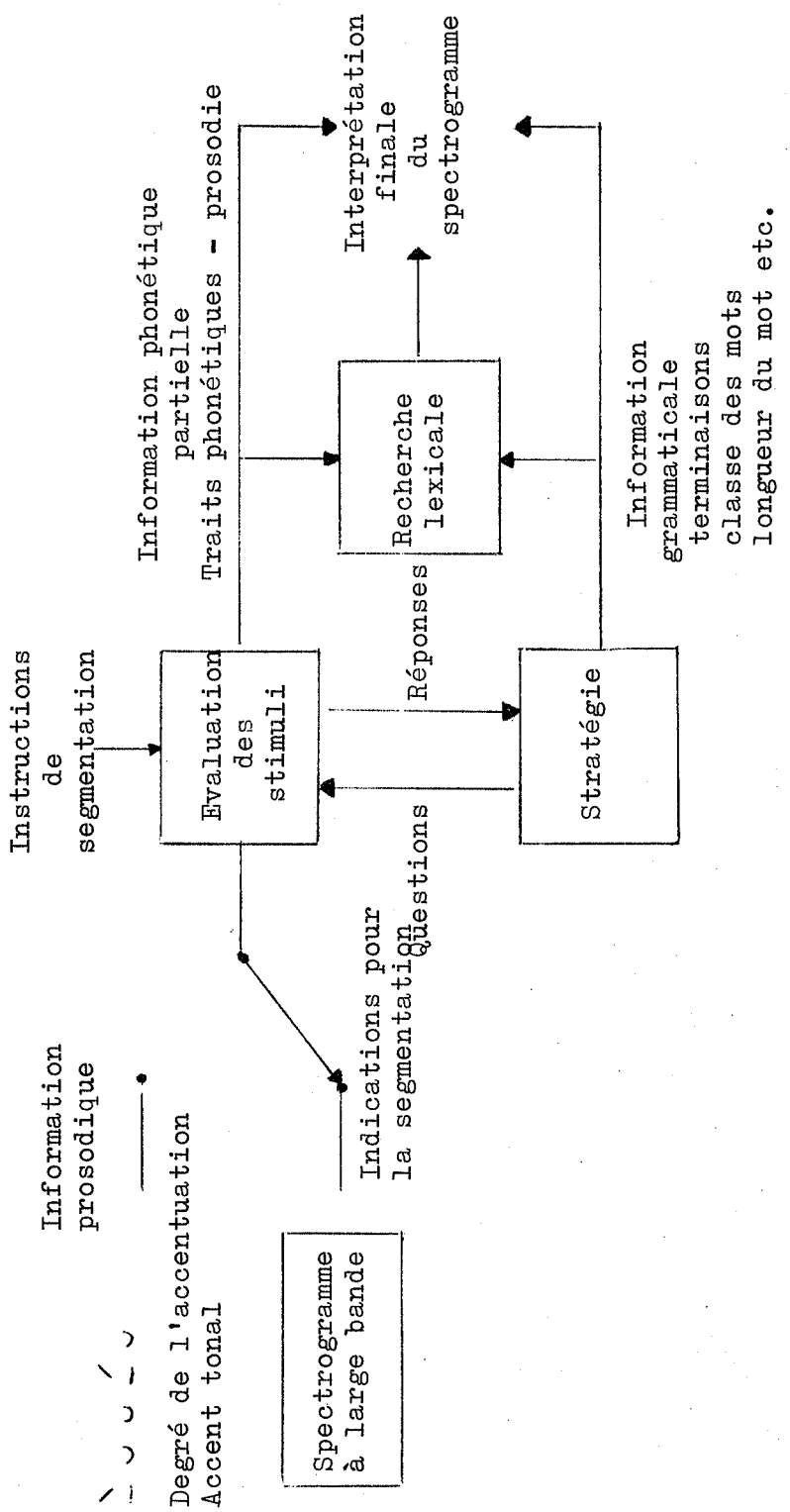
A la sortie on retrouvera précisés les aspects de l'information linguistique et du contenu du message qu'on peut logiquement déduire de cette représentation partielle du signal d'entrée.

Une application expérimentale :

Dans la présente recherche on a essayé d'appliquer ce raisonnement pour définir une stratégie de reconnaissance utilisant les caractéristiques d'entrée/sortie que nous venons de décrire et pour entreprendre une série d'expériences permettant d'évaluer cette stratégie. Une esquisse générale des expériences est indiquée dans la figure 1.

Le travail des sujets humains est de reconnaître le Suédois parlé à partir des spectrogrammes à large bande. Partant de la gauche de la figure on s'aperçoit que deux types d'informations nous sont donnés sur l'expression prononcée à l'entrée : l'information prosodique indiquant si telle syllabe porte l'accent principal et si elle a un accent tonique ou pas et l'information sur les traits phonétiques segmentaux qui est observée sur le spectrogramme visualisé. Pour identifier les stimuli le sujet dispose de deux répertoires d'instructions écrites; l'un des répertoires contient les règles permettant d'interpréter les indications spectrographiques segmentales, l'autre est une stratégie formalisée ou un questionnaire qui sollicite des informations sur les traits prosodiques et segmentaux du stimulus.

A condition de pouvoir obtenir l'information requise, la sortie de la stratégie donnera des indications sur les aspects grammaticaux de


Instructions de segmentation
Evaluation des stimuli
Information phonétique partielle
Traits phonétiques - prosodie
Interprétation finale du spectrogramme
Recherche lexicale
Réponses
Stratégie
Information grammaticale
terminaisons
classe des mots
longueur du mot etc.
Questions
Indications pour la segmentation
Information prosodique
Degré de l'accentuation
Accent tonal
Spectrogramme à large bande

- Figure 1 -

l'expression, tels que la forme phonétique des désinences et des racines la classe du mot, le nombre de syllabes par mot etc... L'interprétation des correspondants spectrographiques des voyelles et des consonnes est parfois très difficile. Bien souvent n'est possible qu'une caractérisation partielle des traits segmentaux. Cependant dans la présente expérience, quand un sujet essaie d'identifier un mot du lexique il cherche à l'intérieur de son propre lexique en se basant non seulement sur les traits segmentaux identifiés mais encore sur l'information prosodique et grammaticale. La réponse finale en découlera.

Examinons plus en détail maintenant les facteurs à contrôler dans une expérience de ce type. Considérons d'abord l'ensemble des phrases. Dans les expériences on a essayé de contrôler la structure grammaticale et phonétique des stimuli. Dans l'ensemble de ces stimuli doit obligatoirement figurer toute consonne qui en Suédois est susceptible de commencer un mot devant une voyelle accentuée, et également toute voyelle en position accentuée, qu'elle soit longue ou brève. Les phrases sont prononcées naturellement mais plutôt distinctement.

La grammaire :

Afin de contrôler la structure grammaticale des phrases de tests on a choisi un sous-ensemble de la grammaire du Suédois. La grammaire de la figure 2 engendre un nombre fini de phrases Suédoises déclaratives et de longueur finie puisque ne peut y figurer aucune procédure récursive. Cependant, le nombre des différentes suites d'unités syntaxiques est très grand et est pratiquement de l'ordre de plusieurs centaines.

De plus conjointement à cette grammaire nous utiliserons un lexique qui n'est pas un ensemble fermé de mots mais seulement un ensemble de mots authentiquement Suédois. Ceci implique que tous les radicaux ont moins de 9 syllabes et que l'accent principal est toujours sur la syllabe initiale de la racine. Par conséquent, en utilisant ces règles on obtient un ensemble fini mais très grand de phrases.

Les Instructions de segmentation :

Les sujets humains ont été choisis parmi un groupe ayant suivi un cours de lecture de spectrogrammes à l'Université de Stockolm pendant le premier semestre de l'année scolaire 1971-72. Pour rendre plus aisée l'identification des segments phonétiques et pour résumer les procédures utilisées pendant le cours, on leur avait donné des instructions sous forme écrite.

Un fragment de la grammaire du Suédois

$$S \quad \left\{ \begin{array}{l} NP_A + \left\{ \begin{array}{ll} Aux & + X \\ V & + Y \end{array} \right\} \\ TA_1 + \left\{ \begin{array}{ll} Aux & + NP_A + X \\ V & + NP_A + Y \end{array} \right\} \end{array} \right\} (+ \text{Prep} + \text{NP})$$

$$X \quad (TA_2 +) \; Aux_0^3 + V(+\text{Part})(+\text{NP})$$

$$Y \quad \left\{ \begin{array}{l} (\text{Pron})(+TA_2) \\ (TA_2+) \quad \left\{ \begin{array}{l} NP_B \\ \text{Part} + \text{NP} \end{array} \right\} \end{array} \right\}$$

$$NP \quad \left\{ \begin{array}{l} \text{Art}(+\text{Adj}) + \text{N} \\ \text{Pron} \\ \text{Npr} \end{array} \right\} \quad (+\text{gen}(+\text{Adj})+\text{N})$$

FIG. 2

Dans ces instructions on trouvait des informations sur la place et la manière d'articulation avec par exemple des indications pour séparer les latérales des nasales, pour identifier les fricatives, etc...
Dans ces instructions on trouvait également des indications sur les fréquences des formants et sur le locus pour les combinaisons du type CV, tels qu'on les observait chez le parleur en question. Ces données avaient été extraites d'une étude de Fant.

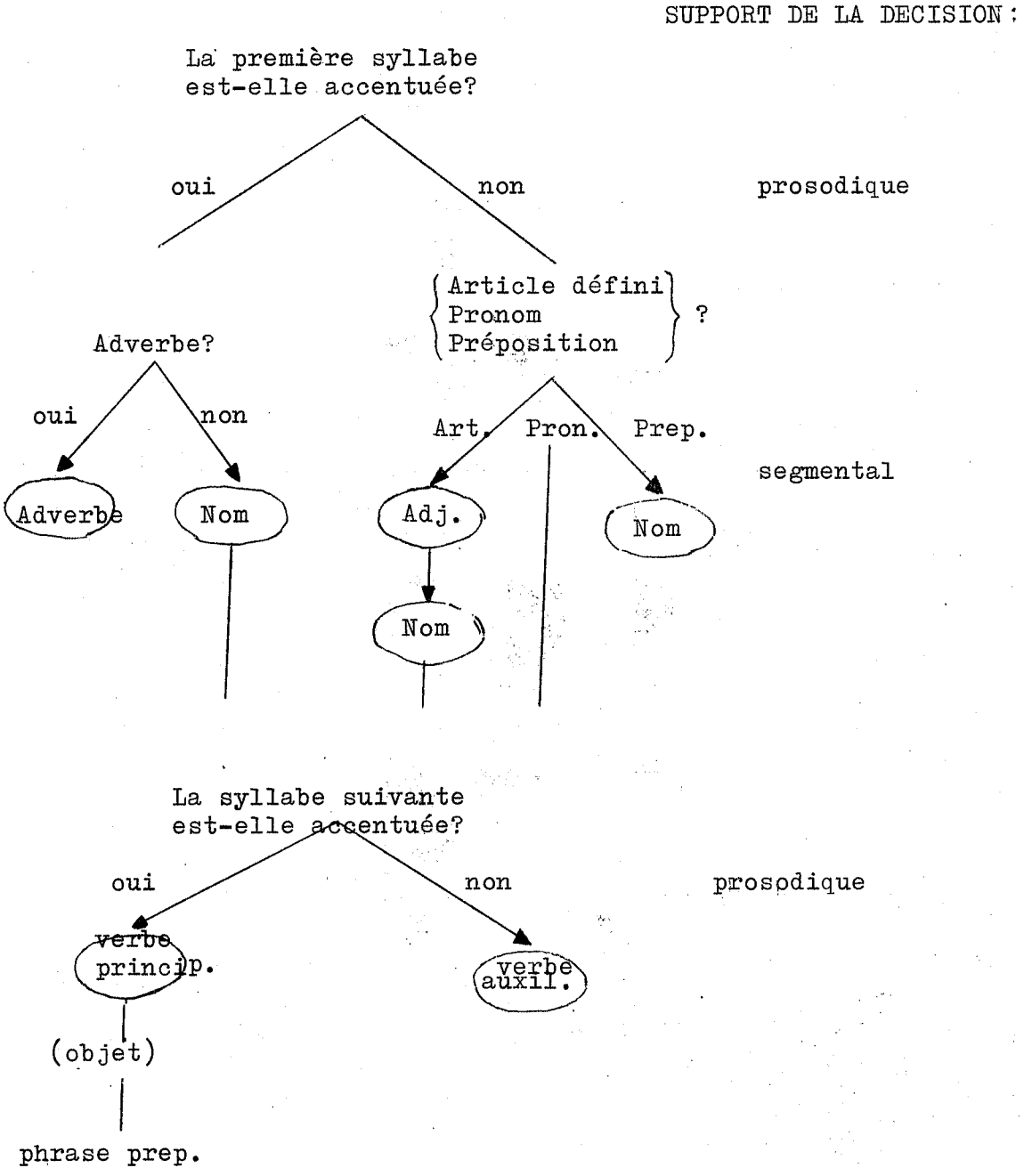
La stratégie :

Nous ferons maintenant quelques commentaires sur la stratégie utilisée pour l'analyse grammaticale et naturellement construite à partir de la grammaire précédemment décrite. Un aperçu général de cette stratégie est indiqué dans la figure 3. Elle est constituée d'un grand nombre de questions qu'on ordonne d'une manière telle qu'on obtient un organigramme semblable à celui d'un programme d'ordinateur. Quelques unes des questions concernent la prosodie, d'autres les segments phonétiques. De cette façon, la stratégie détermine le type de décision à prendre pour le stimulus à chaque point donné de la procédure. Comme on peut le remarquer sur la figure 3, l'ensemble des réponses possibles à donner à une question est très réduit pour la plupart des points où l'on doit prendre une décision.

Ceci est en partie une conséquence des contraintes imposées à la grammaire mais aussi et pour une part plus importante un produit de l'efficacité des traits prosodiques et de certains traits phonétiques qui permettent de réduire le nombre d'alternatives; qu'il soit possible de bâtir une stratégie révélant cette propriété est un fait intéressant.

Pour montrer brièvement comment fonctionne la stratégie nous donnerons un exemple. La figure 4 représente un stimulus typique : le spectre à large bande avec comme indication supplémentaire les accents toniques et d'intensité. La première question de la stratégie est la suivante : "La première syllabe est-elle accentuée ?" Puisque la réponse dans le cas présent est OUI, nous passerons à la question suivante : "Le premier mot est-il un adverbe ?" Les adverbes qui peuvent commencer une expression suédoise constituent un ensemble limité de mots et sont indiqués sur une liste. En regardant le spectrogramme le sujet peut alors conclure pour des raisons évidentes de segmentation qu'il ne peut s'agir d'une adresse de début de phrase. On demande alors au sujet de déterminer l'accent tonique qui dans ce cas est l'accent appelé accent 1. Ceci nous conduit à un noeud où le choix concerne la désinence d'un nom pour laquelle 3 possibilités nous sont offertes : [-en], [-et] ou [-er]; c'est-à-dire la fin de la désinence est-elle une nasale, une plosive sourde ou une fricative [r] ?

SCHEMA DE LA STRATEGIE
BASEE SUR LA PROSODIE



-Figure 3 -

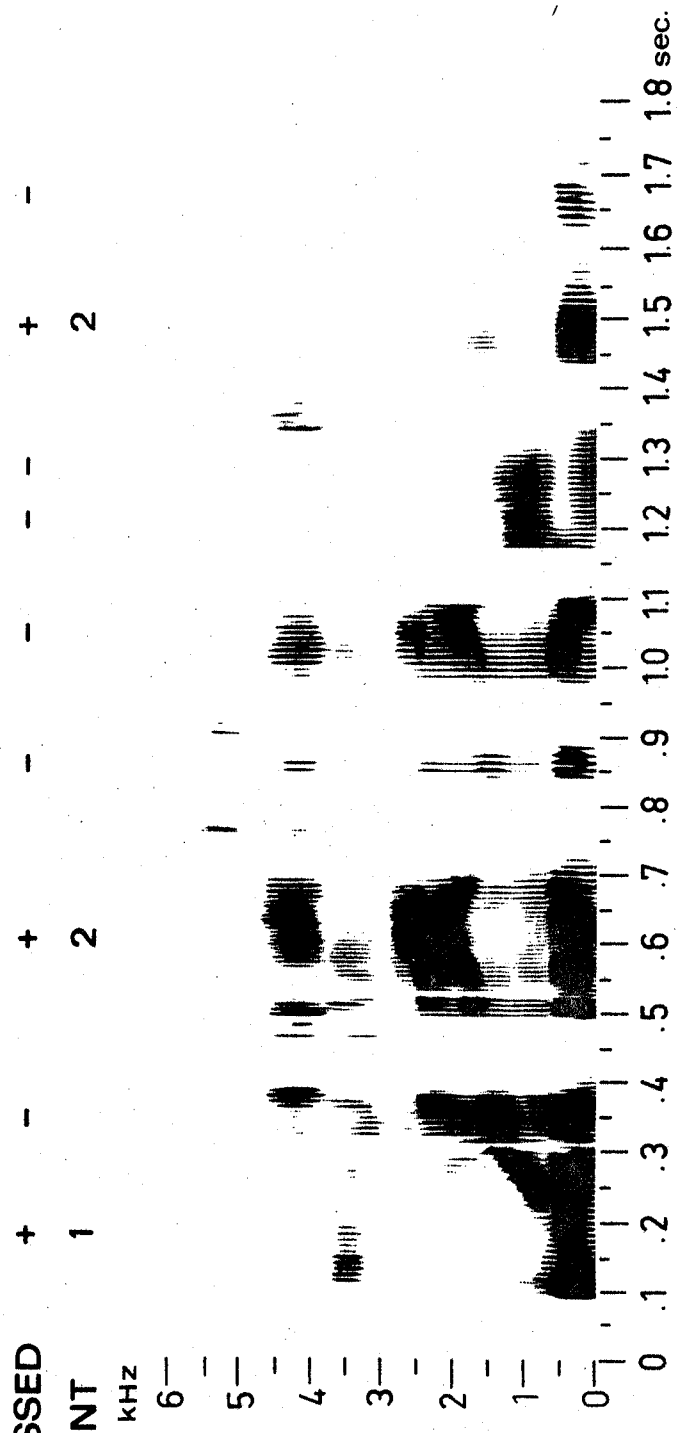

STRESSED + − + − − + −
ACCENT 1 2 2
kHz
6
5
4
3
2
1
0
0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 sec.

Figure 4

Puisque nous choisissons la syllabe qui se termine par une plosive, la stratégie maintenant nous indique clairement que le premier mot est un nom du genre neutre et dans sa forme définie. L'analyste aidé par cette stratégie procède de cette manière et retrouve la structure grammaticale de la phrase toute entière. On remarquera que dans cette stratégie, les informations d'entrée prosodiques et segmentales alternent de façon systématique.

La recherche lexicale :

Pour l'étage que nous avons appelé recherche lexicale dans la figure 1, nous n'avons pas formulé de stratégie explicite même si cela semble possible, en particulier si nous disposons de l'information sur les désinences grammaticales, les classes de mots et sur la prosodie du lexème recherché. Le sujet accomplit cette recherche sans instructions spéciales mais à l'aide de toute l'information qu'il a pu extraire lors des étapes précédentes.

Résultats :

Dans un premier stade de l'évaluation de ce modèle de reconnaissance nous avons décidé de prendre comme sujet un des auteurs de cet article, ceci afin d'assurer une bonne familiarité avec la stratégie et la lecture de spectrogrammes. Les résultats de 9 sessions sont indiqués dans la Table 1. Cette table montre que sur le total des 237 segments de 9 phrases de tests, seuls 2 ont été incorrectement identifiés. Cette expérience malgré tout ne pouvait servir de base pour montrer les contributions correspondant aux différentes informations spectrographiques, prosodiques et grammaticales. Pour cette raison on a entrepris une deuxième expérience plus générale. Pour ce test il y avait deux groupes. Un groupe avait à sa disposition 9 spectrogrammes avec comme indications supplémentaires les traits prosodiques. Ils avaient également accès à la stratégie basée sur la prosodie.

L'autre groupe utilisé comme groupe témoin avait accès aux mêmes spectrogrammes mais pas aux indications de prosodie ni à la stratégie. Cette situation est illustrée dans la figure 5. Aux deux groupes on donnait les mêmes instructions de segmentation. Pour s'assurer que ces deux groupes étaient équivalents en ce qui concerne les traits segmentaux on a réalisé un test d'étalonnage avant le test principal. Lors des sessions relatives à ce premier test on présentait aussi bien aux sujets du groupe témoin qu'à ceux de l'autre groupe 9 spectrogrammes de "pseudo-phrases". Ces phrases avaient des formes prosodiques et des structures phonotactiques identiques aux

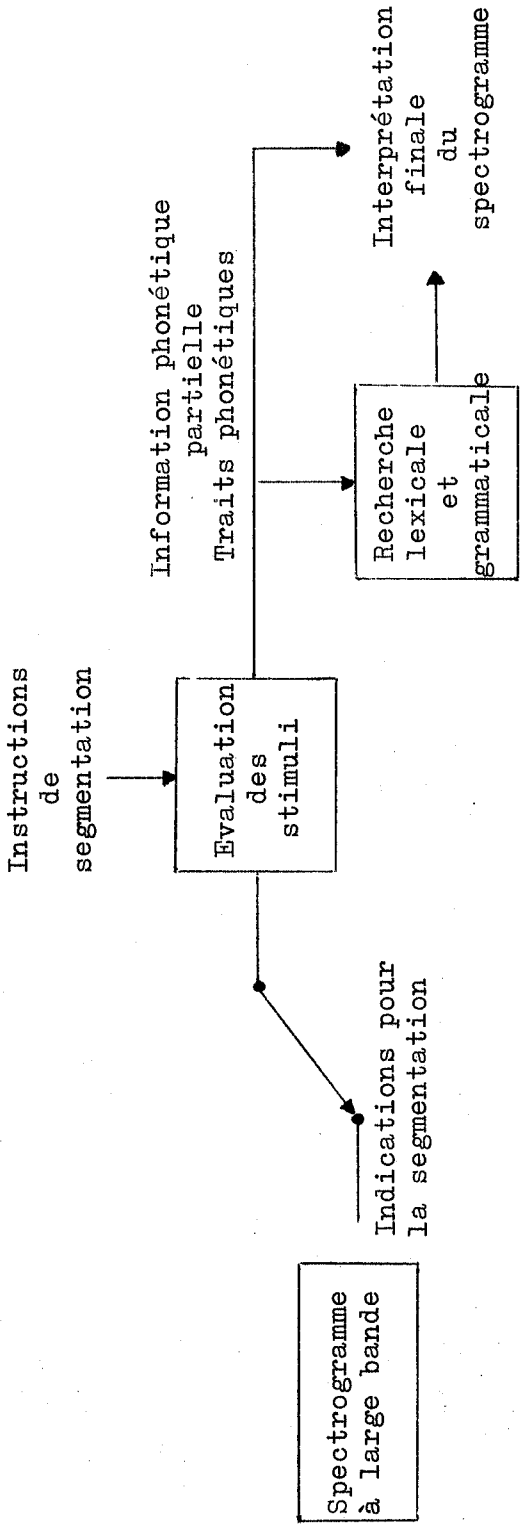
RESULTATS DE LA LECTURE DE SPECTROGRAMMES

A L'AIDE DE LA STRATEGIE (UN SUJET)

	CORRECTEMENT RECONNUS	INCORRECTEMENT RECONNUS	TOTAL	POURCENTAGE DE RECONNAISSANCE CORRECTE
PHRASES	7	2	9	78
SEGMENTS PHONETIQUES	235	2	237	78
FRONTIERES DE MOTS	51	0	51	100
CLASSES DE MOTS	60	0	60	100
DESINENCES	26	1	27	96
LEXEMES	34	1	35	97

TEMPS MOYEN DE LECTURE D'UN SPECTROGRAMME : 35 MINUTES

- TABLE 1 -


Instructions de segmentation
Evaluation des stimuli
Information phonétique partielle
Traits phonétiques
Recherche lexicale et grammaticale
Interprétation finale du spectrogramme
Indications pour la segmentation
Spectrogramme à large bande

- Figure 5 -

phrases Suédoises douées de sens et étaient composées d'un grand éventail de syllabes identique à celui du corpus de phrases sémantiques qui devait suivre. Les phrases étaient grammaticalement anormales parce que l'ordre des syllabes avait été modifié. On trouvera l'illustration de quelques résultats préliminaires dans la figure 6.

Ce diagramme permet de faire une comparaison entre les pourcentages de segments phonétiques correctement reconnus par deux sujets. On peut s'apercevoir que les résultats sont semblables dans le test 1, c'est-à-dire dans le test d'étalonnage. Les stimuli étaient des pseudo-phrases. Seules des instructions de segmentation identique pour les deux sujets, étaient données. Pour les phrases grammaticales, les résultats indiqués dans la partie droite du diagramme montrent que le sujet usant de la stratégie obtenait de moins bons résultats au départ, mais d'un autre côté il obtenait un pourcentage de reconnaissance de 100% lors des 3 dernières sessions. Les résultats du sujet témoin restaient par contre stables aussi bien dans le test 1 que dans le test 2.

Conclusions :

Même si les résultats présentés ne sont que des résultats préliminaires nous essaierons de conclure de la façon suivante :

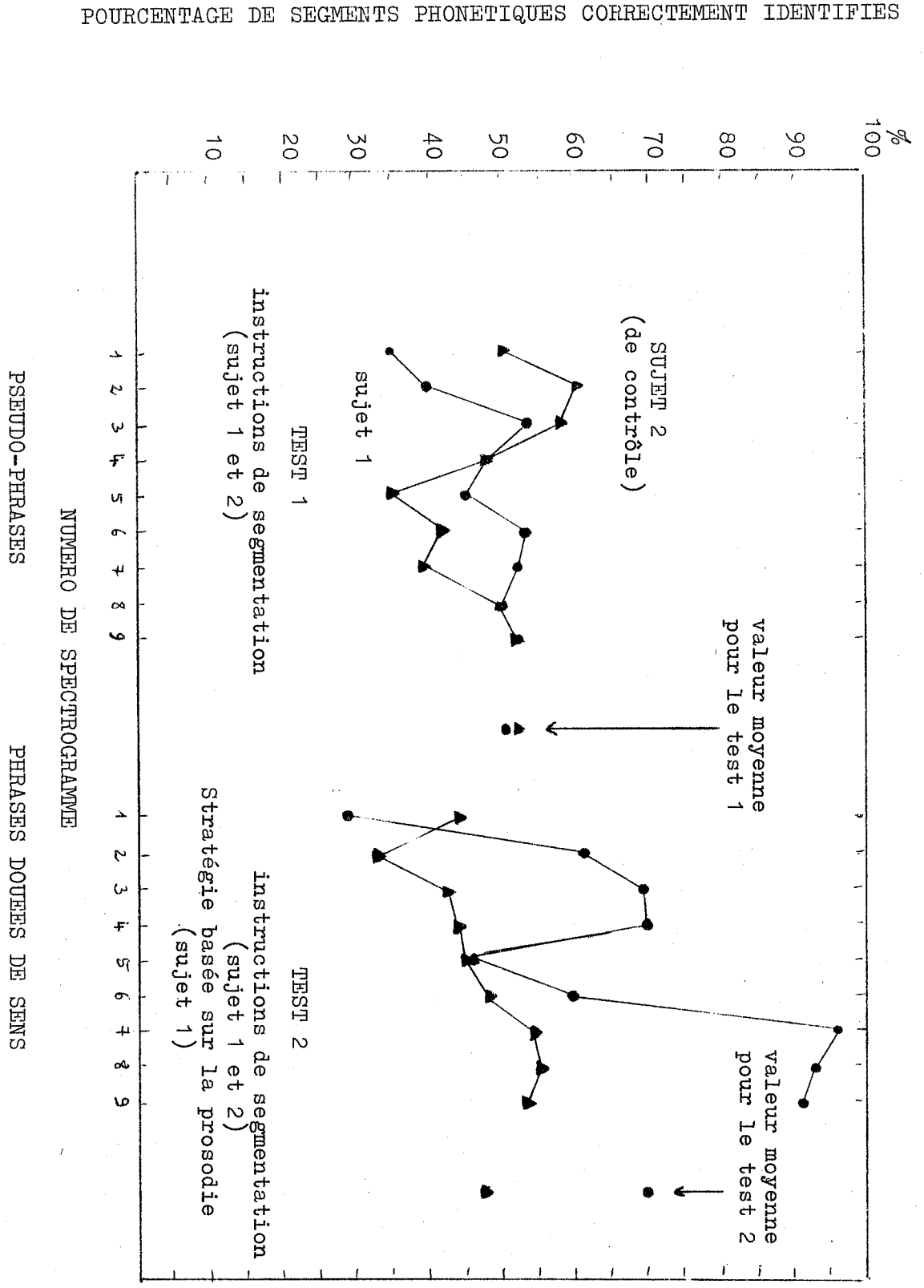
D'abord il est clair que les indications segmentales présentes dans les stimuli spectrographiques se sont avérées suffisantes pour obtenir des scores de reconnaissance élevés dans le contexte expérimental actuel.

D'autre part c'est sans risque semble-t-il qu'on peut attribuer les résultats positifs obtenus à l'introduction de l'information prosodique et à la stratégie de recherche grammaticale dans le processus d'identification. On peut espérer encore une amélioration des résultats en introduisant une procédure automatique de recherche lexicale.

A partir de ces résultats on peut donc faire les recommandations suivantes pour la recherche à venir : par introduction dans les modèles de reconnaissance en plus de l'analyse des traits segmentaux, de l'analyse prosodique et grammaticale automatique on peut probablement améliorer les résultats obtenus même avec des méthodes plutôt grossières pour l'analyse des segments phonétiques. Nos remarques s'appliquent d'abord au Suédois naturellement, mais nous pensons qu'il n'existe pas d'obstacle fondamental pour empêcher le raisonnement précédent de s'appliquer aux autres langues.

Remerciements :

Ce travail a été financé par l'Institut National de la Santé grâce au contrat de Recherche NS 04003-09.


POURCENTAGE DE SEGMENTS PHONETIQUES CORRECTEMENT IDENTIFIES
%
100
90
80
70
60
50
40
30
20
10
SUJET 2
(de contrôle)
sujet 1
valeur moyenne
pour le test 1
valeur moyenne
pour le test 2
TEST 1
instructions de segmentation
(sujet 1 et 2)
TEST 2
instructions de segmentation
(sujet 1 et 2)
Stratégie basée sur la prosodie
(sujet 1)
1 2 3 4 5 6 7 8 9
1 2 3 4 5 6 7 8 9
NUMERO DE SPECTROGRAMME
PSEUDO-PHRASES
PHRASES DOUEES DE SENS

- Figure 6 -

DETERMINATION DE LA FONCTION D'AIRE DU CONDUIT VOCAL

D E T E R M I N A T I O N D E L A

FONCTION D'AIRE DU CONDUIT VOCAL *

Bernard G U E R I N

Laboratoire de la Communication Parlée et de l'Instrumentation de Mesures
E.N.S. d'ELECTRONIQUE et de RADIOELECTRICITE - 23 rue des Martyrs - 38 . GRENOBLE

* Etude entreprise avec l'appui du Comité de Recherches en Informatique.

1. INTRODUCTION

Ces dernières années, les recherches dans le domaine de la parole s'orientent de plus en plus vers une analyse et une synthèse du type articulatoire. Ces recherches sont effectives, en particulier, dans trois directions : acoustique, anatomique et neurophysiologique. Au niveau acoustique, on cherche principalement à étudier les relations entre le signal image de la parole et la configuration du système phonatoire. Au niveau anatomique, on veut décrire les mouvements articulatoires responsables des changements de configuration du conduit vocal en fonction du temps. Enfin, la détermination des règles de coordination de l'articulation, qui pourrait être celles du cerveau, forme le dernier domaine de ces recherches.

La synthèse en terme articulatoire est effectuée en commandant un analogue dynamique du conduit vocal, à partir d'une connaissance de la fonction d'aire du conduit vocal. Cette fonction d'aire est une donnée essentielle dans le cadre de toutes ces études. Par exemple, dans l'optique d'une synthèse par règles, on recherche, grâce aux fonctions d'aires et aux mesures de déplacement des articulations, à déterminer les contraintes et aboutir à un modèle de l'articulation. On peut également espérer faire une reconnaissance de la parole à ce niveau.

Pour déterminer ces fonctions d'aire, plusieurs techniques sont utilisées. La plus ancienne est la cinéradiographie aux rayons X du conduit vocal humain. Si elle a de graves inconvénients, elle a permis de faire de nombreuses et fructueuses études et elle est encore très employée. Mais, ces cinq dernières années, on s'est efforcé de déterminer ces fonctions d'aire de manière indirecte, par le calcul. Les données de base, les méthodes, sont nombreuses et variées. On utilise par exemple les pôles et les zéros de l'impédance aux lèvres du conduit vocal pendant la phonation (SCHROEDER, ZUE, SONDHI et GOPINATH [1,2,3]), le calcul des perturbations en partant d'une fonction d'aire uniforme ou quelconque (MERMELSTEIN, HEINZ [4,5]), la réponse impulsionnelle du conduit vocal (SONDHI et GOPINATH [6]) et enfin une méthode calcul à partir des valeurs des trois premiers formants (KADOKAWA et SUZUKI [7]). Toutes ces méthodes font intervenir un développement mathématique plus ou moins complexe et des contraintes dont le choix est souvent difficile.

Quelques-unes de ces méthodes d'obtention de la fonction d'aire, ainsi que d'autres en cours d'études, seront décrites dans la suite. Nous allons rappeler, en quelques mots, les résultats donnés par la cinéradiographie avant de décrire le calcul de la fonction d'aire à partir des trois premiers formants.

2. DETERMINATION DE LA FORME DU CONDUIT VOCAL PAR CINERADIOGRAPHIE

Pendant ces vingt dernières années, la cinéradiographie à rayons X était le seul moyen utilisé pour déterminer la fonction d'aire du conduit vocal. Cette technique consiste à prendre une série de photographies aux rayons X de l'appareil phonatoire d'un patient, dans un plan parallèle au plan médian dudit conduit et à extrapoler de ces mesures planes l'aire de la section en chacun des points. Pour cela, il faut donc faire une estimation de la forme de la section du conduit vocal en chacun de ses points.

Cette technique souffre de deux inconvénients principaux :
- afin de ne pas exposer le sujet à une dose de rayons X supérieure à celle de sécurité, seul un petit nombre de mesures peuvent être faites sur le même patient,
- l'interprétation des photographies est un art difficile et complexe, et on est alors obligé de faire un certain nombre de suppositions pour en déduire l'aire.

L'exploitation immédiate des mesures est donc impossible et cette méthode ne peut

être utilisé pour une transmission en temps réel. Ceci explique que l'on cherche actuellement à traiter directement le signal acoustique ou les propriétés acoustiques du conduit vocal (impédance, réponse impulsionnelle, etc...), afin d'en déduire le plus rapidement possible et avec le maximum de précision les fonctions d'aire.

Voyons maintenant une de ces méthodes de calcul. Nos données seront les trois premiers formants.

3. DETERMINATION DE LA FONCTION D'AIRE A PARTIR DES FORMANTS

D'après KADOKAWA et SUZUKI [7]

3.1. *Introduction - Contraintes*

On sait résoudre l'équation de propagation à une dimension du son dans le conduit vocal avec des conditions aux limites données, et on peut en déduire les valeurs de fréquence de formant. On peut également obtenir ce résultat en considérant le circuit équivalent formé d'une succession de tubes cylindriques élémentaires.

Cependant, le problème inverse est plus délicat. Pour obtenir une solution unique de la fonction d'aire, il faut connaître par exemple les contraintes physiologiques appropriées ou l'ensemble des formants et antiformants. Comme l'on sait extraire avec précision les trois premiers formants (méthode d'analyse par synthèse), en tenant compte de quelques contraintes, on peut espérer trouver la fonction d'aire.

Négligeant les pertes dans le conduit vocal, le conduit nasal, et en utilisant l'analogie du circuit électrique équivalent pour le calcul, le traitement mathématique de l'équation d'onde sera simplifié (figure 1). On peut alors établir les relations reliant la pression et la vitesse volumique d'entrée et de sortie de la iième section. On obtient les matrices suivantes :

$$\begin{pmatrix} P_i \\ U_i \end{pmatrix} = K_i \begin{pmatrix} P_{i+1} \\ U_{i+1} \end{pmatrix} = \begin{cases} \begin{pmatrix} \cos\theta_N & jZ_i \sin\theta_N \\ jY_i \sin\theta_N & \cos\theta_N \end{pmatrix} \begin{pmatrix} P_{i+1} \\ U_{i+1} \end{pmatrix} & \text{pour } \cos\theta_N = 0 \\ \cos\theta_N \begin{pmatrix} 1 & jZ_i \operatorname{tg}\theta_N \\ jY_i \operatorname{tg}\theta_N & 1 \end{pmatrix} \begin{pmatrix} P_{i+1} \\ U_{i+1} \end{pmatrix} & \text{pour } \cos\theta_N \neq 0 \end{cases}$$

où $\theta_N = \omega \ell_N / c_0 = \frac{\omega}{c_0} \frac{\ell_t}{N}$,

$Z_i = \rho c_0 / A_i$: impédance acoustique de la $i^{ième}$ section ,

$Y_i = A_i / \rho c_0$: admittance acoustique de la $j^{ième}$ section ,

A_i : aire de la section i ,
ρ : densité de l'air ,
c_0 : vitesse du son dans l'air ,
ℓ_t : longueur totale du conduit vocal ,
N : nombre total de sections.

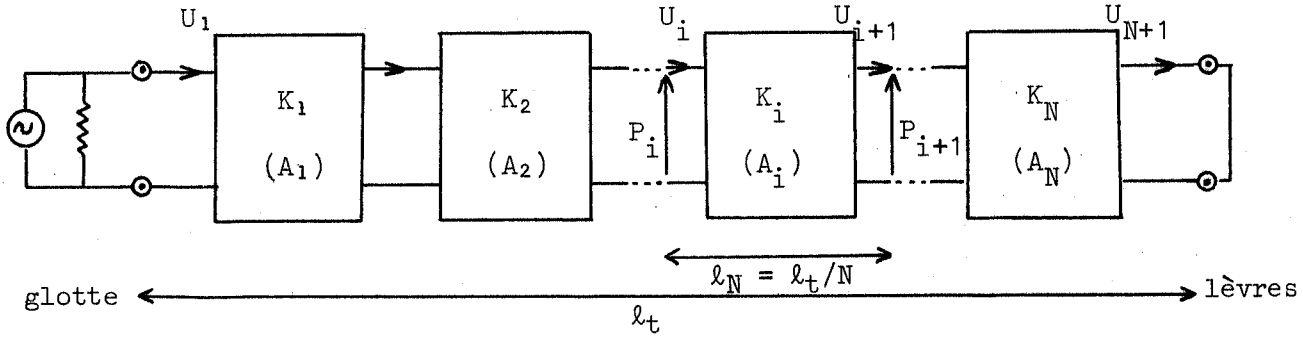
La matrice de transmission du conduit vocal ainsi composée est obtenue en faisant le produit des matrices mises en cascade de la glotte aux lèvres.

On obtient alors :

$$K^N = \prod_{i=1}^{N} \begin{pmatrix} A_i & B_i \\ C_i & D_i \end{pmatrix} = \begin{cases} \begin{pmatrix} A^N & B^N \\ C^N & D^N \end{pmatrix} = \begin{pmatrix} A^{N-1} & B^{N-1} \\ C^{N-1} & D^{N-1} \end{pmatrix} \begin{pmatrix} A_N & B_N \\ C_N & D_N \end{pmatrix} & \text{pour: } \cos\theta_N = 0 \quad N \geq 2 \\ \cos^{(N-1)}\theta_N \begin{pmatrix} a^N & b^N \\ c^N & d^N \end{pmatrix} = \cos^{(N-1)}\theta_N \begin{pmatrix} a^{N-1} & b^{N-1} \\ c^{N-1} & d^{N-1} \end{pmatrix} \begin{pmatrix} a_N & b_N \\ c_N & d_N \end{pmatrix} & \text{pour: } \cos\theta_N \neq 0 \quad N \geq 2 \end{cases}$$

où $A_N = D_N = \cos\theta_N$; $a_N = d_N = 1$.

FIGURE 1 - Chaîne de matrice équivalente au conduit vocal sans perte et sans impédance de rayonnement pour des sons voisés



Etudions les deux cas selon que $\cos\theta_N$ est nul ou non.

3.1.1. $\cos\theta_N = 0$; $N \geq 2$
En effectuant le produit, on trouve $D^N = d^N \cos^N\theta_N$, et la fréquence de coupure passe-bas de la $i^{ième}$ section sera : $f_c = c_0 N/4\ell_t = 1$; 1,5 ; 2 ... kHz pour N = 2, 3, 4, ...

3.1.2. $\cos_N \neq 0$; $N \geq 2$
On peut développer la matrice de transfert sous la forme :

$$\begin{pmatrix} a^N & b^N \\ c^N & d^N \end{pmatrix} = \begin{pmatrix} a^{N-1} & b^{N-1} \\ c^{N-1} & d^{N-1} \end{pmatrix} \begin{pmatrix} a_N & b_N \\ c_N & d_N \end{pmatrix} = \begin{pmatrix} a^{N-1} + b^{N-1} a_N \mathrm{tg}\theta_N & b^{N-1} + a^{N-1} \bar{a}_N \mathrm{tg}\theta_N \\ c^{N-1} + d^{N-1} a_N \mathrm{tg}\theta_N & d^{N-1} + c^{N-1} \bar{a}_N \mathrm{tg}\theta_N \end{pmatrix} \text{ où :}$$

$$a_N = jY_N = j\frac{A_N}{\rho c_0} ; \bar{a}_N = jZ_N = \frac{j\rho c_0}{A_N} ; a_N\bar{a}_N = -1 ; a_M\bar{a}_N = -A_{MN} = a_{MN} = -\frac{A_M}{A_N} ; \bar{a}_M a_N = a_{NM} = -A_{NM} = -\frac{A_N}{A_M} .$$

Le cas particulier où $A_{MN} = 1$ correspond à une configuration de conduit vocal uniforme. Les fréquences de formants seront alors de 500(2n-1) Hz avec n = 1, 2, 3, ...

Etudions le cas intéressant où $A_{MN} \neq 1$. Dans ce cas, la fonction de transfert de la voyelle est donnée par U_{N+1}/U_1, en négligeant l'impédance de rayonnement. Dans ces conditions, les fréquences de formant vérifient l'équation :

$$d^N = d^{N-1} + c^{N-1} a_N \mathrm{tg}\theta_N = 0 .$$

Pour un conduit vocal divisé en un nombre donné de sections : N, on peut déterminer la valeur de l'expression du premier membre en fonction des différents a_i. Comme nous disposons de trois fréquences de formant, nous aurons un système à trois équations simultanées du type ci-dessus à résoudre ; soit :

$$\begin{cases} d_1^N = d^{N-1} + c^{N-1}\, \bar{a}_N\, tg\theta_{N_1} = 0 & \text{pour le premier formant ;} \\ d_2^N = d^{N-1} + c^{N-1}\, \bar{a}_N\, tg\theta_{N_2} = 0 & \text{pour le deuxième formant ;} \\ d_3^N = d^{N-1} + c^{N-1}\, \bar{a}_N\, tg\theta_{N_3} = 0 & \text{pour le troisième formant.} \end{cases}$$

En développant ces équations pour N = 2, 3, 4, ..., on remarque que l'on fait alors intervenir la variable $x_{Nn} = tg^2\theta_{Nn}$. De l'étude des différentes équations obtenues pour N = 2, 3, 4, ..., on voit immédiatement que pour N trop faible, la connaissance de la fréquence d'un ou deux formants détermine celle des autres. D'autre part, comme on suppose qu'il n'y a aucun rayonnement aux lèvres, les aires des sections ne sont connues qu'à un coefficient multiplicatif près.

3.2. *Fonction d'aire avec* N = 6

D'après le système d'équation que l'on a obtenu, on doit pouvoir trouver une solution unique si $N \leq 7$ car, dans ce cas, on a trois équations à résoudre de la forme :

$$1 - \alpha_{61}x_6 + \alpha_{62}x_6^2 - \alpha_{63}x_6^3 = 0 \ ,$$

α_{6n} étant une fonction de A_{NM}. Nous prendrons le cas N = 6 (qui donne l'équation ci-dessus). On peut alors en déduire les équations suivantes :

$$\begin{cases} p\{1+q[1+r(1+s(1+t))]\} + q\{1+r[1+s(1+t)]\} + r[1+s(1+t)] + s(1+t) + t = \beta_1 \\ p\{r[1+s(1+t)]+s(1+t)+t\} + pq[s(1+t)+t] + q[s(1+t)+t] + t(pqr+qr+r) = \beta_2 \\ prt = \beta_3 \end{cases}$$

où
$$\begin{cases} p = A_1/A_2 \ ; \ q = A_2/A_3 \ ; \ r = A_3/A_4 \ ; \ s = A_4/A_5 \ ; \ t = A_5/A_6 \\ \beta_1 = \dfrac{x_1x_2 + x_1x_3 + x_2x_3}{x_1x_2x_3} \ ; \ \beta_2 = \dfrac{x_1 + x_2 + x_3}{x_1x_2x_3} \ ; \ \beta_3 = \dfrac{1}{x_1x_2x_3} \\ x_n = tg^2\theta_n = tg^2\, \dfrac{2\pi F n_0 \ell_t}{6 c_0} \quad \text{pour } n = 1, 2, 3. \end{cases}$$

Nous avons donc maintenant un système à trois équations, avec trois données : β_1, β_2 et β_3, mais avec 5 inconnues. A cette étape du calcul, il faut introduire des contraintes car on sait que les rapports des aires des sections ne sont pas complètement indépendants entre eux. Ces contraintes ont été formulées au moyen de la méthode des perturbations par SCHROEDER et MERMELSTEIN [1,4] et étendues par HEINZ [5]. Ils en ont déduit que l'on pouvait décrire de façon unique la fonction d'aire en prenant l'exponentielle de la somme de fonctions sinusoïdales pondérées lorsque les trois premiers formant et antiformants sont connus. Quand on applique ces contraintes aux cas où l'on ne connaît que les trois premiers formants, on peut obtenir les résultats suivants : $p \simeq t$ et : $q \simeq s$.

Avec ces nouvelles données, on trouve, après quelques arrangements :

$$\begin{cases} p^4\left(-4\beta_2-8\beta_3+(\beta_1-\beta_3)^2\right) + 8p^3\beta_3\left(1+\beta_1+\beta_2+\beta_3\right) - 2p^2\beta_3\left(4\beta_2+\beta_1\beta_2+\beta_1+\beta_3(4\beta_1+\beta_2+9)\right) + \\ \qquad + 8p\beta_3^2\left(1+\beta_2+\beta_3+\beta_1\right) + \beta_3^2\left(-4\beta_3(2+\beta_1)+(\beta_2-1)^2\right) = 0 \\ q = \dfrac{2p^4 - (\beta_1+\beta_3)p^3 + \beta_3(\beta_2+1)p - 2\beta_3^2}{2(\beta_3-p)(p^2+\beta_3)(1+p)} \\ r = \dfrac{\beta_3}{p^2} \end{cases}$$

si $t = p \neq \beta_3$. La première équation du quatrième ordre, nous donnera la valeur de $p = A_1/A_2$ et les deux autres valeurs de q et r s'obtiennent alors directement par les expressions données.

Le cas où $t = p = \beta_3$ donne :

$$\begin{cases} p = \beta_3 \\ q = \dfrac{\sqrt{\beta_3(2+\beta_1) - \beta_3^2}}{1+\beta_3} - 1 \\ r = 1/\beta_3 \end{cases}$$

On considère que le conduit vocal n'a jamais une configuration correspondant à ce cas.

En résumé, pour déterminer la fonction d'aire, on calculera d'abord la valeur des coefficients β_1, β_2 et β_3 à partir de la valeur des fréquences de formant obtenue après correction Fn_0 (correction tenant compte de l'impédance de rayonnement [8]). Ensuite, on cherchera une solution positive à l'équation du quatrième ordre en p donnant également une valeur positive à q. On pourra alors en déduire les cinq rapports d'aire.

3.3. *Lissage de la fonction obtenue*

Pour approcher plus précisément la fonction d'aire de la réalité, la définition à 6 sections n'étant pas suffisante, on va utiliser comme formulation de la fonction d'aire l'exponentielle d'une somme de fonctions sinusoïdales pondérées (d'après MERMELSTEIN et SCHROEDER). Avec les six sections, nous avons :

$A(x) = \exp \sum_{N=1}^{6} m_N \cos\left((N-1)\pi x/\ell_t\right)$ où x est la distance à la glotte.

Connaissant 6 valeurs de A(x), on peut en déduire les coefficients m_N et calculer la valeur des aires des sections d'un conduit vocal divisé en 18 tubes élémentaires par exemple. Elles seront données par :

$$A(x) = A\left(x = \frac{(2N-1)\ell_t}{36}\right) = \exp\left(\sum_{i=1}^{6} m_i \cos\left((i-1)(2N-1)\frac{\pi}{36}\right)\right) \quad \text{pour } N = 1, 2, \ldots 36.$$

Dans ce cas, la fréquence de coupure du tube acoustique est de 9 kHz.

3.4. *Application - Discussion*

Pour vérifier la validité de cette méthode, KADOKAWA et SUZUKI ont comparé leurs résultats de calcul à ceux obtenus à partir des rayons X par CHIBA et KAJIYAMA (cinq voyelles japonaises), FANT (cinq voyelles russes) et HEINZ (six voyelles anglaises). Il faut que les fréquences de formants de départ soient inférieures à 3 kHz, fréquence de coupure de chacun des 6 tubes élémentaires.

Le meilleur résultat est obtenu pour la voyelle neutre [e], la configuration se rapprochant beaucoup de celle du tube uniforme. Plus la constriction maximale est avancée et étroite, plus la différence est grande entre l'original et le modèle calculé. La valeur des fréquences de formant calculé d'après la fonction d'aire obtenue après lissage est en général plus élevée que celle de départ. Les différences entre la forme du conduit vocal calculé et celui obtenu par mesures aux rayons X sont plus importantes que les variations sur les valeurs des formants déterminés d'après la fonction d'aire calculée.

Les auteurs pensent que l'on pourrait améliorer les résultats en choisissant convenablement les coefficients k_1 et k_2 qui déterminent les relations $p = k_1 t$ et $q = k_2 s$. On pourrait ainsi obtenir des fonctions d'aire plus proches des données aux rayons X dans le cas de conduit très perturbé comme pour le [u].

L'application aux voyelles françaises a donné les résultats représentés figure 2. Pour les quatre voyelles [ɔ], [a], [e] et [ɛ], on a obtenu des fonctions d'aire assez

proche de celles données par les rayons X. En effectuant les calculs, on a remarqué que la valeur de la longueur du conduit vocal devait être précise.

En conclusion, on peut dire que cette méthode est simple à mettre en oeuvre, mais elle demande une connaissance précise des trois premiers formants ainsi que de la longueur du conduit vocal.

FIGURE 2 - Fonctions d'aires de quatre voyelles françaises ; en trait plein, la fonction d'aire obtenue par les rayons X et en pointillé, celle obtenue par le calcul.

[ɔ]

	1	2
F_{10}	574	532
F_{20}	950	1042
F_{30}	2610	2350

[a]

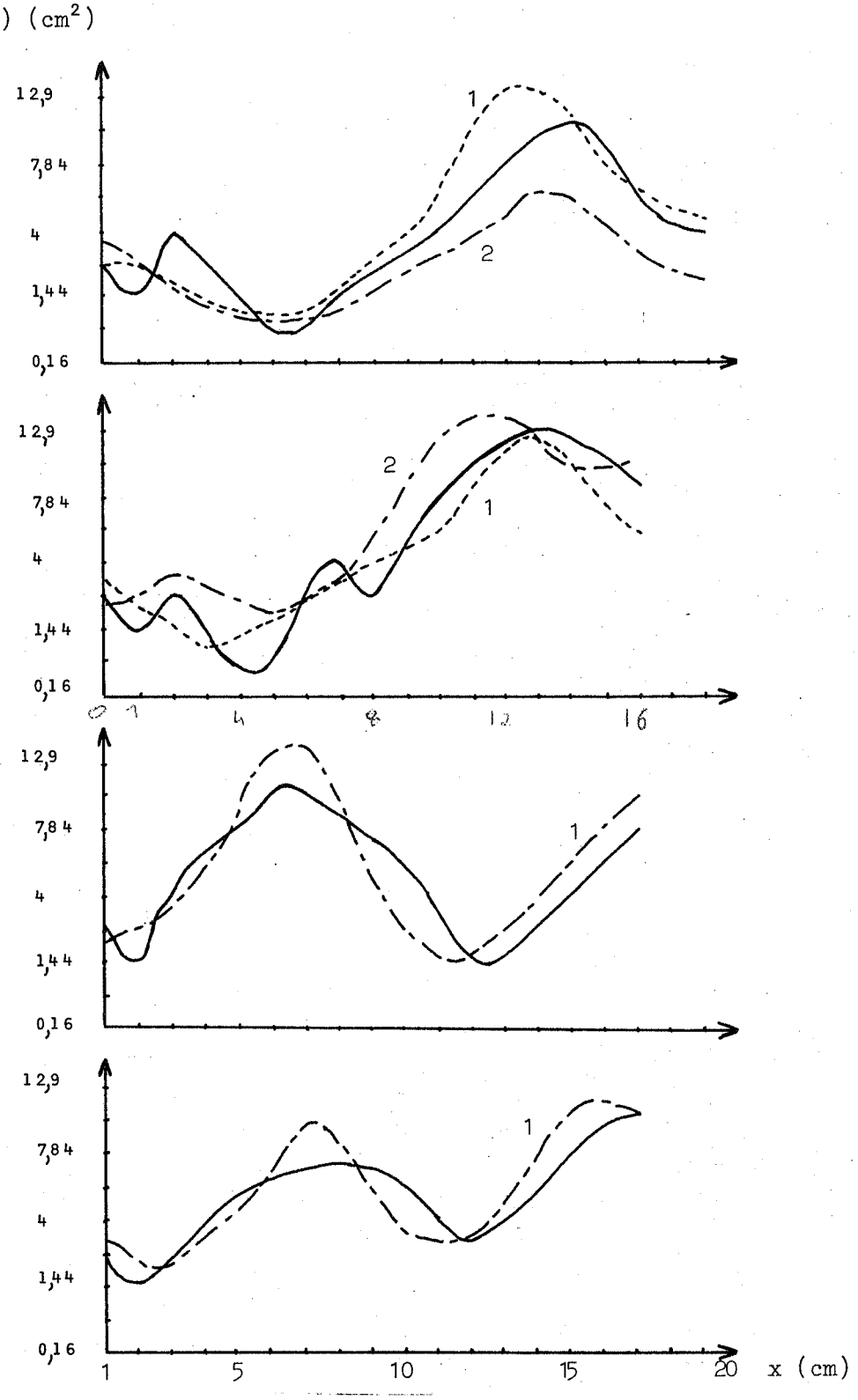
F_{10}	750	756
F_{20}	1320	1360
F_{30}	2500	2760

[e]

F_{10}	393
F_{20}	2250
F_{30}	2620

[ɛ]

F_{10}	575
F_{20}	1885
F_{30}	2300



BIBLIOGRAPHIE

[1] M.R. SCHROEDER
Détermination of the geometry of the human vocal tract by acoustic measurements
J.A.S.A., 41 n° 4, 1002-1010 (1967) ;

[2] V.W. ZUE, A. PAIGE
Computation of vocal tract area function
I.E.E.E. Trans. AU.18 n° 1, 7-18 (1970) ;

[3] B. GOPINATH, M. SONDHI
Determination of the shape of the human vocal tract from acoustical measurements
B.S.T.J., 49 n° 6, 1195-1214 (1970) ;

[4] P. MERMELSTEIN
Determination of the vocal tract shape from measured formant frequencies
J.A.S.A., 41 n° 5, 1283-1294 (1967) ;

[5] J. HEINZ
Perturbation functions for the determination of vocal tract area function from vocal tract eigenvalues
Q.P.S.T.R. Stockholm, 1-14 (april 1967) ;

[6] M. SONDHI, B. GOPINATH
Determination of vocal tract shape from impulse response at the lips
J.A.S.A., 49 n° 6, 1867-1873 (1970)

[7] Y. KADOKAWA, J. SUZUKI
Simple calculation of the vocal tract configuration from three formant frequencies
J. of the Radio Research Lab., 15 n° 79, 147-164 (1968) .

INTERVENTION

M. RISSET

Il y a une autre méthode numérique de synthèse par simulation du conduit vocal : résoudre par ordinateur l'équation de WEBSTER pour obtenir les pôles et utiliser les valeurs obtenues pour commander un synthétiseur à formants numérique.
Cette méthode détournée a été utilisée par COKER ; elle donne un temps de calcul réduit (rapport temps de calcul sur temps de parole voisin de 1 avec un mini-ordinateur).

FORME APPROCHEE DU CONDUIT VOCAL DEDUITE DES FREQUENCES DE RESONANCE

THEORIE DES PERTURBATIONS ET METHODE VARIATIONNELLE

P. JOSPA

Institut de Phonétique
Université de BRUXELLES

Pages

FORME APPROCHEE DU CONDUIT VOCAL DEDUITE DES FREQUENCES DE RESONANCE THEORIE DES PERTURBATIONS ET METHODE VARIATIONNELLE

Résumé.

Nous rappelons d'abord les relations entre fréquences de résonance et fonction d'aire du conduit vocal établies par Ungeheuer, Schroeder, Mermelstein et Heinz à l'aide de la théorie des perturbations. Une méthode est ensuite développée, permettant de calculer, dans le cadre d'un modèle articulatoire donné, la fonction d'aire à partir d'un nombre limité de fréquences de résonance. Cette méthode est basée sur une formulation variationnelle des lois de propagation de l'onde sonore dans le conduit vocal assimilé à un tube acoustique non uniforme et non absorbant. La procédure de Rayleigh-Ritz est utilisée pour obtenir des équations algébriques non-linéaires approximatives reliant paramètres articulatoires et fréquences propres.

Abstract.

First of all we are reminded of the results concerning the relation between resonance frequencies and vocal tract functions obtained by Ungeheuer, Schroeder, Mermelstein and Heinz from the use of the perturbation theory. A method is then described which allows the area function to be computed from a finite number of resonances within the frame of some given articulatory model. This method is based on the variational formulation of the sound wave propagation laws in the vocal tract modelled as a lossless nonuniform acoustic tube. The Rayleigh-Ritz procedure is used to derive the approximate non-linear algebraic equations which link articulatory parameters and eigenfrequencies.

I. INTRODUCTION

Depuis les travaux de Schroeder [1,3], Mermelstein [1,2] et Heinz [4] dans les années 1965-67, la question de la détermination, à partir des données acoustiques, des formes que prend le conduit vocal pendant la phonation fait l'objet d'un intérêt croissant.

L'équation de Webster [5,6] a souvent été utilisée, conjointement à certaines conditions aux limites, comme modèle approché reliant l'aire des sections droites du conduit vocal (fonction d'aire) aux fréquences formantiques des sons vocaux. Selon ce modèle, le conduit vocal est assimilé à un tube acoustique non uniforme, étroit et non absorbant. Par étroit, nous entendons une largeur du tube petite par rapport à la longueur d'onde du son. Cette condition permet de négliger les variations des grandeurs physiques dans les plans transversaux et de traiter la propagation sonore selon la seule dimension longitudinale. La variation suivant l'axe du tube de l'aire des sections droites doit encore être faible sur des distances de l'ordre de la largeur du tube. Enfin, la variation au cours du temps de la forme du conduit est supposée négligeable sur des durées de l'ordre des périodes de vibration des modes de résonance.

Les conditions aux limites généralement adoptées correspondent à une impédance infinie à la glotte et à l'absence de rayonnement aux lèvres. Pendant la phonation, cependant, le rayonnement est faible mais non nul. Ce rayonnement a pour effet de déplacer légèrement les formants. A condition d'effectuer la correction inverse sur les fréquences formantiques mesurées, le rayonnement aux lèvres peut être considéré comme nul.

Malgré l'approximation certainement considérable que représente ce modèle pour le conduit vocal, il est remarquable que des résultats relativement précis, tout au moins dans le cas des voyelles, aient pu être obtenus.

Soit x la distance à la glotte, L la longueur du conduit, A(x) la fonction d'aire et f_n une fréquence de résonance. Pour une onde de pression sonore monochro-

matique, de fréquence f_n et d'amplitude $p_n(x)$, l'équation de Webster s'écrit :

$$\frac{d}{dx}\left(A(x)\frac{d}{dx}p_n(x)\right) - \lambda_n A(x)\, p_n(x) = 0 \qquad (1a)$$

avec : $\lambda_n = \left(\frac{2\pi f_n}{c}\right)^2$

c étant la vitesse du son et A(x) étant supposé positif de la glotte aux lèvres :

$A(x) > o$ pour $o \leq x \leq L$.

En développant le premier terme de (1a) et en divisant par A(x), cette équation devient :

$$p''_n(x) + \frac{A'(x)}{A(x)} p'_n(x) + \lambda_n p_n(x) = o \qquad (1b)$$

avec : $f'(x) \equiv \frac{d}{dx} f(x)$

Les conditions aux limites que nous envisagerons en premier lieu et auxquelles sont associées les fréquences formantiques s'écrivent :

$$\begin{cases} p'_n(o) = o & \text{(conduit fermé à la glotte)} \\ p_n(L) = o & \text{(conduit ouvert aux lèvres en l'absence de rayonnement)} \end{cases} \qquad (3)$$

La détermination de la fonction d'aire à partir des fréquences de résonance du conduit vocal se présente ainsi comme l'inverse d'un problème aux valeurs propres classique. Ici, en plus des conditions aux limites, les valeurs propres sont connues et c'est la fonction distribuée A(x) qui est recherchée*.

Le problème étant ainsi posé, nous rappellerons d'abord les résultats auxquels a conduit l'usage de la théorie des perturbations du premier ordre. Nous exposerons ensuite les principes d'une méthode basée sur une formulation variationnelle du problème et permettant le calcul de la fonction d'aire dans le cadre d'un modèle paramétrique donné.

---

* Le lien entre grandeurs acoustiques et fonction d'aire A(x) établi par l'équation de Webster reste invariant si A(x) est multiplié par une constante non nulle. Aussi, à l'aide de cette seule équation, ne peut-on espérer déterminer la fonction d'aire qu'à un facteur multiplicatif près.

II. METHODE DES PERTURBATIONS

1. C'est dans les années 1958-60 que Ungeheuer [6] a, le premier, utilisé la méthode des perturbations [6,7,8] pour l'étude quantitative des déplacements des fréquences formantiques résultant de faibles déformations du conduit vocal uniforme. Partant de l'équation de Webster (1b), l'auteur interprête l'expression :

$$\frac{A'(x)}{A(x)}\, p_n'(x) \equiv \left(\log A(x)\right)' p_n'(x)$$

qui est identiquement nulle dans le cas d'un conduit uniforme, comme un terme de perturbation, c'est-à-dire petit par rapport aux autres termes.

L'équation "non perturbée" (associée au conduit uniforme) s'écrit donc :

$$p^{\circ\prime\prime}(x) + \lambda^{\circ} p^{\circ}(x) = 0 \qquad (4)$$

l'indice $^\circ$ caractérisant les grandeurs non perturbées. La solution générale de cette équation s'écrit : $p^\circ(x) = C \cos \lambda^\circ x + D \sin \lambda^\circ x$
où C et D sont des constantes définies par les conditions aux limites particulières. Avec les conditions aux limites (3), la solution n'existe que pour certaines valeurs, $\{\lambda_n^\circ\}$, du paramètre λ° :

$$\lambda_n^\circ = \left((2n-1)\pi/2L\right)^2 \quad ; \; n = 1, 2, \ldots \qquad (5)$$

ou encore, pour les fréquences (formantiques) $F_n^\circ = (2n-1)c/4L$, en se référant à (2).

A ces valeurs propres sont associées les "fonctions propres" :

$$p_n^\circ(x) = \cos \sqrt{\lambda_n^\circ}\, x \qquad (6)$$

solutions du problème aux limites "non perturbé".

Ces fonctions propres, définies à un facteur multiplicatif près, jouissent des propriétés habituelles suivantes :

1°) Elles sont mutuellement orthogonales :

$$\int_0^L p_n^0(x)\, p_m^0(x)\, dx = \begin{cases} 0 & \text{pour } n \neq m \\ \neq 0 \ \left(= \frac{L}{2}\right) & \text{pour } n = m \end{cases} \qquad (7)$$

2°) Elles forment un ensemble complet de fonctions. Ceci signifie que toute fonction* $f(x)$ définie sur l'intervalle $0 \leq x \leq L$ peut être exprimée à l'aide d'une série de la forme :

$$f(x) = \sum_{n=1}^{\infty} c_n\, p_n^0(x) \qquad \text{pour } 0 \leq x \leq L \qquad (8)$$

En vertu des relations d'orthogonalité (7) nous avons :

$$c_n = \frac{1}{N} \int_0^L f(x)\, p_n^0(x)\, dx \qquad \text{avec} \quad N = \int_0^L [p_n^0(x)]^2\, dx = \frac{L}{2}$$

Considérons à présent l'équation perturbée (1b). En raison des propriétés que nous venons d'énoncer, nous pouvons exprimer la solution $p_n(x)$ sous forme de la série :

$$p_n(x) = p_n^0(x) + \sum_{k \neq n}^{\infty} c_k^{(n)}\, p_k^0(x) \qquad (9)$$

Nous poserons également :

$$\lambda_n = \lambda_n^0 + \delta\lambda_n \qquad (10)$$

La méthode des perturbations du premier ordre suppose le terme de perturbation suffisamment petit pour que soit valable l'approximation suivante :

$$\int_0^L (\log A(x))'\, p_n'(x)\, p_n^0(x)\, dx \simeq \int_0^L (\log A(x))'\, p_n^{0\prime}(x)\, p_n^0(x)\, dx \qquad (11)$$

---

* $f(x)$ doit être de "carré sommable" c'est-à-dire vérifier la condition :

$$\int_0^L f^2(x)\, dx < \infty$$

Introduisons (9) et (10) dans (1b). Puisque les $p_k^\circ(x)$ (avec k = 1, 2, ...) satisfont à l'équation non perturbée (4), il vient :

$$\sum_{k\neq n}^{\infty} c_k^{(n)} \left(\lambda_n - \lambda_n^\circ\right) p_n^\circ(x) + \left(\log A(x)\right)' p_n^{\circ\prime}(x) + \delta\lambda_n\, p_n^\circ(x) = 0$$

Multiplions cette équation par $p_n^\circ(x)$ et intégrons sur x. En vertu de l'orthogonalité des fonctions $p_n^\circ(x)$, le premier terme s'annule. Tenant compte de l'approximation (11), nous obtenons :

$$\int_0^L \left(\log A(x)\right)' p_n^{\circ\prime}(x)\, p_n^\circ(x)\, dx + \delta\lambda_n \int_0^L \left(p_n^\circ(x)\right)^2 dx \simeq 0$$

soit :

$$\delta\lambda_n \simeq - \frac{\int_0^L \left(\log A(x)\right)' p_n^\circ(x)\, p_n^\circ(x)\, dx}{\int_0^L \left(p_n^\circ(x)\right)^2 dx} \qquad (12a)$$

Intégrons par parties :

$$\int_0^L \left(\log A(x)\right)' p_n^{\circ\prime}(x)\, p_n^\circ(x)\, dx = - \int_0^L \log A(x) \left(p_n^{\circ\prime}(x)\, p_n^\circ(x)\right)' dx + \left[\log A(x)\, p_n^{\circ\prime}(x)\, p_n^\circ(x)\right]_0^L$$

Le terme aux limites est nul (voir (2)). Dès lors :

$$\delta\lambda_n \simeq \frac{\int_0^L \log A(x) \left(p_n^{\circ\prime}(x)\, p_n^\circ(x)\right)' dx}{\int_0^L \left(p_n^\circ(x)\right)^2 dx} \qquad (12b)$$

Comme $p_n^\circ(x) = \cos\left((2n-1)\,\pi x/2L\right)$, il est aisé de vérifier que :

$$\delta\lambda_n \simeq - \lambda_n^\circ\, \alpha_{2n-1} \qquad (13a)$$

avec : $$\alpha_{2n-1} = \frac{2}{L} \int_0^L \log A(x) \cos \frac{(2n-1)\pi x}{L}\, dx$$

On reconnaît dans cette expression de a_{2n-1} la définition même du coefficient d'ordre impair : 2n-1, du développement en série cosinus du logarithme de la fonction d'aire *:

$$\log A(x) = \log A_0 + \sum_{n=1}^{\infty} a_n \cos \frac{n\pi x}{L} \qquad ** \qquad (14)$$

Puisque $\lambda_n = (2\pi F_n / c)^2$, F_n étant la nième fréquence formantique, la relation (13a) établit une relation univoque entre le déplacement du $n^{è}$ formant et le seul coefficient impair a_{2n-1} de la série cosinus du logarithme de l'aire.

En admettant un déplacement des formants $\delta F_n = F_n - F_n^o$ suffisamment petit pour que l'approximation : $F_n \delta F_n \simeq F_n^o \delta F_n$ soit valable - approximation cohérente avec l'application de la théorie des perturbations du premier ordre -, la relation (13a) peut s'écrire :

$$F_n - F_n^o \simeq -\frac{1}{2} F_n^o a_{2n-1} \qquad (13b)$$

C'est cette relation obtenue tout d'abord par Ungeheuer, que Schroeder et Mermelstein ont inversée pour définir le coefficient a_{2n-1}.

2. Le calcul des coefficients impairs de la série cosinus ne suffit pas pour approcher de manière satisfaisante la fonction d'aire. Mermelstein et Schroeder ont alors considéré le problème aux valeurs propres associé au conduit fermé aux deux extrémités. Les conditions aux limites s'écrivent dans ce cas :

$$p_n'(0) = p_n'(L) = 0 \qquad (15)$$

Les fonctions propres du conduit uniforme (non perturbé) sont de la forme :

---

* Schroeder [3] en 1967 a obtenu le même résultat à partir d'un théorème d'Ehrenfest, sans faire appel à l'équation de Webster.

** A_o est l'aire d'une section droite du conduit uniforme. Nous pouvons poser A_o=1. Dès lors : log A_o = 0.

$$p_n^\circ(x) = \cos\sqrt{\lambda_n^\circ}\, x \qquad n = 1, 2, \ldots \qquad (16)$$

avec : $\lambda_n^\circ = (n\pi / L)^2$ représentant les valeurs propres associées.

Le calcul de perturbation, tel que nous l'avons développé dans le cas précédant, peut être identiquement reproduit ici et conduit au résultat suivant :

$$f_n - f_n^\circ \simeq -\frac{1}{2} f_n^\circ a_{2n} \qquad (17)$$

avec : $$a_{2n} = \frac{2}{L} \int_0^L \log A(x) \cos \frac{2n\pi x}{L}\, dx$$

et où f_n désigne la nième fréquence de résonance du conduit perturbé fermé aux deux extrémités.

Ainsi se trouve établie la relation permettant d'approcher les coefficients pairs de la série (14) à partir des fréquences f_n.

Insistons sur le fait que ces relations (13) et (17), très simples,ne sont valables que dans le cas où la fonction d'aire A(x) ne s'écarte pas trop de celle associée au conduit uniforme, ou, ce qui revient au même, lorsque les inégalités :

$$\left|\frac{F_n - F_n^\circ}{F_n}\right| \ll 1 \quad \text{et} \quad \left|\frac{f_n - f_n^\circ}{f_n^\circ}\right| \ll 1$$

sont satisfaites pour tout n.

3. Le calcul effectif de la fonction d'aire à l'aide des relations (13) et (17) et de la série (14) pose évidemment le problème de la mesure des fréquences propres F_n et f_n. Deux difficultés se présentent. Premièrement, les fréquences f_n ne peuvent pas être extraites du signal de parole puisqu'elles appartiennent à la situation d'un conduit fermé aux lèvres qui n'est pas celle de la production des sons vocaux. Mermelstein et Schroeder ont montré que ces fréquences peuvent être estimées à partir de la mesure de l'admittance aux lèvres. Cette admittance est propor-

tionnelle à : $p'(L,f)/p(L,f)$ où $p(L,f)$ est l'amplitude d'une onde de pression sonore monochromatique de fréquence f mesurée aux lèvres et où

$$p'(L,f) \equiv \left[\frac{d}{dx}\, p(x,f)\right]_{x=L}$$ est proportionnel à l'amplitude de la

vitesse volumique aux lèvres et à la fréquence f. Les pôles (fréquences auxquelles le dénominateur $p(L,f)$ s'annule) et les zéros de cette admittance (fréquences auxquelles le numérateur $p'(L,f)$ s'annule) jouissent des propriétés suivantes. En admettant une impédance infinie à la glotte et en l'absence de rayonnement aux lèvres, les pôles coïncident avec les formants. Les zéros de l'admittance aux lèvres coïncident avec les fréquences propres associées au conduit fermé aux lèvres. Ainsi la donnée des zéros de l'admittance mesurée aux lèvres permet-elle de définir les coefficients pairs du développement (14). Un appareillage spécialement destiné à la mesure de l'admittance aux lèvres est décrit dans [3].

La seconde difficulté résulte du modèle acoustique simple adopté, caractérisé par l'équation de Webster. Ce modèle n'est valable que si la longueur d'onde du son est grande devant la largeur du conduit vocal, soit pour des fréquences inférieures ou de l'ordre de 4000 Hz. Les fréquences F_n et f_n supérieures à 4000 Hz ou 5000 Hz, quand bien même elles auraient été mesurées, ne sauraient être correctement interprètées dans le cadre du modèle. Aussi, les auteurs se sont-ils limités à la mesure des six premières fréquences F_n et f_n (n = 1, 2, 3) et ont approché la fonction d'aire à l'aide des six premiers termes de la série (14) :

$$\log A(x) \simeq \sum_{n=1}^{6} a_n \cos \frac{n\pi x}{L} \qquad (18)$$

A l'appui de cette approximation, notons le fait que les coefficients a_n d'une fonction d'aire réelle (mesurée par rayons X) pour $n > 6$ sont souvent petits devant les coefficients a_n pour $n < 6$. D'autre part, il est bien connu qu'en ce qui concerne la caractérisation phonétique des sons vocaux (sur le plan perceptif en particulier) ce sont les trois premiers formants F_1, F_2, F_3 et eux seuls qui sont déterminants. Or ceux-ci, en vertu des relations (13), ne sont pas, en première approximation, corrélés aux coefficients a_n d'ordre supérieur ou égal à 6.

4. En résumé, il ressort des travaux de Schroeder et Mermelstein les propriétés suivantes, établies dans le cas de perturbations du premier ordre autour d'un conduit uniforme :

1°) la fonction d'aire n'est pas définie univoquement par la donnée même complète des fréquences des formants,

2°) la fonction d'aire est définie univoquement, à une constante multiplicative près, par la donnée de tous les pôles et zéros de l'admittance aux lèvres si la longueur du conduit est connue,

3°) le logarithme de la fonction d'aire peut-être exprimé en terme d'une série de fonction chacune d'elle affectant, en première approximation, un et un seul pôle ou zéro de l'admittance aux lèvres.

Soit $\lambda_1, \lambda_3, \cdots, \lambda_{2n-1}, \ldots$ la suite croissante des valeurs propres reliée aux pôles de l'admittance (formants) et soit $\lambda_2, \lambda_4, \ldots, \lambda_{2n}, \ldots$ une telle suite relative aux zéros de l'admittance. Nous conviendrons, dans la suite de l'exposé, d'affecter d'un indice impair les grandeurs associées aux conditions aux limites (3), et d'un indice pair les grandeurs associées aux conditions (15). Avec ces notations, les propriétés qui viennent d'être énoncées s'expriment par la série :

$$\delta \log A(x) \simeq -\sum_{k=1}^{\infty} \frac{\delta\lambda_m}{\lambda^0_m} \cos(m\pi x/L) \qquad \text{si } \left|\frac{\delta\lambda_m}{\lambda^0_m}\right| \ll 1 \qquad (14b)$$

avec :

$$\delta \log A(x) \equiv \log A(x) - \log A_0(x)$$

Ces propriétés ont été étendues par Heinz [4] au cas général de faibles perturbations autour d'une configuration quelconque.
Pour Heinz, le noeud du problème a été de définir des fonctions $\hat{\psi}(x)$, dites fonctions de perturbation, qui généralisent les $\cos(n\pi x/L)$ de la série (14) lorsque l'aire non perturbée $A_o(x)$ est quelconque :

$$\delta \log A(x) = - \sum_{n=1}^{\infty} \frac{\delta\lambda_n}{\lambda^o_n} \hat{\psi}^o_n(x) \qquad (19)$$

Ces fonctions $\hat{\psi}(x)$ se caractérisent par un effet sélectif sur une et une seule résonance en première approximation. Pour des conditions aux limites données, elles dépendent essentiellement de $A_o(x)$, donc des grandeurs non perturbées. Nous exposons brièvement en annexe, la méthode établie par Heinz pour construire les $\hat{\psi}(x)$.

Ces fonctions de perturbation permettent d'approcher la fonction d'aire à partir de la donnée des premières singularités de l'admittance mesurée aux lèvres et d'une première approximation de la fonction d'aire. Cette première approximation, qui joue le rôle de fonction d'aire non perturbée, est généralement éloignée de la solution cherchée. Aussi est-il nécessaire de procéder par étapes.

Partant des valeurs propres associées à la première approximation de la fonction d'aire, ces valeurs propres sont modifiées par petites unités jusqu'aux valeurs voulues. A chaque étape, les fonctions de perturbation sont construites et la fonction d'aire perturbée est obtenue simplement à l'aide de la série (19) limitée à un nombre fini de termes. La fonction d'aire obtenue joue ensuite le rôle de fonction non perturbée à l'étape suivante.

Heinz a ainsi calculé la fonction d'aire associée aux données de G. Fant pour la voyelle /a/ au départ du conduit uniforme et des 6 premières singularités de l'admittance aux lèvres (fig.1).

La découverte des fonctions de perturbation est l'aboutissement auquel l'application de la théorie des perturbations du premier ordre a donné lieu. Outre leur usage pour le calcul approché des fonctions d'aire, ces fonctions de perturbation constituent probablement un outil important d'investigation des formes compensatoires dans l'articulation, puisque les fonctions d'indice pair n'ont pas d'effet du premier ordre sur les fréquences des formants.

III. METHODE VARIATIONNELLE

1. L'équation de Webster avec des conditions aux limites du type envisagé précédemment, peut être regardée comme la condition d'Euler-Lagrange à laquelle la fonction $p_n(x)$ doit satisfaire pour que la fonctionnelle $W(p_n)$:

$$W(p_m) = \int_0^L A(x)\left\{ (p_m'(x))^2 - \lambda_m p_m^2(x) \right\} dx$$

prenne une valeur stationnaire (minimum dans ce cas).

Le problème est alors posé en ces termes : connaissant N valeurs propres λ déterminer une fonction d'aire A(x) telle que les variations premières des fonctionnelles $W(p_n)$, n variant de 1 à N, s'annulent, les conditions aux limites étant satisfaites. Autrement dit, la fonction A(x) à déterminer doit permettre l'existence d'une solution, soit $\{p_n(x)\} = \{\bar{p}_n(x)\}$, au système d'équations :

$$\delta W(p_m(x); \lambda_m) = 0 \quad ; \qquad m = 1, \ldots, N \qquad (20)$$

les fonctions à varier: les $\{p_n(x)\}$, étant soumises aux conditions aux limites prescrites et les λ_n étant donnés.

Une telle formulation du problème présente certains avantages. Il n'est pas fait appel à un problème non perturbé et la première approximation de la fonction d'aire peut être relativement éloignée de la solution. Ensuite, le fait que fonction d'aire et amplitude de pression sonore apparaissent sous un signe d'intégration, conduit à des équations liées bien plus au caractère général qu'au comportement détaillé de ces grandeurs; de ce fait, cette formulation intégrale se prête bien à l'usage d'expressions paramétriques pour représenter ces grandeurs. Enfin, les méthodes variationnelles permettent de traiter des problèmes extrêmement variés, dans un même formalisme et selon des procédures de calcul relativement simples et formellement semblables.

Il est commode d'effectuer le changement de variable suivant :

$$z = \frac{\pi}{L} x \qquad (21)$$

qui ramène l'intervalle [0,L] à [0,π] . L'équation de Webster devient alors, puisque $dx = \frac{L}{\pi} dz$:

$$\left(A(z)\, p_n'(z)\right)' + \tilde{\lambda}_n A(z)\, p_n(z) = 0$$

avec :
$$\tilde{\lambda}_n = \left(\frac{L}{\pi}\right)^2 \lambda_n \qquad (22)$$

soit :
$$\tilde{\lambda}_n = \left(\frac{2L}{c} f_n\right)^2$$

La fonctionnelle $W(p_n)$ devient :

$$W(p_n) = \int_0^L A(z) \left\{ \left(p_n'(z)\right)^2 - \tilde{\lambda}_n p_n(z) \right\} dz \qquad (23)$$

Admettons maintenant que la fontion d'aire A(z) puisse être définie au moyen d'une expression paramétrique donnée : $A(\{\beta_k\}; z)$ dépendant de z et de K paramètres "articulatoires" : β_k, k = 1, 2, ..., K. Cette expression, avec d'éventuelles contraintes imposées aux paramètres β_k, définit un modèle paramétrique des fonctions d'aire. Le problème consiste dès lors à déterminer les valeurs de ces paramètres à partir des conditions : $\delta W(p_n; \tilde{\lambda}_n) = 0$.

2. Pour résoudre ce problème, nous avons appliqué la méthode de Rayleigh-Ritz dont l'usage est par ailleurs très largement répandu pour le calcul approché des valeurs propres des opérateurs auto-adjoints [7] .

Selon cette méthode, nous choisissons un ensemble de fonctions linéairement indépendantes satisfaisant les conditions aux limites requises pour $p_n(z)$ et tel que $p_n(z)$ puisse être approché par une combinaison linéaire d'un nombre limité de fonctions de cet ensemble. Soit I_n ce nombre et soit $\{p^o_{n,1}, p^o_{n,2}, \ldots, p^o_{n,i}, \ldots\}$ cet ensemble. Nous avons donc :

$$p_n(z) \simeq \sum_{i=1}^{I_n} C_i^{(n)} p^o_{n,i}(z) \qquad (24)$$

Cette approximation est une fonction (fonction d'essai) linéaire et homogène en ses paramètres $C_i^{(n)}$. Introduisons cette approximation, ainsi que l'expression paramétrique

de la fonction d'aire, soit $A(\{\beta_k\}, z)$, dans (23). Il vient :

$$W(p_n) \simeq \sum_{i,j=1}^{I_n} c_i^{(n)} c_j^{(n)} W_{n,ij} \qquad (25)$$

avec :

$$W_{n,ij} = U_{n,ij} - \tilde{\lambda}_n V_{n,ij} \qquad (26a)$$

$$U_{n,ij} = \int_0^n A(\{\beta_k\}; z)\, p'^{\,o}_{n,i}(z)\, p'^{\,o}_{n,j}(z)\, dz \qquad (26b)$$

$$V_{n,ij} = \int_0^n A(\{\beta_k\}; z)\, p^o_{n,i}(z)\, p^o_{n,j}(z)\, dz \qquad (26c)$$

Remarquons que la matrice $(I_n \times I_n)$: W_n d'élément $W_{n,ij}$ est symétrique, puisque $U_{n,ij} = U_{n,ji}$ et $V_{n,ij} = V_{n,ji}$.

La fonctionnelle $W(p_n)$ est ainsi remplacée par une fonction quadratique ordinaire des paramètres $C_i^{(n)}$. La condition $\delta W(p_n) = 0$ est alors remplacée par :

$$\sum_{i=1}^{I_n} \delta c_i^{(n)} \sum_{j=1}^{I_n} c_j^{(n)} W_{n,ij} = 0$$

Puisque les fonctions $\{p^o_{n,i}(z)\}$ vérifient les conditions aux limites, les accroissements $\delta C_i^{(n)}$ sont arbitraires. Nous devons donc avoir :

$$\sum_{j=1}^{I_n} W_{n,ij}\, c_j^{(n)} = 0 \qquad \text{pour } i = 1, \ldots, I_n \qquad (27)$$

(27) définit un système d'équations algébriques linéaires et homogènes en les $C_i^{(n)}$. Pour que ce système ait une solution non triviale, il faut imposer aux coefficients $W_{n,ij}$, la condition bien connue :

$$\det\left[W_{n,ij}\right] \equiv \det\left[U_{n,ij} - \tilde{\lambda}_n V_{n,ij}\right] = 0 \qquad *$$

---

* L'équation : $\det\left|U_{n,ij} - \lambda V_{n,ij}\right| = 0$ est un polynôme de degré I_n en λ. On démontre [8] que les I_n racines de cette équation "séculaire" : $\lambda^{(1)}, \lambda^{(2)}, \ldots$ sont toutes réelles et fournissent les valeurs approchées par excès des I_n premières valeurs propres associées au problème aux limites considéré.

Une matrice W_n peut ainsi être construite pour chaque valeur propre $\tilde{\lambda}_n$ associée à chaque fonction propre $p_n(z)$, n variant de 1 à N. Nous obtenons finalement un système de N équations algébriques non linéaires :

$$D(\{\beta_k\}; \tilde{\lambda}_m) \equiv \det[W_{m,ij}] = 0 \quad ; \quad n = 1, \ldots, N \qquad (28)$$

chacune de ces équations reliant une et une seule valeur propre λ_n aux K paramètres articulatoires $\{\beta_k\}$.

3. Ce système d'équations (28) définit un ensemble de contraintes acoustiques imposées aux paramètres articulatoires $\{\beta_k\}$; il est le point de départ du calcul de ces derniers.

Notons, au passage, que les expressions $D(\{\beta_k\}, \tilde{\lambda}_n)$ ne font pas intervenir les coefficients $C_i^{(n)}$ du développement (24) et, par conséquent, la méthode envisagée ici n'exige pas le calcul de la pression sonore $p_n(z)$ dans le conduit vocal.

Selon les valeurs respectives de K et de N, trois cas sont à envisager.

a) Si K < N, une solution optimale, dans le cadre du modèle paramétrique choisi, peut être obtenue par minimisation d'une somme pondérée des carrés des seconds membres (résidus) de (28).

b) Si K = N et si le modèle paramétrique choisi est compatible avec les contraintes acoustiques, une solution vérifiant le système (28) existe.
Si le système (28) est incompatible, cela signifie que le modèle paramétrique est inapte à générer une fonction d'aire vérifiant les contraintes acoustiques données. Dans ce cas, à défaut d'une modification du modèle, une solution optimale, au sens indiqué plus haut, peut néanmoins être recherchée.

c) Si K > N, des contraintes supplémentaires doivent être imposées aux $\{\beta_k\}$ si une solution unique* est souhaitée. Ces contraintes peuvent appartenir au modèle lui-même ou peuvent résulter de mesures directes, mais partielles de la fonction d'aire (aux lèvres et à la glotte par exemple).

---

* Nous n'avons pas tenté d'établir de théorème d'unicité de la solution en raison de la complexité que cela représente sur le plan mathématique. Il est néanmoins certain, qu'en raison du volume limité d'information acoustique accessible, l'unicité ne peut résulter que des contraintes imposées à la solution au travers du modèle paramétrique choisi.

Convenons de la notation abrégée :

$$D_n \equiv D(\{\beta_k\}; \tilde{\lambda}_n)$$

et considérons d'abord le cas général où une solution optimale, au sens des moindres carrés, est recherchée. Pour la clarté de l'exposé, nous supposons l'absence de contraintes explicites autres que les contraintes acoustiques (28)*. Nous supposons également : $K \leq N$.

Le problème consiste à minimiser la somme des expressions D_n^2 pondérée par des poids $P_n > 0$:

$$\sum_{n=1}^{N} P_n D_n^2 \qquad (29)$$

Ceci revient à résoudre le système de K équations algébriques non linéaires à K inconnues:

$$\sum_{n=1}^{N} P_n D_n \frac{\partial D_n}{\partial \beta_k} = 0 \qquad (30)$$

Le choix des poids P_n mérite une attention particulière. Soit $\{\bar{\beta}_k\}$ un ensemble de valeurs plus ou moins proches de la solution et considérons l'expression :

$$\bar{D}_n = \left(\frac{\partial D_n}{\partial \tilde{\lambda}_n}\right)^{-1}_{\{\beta_k\}=\{\bar{\beta}_k\}} \cdot D_n(\{\beta_k\}; \tilde{\lambda}_n)$$

Nous supposons :

$$\left(\frac{\partial D_n}{\partial \tilde{\lambda}_n}\right)_{\{\beta_k\}=\{\bar{\beta}_k\}} \neq 0$$

ce qui a toujours été vérifié pour des valeurs $\{\bar{\beta}_k\}$ réalistes.

---

* Les éléments de procédure que nous présentons ici s'étendent sans difficulté aux cas où des contraintes supplémentaires sont présentes. Un exemple comportant une contrainte résultant de la mesure de l'aire à la glotte et aux lèvres est donné en annexe II.

Si $\tilde{\lambda}_n(\{\beta_k\})$ désigne la nè valeur propre associée à la fonction d'aire $A(\{\beta_k\},z)$ il est aisé de voir que $\overline{D}_n$ est une approximation linéaire autour de de la différence entre la valeur propre imposée $\tilde{\lambda}_n$ et $\tilde{\lambda}_n(\{\beta_k\})$:

$$D(\{\beta_k\};\tilde{\lambda}_m) \simeq 0 + \left(\frac{\partial D_m}{\partial \lambda_m}\right)_{\{\beta_k\}=\{\overline{\beta}_k\}} \cdot \left(\tilde{\lambda}_m - \tilde{\lambda}_m(\{\beta_k\})\right)$$

donc : $$\overline{D}_m \simeq \tilde{\lambda}_m - \tilde{\lambda}_m(\{\beta_k\})$$

La signification de la contrainte acoustique $\overline{D}_n = 0$ (ou $D_n = 0$) est ainsi clairement explicitée : la valeur propre $\tilde{\lambda}_m(\{\beta_k\})$ associée à la fonction d'aire calculée doit être égale à la valeur mesurée $\tilde{\lambda}_n$.
La valeur de :

$$\sum_{m=1}^{N} \overline{D}_m^2 \simeq \sum_{m=1}^{N} \left(\tilde{\lambda}_m - \tilde{\lambda}_m(\{\beta_k\})\right)^2$$

pour une fonction d'aire donnée est une mesure de la précision avec laquelle cette fonction d'aire satisfait aux contraintes acoustiques. Dès lors, un choix pertinent de poids P_n consiste à prendre :

$$P_m = \left(\frac{\partial D_m}{\partial \tilde{\lambda}_m}\right)^{-2}_{\{\beta_k\}=\{\overline{\beta}_k\}}$$

Un autre choix possible qui favorise les contraintes acoustiques relatives aux basses fréquences est le suivant :

$$P_m = \left\{ \tilde{\lambda}_m^0 \left(\frac{\partial D_m}{\partial \tilde{\lambda}_m}\right)_{\{\beta_k\}=\{\overline{\beta}_k\}} \right\}^{-2}$$

avec $\tilde{\lambda}_n^0$, la valeur propre correspondante associée au conduit uniforme.
L'expression à minimiser devient alors :

$$\sum_{m=1}^{N} \left(\frac{1}{\tilde{\lambda}_m^0} \overline{D}_m\right)^2 \simeq \sum_{m=1}^{N} \left(\frac{\tilde{\lambda}_m - \tilde{\lambda}_m(\{\beta_k\})}{\tilde{\lambda}_m^0}\right)^2$$

De tels choix de P_n nécessitent le calcul des dérivées partielles : $\frac{\partial D_m}{\partial \tilde{\lambda}_m}$, D_n étant, rappelons le, défini par le déterminant de la matrice W_n (voir (28)).

Si $\hat{W}$ désigne la matrice adjointe d'une matrice carrée $W(\lambda)$ dépendant d'un paramètre λ, il est bien connu que :

$$\frac{d}{d\lambda} \det[W(\lambda)] = \sum_{i,j} \hat{W}_{ji} \frac{dW_{ij}}{d\lambda}$$

Par conséquent, compte tenu de la définition (26) des $W_{n,ij}$ nous obtenons :

$$\frac{\partial D_m}{\partial \tilde{\lambda}_m} = - \sum_{i,j=1}^{I_m} \hat{W}_{m,ji} V_{m,ij}$$

Ainsi notre choix des P_n nécessite le calcul des matrices adjointes aux W_n. Ce fait n'introduit pas de difficulté pratique supplémentaire, étant donné que ces matrices adjointes sont nécessaires au calcul des $\frac{\partial D_m}{\partial \beta_k}$ dans (30) :

$$\frac{\partial D_m}{\partial \beta_k} = \sum_{i,j=1}^{I_m} \hat{W}_{m,ij} \frac{\partial W_{m,ij}}{\partial \beta_k} \qquad (31)$$

4. Venons en maintenant au cas particulier où N = K lorsqu'une solution exacte de (28) existe (autrement dit, lorsque le modèle paramétrique est compatible avec le système de contraintes acoustiques). Ce cas nous semble intéressant du point de vue du calcul. En effet, nos premières applications numériques donnent à penser que, sous ces conditions, le système (28) peut être résolu par la méthode classique de Newton-Raphson, et ce au départ d'une première approximation relativement éloignée de la solution; cette première approximation pouvant, par exemple, correspondre au conduit uniforme. Par contre, cette méthode s'est avérée inapplicable au système (30) à moins d'une première approximation très proche de la solution.

La méthode de Newton-Raphson de résolution d'un système d'équations algébriques non linéaires est l'une des plus simples et des plus rapides, mais elle exige que le système soit linéarisable autour de la première approximation. S'il n'en est pas ainsi, la suite des "approximations" que fournit cette méthode diverge. Soit donc $\{\beta_k^{(0)}\}$ la première approximation choisie. La méthode consiste à remplacer

le système non linéaire (28) par une succession de systèmes linéarisés :

$$D_n^{(m)} + \sum_{k=1}^{N} G_{nk}^{(m)} \, \delta\beta_k^{(m)} = 0 \qquad ; \; n = 1, \ldots, N$$

$$\beta_k^{(m+1)} = \beta_k^{(m)} + \delta\beta_k^{(m)}$$

pour $m = 0, 1, 2, \ldots$, avec :

$$D_n^{(m)} = D\left(\{\beta_k^{(m)}\} \; ; \lambda_n\right)$$

$$G_{nk}^{(m)} = \left(\frac{\partial D_n}{\partial \beta_k}\right)_{\{\beta_k\} = \{\beta_k^{(m)}\}}$$

La solution $\{\beta_k^{(m+1)}\}$ obtenue à la m$^{\text{è}}$ itération sert de première approximation à l'itération suivante. La correction $\delta\beta_k^{(m)}$ est donnée par :

$$\delta\beta_k^{(m)} = -\sum_{n=1}^{N} \left[G^{(m)}\right]^{-1}_{kn} D_n^{(m)}$$

Il faut encore que les matrices carrées $G^{(m)}$ calculées successivement ne soient pas singulières ($\det(G^{(m)}) \neq 0$) pour que les matrices inverses $[G^{(m)}]^{-1}$ existent.

Le processus itératif est arrêté lorsque les conditions :

$$|\delta\beta_k^{(m)}| < \varepsilon_1 \qquad \text{et} \qquad |\bar{D}_n^{(m)}| \text{ ou } |\hat{\lambda}_n(\{\beta_k^{(m)}\}) - \tilde{\lambda}_n| < \varepsilon_2 \qquad \text{pour } k, n = 1, \ldots, N$$

sont remplies, les nombres positifs ε_1 et ε_2 étant choisis suffisamment petits.

Pour clôturer ces considérations générales, signalons que nous avons rassemblé en annexe quelques brèves indications concernant les méthodes de résolution applicables au système (30), et décrites dans la littérature [9].

5. Reportons-nous au problème traité par la théorie des perturbations et proposons-nous de lui appliquer les formules générales et la procédure qui découlent de la méthode variationnelle.

Comme fonctions $\{p^0_{m,i}(t)\}$ servant de base aux séries (24), nous choisirons

les fonctions propres du conduit uniforme vérifiant les conditions aux limites :

$p^{o\,\prime}_{n,i}(0) = p^{o}_{n,i}(\pi) = 0$ pour n impair et $p^{o\,\prime}_{n,i}(0) = p^{o\,\prime}_{n,i}(\pi) = 0$ pour n pair. Ainsi :

$$p^{o}_{n,i}(z) = \cos(J_{ni}\, z) \qquad ; \; n = 1, 2, \dots N \; ; \; i = 1, 2, \dots, I_n \qquad (32)$$

avec :

$$J_{ni} = \begin{cases} i - 1/2 & \text{pour } n \text{ impair} \\ i & \text{pour } n \text{ pair} \end{cases}$$

A l'aide des relations trigonométriques :

$$\begin{aligned} \cos\alpha \cos\beta &= \tfrac{1}{2}\left[\cos(\alpha-\beta) + \cos(\alpha+\beta)\right] \\ \sin\alpha \sin\beta &= \tfrac{1}{2}\left[\cos(\alpha-\beta) - \cos(\alpha+\beta)\right] \end{aligned} \qquad (33)$$

il est alors aisé d'expliciter, en termes de coefficients de série cosinus de la fonction d'aire :

$$a(l) = \int_0^{\pi} A(1/\beta_k 3; z) \cos l z \, dz \qquad (34)$$

les éléments des matrices définis par (26):

$$U_{n,ij} = \tfrac{1}{2}\, J_{n,i}\, J_{nj}\left[a(J_{ni} - J_{nj}) - a(J_{ni} + J_{nj})\right] \qquad (35a)$$

$$V_{n,ij} = \tfrac{1}{2}\left[a(J_{ni} - J_{nj}) + a(J_{ni} + J_{nj})\right] \qquad (35b)$$

$$W_{n,ij} = \tfrac{1}{2}\left(J_{ni} J_{nj} - \tilde{\lambda}_n\right) a(J_{ni} - J_{nj}) - \tfrac{1}{2}\left(J_{ni} J_{nj} + \hat{\lambda}_n\right) a(J_{ni} + J_{nj}) \qquad (35c)$$

Si nous voulons expliciter les dérivées partielles $\frac{\partial D_n}{\partial \beta_k}$ exprimées en termes de $\frac{\partial W_{n,ij}}{\partial \beta_k}$ (voir (31)), il nous faut préciser le modèle articulatoire :

$$\frac{\partial W_{n,ij}}{\partial \beta_k} = \tfrac{1}{2}\left(J_{ni} J_{nj} - \tilde{\lambda}_n\right) \frac{\partial}{\partial \beta_k}\, a(J_{ni} - J_{nj}) - \tfrac{1}{2}\left(J_{ni} J_{nj} + \hat{\lambda}_n\right) \frac{\partial}{\partial \beta_k}\, a(J_{ni} + J_{nj}) \qquad (36)$$

avec : $$\frac{\partial a(l)}{\partial \beta_k} = \int_0^{\pi} \frac{\partial}{\partial \beta_k} A(\{\beta_k\};z) \cos lz \, dz \qquad (37)$$

Considérons la "classe" de modèles paramétriques :

$$A(\{\beta_k\};z) = A_0(z) \exp\left[\sum_{k=1}^{K} \beta_k \varphi_k(z)\right] \qquad (38)$$

où $A_0(z)$ est une fonction donnée de z positive sur $[0,\pi]$ et où les $\varphi_k(z)$, données à une constante additive près, sont linéairement indépendantes sur $[0,\pi]$. Nous avons :

$$\frac{\partial A}{\partial \beta_k} = A\varphi_k \qquad \text{d'où} \qquad \frac{\partial a(l)}{\partial \beta_k} = \int_0^{\pi} A\,\varphi_k(z) \cos lz \, dz \qquad (39)$$

Il est commode de représenter les fonctions φ_k par leur développement en série cosinus :

$$\varphi_k(z) = \sum_{m=1} \alpha_{k,m} \cos mz \qquad (40)$$

En introduisant cette expression dans (39) on obtient aisément :

$$\frac{\partial a(l)}{\partial \beta_k} = \frac{1}{2} \sum_{m=1} \alpha_{k,m} \left[a(l-m) + a(l+m)\right] \qquad (41)$$

Ainsi, dans le cas d'un modèle articulatoire appartenant à (38), le second membre de (36) et, par suite, les $\frac{\partial D_n}{\partial \beta_k}$ peuvent être exprimés en fonction des coefficients de la série cosinus de la fonction d'aire, uniquement, tout comme les D_n eux-mêmes.

Un choix de la fonction $A_0(z)$ intervenant dans l'expression (38) nous est suggéré par Gopinath et Sondhi [10]. Il consiste à prendre une fonction comportant les discontinuités propres à la géométrie du conduit vocal, (au niveau de l'épiglotte essentiellement), discontinuités indépendantes de l'articulation (fig. 6).

6. Dans un but exploratoire, nous avons appliqué cette méthode au problème traité par Schroeder et Mermelstein. Des six valeurs propres données, trois correspondent aux trois premiers pôles de l'admittance aux lèvres (c'est-à-dire aux formants) et les trois autres aux trois premiers zéros de cette admittance.

Nous avons utilisé successivement les expressions paramétriques suivantes de la fonction d'aire :

$$A(\{\beta_k\}; z) = \exp\left[\sum_{k=1}^{6} \beta_k \cos kz\right] \quad \text{(introduite par Mermelstein [2])} \qquad (42a)$$

et $$A(\{\beta_k\}; z) = \exp\left[\sum_{k=1}^{6} \beta_k T_k\left(\frac{2z}{\pi} - 1\right)\right] \quad , 0 \leq z \leq \pi \qquad (42b)$$

où T_k (y) est le polynôme de Tchebycheff de degré k défini pour $-1 \leq y \leq +1$ (voir fig. 2b).

Avec chacun des deux modèles nous avons obtenu, pour les données de G.Fant relatives à la voyelle russe /e/, une solution vérifiant exactement les six équations (K = 6) du système (28) (voir fig. 3 et 4).
Dans les deux cas, la solution a été obtenue par la méthode classique de Newton-Raphson au départ du conduit uniforme. Des précisions concernant les calculs sont données en annexe II.

Le second modèle donne, pour cette voyelle /e/, un meilleur accord avec les mesures effectuées par rayons X. Il ne semble cependant pas que cette supériorité du modèle (42b) soit vérifiée pour toutes les configurations vocales ni même pour la plupart. Nous n'avons pas tenté d'approfondir la question, ces deux modèles n'ayant qu'une valeur d'exemple et devant être considérés aujourd'hui comme dépassés.

Lorsque la configuration vocale s'écarte trop du conduit uniforme, une fonction d'aire vérifiant exactement les six premières équations du système (28) ne peut plus être exprimée à l'aide de l'une des deux expressions (42 a ou b). La modification du modèle paramétrique ou la recherche d'une solution optimale au sens indiqué plus haut est alors nécessaire. Ainsi pour les données de G. Fant relatives à la voyelle russe /a/, nous avons obtenu une solution optimale dans le cadre d'un modèle paramétrique quelque peu modifié (voir annexe II).

Les difficultés pratiques qui accompagnent l'application de la méthode variationnelle sont essentiellement liées à la résolution du système d'équations algébriques non linéaires. Lorsqu'une solution optimale est recherchée (système (30)),

les méthodes itératives généralement utilisées ne garantissent pas nécessairement une convergence vers la solution réalisant le minimum absolu du critère (29) : un minimum local peut être atteint. L'usage d'un modèle paramétrique apte à générer les fonctions d'aire du conduit vocal, et non n'importe quelle fonction, devrait faciliter la tâche et accélérer la convergence vers la solution souhaitée. Comme, de surcroît, la précision de l'image ainsi construite dépend très largement du modèle paramétrique choisi, la définition d'un tel modèle doit faire l'objet d'une attention toute particulière.

7. Nous pouvons envisager deux catégories de modèle. Dans la première, nous rangeons les modèles construits à partir de considérations purement articulatoires : les paramètres du modèle définissant la position et la forme des diverses parties articulées du conduit vocal : les lèvres, la langue, etc.... Citons comme exemples les modèles simples à trois paramètres de Stevens et House [11] et de G. Fant [12] . Dans la seconde catégorie, nous rangeons les modèles qui expriment une fonction d'aire, ou, de préférence, une transformation de celle-ci (telle que logA(x) par exemple) sous forme d'une combinaison linéaire d'un nombre fini de fonctions données. Ces fonctions jouent en quelque sorte le rôle de coordonnées d'un "espace" et sont définies sur toute la longueur du conduit vocal.

Les modèles que nous avons considérés ici sont de ce type. Pour la "classe" de modèle définie par (38), ce sont les fonctions linéairement indépendantes $\{\varphi_k\}$ qui jouent le rôle de fonctions "coordonnées" :

$$\log\left(\frac{A(z)}{A_o(z)}\right) = \sum_{k=1}^{k} \beta_k \varphi_k(z)$$

Dans les exemples traités numériquement, ce rôle est tenu respectivement par les fonctions cosinus et par les polynômes de Tchébycheff.

Un modèle de cette catégorie, mieux adapté aux configurations vocales, pourrait être construit selon une voie inspirée du modèle de Öhman [14].
Les fonctions coordonnées seraient construites à partir des données expérimentales relatives à des configurations particulières du conduit telles celles qui accompagnent l'articulation des voyelles /u/, /a/ et /i/ par exemple.

8. Une qualité importante de la méthode réside dans le fait qu'elle autorise, au travers du modèle paramétrique choisi, d'incorporer au processus même du calcul des propriétés géométriques et des contraintes propres au conduit vocal et à l'articulation.

Enfin, en vertu du caractère très général que présentent les méthodes variationnelles, les principes de la procédure présentés ici ne devraient pas être profondément modifiés si l'on adopte d'autres conditions aux limites associées à d'autres données acoustiques. Il en serait encore de même si, pour décrire le comportement acoustique du conduit vocal, le modèle de la ligne de transmission absorbante était adopté. Ce modèle est en effet plus général que celui décrit par l'équation de Webster parce qu'il permet de rendre compte, dans une certaine mesure, des pertes d'énergie non négligeables intervenant dans le conduit vocal. Dans la perspective d'un tel élargissement du problème, il y a lieu de noter que la généralisation formelle de la fonctionnelle $W(p_n)$ n'est pas immédiate. L'extension de la méthode variationnelle au conduit absorbant est actuellement à l'étude.

Remerciements

Je tiens à remercier M. P.V. Paquet de l'Université de Bruxelles avec qui j'ai eu de fructueuses discussions et le Professeur M. Wajskop qui a porté à ma connaissance les articles [1, 2, 3, 4] provoquant ainsi mon intérêt sur la question.

ANNEXE 1 : Définition des fonctions de perturbation.

Pour une fonction d'aire non perturbée arbitraire $A_0(x)$, la formule de perturbation généralisant (12a) s'écrit :

$$\delta\lambda_n \simeq - \frac{\int_0^L (\delta \log A)' A_0 p_n^{\prime 0} p_n^0 \, dx}{\int_0^L A_0 (p_n^0)^2 \, dx} \qquad \text{(A1;1)}$$

Posons : $\psi_n^0(x) = A_0(x) p_n^{0\prime}(x) p_n^0(x) = v_n^0(x) p_n^0(x)$

avec $v_n^0 = A_0 p_n^{0\prime}$

La fonction $v_n^0(x)$ est proportionnelle à la vitesse volumique et satisfait à l'équation de Webster "duale" :

$$\left(\frac{1}{A_0} v_n^{0\prime}\right)' + \lambda_n^0 \frac{1}{A_0} v_n^0 = 0$$

En remplaçant dans (A1,1) $\delta \log A(x)$ par son développement formel (19), il vient :

$$\delta\lambda_n = \sum_{m=1}^{\infty} \frac{\delta\lambda_m}{\lambda_m^0} \frac{\int_0^L \hat{\psi}_m^{0\prime}(x) \psi_n^0(x) \, dx}{\int_0^L A_0 p_n^{0\,2} \, dx}$$

Comme les $\delta\lambda_m$ sont arbitraires, nous devons avoir :

$$\int_0^L \hat{\psi}_m^{0\prime} \psi_n^0 \, dx = \begin{cases} 0 & \text{si } m \neq n \\ \lambda_n^0 \int_0^L A_0 p_n^{0\,2} \, dx & \text{si } m = n \end{cases} \quad ; \; m = 1, 2, \cdots \qquad \text{(A1;2)}$$

Se référant à des résultats obtenus par Borg [14], Heinz a montré qu'un ensemble de fonctions $\{\hat{\psi}_m^0(x)\}$ vérifiant les conditions (A1;2) peut-être obtenu si :

$$\hat{\psi}_m^0(x) = \hat{v}_m^0(x) \hat{p}_m^0(x) \quad ,$$

où $\hat{p}^{o}_{m}(x)$ est solution de l'équation de Webster :

$$\left(A_o \hat{p}^{o\prime}_m\right)' + \lambda^{o}_m A_o \hat{p}^{o}_m = 0 \qquad \text{(A1,3)}$$

et où $\hat{v}^{o}_{m}(x)$ est solution de l'équation duale :

$$\left(\frac{1}{A_o} \hat{v}^{o\prime}_m\right)' + \lambda^{o}_m \frac{1}{A_o} \hat{v}^{o}_m = 0 \qquad \text{(A1;4)}$$

lorsque les conditions aux limites suivantes sont imposées :

$$\text{pour } m = 1, 3, 5 \ldots \quad \begin{cases} \hat{p}^{o}_m(L) = \hat{p}^{o\prime}_m(0) = 0 \\ \hat{v}^{o}_m(L) = 0 \end{cases}$$

$$\text{pour } m = 2, 4, 6 \ldots \quad \begin{cases} \hat{p}^{o}_m(L) = 0 \\ \hat{v}^{o}_m(L) = \hat{v}^{o}_m(0) = 0 \end{cases} \qquad \text{(A1;5)}$$

$$\text{et pour } m = 1, 2, 3, 4, \ldots \begin{cases} \hat{p}^{\prime o}_m(L) = (-1)^{m-1}\sqrt{\lambda^{o}_m} \\ \hat{v}^{o\prime}_m(L) = \ell\sqrt{\lambda^{o}_m} \end{cases}$$

La construction des fonctions de perturbation est ainsi ramenée au calcul des solutions des équations de Webster (A1;3) et (A1;4), pour les conditions aux limites (A1;5).

On remarque que $\hat{p}^{o}_{2n-1}(x)$ est proportionnel à $p^{o}_{2n-1}(x)$: la $n^{\text{è}}$ fonction propre de la pression sonore dans le conduit ouvert aux lèvres, et que $\hat{v}^{o}_{2n}(x)$ est proportionnel à $v^{o}_{2n}(x)$: la $n^{\text{è}}$ fonction propre de la vitesse volumique dans le conduit fermé aux lèvres.

ANNEXE 2.

1. Pour le calcul des fonctions d'aire relatives à la voyelle /e/, dans le cadre des modèles (42 a et b) (voir fig. 3 et 4), nous avons utilisé les fonctions d'essai suivantes ($I_n = 3$ pour n = 1, ..., 6) :

$$p_n(z) = \begin{cases} C_1^{(n)} \cos(z/\ell) + C_2^{(n)} \cos\frac{3z}{\ell} + C_3^{(n)} \cos\frac{5z}{\ell} & \text{pour } n = 1, 3, 5 \\ C_1^{(n)} \cos z + C_2^{(n)} \cos 2z + C_3^{(n)} \cos 3z & \text{pour } n = 2, 4, 6 \end{cases}$$

Les matrices W_n sont ainsi de dimension (3 x 3) et les déterminants : $D_n = \det [W_n]$ se calculent rapidement.

Pour cette voyelle en particulier, nos calculs faisant usage de fonctions d'essai comportant plus de trois termes (jusqu'à $I_n = 8$), n'ont pas révélés, quant au résultat, de différences notables vis-à-vis de la fonction d'aire obtenue pour $I_n = 3$.

Dans le cas du modèle (42a), nous avons (voir (38) et (40)) :

$$\varphi_k(z) = \cos kz \quad ; \quad \alpha_{km} = \delta_{k,m} = \begin{cases} 0 & \text{si } k \neq m \\ 1 & \text{si } k = m \end{cases}$$

et l'égalité (41) se réduit à :

$$\frac{\partial a(\ell)}{\partial \beta_k} = \frac{1}{2} \left[a(\ell-k) + a(\ell+k) \right]$$

Le calcul des dérivées partielles $\frac{\partial D_n}{\partial \beta_k}$, (31), est donc relativement simple (voir (36)).

Dans le cas du modèle (42b), nous n'avons pas exprimé les $\varphi_k(z)$:

$$\varphi_k(z) = T_k\left(\frac{2z}{\pi} - 1\right)$$

sous la forme (40) : nous avons fait usage des formules d'interpolation de Tchebytcheff pour calculer directement l'intégrale (39) définissant $\frac{\partial a(\ell)}{\partial \beta_k}$.

A l'aide de la méthode de Newton-Raphson, la solution du système (28) a été obtenue après 6 itérations au départ de $\{\beta_k^o = o\}$ (conduit uniforme) pour une précision :

$$\sum_{n=1}^{6} \left(\frac{\overline{D}_n}{\lambda_n^o}\right)^2 < 10^{-5} ,$$

et ceci dans le cadre de chacun des deux modèles.
Le temps de calcul, sur CDC 6400, d'une itération est inférieur à 1 seconde. Une grande partie de ce temps est occupée par l'inversion de la matrice (6 x 6) G :

$$G_{n,k} = \frac{\partial D_n}{\partial \beta_k}$$

Ce temps peut être réduit si, après une première estimation de la matrice inverse G^{-1}, celle-ci est réajustée à chaque itération selon la méthode de C.G. Broyden [9], sans plus devoir effectuer d'inversion.

2. Le modèle paramétrique, dans le cadre duquel nous avons calculé la fonction d'aire relative à la voyelle /a/ (fig.5), comprend 7 paramètres et une contrainte explicite résultant de la mesure de l'aire aux lèvres ($z=\pi$) et à la glotte ($z=o$) :

$$A(\{\beta_k\}; z) = A_o(z) \exp\left[\sum_{k=1}^{7} \beta_k T_k\left(\frac{2z}{\pi} - 1\right)\right] \qquad (A2;1)$$

$$A(\pi)/A(o) = 1{,}75$$

$T_k(y)$ étant le polynôme de Tchébytcheff de degré k, et avec

$$A_o(z) = \begin{cases} \exp[-0{,}9 \cdot P] & \text{pour } 0 \le z < 0{,}1\,\pi \\ \exp[0{,}1\,P] & \text{pour } 0{,}1\pi \le z \le \pi \end{cases}$$

où P est la valeur du saut de $\log A_o(z)$ en $z = 0{,}1$, c'est-à-dire au niveau de l'épiglotte (voir fig. 6). Nous avons choisi P = 1,5.

Sachant que : $T_k(-1) = (-1)^k$ et $T_k(1) = 1$, la contrainte s'exprime aisément en fonction des $\{\beta_k\}$:

$$\log(1{,}75) - P = 2(\beta_1 + \beta_3 + \beta_5 + \beta_7)$$

Nous avons utilisé ici des fonctions d'essai à 5 termes ($I_n = 5$) qui conduisent à des matrices W_n de dimensions 5x5. La fonction d'aire calculée a été obtenue par minimisation de l'expression :

$$\sum_{n=1}^{6} \bar{D}_n^2 + w\left[2(\beta_1+\beta_3+\beta_5+\beta_7) - \log(1{,}75) + P\right]^2$$

avec un poids w = 5.
Les valeurs recherchées des paramètres $\{\beta_k\}$ sont donc solution du système :

$$\sum_{n=1}^{6} \bar{D}_n \frac{\partial \bar{D}_n}{\partial \beta_k} + w\left\{2\left[\beta_1+\beta_3+\beta_5+\beta_7\right) - \log(1{,}75) + P\right\}\left(1-(-1)^k\right) = 0 \; ; \; k = 1, \ldots, 7$$

Ce système a été résolu à l'aide du programme FORTRAN écrit par M.J.D. Powell [9]. La solution vérifie les 6 premières équations : $\bar{D}_n = o$ (n = 1, ..., 6), avec une précision de l'ordre de

$$\sum_{n=1}^{6} \left(\frac{\bar{D}_n}{\lambda^o_n}\right)^2 < 2.10^{-2}$$

ce qui indique que le modèle (A2;1) est capable de générer une fonction d'aire vérifiant relativement bien les contraintes acoustiques données.

3. Outre la méthode de Powell, il existe évidemment d'autres méthodes applicables à la recherche d'une solution optimale du système (28) au sens des moindres carrés. Signalons en particulier la méthode de K. Levenberg (méthodes des moindres carrés amortis) [9,15] , qui se traduit par un algorithme plus simple que celui de Powell, mais qui exige un nombre d'itérations généralement plus élevé .

Quelle que soit la méthode itérative utilisée pour résoudre le système (30) ou (28), il faut partir d'une première approximation. Si celle-ci est éloignée de la solution, il peut se faire que le processus itératif diverge ou converge vers une solution qui n'est pas celle souhaitée (il en est généralement ainsi lorsque le modèle paramétrique est mal adapté à générer une fonction d'aire du type recherché). Il convient alors de procéder par étapes (comme c'était le cas pour la méthode des perturbations [Heinz]).

Calculons les valeurs propres relatives à la première approximation, soit $\{\lambda_k^o\}$. Donnons leur des accroissements $\{\Delta\lambda_n\}$, dans le sens désiré et résolvons le problème associé à ces valeurs propres intermédiaires. La solution obtenue sert ensuite de première approximation à l'étape suivante définie par un nouvel accroissement des valeurs propres. En prenant des $\Delta\lambda_n$ suffisament petits, on espère ainsi posséder, pour chaque problème intermédiaire, une première approximation proche de la solution.

Dans le cas de la voyelle /a/, nous avons procédé de la sorte en deux étapes, au départ des $\{\beta_k^o = o\}$.

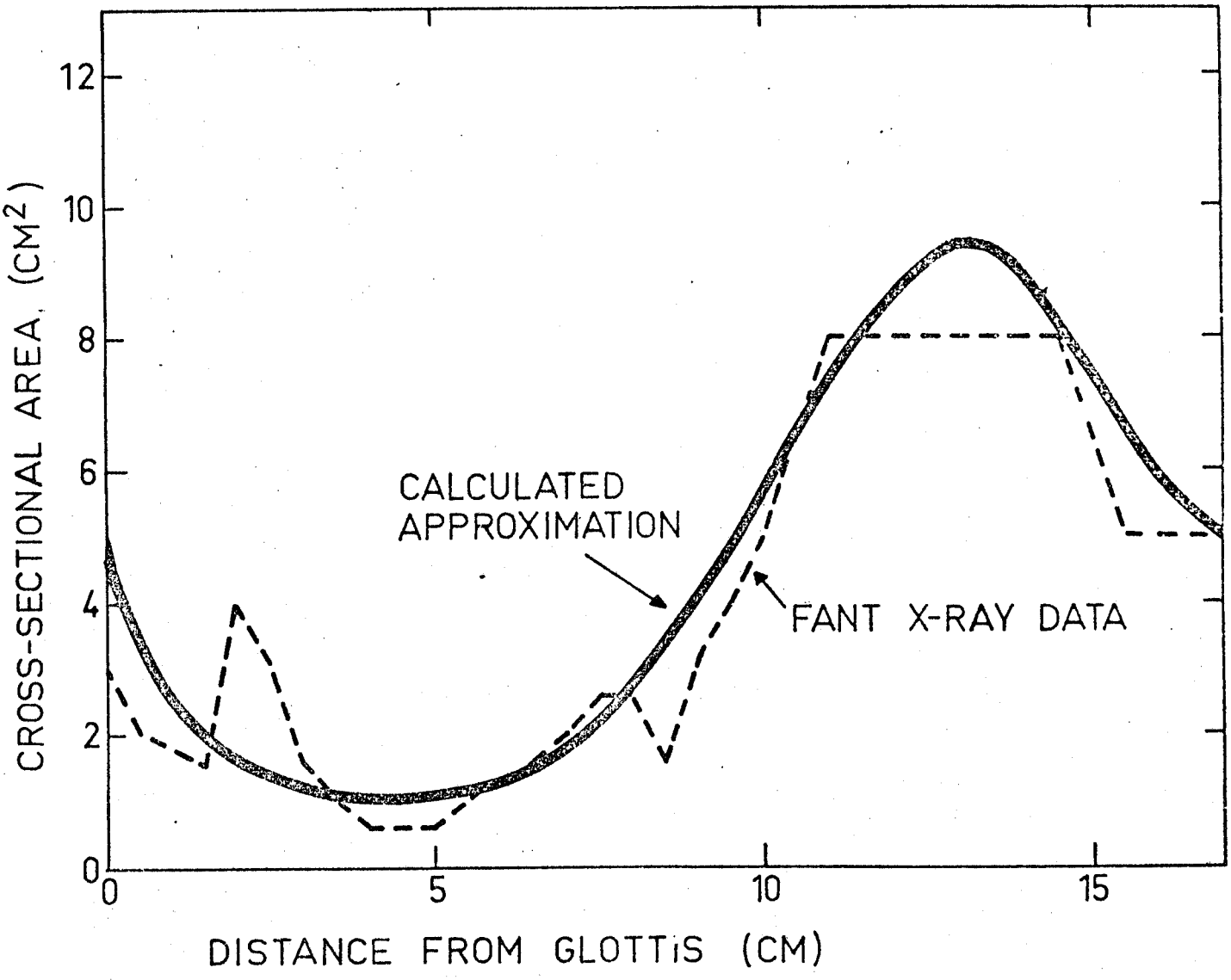


Fig. 1 (d'après Heinz [4]) : Fonctions d'aire relatives aux données de Fant [11] pour la voyelle /a/.
Trait plein : solution obtenue par la méthode des perturbations [Heinz].

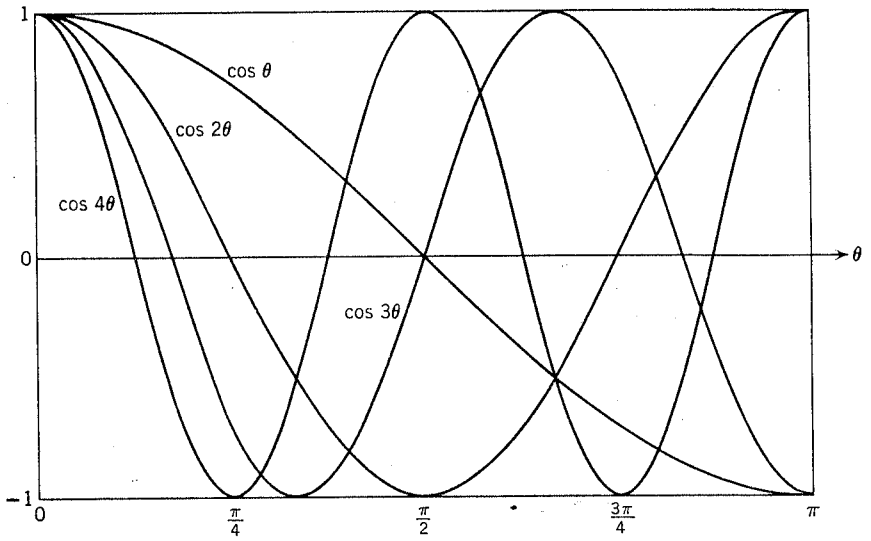


Fig. 2a : cos kθ, k = 1, 2, 3, 4, pour $0 \leq \theta \leq \pi$

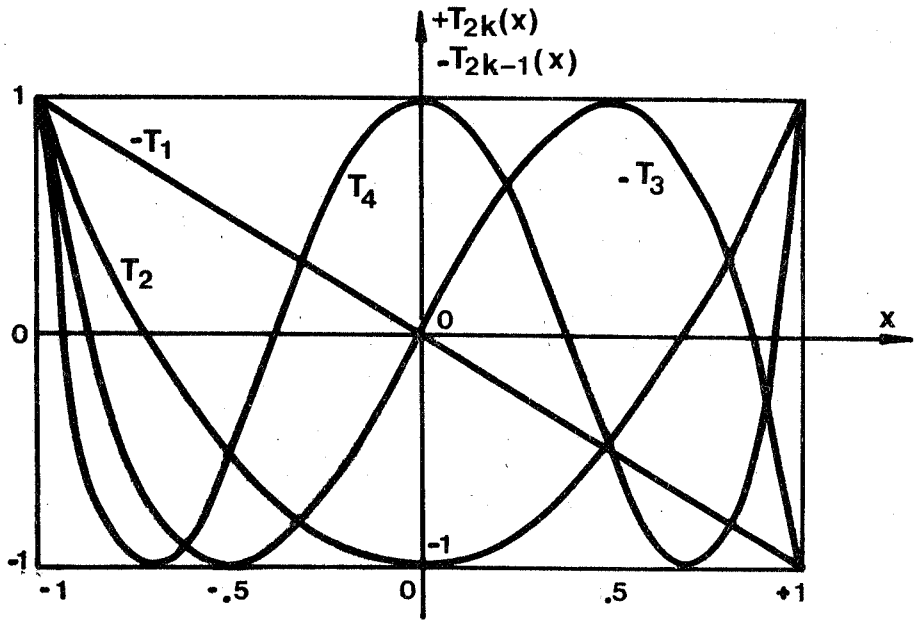


Fig. 2b : Polynômes de Tchebytcheff : $-T_1(x)$, $+T_2(x)$, $-T_3(x)$, $+T_4(x)$; $-1 \leq x \leq 1$.

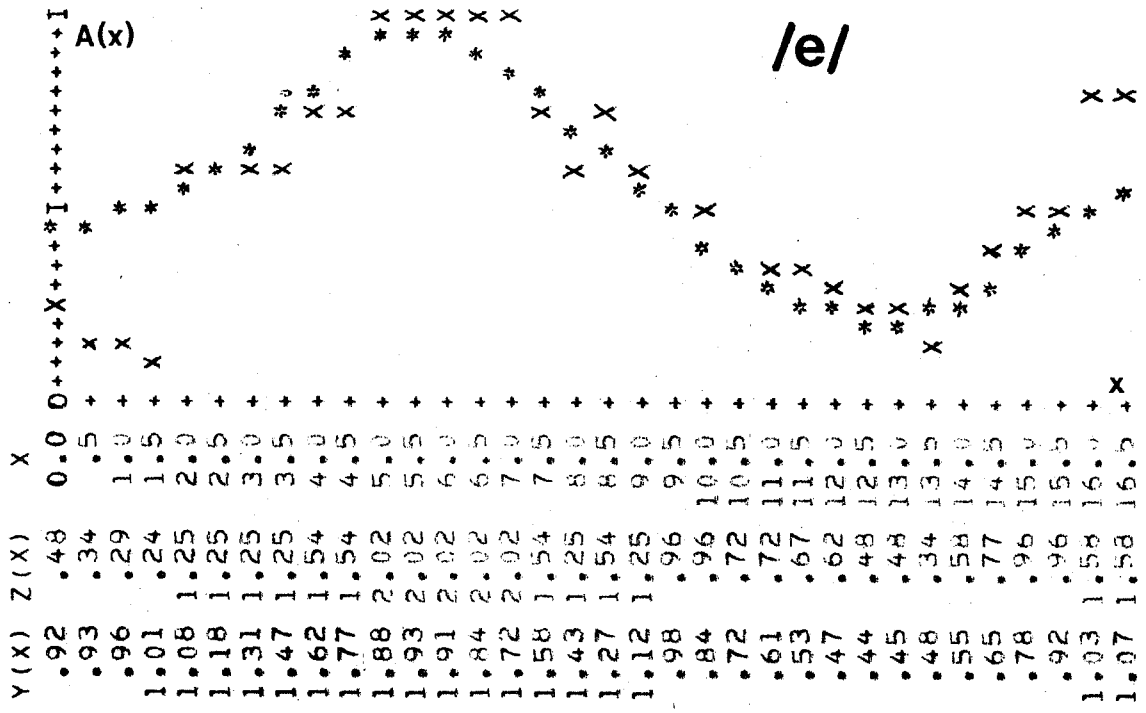


Fig. 3 : Fonctions d'aire normalisées relatives aux données de Fant pour la voyelle /e/ -
× : valeurs mesurées par Rayons X [Fant] .
* : " calculées dans le cadre du modèle 42a : $\varphi_k^{(x)} = \cos \frac{k\pi x}{L}$,

$o \leq x \leq L$.

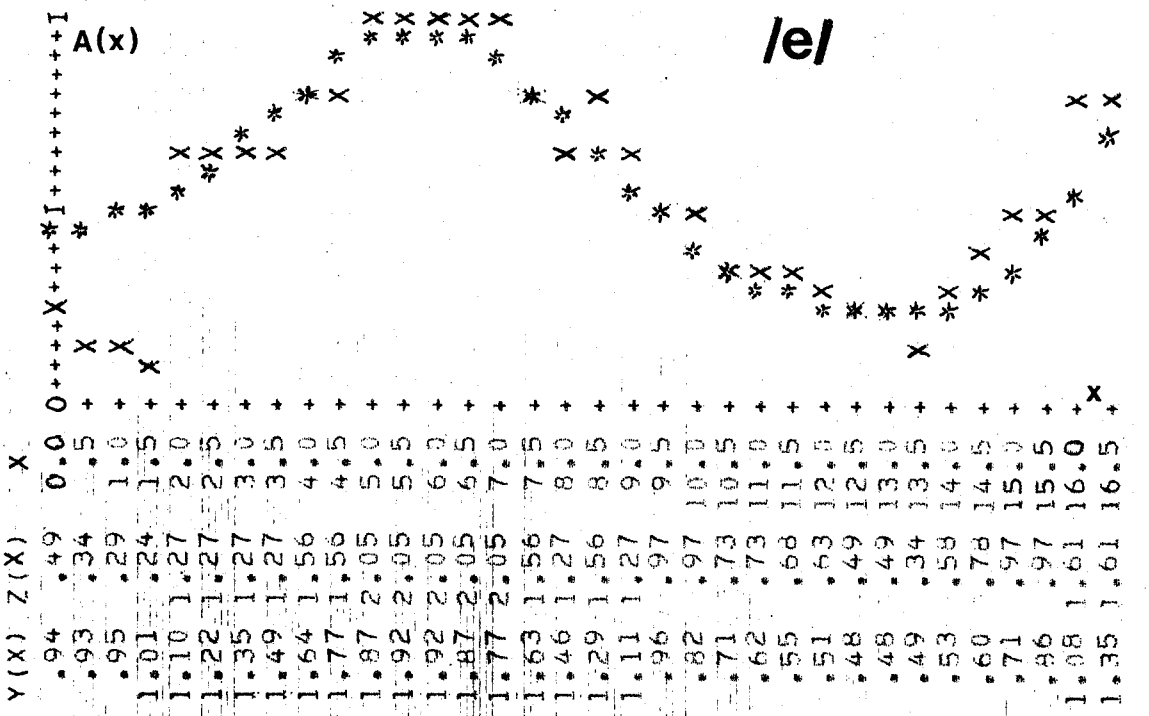


Fig. 4 : Fonctions d'aire normalisées relatives aux données de Fant pour la voyelle /e/.
× : valeurs mesurées par Rayons X [Fant].
* : " calculées dans le cadre du modèle 42b : $\varphi_k^{(x)} = T_k(\frac{2x}{L} - 1)$;

$o \leq x \leq L$.

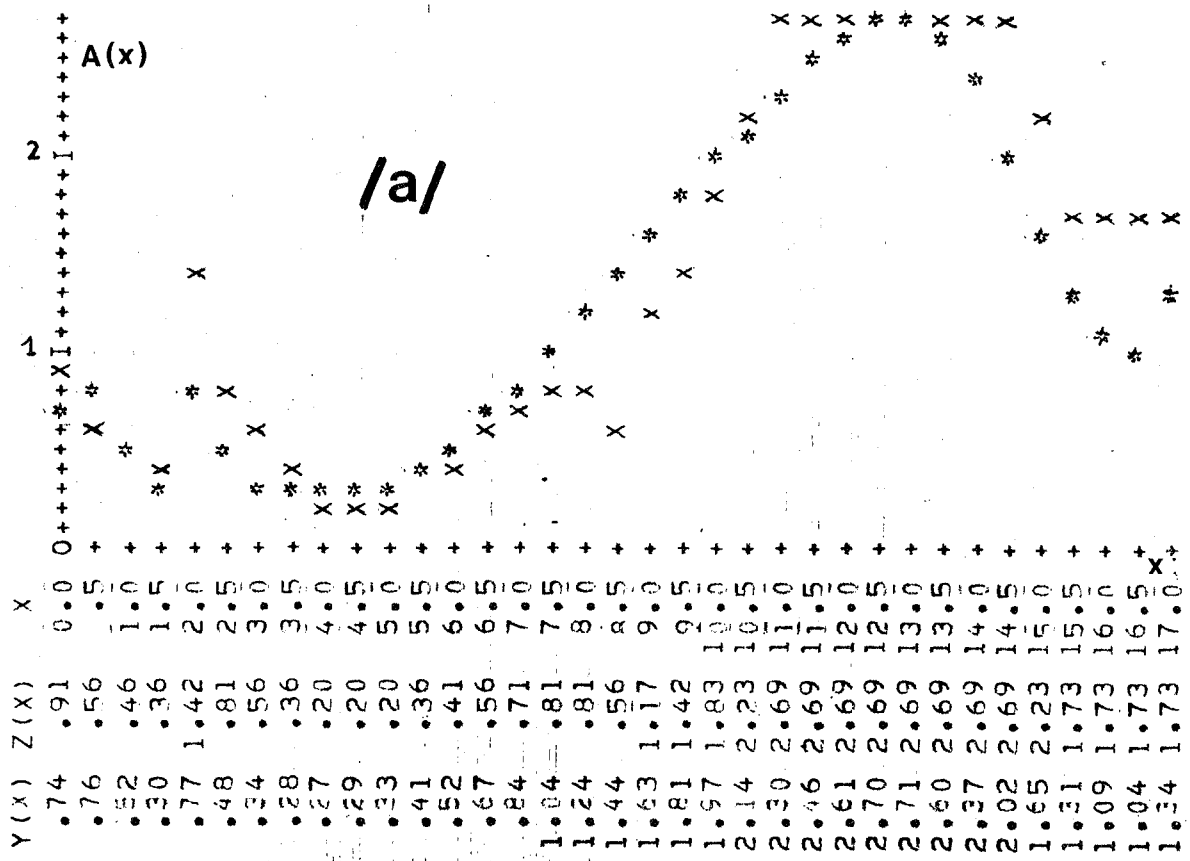


Fig. 5 : Fonctions d'aire normalisées relatives aux données de Fant pour la voyelle /a/.
✕ : valeurs mesurées par Rayons X [Fant];
✱ : " calculées dans le cadre du modèle A2.1.

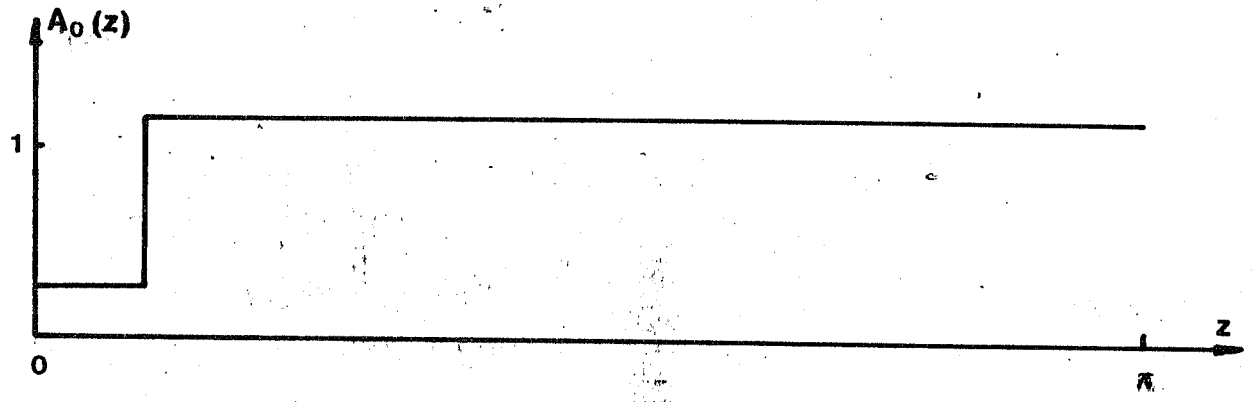


Fig. 6 : Fonction $A_o(x)$ du modèle A2.1.

BIBLIOGRAPHIE

1. M.R. Schroeder and P. Mermelstein : "Determination of Smoothed Cross-Sectional Area Function of the Vocal Tract from Formant Frequencies ", 5e Congrès Int. d'Acoustique, vol. 1a, papier A-24 (Liège, 1965, D.E. Commins, Editeur).
2. P. Mermelstein : "Determination of the Vocal-Tract Shape from Measured Formants Frequencies", J. Acoust. Soc. Am. 41, 1283-1294, (1967)
3. M.R. Schroeder : "Determination of the Geometry of the Human Vocal Tract", J. Acoust.Soc. Am. 41, 1002-1010 (1967).
4. J. Heinz : "Perturbation Functions for the Determination of Vocal Tract Area Functions from Vocal Tract Eigen values", S.T.L.-Q.P.S.R., (April 1967), 1-14.
5. P.M. Morse : Vibration and Sound (Mc Graw-Hill, New York, 1948).
6. G. Ungeheuer : "Elemente liner akustischen theorie der vokalartikulation". (Springer-Verlag, Berlin 1962).
7. B. Friedman Principles and Techniques of Applied Mathematics. (John Willey, New York 1956).
8. B.L. Moiseiwitsch : "Variational Principles" (Interscience publ. John Willey, London 1966).
9. "Numerial Methods for Nonlinear Algebraic Equations" edited by P. Rabinovitz (Gordon and Breach, London 1970).
10. B. Gopinath and M.M. Sondhi : "Determination of the Shape of the Human Vocal Tract from Acoustical Measurements", The Bell Sys. Techn. Journal, (juillet-août 1970) pp. 1195-1214.
11. G. Fant : Acoustic Theory of Speech Production (Mouton & Co., The Hague, 1960).
12. K.N.S. Stevens and A.S. House, J. Acoust. Soc. Am. 27, 484-493 (1955)
13. S.E.G. Öhman, "Numerical Model of Coarticulation", J. Acoust. Soc. Am. 41, 310-320 (1967).
14. G. Borg Acta Math. 78, 1-96, (1946).
15. A. Hocquenghem :"Méthode des moindres carrés amortis", Mathematica, vol. 10 (33) 1, (1968), 75-83.

DETERMINATION DE LA FONCTION D'AIRE DU CONDUIT VOCAL A PARTIR DE LA REPONSE IMPULSIONNELLE AUX LEVRES

Monsieur R. DESCOUT
Département E.T.A. - C.N.E.T. - LANNION -

I. INTRODUCTION

Nous proposons ici une méthode permettant la détermination de la fonction d'aire $\mathcal{A}(x)$ du conduit vocal à partir de mesures acoustiques relevées aux lèvres.

En résumé, si l'on envoie à l'extrémité d'un tuyau acoustique non uniforme, une impulsion connue, la mesure de l'onde retour doit nous permettre de déterminer les différentes sections de ce tuyau.

Pour la résolution de ce problème, deux types d'approches ont déjà été effectués :

- Dans le domaine fréquenciel, à partir des fréquences de formants (Kadokawa, Mermelstein) ; à partir de la mesure de l'impédance aux lèvres (Schröeder, Paige et Zue) ; ou en utilisant une méthode de perturbations (Mermelstein, Heinz, Jospa) ;

- Dans le domaine temporel : Sondhi et Gopinath déterminent $\mathcal{A}(x)$ à partir de la mesure de la réponse impulsionnelle du conduit vocal, mais aucun résultat d'expérience n'a encore été fourni.

Nous proposons ici une méthode dans le domaine temporel qui présente les avantages suivants :

- Rapidité des mesures ;
- Les conditions aux limites (à la glotte en particulier), ne sont pas imposées ;
- On s'affranchit totalement de la source vocale, puisque le sujet ne parle pas.

II. METHODE

1. Modèle de Kelly et Lochbaum

Nous nous placerons dans l'analogie acoustique/électrique du système :

- Pression $\longleftrightarrow$ tension
- Vitesse $\longleftrightarrow$ courant

Le modèle utilisé est celui proposé par Kelly et Lochbaum (Stockholm - 1962) : celui - ci est composé d'une succession de tubes cylindriques uniformes d'égale longueur et d'aire variable.

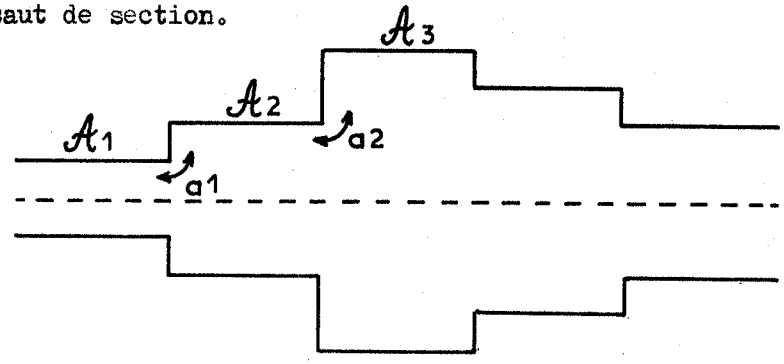
Le modèle est supposé sans pertes et ne fait pas intervenir de conduit nasal.

Les coefficients caractéristiques dans ce modèle sont les coefficients de réflexion a_i, liés aux rapports entre deux aires adjacentes $\mathcal{A}_i$ et $\mathcal{A}_{i+1}$ par la relation :

$$a_i = \frac{\mathcal{A}_i - \mathcal{A}_{i+1}}{\mathcal{A}_i + \mathcal{A}_{i+1}}$$

(Ce coefficient étant l'analogue électrique du coefficient de réflexion aux bornes d'un quadripôle relié à une impédance de charge).

Le problème consiste à déterminer les coefficients a_i caractéristiques de chaque saut de section.



Nous supposerons aussi que la propagation du son se fait en mode plan.

Considérons une onde incidente, sous réserve que sa fréquence soit inférieure à 5 kHz (à ces fréquences la propagation en mode transverse n'est pas possible), cette onde traverse une section sans affaiblissement, car le conduit est supposé sans perte, et elle subit un retard égal à l/c si l est la longueur de la section, et c la vitesse du son. Pression et vitesse dans le tube sont exprimés en fonction d'une onde aller A(x,t) se dirigeant des lèvres vers la glotte, et d'une onde retour B(x,t) se dirigeant de la glotte vers les lèvres. Au moment de la discontinuité entre deux sections l'onde subit des réflexions et transmissions multiples dont les coefficients sont liés au rapport entre deux aires adjacentes.

Les amplitudes de A et de B sont calculées pour chaque section par :

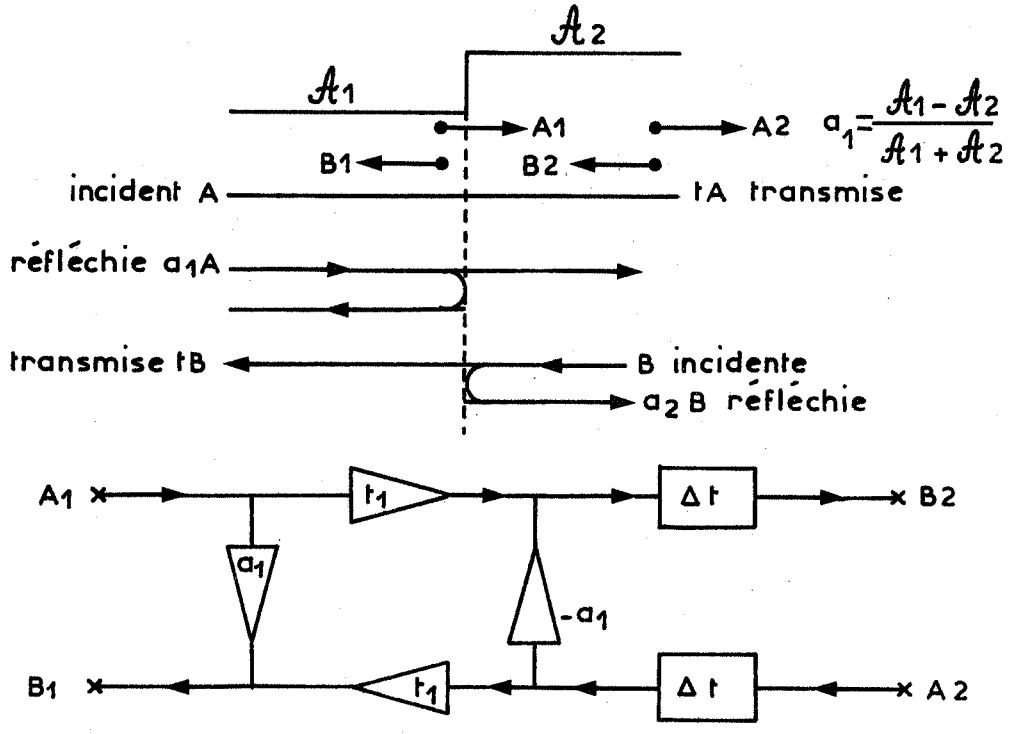
$A = \frac{1}{2}(p + ZV)$ p = pression acoustique $p = A + B$

$B = \frac{1}{2}(p - ZV)$ V = Vitesse $V = \frac{1}{Z}(A - B)$

Z étant l'impédance acoustique du tube au point considéré :

$Z = \rho c/A$

A = Aire de la section
ρ = masse volumique de l'air
c = Vitesse du son



Posons :

$$A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} \qquad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}$$

On appelle S, la matrice de répartition telle que :

$$B = SA$$

L'expression de S est caractéristique de la section, comme celles-ci sont réciproques : $S_{21} = S_{12}$.

De plus, le conduit étant supposé sans pertes :

$$S_{11} \cdot S_{22} + S_{12} S_{21} = 1$$

Conséquence :

$$\begin{pmatrix} B_1 \\ B_2 \end{pmatrix} = \begin{pmatrix} a & t \\ t & a \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}$$

où a est le coefficient de réflexion au passage entre deux sections, et $t = \sqrt{1 - a^2}$, le coefficient de transmission correspondant.

2. Cas où l'onde incidente est une impulsion infiniment brève

Supposons maintenant que :

$$\begin{cases} A(o,t) = 1 & \text{pour } t = o \\ A(o,t) = o & \text{pour } t \neq o \end{cases}$$

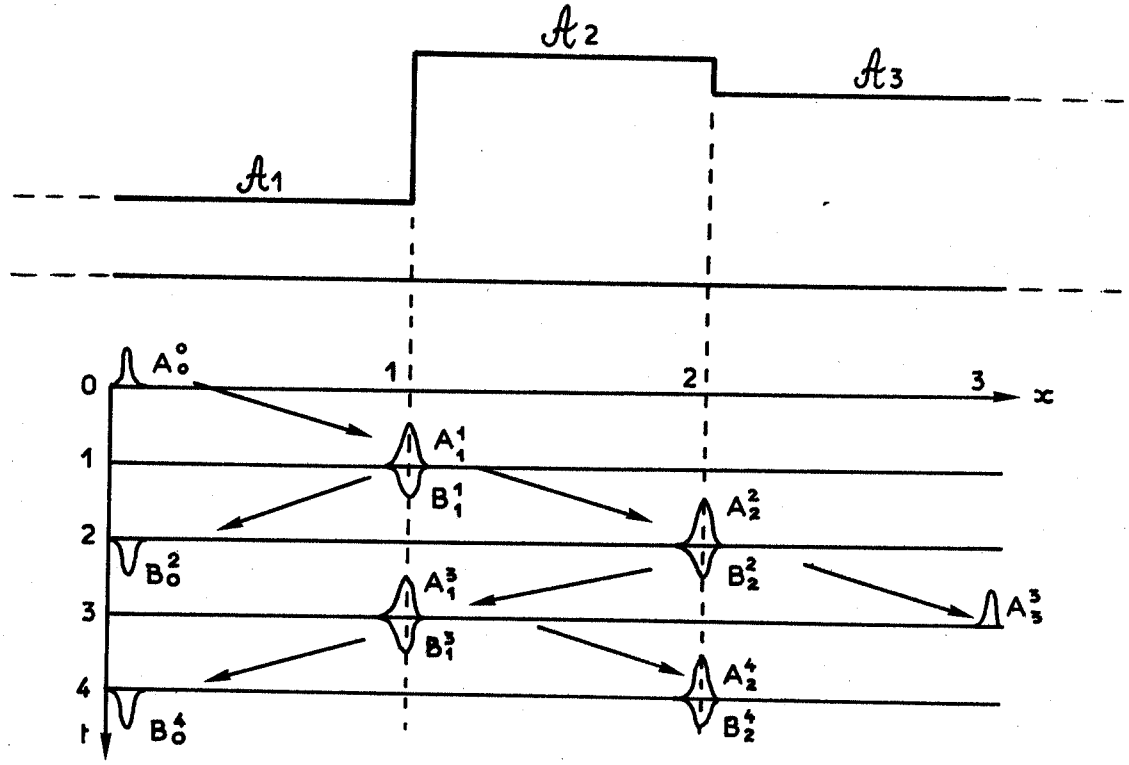
Soit une onde incidente constituée par une impulsion égale à l'unité et considérons l'onde progressive dans chaque section.

Posons :

A_i^j = Valeur de l'onde aller, en terme de pression, à la section $i\Delta x$ et au temps $j\Delta t$.

B_i^j = Valeur de l'onde retour au même point et au même instant.

Le schéma de répartition dans l'espace et dans le temps est alors le suivant :



Dans ces conditions, à la traversée d'une section, on peut écrire :

$$(1)\quad \begin{cases} A_i^j = t_i\ A_{i-1}^{j-1} - a_i\ B_{i+1}^{j-1} & (1a) \\ B_i^j = t_i\ B_{i+1}^{j-1} + a_i\ A_{i-1}^{j-1} & (1b) \end{cases}$$

Considérons le conduit vocal initialement au repos. A l'instant $i\Delta t$ toutes les section au delà de la section i restent non perturbées (puisque l'onde n'en est qu'au point $i\Delta x$).

Par conséquent : $B_i^i = a_i\ A_{i-1}^{i-1}$

et donc $$a_i = \frac{B_i^i}{A_{i-1}^{i-1}}$$

Supposons que l'on puisse calculer à chaque instant les valeurs des ondes A et B pour toutes les sections précédant la section d'analyse i, on en déduira :

$$A_{i-1}^{i-1}$$

B_i^i sera connue à partir de B_o^{2i}, valeur de l'onde retour aux lèvres au temps $2i\Delta t$: on ne peut en effet connaître les caractéristiques de la section i qu'au bout du temps $2i\Delta t$ mis par l'onde pour aller en i et en revenir.

On en déduit un algorithme général de calcul :
A partir des équations (1) on aboutit à l'expression de B_i^j suivante

$$(2)\quad B_K^{2i-K} = \frac{1}{t_{k-1}} \left[B_{K-1}^{2i-K+1} - a_{K-1} \; A_{K-2}^{2i-K} \right] \quad \text{pour } 0 \leqslant K \leqslant 2i$$

Dans (2) on ne fait intervenir que les sections précédemment étudiées aux pas $(k-1)\Delta t$ et $(k-2)\Delta t$ aux instants :

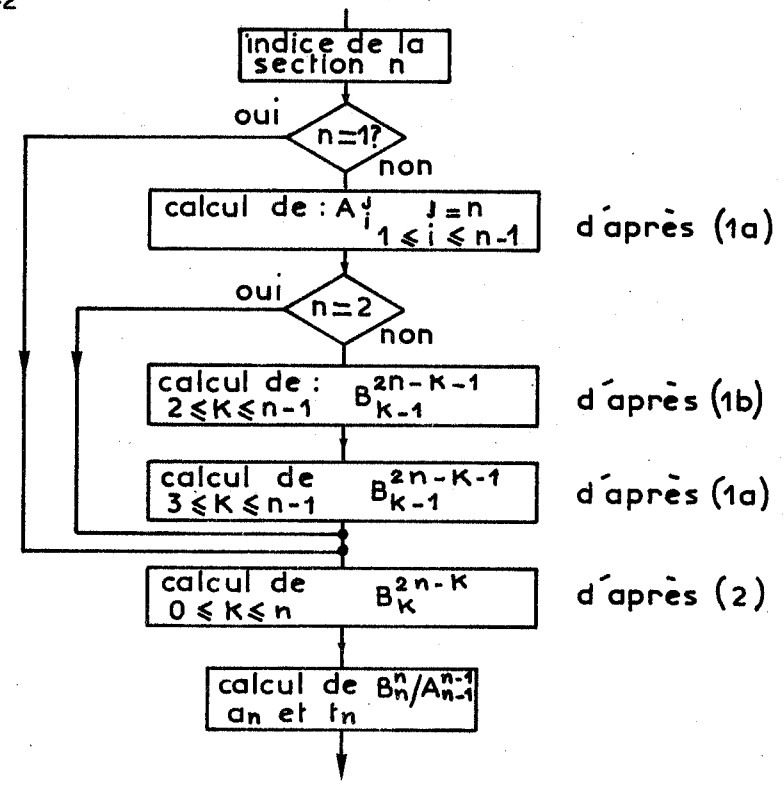
$$(2i - k + 1)\Delta t \quad \text{et} \quad (2i - k)\Delta t.$$

Ce qui nécessite le calcul préliminaire de :

B_{k-1}^{2i-k+1} à partir de (1a)

et

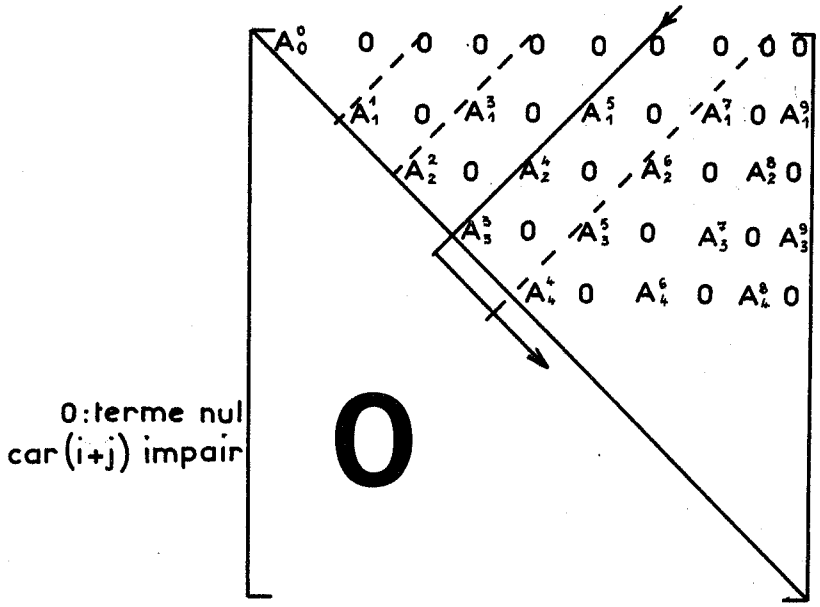
A_{k-2}^{2i-k} à partir de (1b)



Remarquons que $A_i^j = 0$ et $B_i^{j'} = 0$ pour (i+j) impair

0 i j

La matrice des A_i^j (ou B_i^j) se présente sous la forme suivante :



L'algorithme se déroule de la façon indiquée sur le schéma.

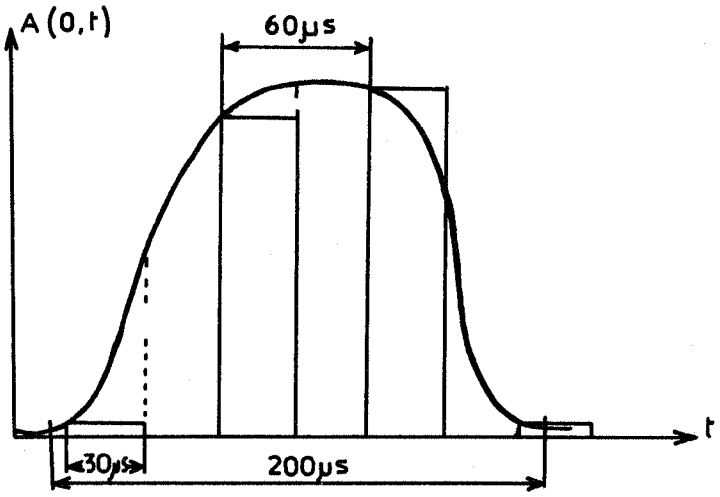
Enfin, $A_{i-1}^{i-1} = \prod_{k-1}^{i-1} t_k$

Dans cette hypothèse on peut calculer a_i pour toutes les sections d'étude.

3. Cas où l'onde n'est pas une impulsion de courte durée

Dans le cas ou l'onde aller est une impulsion de courte durée, cela représente à l'entrée adaptée du conduit, une impulsion de durée très inférieure au temps de transfert dans une cellule $\Delta t = \ell/c$ (30 µs dans le cas où ℓ = 1 cm).

Or nous avons vu que si la fréquence de l'onde incidente est inférieure à 5 kHz seul le mode plan peut se propager. Ceci représente une onde dont la durée est supérieure à 200 µs ; nous allons donc être amenés à échantillonner cette onde incidente, et si la durée de chaque échantillon reste inférieure à 30 µs, pour chacun d'entre eux, l'algorithme précédent sera valable.



Nous supposerons que le système linéaire, et que chacun des échantillons se propage indépendemment des autres dans le conduit.

Dans la mesure ainsi recueillie, les données d'onde retour comprendront les différentes réflexions dues à chacun des échantillons. Soit $s(i)$ l'onde retour mesurée à l'instant $i\Delta t$, $x(i)$ étant la réponse du système à l'instant $i\Delta t$ à une onde incidente unité "infiniment courte" au temps $t = 0$.
Supposons le système déterminé par les 4 échantillons :

α_o, α_1, α_2, α_3 aux instants o, $2\Delta t$, $4\Delta t$, $6\Delta t$.

Les mesures recueillies s'écriront :

$$s(2) = \alpha_o\, x(2)$$

$$s(2k) = \alpha_3\, x(2k-6) + \alpha_2\, x(2k-4) + \alpha_1\, x(2k-2) + \alpha_o\, x(2k)$$

Conséquence : $B_o^2 = x(2) = \dfrac{s(2)}{\alpha_o}$

$$B_o^4 = x(4) = \frac{s(4) - \alpha_1\, x(2)}{\alpha_o}$$

$$B_o^{2K} = x(2K) = \left[s(2K) - \alpha_3\, x(2K-6) - \alpha_2\, x(2k-4) - \alpha_1\, x(2k-2) \right] / \alpha_o$$

Connaissant ainsi les B_o^{2K} de proche en proche, il est possible de déterminer tous les a_k successifs de façon exacte, quelle que soit la forme de l'onde incidente et d'en déduire $\mathcal{A}_k$:

$$\mathcal{A}_{k+1} = \mathcal{A}_K \frac{1 - a_K}{1 + a_K}$$

III. MOYENS D'ETUDE

1. Temps de Mesure et de Calcul

Si l'on suppose que la longueur du conduit vocal est environ de L = 17 cm, avec des sections de ℓ= 1 cm, la durée de la mesure sera :

$$2T = \frac{2 \times 17.10^{-2}}{340} = 1 \text{ ms}$$

On peut considérer que pendant ce temps, même au cours d'une élocution continue, la configuration du conduit vocal reste stable.

La fréquence d'échantillonnage conditionne la précision avec laquelle on veut connaître chaque section. En effet, f doit être telle que :

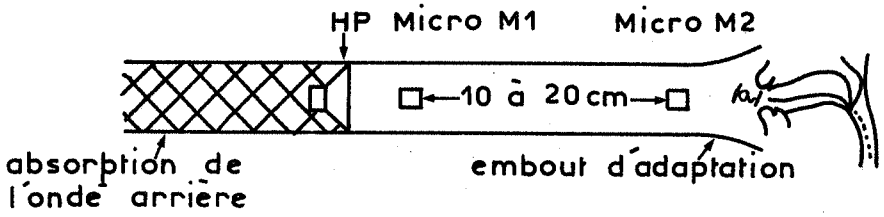
$$f = \frac{1}{2\Delta t}$$

où Δt est le temps de transfert de l'onde sonore dans une cellule. Si l'on veut connaître les sections tous les cm : f = 16 kHz.

Le temps de calcul pour la détermination des 17 paramètres (calculé pour le CII 10070 en simple précision) est environ de 20 ms.

2. Mesures acoustiques

Le système à mettre en place doit s'inspirer du schéma suivant :



Le diamètre du tube doit être tel que seule la propagation en mode plan puisse avoir lieu : soit environ 3, 4 cm pour des fréquences inférieures à 5,7 kHz.

La détermination des α_i, échantillons de l'onde incidente, se fait à partir du micro M1, convenablement synchronisé, le micro M2 est destiné à la mesure de l'onde retour.

L'embout au niveau de la bouche doit aussi apporter des termes d'erreur.

On détermine l'onde aller et retour en mesurant la pression par un micro de faible taille (Ø 6 mm) qui ne doit pas perturber le champ acoustique dans le tube.

IV. RESULTATS ET PLAN D'ETUDE

Une première expérience sur un modèle électrique statique à 3 sections de 2 cm, dans laquelle les mesures étaient faites sur un oscilloscope, permettait de déterminer le premier coefficient de réflexion, et le second avec une précision beaucoup plus faible. Mais on se heurte dans l'analogie électrique au fait qu'il faut connaître l'amplitude de l'onde retour, et le temps de transfert de chaque cellule L - C avec beaucoup de précision. Une seconde expérience est en cours sur un modèle acoustique, on a réalisé par moulage 8 sections correspondant à la voyelle /a/, avec le dispositif de mesure décrit précédemment.

Enfin, dans une première phase des travaux, on s'attachera à étudier des voyelles soutenues, puis dans une seconde étape, une élocution continue.

V. CONCLUSION

Un inconvénient de cette méthode (comme beaucoup d'autres dans ce domaine) est de ne pas faire intervenir les pertes dans le conduit vocal, ni l'influence du conduit nasal, bien que l'on connaisse ainsi la partie la plus mobile jusqu'au voile du palais.

Elle présente en outre les avantages suivants :

- Etude du phénomène dans le domaine temporel
- le temps de mesure est très court
- Il n'est pas nécessaire de connaître les conditions aux limites.

Il reste maintenant à juger de l'application pratique de cette méthode et des résultats.

Questions et réponses faisant suite à la communication

de R. DESCOUT

M. PAILLE (ENSERG)

Quelle espérance raisonnable mettez-vous dans ce modèle ? je précise ma question :

Vous avez dit que toutes les interventions précédentes négligeaient les pertes, comment pensez-vous les introduire dans ce modèle ? Et si vous pouvez obtenir d'assez bons résultats pour les voyelles, ne pensez-vous pas que cette méthode d'approche sera totalement inopérante dans le cas des consonnes et des fricatives en particulier ?

M. DESCOUT (CNET)

Pour les consonnes comme pour les fricatives, on doit pouvoir obtenir, par cette méthode, les valeurs des sections jusqu'à la constriction, ensuite on n'aura plus aucun résultat valable puisque pas une fraction de l'onde aller ne pourra passer. Mais est-ce si important ? Je pense que non, car avant la constriction, on est en présence d'un générateur de courant, et la forme du conduit vocal en amont de ce point n'importe pas. Ce qu'il faut connaître, c'est la position exacte de cette constriction et la forme du résonnateur en aval.

M. CARRE (ENSERG)

Je ne pense pas que ceci soit tout à fait juste, car la cavité qui est en amont va apporter des zéros à la fonction de transfert de l'ensemble. On n'a pas en amont un générateur d'air d'impédance constante.

.../...

M. GENIN (CNET)

Je voudrais préciser mon point de vue au sujet des constrictions : la constriction, et en particulier la fricative, va apporter des phénomènes non linéaires et des pertes, mais on oublie que dans le modèle proposé par M. DESCOUT, le locuteur ne parle pas, donc il n'y aura pas de courant d'air dans la zone rétrécie et par conséquent pas de pertes dynamiques au niveau de la constriction. Mais il faudra quand même tenir compte de celles dûes au facteur de forme !

Néanmoins, il y a un revers à la médaille : si le locuteur ne parle pas, on ne sait pas ce qu'il aurait dit s'il avait parlé, car il y a bien entendu un phénomène de feed-back auditif dans la parole qui n'est pas à négliger, et c'est justement au niveau des constrictions que l'on aura la plus faible précision.

D'autre part, au sujet des deux facteurs de perte de Flanagan, toutes les méthodes de détermination de la fonction d'aire font une hypothèse sur la loi de répartition des pertes, (par exemple : résistance linéique proportionnelle à la self ou à la capacité). Ici, on sera obligé de faire une autre hypothèse : une formulation simple dans notre cas est de dire que l'énergie ne se conserve pas au passage entre deux sections (ce qu'avait déjà proposé KELLY en 1962 sous la forme d'un facteur d'affaiblissement de l'onde au passage entre deux sections).

M. GUEGEN (E.N.S.T.)

Etant donné que le signal d'entrée et le signal de retour sont deux paramètres inconnus qu'il faut mesurer en même temps, comment comptez-vous séparer les deux signaux reçus par les micros M_1 et M_2 ?

M. DESCOUT (CNET)

Si les deux micros sont assez éloignés, et si l'on fait fonctionner le micro M_1 pendant les 200 premières micro secondes, on ne doit théoriquement mesurer que l'onde aller. Si le micro M2

.../...

est séparé de 10 à 20 cm du premier, ce dernier ne va recevoir que l'onde retour, si on le synchronise convenablement par rapport à M1.

M. GUEGEN (E.N.S.T.)

Il me semble que l'on est là en présence d'un problème d'identification de système à paramètres répartis, et il a toujours été reconnu que ce genre d'identification était très difficile. Vous envisagez l'un des résultats qui est connu pour résoudre ce genre de problème, mais il existe une forte sensibilité du modèle à la forme de l'entrée : vous avez choisi une entrée impulsionnelle, ne pensez-vous pas que les résultats vont être très sensibilisés par le fait que c'est une réponse impulsionnelle que vous calculez ? En effet, la forme de l'entrée est très importante : soit par exemple une sinusoïde, on va trouver sur le modèle un certain nombre de pôles, si on met une entrée différente, on va trouver des pôles qui sont parfois assez éloignés des premiers ; ne pensez-vous pas que ce sera un des problèmes qui se poseront ?

M. DESCOUT (CNET)

Dans notre raisonnement, nous avons supposé le système linéaire dans la bande de fréquences de 0 à 5 k Hz, c'est-à-dire que toutes les fréquences sont transmises de façon indépendante les unes des autres et sans affaiblissement, on peut donc supposer que la forme de l'entrée importe peu. Si à la place de l'impulsion, on envoyait un bruit pseudo-aléatoire, ne pensez-vous pas que nous aurions la même réponse ?

M. GUEGEN (ENST)

Pas de façon évidente ; en fait cela tient beaucoup aussi à la méthode de calcul, on peut proposer plusieurs méthodes : une déconvolution, un calcul inverse comme vous l'avez fait, une méthode du modèle où l'on met en parallèle le système et un modèle, de façon à minimiser une fonction d'erreur, ... en comparant toutes ces méthodes, on s'aperçoit de la sensibilité du système par rapport à l'entrée.

.../...

M. GENIN (CNET)

Je réponds à M. GUEGEN : Si l'on compare une impulsion et un signal pseudopériodique à large bande, ces deux signaux peuvent avoir le même spectre, ils différent essentiellement par leur relation de phase. Donc je ne comprends pas pourquoi on devrait s'attendre à des résultats différents puisque le système tel qu'on l'a proposé est linéaire.

M. GUEGEN (ENST)

C'est en théorie ! en fait le calcul fait apparaître des imprécisions : ce n'est pas dans le modèle qu'elles interviennent, mais dans la résolution. Si l'on utilisait une méthode du modèle, la méthode de minimisation de l'erreur intervient ; ici, on pourrait certainement poser le système d'équations différemment et au lieu d'avoir une matrice triangulaire, on pourrait avoir une matrice de Henkel qu'il faudrait inverser, c'est à ce niveau qu'interviendra la limitation, pas au niveau du modèle.

DETERMINATION DE LA FONCTION D'AIRE DU CONDUIT VOCAL PAR CODAGE PREDICTIF

M. I. I. EL-MALLAWANY
Département C.S.I. - C.N.E.T. - LANNION

Pages

DETERMINATION DE LA FONCTION D'AIRE DU CONDUIT VOCAL PAR CODAGE PREDICTIF

I. I. EL-MALLAWANY

CNET - LANNION

Dépt C.S.I.

I - INTRODUCTION.

L'intérêt porté aux études sur la compression de la parole tient à des considérations d'ordre technique et économique liées au problème d'écoulement d'un trafic toujours plus intense sur le réseau téléphonique. L'orientation initiale prise dans cette voie était dirigée vers une compression du signal capté; abstraction faite du processus articulatoire. Les méthodes [1] élaborées à cette fin avaient pour dénominateur commun une analyse spectrale. La raison en était la variation relativement faible du spectre du signal de parole en fonction du temps. Néanmoins, ces méthodes se sont révélées imprécises en raison de la nonstationnarité du signal et du caractère pseudo-périodique de la source d'excitation dans le cas des sons sonores. Une analyse à résolution spectrale suffisante d'un signal à l'aide de la transformée de Fourier doit, nécessairement, porter sur un intervalle de temps relativement long. Le prix en est l'impossibilité de suivre avec précision des évolutions rapides dans l'élocution. De l'analyse spectrale, dans le cas de sons sonores, il n'est obtenu des indications qu'aux fréquences harmoniques de la fondamentale (pitch), la quantité d'information recueillie devenant fort réduite pour des fréquences fondamentales élevées. Les méthodes d'analyse en synchronisme avec le pitch [2] n'apportent que des solutions partielles à ces défauts, et au prix de calcul complexe nécessitant un temps de calcul considérable.

Une nouvelle approche s'imposait. Les méthodes de compression par plages [3] étaient inapplicables; le signal de parole variant trop rapidement dans le temps. L'alternative résidait dans l'élaboration d'un modèle dont les paramètres ajustables sont liés à la fonction de transfert du conduit vocal (C.V.) et aux caractéristiques de la fonction d'excitation. Les applications d'une telle méthode dans l'étude de la phonation et de la perception de la parole sont multiples. Par conséquent, l'intérêt s'est orienté vers une meilleure connaissance du système articulatoire.

.../...

II - OBJET DE L'ETUDE.

L'objet de l'étude est de déterminer des paramètres articulatoires à partir du signal enregistré. Un moyen d'y parvenir comprend une étape intermédiaire, à savoir, la détermination de la fonction d'aire du C.V. Les paramètres articulatoires intéressent les études sur la synthèse par règles et la reconnaissance de la parole. La fonction d'aire permet d'en déduire des paramètres pour la commande d'un simulateur de conduit vocal. Cette dernière fonction peut être obtenue soit directement par cinéradiographie (rayon X), ou à partir de la réponse impulsionnelle aux lèvres (Acoustique), soit à partir du traitement du signal de parole (électrique). Seule cette dernière catégorie de méthode s'insère dans le cadre de notre étude et, plus précisément, les méthodes à base de codage prédictif. Cette méthode, proposée par ATAL [4 - 7], permet de décrire le signal de parole en fonction d'un modèle dont les paramètres sont ajustés périodiquement. Ces paramètres sont liés à la fonction de transfert du Conduit Vocal et aux caractéristiques de la source d'excitation.

III - LE MODELE.

L'analyse de la structure d'un signal est un préalable à toute définition d'un modèle. Deux cas sont à distinguer dans un signal de parole :

- Les sons sonores : L'excitation du C.V. par les cordes vocales prend la forme d'une suite de créneaux presque périodiques.
- Les sons sourds : L'excitation est engendrée par le passage turbulent de l'air à travers des constrictions du C.V.

Le modèle de C.V. le plus simple est un filtre numérique linéaire à coefficients variables. Dans la mesure où les variations du conduit vocal peuvent être approximées par une succession de configurations stationnaires, il est possible de définir une fonction de transfert en la variable complexe z pour le conduit. Dans le cas de sons sonores non nasalisés cette fonction ne comprendra que des pôles. Dans le cas contraire (sons nasalisés et sons sourds), il y aura en plus des antirésonances. Ces zéros, situés à l'intérieur du cercle unité, peuvent être remplacés par des pôles (dont le nombre dépendra de la précision requise), ce qui réduit le modèle à un filtre numérique linéaire constitué de pôles exclusivement.

La transformation en z de la fonction d'excitation au niveau de la glotte pendant une période de pitch peut être approchée par un modèle à deux pôles de la forme :

$$U_g = \frac{K_1}{(1-z_a z^{-1})\,(1-z_b z^{-1})} \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

où K_1 est une constante

et z_a, z_b sont des pôles réels à l'intérieur du cercle unité.

.../...

Dans l'hypothèse où le rayonnement aux lèvres peut être assimilé à celui d'une source sphérique, il vient que le rapport pression au capteur/vitesse volumique aux lèvres est égale à $K_2 (1-z^{-1})$, où K_2 est une constante fonction de l'amplitude de l'onde aux lèvres et de la distance lèvres-capteur.

La part de l'excitation et du rayonnement dans la fonction de transfert globale est de la forme :

$$\frac{K_1 K_2 (1-z^{-1})}{(1-z_a z^{-1})(1-z_b z^{-1})}$$

ou par approximation

$$\frac{K_1 K_2}{[1+(1-z_a) z^{-1}] [1-z_b z^{-1}]} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

l'erreur ainsi introduite étant

$$\frac{K_1 K_2 z^{-2}(1-z_a)}{(1-z_a z^{-1}) [1+(1-z_a)z^{-1}](1-z_b z^{-1})}$$

z_a étant voisin de l'unité, la valeur de l'erreur demeure faible sur la bande de fréquence considérée.

Il vient que le modèle retenu pour représenter l'ensemble C.V., rayonnement et excitation est un filtre récursif unique. Par conséquent, le problème de la séparation de la contribution du C.V. de l'influence de la fonction d'excitation ne se pose pas au niveau de la détermination des paramètres.

Le modèle de génération de la parole est représenté sur la figure : 1. Deux générateurs font fonction de source d'excitation, l'un émettant des impulsions d'amplitude et de périodicité variables pour les sons sonores, l'autre un bruit blanc pour les sons sourds. Le filtre récursif ce décompose en deux opérations : une prédiction, $\hat{s}_n$, obtenue à l'aide d'un filtre transversal à p retards unitaires donc portant sur les p derniers échantillons du signal de sortie s_{n-1} et une addition $(s_n + \delta_n)$, où δ_n est la valeur de la source d'excitation à l'instant nT. La sortie du filtre récursif à l'instant nT est de la forme

$$s_n = \sum_{k=1}^{p} a_k s_{n-k} + \delta_n \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3)$$

où les coefficients de prédiction a_k sont ceux du filtre récursif.

.../...

La fonction de transfert du filtre linéaire de la figure : 1 est

$$H(z) = 1/ (1- \sum_{k=1}^{p} a_k z^{-k}) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (4)$$

Le filtre numérique sera stable, si ces pôles sont situés à l'intérieur du cercle unité du plan Z.

Le nombre de coefficients nécessaire à une représentation suffisamment précise d'un intervalle de parole dépend du nombre de résonances et d'antirésonances du C.V. dans la bande de fréquence considérée et de l'influence de l'excitation et du rayonnement. Il a été démontré, ci-avant, qu'une représentation à base de deux pôles suffisait pour tenir compte de cette dernière influence. La condition d'une représentation adéquate des poles de la fonction de transfert du C.V. est précisée dans l'Annexe. Il y est démontré que la durée de la mémoire du filtre de prédiction doit être égale à deux fois le temps de propagation d'une onde de la glotte aux lèvres . A titre indicatif, si l'on échantillonne à une fréquence $f_e = 8KH_z$ ($T = 1/f_e$) un signal émis par un C.V. de longueur égale à 17 cm, la mémoire du filtre de prédiction doit être d'environ 1ms, soit p=8. En y ajoutant les deux pôles d'excitation et de rayonnement on obtient p = 10. La valeur de p ainsi calculée n'est qu'une estimation. En général, p varie en fonction de f_e, du locuteur et des sons prononcés.

Les paramètres à déterminer dans un intervalle de temps pendant lequel le conduit vocal est presque stationnaire sont :

- les coefficients a_k de l'extrapolateur.
- la période du pitch
- la valeur efficace du signal
- un indicateur de sons sonores ou non.

L'intervalle en question peut être la période de pitch, mais d'une durée maximum qui peut être fixée à 5ms ou 10ms; période pendant laquelle le C.V. est sensiblement stationnaire.

IV - L'ANALYSE DE LA PAROLE.

IV-A : La Détermination des Coefficients de Prédiction :

La suite d'échantillons s_n du signal de parole d'une période de pitch ou d'une période d'adaptation (5 ou 10 ms) constitue la réponse impulsionnelle d'un filtre numérique linéaire, selon notre hypothèse. Ce sont les coefficients de ce filtre qui sont à déterminer. Or pour une période de pitch, exception faite du premier échantillon de l'intervalle, les valeurs des échantillons peuvent être obtenues par une prédiction portant sur les p échantillons précédents. L'erreur de prédiction e_n est,

.../...

par conséquent, la différence entre l'échantillon de parole s_n et l'approximation $\hat{s}_n$.

$$\hat{s}_n = \sum_{k=1}^{p} a_k \, s_{n-k} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (5)$$

et

$$e_n = s_n - \hat{s}_n = s_n - \sum_{k=1}^{p} a_k \, s_{n-k} \qquad \ldots\ldots\ldots\ldots\ldots \quad (6)$$

L'erreur moyenne quadratique $\overline{<e_n^2>}$ sera définie par la moyenne de e_n^2 sur la totalité de l'intervalle ($n = 1 \rightarrow m$), exception faite du premier échantillon ($n = o$) d'une période de pitch.

$$\overline{<e_n^2>} = \overline{<(s_n - \sum_{k=1}^{p} a_k \, s_{n-k})^2>} \qquad \ldots\ldots\ldots\ldots\ldots \quad (7)$$

ou

$$\sum_{n=1}^{m} e^2 = \sum_{n=1}^{m} s_n^2 + \sum_{n=1}^{m} \left(\sum_{k=1}^{p} a_k \, s_{n-k} \right)^2 - 2 \sum_{n=1}^{m} s_n \sum_{k=1}^{p} a_k \, s_{n-k}$$

Les coefficients a_k recherchés sont ceux qui rendent minimum $\overline{<e_n^2>}$, et sont déterminés en posant la dérivée partielle de $\overline{<e_n^2>}$ par rapport à chaque a_k égale à zéro, ce qui donne :

$$2 \sum_{n=1}^{m} s_{n-i} \sum_{k=1}^{p} a_k \, s_{n-k} - 2 \sum_{n=1}^{m} s_n \, s_{n-i} = 0 \quad ; \; i = 1,2,\ldots,p$$

ou

$$\sum_{k=1}^{p} \left(\sum_{n=1}^{m} s_{n-i} \, s_{n-k} \right) a_i = \sum_{n=1}^{m} s_n \, s_{n-i} \quad ; \; i = 1,2,\ldots,p$$

ou encore

$$\sum_{k=1}^{p} \varphi_{ik} \, a_k = \varphi_{io} \quad ; \; i = 1,2,\ldots,p \qquad \ldots\ldots\ldots\ldots \quad (8)$$

où

$$\varphi_{ik} = \overline{<s_{n-i} \, s_{n-k}>}$$

Le système d'équations linéaires (8) peut être mis sous forme matricielle

.../...

$$\Phi . A = \Psi \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (9)$$

où

Φ est une matrice symétrique définie positive

A est le vecteur colonne des coefficients a_k

Ψ est le vecteur colonne dont $\psi_i = \varphi_{io}$,

La détermination des paramètres a_k passe par la résolution de ce système d'équations linéaires. Il reste à vérifier la stabilité du filtre numérique calculé, plus précisément, que les pôles se situent à l'intérieur du cercle unité du plan Z (il existe un algorithme pour réaliser cette vérification [4]). Les pôles instables devront être corrigés. Dans le cas de sons sourds, la procédure de calcul ne diffère pas, sinon sur l'intervalle d'analyse qui est fixe, soit 5 ou 10 ms.

IV-B : Détection du PITCH :

Du modèle retenu (voir figure : 1) il vient que

$$S(z) = H(z).\ \Delta(z)$$

$$= \frac{1}{1-P(z)} .\ \Delta(z) \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (10)$$

où

$$P(z) = \sum_{k=1}^{p} a_k\, z^{-k}$$

et a_k les coefficients précis du filtre.

Le signal de prédiction à l'instant nT est lié aux coefficients b_i déterminés par la méthode décrite dans § IV-A, par la relation

$$\hat{s}_n = \sum_{i=1}^{p} b_i\, s_{n-i}$$

ou

$$\hat{S}(z) = B(z).\ S(z) \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (11)$$

.../...

L'erreur de prédiction est, par conséquent,

$$E(z) = S(z) - B(z).\ S(z)$$

$$= \frac{1-B(z)}{1-P(z)} .\ A(z) \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (12)$$

Si $B(z) \sim P(z)$, le signal d'erreur e_n obtenu après filtrage par un passe-bas (coupure 1KHz) est une approximation du train d'excitation, δ_n , et peut servir à la détection du pitch. Dans un intervalle de sons sonores seul le premier échantillon d'une période de pitch ne peut être approximé par prédiction, d'où une pointe marquée dans la séquence $\{e_n\}$. C'est la détection de cette pointe qui donne la période du pitch. La figure : 2 illustre la méthode utilisée.

La valeur du bit-indicateur de sons sonores ou non dépend de la valeur du rapport (valeur efficace du signal de parole/valeur efficace du signal d'erreur de prédiction). Ce rapport étant bien plus faible (de l'ordre de 10) pour les sons sourds.

V - DETERMINATION DE LA FONCTION D'AIRE DU C.V.

Le C.V. peut être représenté par une succession de N sections cylindriques d'égale longueur et d'aire de section S_i (voir figure : 3). Si la longueur Δ est bien plus petite que la longueur d'onde, l'erreur introduite par cette approximation sera faible. A l'entrée de la n ième section, les composantes de la vitesse volumique sont $p_n(t)$ et $v_n(t)$ résultant de l'onde aller (glotte-lèvres) et de l'onde retour respectivement. Considérons la transmission du son entre deux sections adjacentes n et n+1, en faisant l'hypothèse que la source à une vitesse volumique constante à l'entrée de la première section. En appliquant les conditions de continuité de vitesse volumique et de pression, on obtient :

$$p_n\left(t-\frac{\Delta}{c}\right) - v_n\left(t+\frac{\Delta}{c}\right) = p_{n+1}(t) - v_{n+1}(t) \qquad \ldots\ldots\ldots \quad (13)$$

$$\left[p_n\left(t-\frac{\Delta}{c}\right) + v_n\left(t+\frac{\Delta}{c}\right)\right]\frac{\rho c}{S_n} = \left[p_{n+1}(t) + v_{n+1}(t)\right]\frac{\rho c}{S_{n+1}} \qquad \ldots \quad (14)$$

où ρc est l'impédance caractéristique de l'air.

.../...

A partir des équations (13) et (14) on obtient

$$p_{n+1}(t) = \frac{1}{1+r_n}\; p_n\left(t-\frac{\Delta}{c}\right) - r_n\, v_n\left(t+\frac{\Delta}{c}\right) \qquad \ldots\ldots\ldots \quad (15)$$

$$v_{n+1}(t) = \frac{1}{1+r_n}\left[-r_n p_n\left(t-\frac{\Delta}{c}\right) + v_n\left(t+\frac{\Delta}{c}\right)\right] \qquad \ldots\ldots\ldots \quad (16)$$

où $$r_n = \frac{S_n - S_{n+1}}{S_n + S_{n+1}}$$

L'application de la transformée en z aux équations (15) et (16) conduit à

$$\begin{bmatrix} P_{n+1}(z) \\ V_{n+1}(z) \end{bmatrix} = \frac{1}{1+r_n} \begin{bmatrix} z^{-1/2} & -r_n z^{1/2} \\ -r_n z^{-1/2} & z^{1/2} \end{bmatrix} \begin{bmatrix} P_n(z) \\ V_n(z) \end{bmatrix} \qquad \ldots\ldots\ldots \quad (17)$$

où $P_n(z)$ et $V_n(z)$ sont les transformées en z de $p_n(t)$ et $v_n(t)$ respectivement avec $z = \exp\left[j\,\omega(2\Delta/c)\right]$. La relation inverse de l'équation (17) s'écrit :

$$\begin{bmatrix} P_n(z) \\ V_n(z) \end{bmatrix} = \frac{1}{1 - r_n} \begin{bmatrix} z^{1/2} & r_n z^{1/2} \\ r_n z^{-1/2} & z^{-1/2} \end{bmatrix} \begin{bmatrix} P_{n+1}(z) \\ V_{n+1}(z) \end{bmatrix} \qquad \ldots\ldots\ldots \quad (18)$$

ou encore

$$H_n(z) = Q_n(z)\, H_{n+1}(z) \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (19)$$

Il découle de l'équation (19) que

$$H_1(z) = \prod_{k=1}^{N} Q_k(z)\, H_{N+1}(z) \qquad \ldots\ldots\ldots\ldots\ldots \quad (20)$$

$$= W_N(z)\, H_{N+1}(z) \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (21)$$

.../...

Si

$$W_n(z) = \prod_{k=1}^{n} Q_k(z) = \begin{bmatrix} w_{11}^{(n)}(z) & w_{12}^{(n)}(z) \\ w_{21}^{(n)}(z) & w_{22}^{(n)}(z) \end{bmatrix} \quad \ldots\ldots\ldots \quad (22)$$

Il vient que

$$W_{n+1}(z) = W_n(z)\, Q_{n+1}(z) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (23)$$

On peut démontrer à partir de l'équation (22) que

$$J\,[W_n^t(z^{-1})]^{-1}\,J = W_n(z) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (24)$$

où

$$J = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

De l'équation (24) il vient

$$\begin{bmatrix} w_{11}(z) & w_{12}(z) \\ w_{21}(z) & w_{22}(z) \end{bmatrix} = \begin{bmatrix} w_{22}(z^{-1}) & w_{21}(z^{-1}) \\ w_{12}(z^{-1}) & w_{11}(z^{-1}) \end{bmatrix} \quad \ldots\ldots\ldots\ldots \quad (25)$$

ou

$$W_n(z) = \begin{bmatrix} w_{11}^{(n)}(z) & w_{21}^{(n)}(z^{-1}) \\ w_{21}^{(n)}(z) & w_{11}^{(n)}(z^{-1}) \end{bmatrix} \quad \ldots\ldots\ldots\ldots \quad (26)$$

Si le tube est terminé par une résistance acoustique unitaire il en découle

$$p_{N+1}(t) + v_{N+1}(t) = p_{N+1}(t) - v_{N+1}(t) \quad \ldots\ldots\ldots\ldots \quad (27)$$

Il en découle que $v_{N+1}(t) = 0$. La vitesse volumique à l'entrée du tube est égale à $p_1(t) - v_1(t)$. Définissons :

$$C_N(z) = \frac{\text{vitesse volumique à l'entrée}}{\text{vitesse volumique à la sortie}} = \frac{P_1(z)\, V_1(z)}{P_{N+1}(z)} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (28)$$

.../...

Il est possible de vérifier à partir de l'équation (21) que

$$C_N(z) = w_{11}^{(N)}(z) - w_{21}^{(N)}(z) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (29)$$

Définissons pour chaque n entre 1 et N,

$$C_n(z) = w_{11}^{(n)}(z) - w_{21}^{(n)}(z) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (30)$$

En multipliant l'équation (23) par le vecteur $[1 \quad -1]$, et en utilisant la valeur de $W_n(z)$ dans l'équation (26), on obtient :

$$[C_{n+1}(z) - C_{n+1}(z^{-1})] = \frac{1}{1-r_{n+1}} [C_n(z) - C_n(z^{-1})] \times \begin{bmatrix} z^{1/2} & r_{n+1} z^{1/2} \\ r_{n+1} z^{-1/2} & z^{-1/2} \end{bmatrix} \quad (31)$$

d'où

$$C_{n+1}(z) = \frac{1}{1-r_{n+1}} [z^{1/2} C_n(z) - z^{-1/2} r_{n+1} C_n(z^{-1})]$$

$$= \frac{z^{1/2}}{1-r_{n+1}} [C_n(z) - r_{n+1} C_n(z^{-1}) z^{-1}] \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (32)$$

Exception faite du facteur $z^{n/2}$, chaque $C_n(z)$ est un polynôme d'ordre n. Par conséquent, la fonction de transfert qui est l'inverse de $C_N(z)$, comprend un facteur $z^{-N/2}$ divisé par un polynôme de degré N. Le facteur $z^{-N/2}$ représente le temps de transmission dans le tube. La fonction de transfert a N pôles qui sont les zéros de $C_N(z)$. Ces pôles sont situés à l'intérieur du cercle unité si r_n satisfait à la condition.

$$|r_n| < 1 \quad ; \quad 1 \leq n \leq N \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (33)$$

ou alternativement

$$S_n > 0 \quad ; \quad 1 \leq n \leq N$$

.../...

Il est possible de démontrer à partir de l'équation (31) que

$$C_n(z) = \frac{z^{-1/2}}{1+r_{n+1}} \left[C_{n+1}(z) + r_{n+1} \; C_{n+1}(z^{-1})\right] \quad \ldots\ldots\ldots \quad (34)$$

Etant donné $C_N(z)$, il est possible d'en déduire $C_n(z)$ pour des valeurs décroissantes de n en commençant par n=N à partir de l'équation (34). Dans tous les cas le coefficient r_n est déterminé par le rapport des coefficients de $z^{-n/2}$ et $z^{n/2}$. Une séquence de valeurs r_1, r_2, ..., r_N, ainsi calculée permet d'en déduire un tube dont les aires de section sont S_1, S_2, ..., S_N, à condition que les aires soient positives ou $|r_n| < 1$ pour $1 \leq n \leq N$. Si les pôles du polynôme $C_N(z)$ sont à l'intérieur du cercle unité la condition est satisfaite.

VI- ANALYSE DES FORMANTS.

L'objectif d'une analyse de formants est la détermination des fréquences complexes naturelles du C.V. et l'étude de leurs évolutions pendant l'élocution. Or le signal de parole est le produit de trois composantes : le signal source, la réponse impulsionnelle du C.V., et la propagation dans la cavité nasale. Si le spectre de la source a un zéro voisin d'une fréquence naturelle du C.V., il sera pratiquement impossible de déterminer la fréquence exacte et la largeur de bande de ce formant. Le conduit nasal peut également poser le même problème. Par conséquent, il n'est possible de déterminer que les fréquences et largeurs de bande du signal qui ne soient pas masquées par l'effet de la source ou le conduit nasal.

Le filtre numérique linéaire obtenu par codage prédictif tient compte de l'effet conjugué de la source d'excitation et du C.V. Par conséquent, les pôles de la fonction de transfert n'appartiennent pas qu'au C.V., et il convient d'effectuer la séparation. Or les pôles de la source sont soit réels soit complexes conjugués. Dans le dernier cas ces pôles se caractérisent par une pointe de faible amplitude dans l'enveloppe spectrale. Disposant d'un seuil minimum (1,7 selon Atal), il est possible de séparer les pôles du C.V. de ceux de la source. La méthode de calcul est la suivante :

Les pôles de la fonction de transfert sont les racines de l'équation

$$\sum_{k=1}^{p} a_k \, z^{-k} = 1 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (35)$$

On obtient n pôles complexes conjugués z_1, z_1^*; ...; z_n, z_n^*.

.../...

La fonction de transfert prend, par conséquent, la forme

$$V(z) = \prod_{i=1}^{n} (1-z_i)\,(1-z_i^*) \Big/ \prod_{i=1}^{n} (z-z_i)\,(z-z_i^*) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (36)$$

Les facteurs supplémentaires introduits dans le numérateur font que $V(z) = 1$ dans le cas continu ($z=1$). La pointe spectrale caractérisant le pôle k a une amplitude

$$A_k = \left| \frac{(1-z_k)\,(1-z_k^*)}{(z-z_k)\,(z-z_k^*)} \right|^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (37)$$

où $z = \exp(2\,\pi j\, f_k T)$

$$z_k = |z_k| \exp(2\,\pi j f_k T)$$

et T = période d'échantillonnage.

Si $A_k <$ seuil ; z_k est un pole de la fonction d'excitation.

Par ailleurs, on obtient :

$$F_k = (1/2\,\pi T)\ \mathrm{Im}\ (\mathrm{Ln}\ z_k) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (38)$$

$$B_k = \left|(\pi T)\ \mathrm{Re}\ (\mathrm{Ln}\ z_k)\right| \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (39)$$

VII - CODAGE PREDICTIF A L'AIDE DU CEPSTRUM.

L'intérêt du Cepstrum [8] réside dans la séparation du signal source $e(nT)$ de la réponse impulsionnelle du C.V., $v(nT)$. Nous décrivons très brièvement la méthode. Le signal de parole s'exprime comme suit :

$$s(nT) = e(nT) \otimes v(nT) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (40)$$

où $\otimes$ signifie produit de convolution.

Ce signal est pondéré à l'aide d'une fenêtre $w(nT)$ afin de garantir la convergence de sa transformation de Fourier. Il vient

$$s_1(nT) = s(nT)\,.w(nT)$$
$$= [e(nT) \otimes v(nT)]\,.w(nT) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (41)$$

.../...

Si de surcroit w(nT) varie peu pendant la durée de v(nT), l'approximation suivante devient légitime :

$$s_1(nT) = e_1(nT) \otimes v(nT) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (42)$$

où $e_1(nT) = e(nT) \ .w(nT)$

Si $E_1(j\omega)$ et $V(j\omega)$ représentent les spectres de $e_1(nT)$ et $v(nT)$ respectivement, il s'en suit

$$S_1(j\omega) = E_1(j\omega).\ V(j\omega)$$

$$\text{Log } S_1(j\omega) = \text{Log } E_1(j\omega) + \text{Log } V(j\omega) \quad \ldots\ldots\ldots\ldots\ldots \quad (43)$$

Pour l'essentiel, Log $V(j\omega)$ varie lentement en fréquence, tandis que Log $E_1(j\omega)$ tend à varier plus rapidement et périodiquement en fréquence. Par conséquent, l'application de la transformée inverse de Fourier à Log $S_1(j\omega)$, se traduit par la concentration de la contribution de la réponse impulsionnelle du C.V. dans le voisinage de l'origine et l'influence de la source d'excitation à des multiples de la période de pitch. Cette tendance du signal d'excitation est à la base de la méthode de détection du pitch [9]. Il vient qu'en ne retenant que les valeurs dans le voisinage de l'origine [à l'aide d'un filtrage idéal dans le temps], on en déduit la contribution du C.V. Il reste à entreprendre une transformée de Fourier, suivie d'une exponentiation avant d'exécuter une transformée inverse de Fourier afin d'obtenir la réponse impulsionnelle du C.V. Le schéma fonctionnel de cette procédure est donné sur la figure : 4.

Disposant de cette réponse impulsionnelle du C.V., il devient possible d'appliquer le codage prédictif, soit déterminer le filtre numérique équivalent. L'analyse des formants se trouve simplifiée, du fait qu'il n'est tenu compte que de la contribution du C.V.

Cette méthode d'analyse n'est pas en synchronisme avec la période du pitch. Néanmoins, compte tenu de la contrainte imposée à la fenêtre w(nT) l'intervalle d'analyse est relativement important. Par conséquent, l'hypothèse que la configuration du C.V. est presque stationnaire dans l'intervalle d'analyse est moins légitime. En d'autres termes, il ne sera pas possible de suivre avec précision des évolutions rapides dans l'élocution. Par ailleurs, le raisonnement, qui a conduit au critère de séparation entre la contribution du C.V. et l'influence de l'excitation, est basé, uniquement, sur des tendances générales ; l'erreur introduite relativement faible est difficilement chiffrable. En conclusion, la fonction d'aire du C.V. déterminée au moyen de cette méthode ne sera pas très représentative de la configuration réelle du C.V.

.../...

VIII - CODAGE PREDICTIF ET FILTRE NUMERIQUE A ECHELLE.

Une méthode proposée par Ittakura et Saito [10]permet le calcul d'un filtre numérique à structure en échelle à partir d'une réponse impulsionnelle connue d'un filtre numérique. Ce calcul a pour base la théorie des polynômes orthogonaux et conduit à un filtre numérique linéaire ne comprenant que des pôles.

Base Mathématique :

Posons $z = \exp(-j\,\omega T) = \exp(-j\,\lambda)$

$f(\lambda)$ est la densité spectrale d'un signal échantillonné s(nT). Si X(z) et Y(z) sont des polynômes arbitraires en z, on définira le produit scalaire par

$$[X(z), Y(z)] = \int_{-\pi}^{\pi} X[\exp(-j\lambda)]\,\overline{Y}[\exp(-j\lambda)].f(\lambda)\,d\lambda \quad \ldots\ldots\ldots (44)$$

Si, par ailleurs, $X(z) = z^m$, et $Y(z) = z^n$, il vient

$$[z^m, z^n] = \int_{-\pi}^{\pi} \exp[-j(m-n)\lambda]\, f(\lambda)\, d\lambda$$

$$= v_{m-n} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (45)$$

où $\{ v_n \}$ est égale à l'autocovariance de s(nT).

On introduit le déterminant $\Delta_k = | v_{i-j} |_1^k$, (k=0,1, ...) et les polynômes $\varphi_n(z)$, $\varphi_n^*(z)$ où

$$\varphi_n(z) = \frac{1}{\Delta_{n-1}} \begin{vmatrix} v_o & v_1 & \cdots & v_{n-1} & v_n \\ v_1 & v_o & & \cdot & \cdot \\ \cdot & & \cdot & \cdot & \cdot \\ \cdot & & & \cdot & \cdot \\ \cdot & & & \cdot & \cdot \\ v_{n-1} & \cdot & \cdots & v_o & v_1 \\ 1 & z & \cdots & z^{n-1} & z^n \end{vmatrix} = z^n + \ldots \quad ;(n=0,1,\ldots) \ldots\ldots (46)$$

avec $\Delta_{-1} = 1$

$$\varphi_n^*(z) = z^n \varphi_n(1/z) = 1 + \alpha_1^n z + \ldots + \alpha_n^n z^n \quad \ldots\ldots\ldots\ldots\ldots\ldots (47)$$

.../...

Propriétés des polynômes $\varphi_n(z)$

i) Orthogonalité :

$$[\varphi_m(z), \varphi_n(z)] = \begin{cases} 0 & , m \neq n \\ \Delta_n / \Delta_{n-1} & , m = n \end{cases} \quad \ldots\ldots\ldots\ldots\ldots (48)$$

ii) Equations de Récurrence :

$$\varphi_{n+1}(z) = z\varphi_n(z) - k_{n+1}\varphi_n^*(z) \quad \ldots\ldots\ldots\ldots\ldots (49)$$

$$\varphi_{n+1}^*(z) = \varphi_n^*(z) - k_{n+1}\, z\varphi_n(z)$$

où

$$\varphi_n^*(z) = z^n \varphi_n(1/z)$$

$$k_{n+1} = -\varphi_{n+1}(0) \quad , |k_n| < 1 \quad \ldots\ldots\ldots\ldots\ldots (50)$$

iii) Récurrence sur Δ_n/Δ_{n-1},

$$\sigma_n^2 = \Delta_n/\Delta_{n-1} = \sigma_{n-1}^2 (1-k_n^2) \quad \ldots\ldots\ldots\ldots\ldots (51)$$

iV) Condition de Stabilité : Tous les zéros de $\varphi_n(z)$ se situent à l'extérieur du cercle unité.

V) Si $\int_{-\pi}^{\pi} |\text{Log } f(\lambda)|\, d\lambda < \infty$, il vient

$$\sigma_n^2 \, |\varphi_n^*[r.\exp(-j\lambda)]|^{-2} \simeq f(\lambda) \text{ uniformément pour } r \leq 1+\delta \ldots\ldots (52)$$

Détermination des paramètres :

Le signal de parole est un processus presque stationnaire dans un interval de temps relativement court (5 à 10 ms ou une période de pitch).

.../...

La fonction de transfert du filtre numérique est de la forme :

$$H(z) = \frac{\sigma_n}{\varphi^*_n(z)} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (53)$$

Or σ_n et $\varphi^*_n(z)$ dépendent des valeurs v_o et $\{k_n\}_1^n$

La définition des $\{k_n\}_1^n$ est

$$k_{n+1} = \sum_{i=0}^{n} \alpha_i^n v_{n+1-i} \Big/ \sigma^2_{n-1} (1-k^2_n)$$

$$= \left\{ \frac{[\varphi^*_n(z),\ z\varphi_n(z)]}{[\varphi^*_n(z), \varphi^*_n(z)] \cdot [z\varphi_n(z), z\varphi_n(z)]} \right\}^{1/2} \qquad \ldots\ldots\ldots (54)$$

Les k_{n+1} sont des coefficients d'intercorrélation entre deux signaux, qui sont les réponses de deux filtres numériques $\varphi^*_n(z)$ et $z\varphi_n(z)$ à un signal d'entrée $s(nT)$. Le schéma " d'extraction " de ces paramètres est décrit sur la figure : 5. Le calcul se fait par étape et les paramètres sont réajustés périodiquement. Trois structures en échelle de filtre numérique équivalent sont présentées sur la figure : 6. Ces modèles permettent la synthèse de la parole à partir des coefficients $\{k_n\}$.

La méthode de détection de pitch sera la même que pour le codage prédictif, à titre indicatif, détection d'une pointe dans le signal d'erreur de prédiction. Les coefficients du filtre de prédiction sont ceux de $\varphi^*_n(z)$.

L'analyse des formants sera identique à celle du codage prédictif ; le principe de séparation des pôles du C.V. de ceux de la source demeurant le même.

IX - CONCLUSION.

A partir de la méthode de codage prédictif on peut obtenir un ensemble assez complet de paramètres, à savoir :

- Les aires des sections du C.V.
- Les formants et leurs largeurs de bandes.
- La période de pitch.
- La fonction d'autocorrélation.

.../...

En plus, il est possible de séparer les pôles de la source d'excitation de ceux du C.V.

Les inconvénients de la méthode sont les suivants :

- La complexité des calculs.

- L' absence de différenciation entre l'influence du conduit nasal de celle du C.V.

- L'obligation de faire une hypothèse sur l'impédance de rayonnement pour la fonction d'aire.

Annexe

Lien entre Longueur du C.V. et Nombre de Coefficients de Prédiction :

A des fréquences inférieures à 5KHz, le C.V. peut être représenté, à des fins d'analyse, par un tube acoustique à aire de section variable. La relation reliant d'une part la pression P_g et la vitesse volumique U_g à la glotte et les quantités correspondantes P_1, U_1 aux lèvres est exprimée, le plus simplement, à l'aide des paramètres de la matrice (matrice de chaîne) ABCD du tube acoustique. Ces paramètres sont définis par l'équation matricielle [voir figure : 7]:

$$\begin{bmatrix} P_g \\ U_g \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} P_1 \\ U_1 \end{bmatrix} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(A1)}$$

Il nous reste à démontrer que les transformées inverse de Fourier de ces paramètres dans le domaine du temps ont une durée finie $\tau = 2l/c$, où l est la longueur du tube et c la vitesse du son. Prenons S(x) pour représenter la fonction d'aire du C.V., où x est la distance de la glotte au point où l'aire de Section est évaluée. Prenons un petit élément du tube de longueur dx à une distance x de la glotte. Les paramètres de la matrice ABCD du tube élémentaire dx sont

$$\begin{aligned} A = D &= \cosh \Gamma dx = 1/2\ [\exp(\Gamma dx) + \exp(-\Gamma dx)\ ], \\ B &= -Z_o \sinh \Gamma dx = -Z_o\ [\exp(\Gamma dx) - \exp(-\Gamma dx)]/2, \qquad \ldots\ldots\ldots \text{(A2)} \\ C &= -\sinh \Gamma dx/Z_o = -[\exp(\Gamma dx) - \exp(-\Gamma dx)]/2Z_o, \end{aligned}$$

où Z_o est l'impédance caractéristique du tube élémentaire dx et égale à $\rho c/S(x)$, Γ est la constante de propagation = $j\omega/c$, ρ étant la densité de l'air et ω la fréquence angulaire en radians.

La matrice ABCD du tube est obtenue à partir du produit des matrices ABCD des tubes élémentaires de longueur dx et situés à des distances dx l'un de l'autre sur toute la longueur du tube l= ndx. Il est possible de démontrer que chacun des paramètres ABCD du tube peut être exprimé sous forme d'une série de puissances en exp (Γdx), à titre indicatif.

$$\sum_{k=-n}^{n} \alpha_k . \exp(k \Gamma dx) \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(A3)}$$

.../...

Les paramètres ABCD sont, par conséquent, les transformées de Fourier de fonctions de temps, dont chacune est de durée $\tau = 2n.dx/c$. En passant à la limite quand $dx \longrightarrow 0$, $n \longrightarrow \infty$, et $n.dx = l$, on obtient $\tau = 2l/c$.

De l'équation (A1) on obtient

$$U_g = CP_l + DU_g \qquad \text{(A4)}$$

Etant donné que $P_l \simeq j\omega KU_l$, où K est une constante fonction de l'aire aux lèvres, l'équation (A4) peut s'écrire

$$U_g \simeq (j\omega KC+D)\, U_l \qquad \text{(A5)}$$

La durée de la mémoire du filtre de prédiction est, par définition, égale à la durée de la transformée inverse de Fourier de la fonction de transfert inverse définie par le rapport entre les vitesses volumiques aux lèvres et à la glotte. Par conséquent, tenant compte de l'équation (A5), la mémoire du filtre de prédiction a une durée de $\tau = 2l/c$.

Références :

[1] J.L. Flanagan : "Speech Analysis, Synthesis and Perception" ; Springer-Verlag Berlin, Heidelberg, New York (1965).

[2] E.N. Pinson : " Pitch-Synchronous Time-Domain Estimation of Formant Frequencies and Bandwidths" J. Acoust. Soc. Amer. 35, 1264-1273 (1963).

[3] L.C. Wilkins and P.A. Wintz : "Bibliography on Data Compression, Picture Properties, and Picture Coding" ; IEEE Trans. on Information Theory, Vol. IT-17, no.2, 180-199, March (1971).

[4] B.S. Atal and S.L. Hanauer : "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave" ; J. Acoust. Soc. Amer. 50, no.2 (Part2), 637-655, (1971).

[5] B.S. Atal and M.R. Schroeder : "Adaptive Predictive Coding of Speech Signals", Bell Syst. Tech. J. 49, 1973-1986 (1970).

[6] B.S. Atal : "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", J. Acoust. Soc. Amer. 47,65 (A)(1970).

[7] B.S. Atal : "Sound Transmission in the Vocal Tract with Applications to Speech Analysis and Synthesis," Proc. Int. Congr. Acoust., 7th, Budapest, Hungary (Aug. 1971).

[8] C.J. Weinstein, A.V. Oppenheim : "Predictive Coding in a Homomorphic Vocoder". IEEE Trans. on Audio and Electroacoustics Vol. AU-19, no.3, 243-248, (Sept. 1971).

[9] A.M. Noll : "Cepstrum Pitch Determination". J. Acoust. Soc. Amer. 41, 293-309 (1967).

[10] F. Ittakura - S. Saito : "Digital Filtering Techniques for Speech Analysis and Synthesis". Proc. Int. Congr. Acoust., 7th, Budapest, Hungary(Aug. 1971).

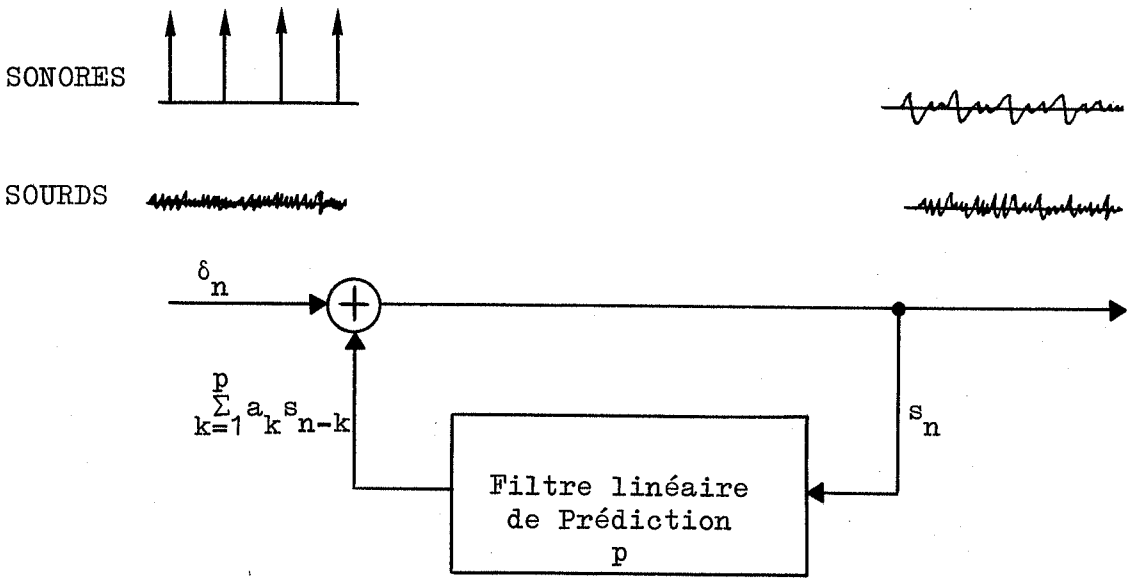


Figure 1 - Schéma du modèle de production de la parole

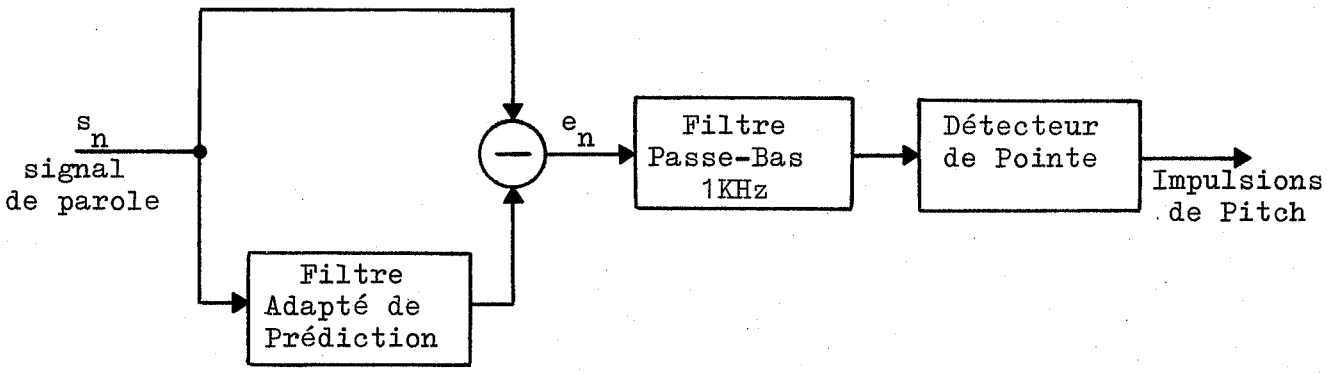


Figure 2 - Schéma Fonctionnel du Détecteur de la pointe du Pitch

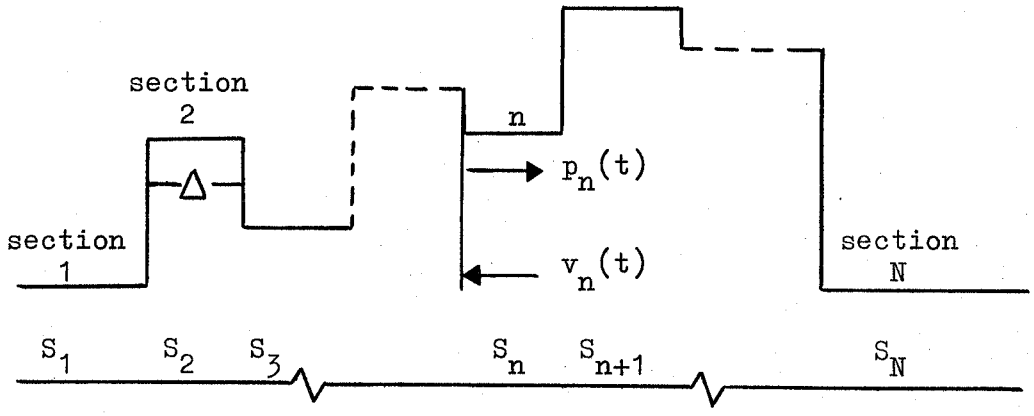


Figure 3 - Succession de tubes cylindriques d'égale longueur

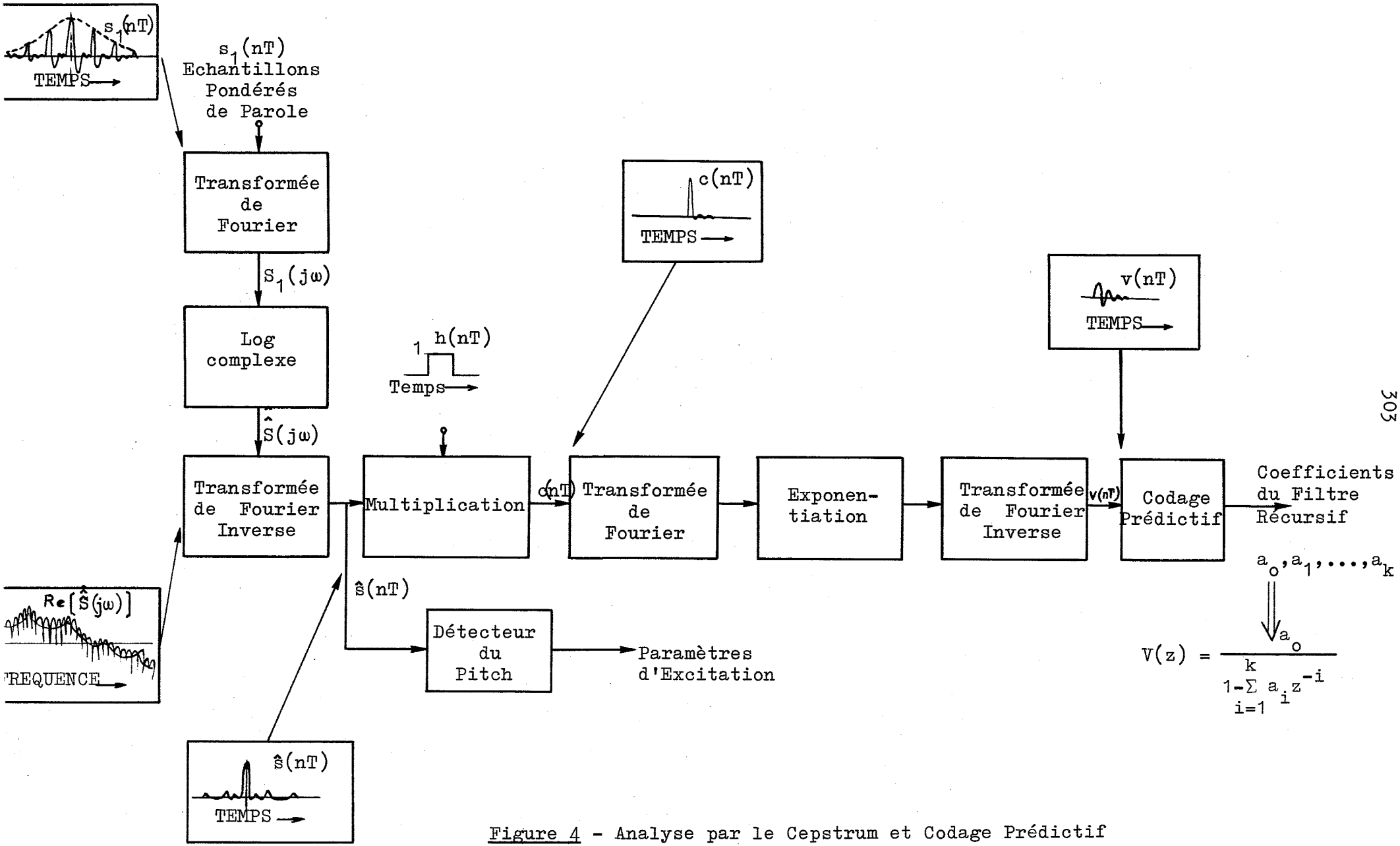


Figure 4 - Analyse par le Cepstrum et Codage Prédictif

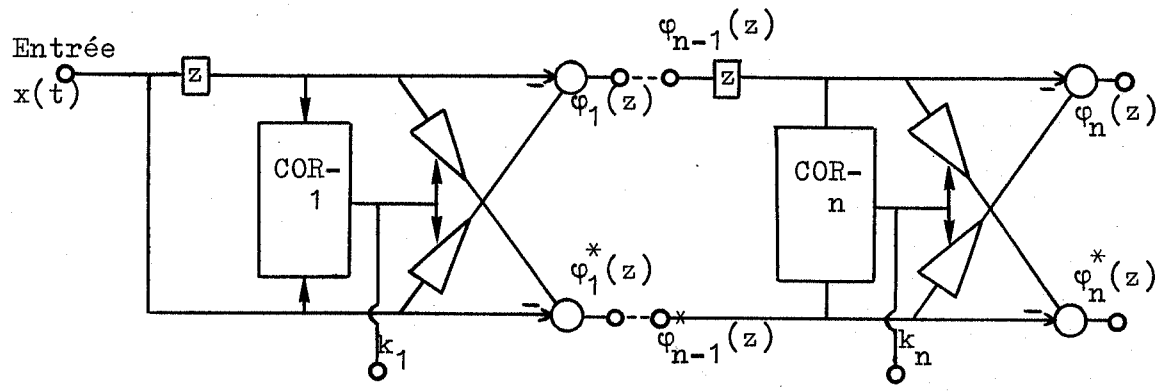
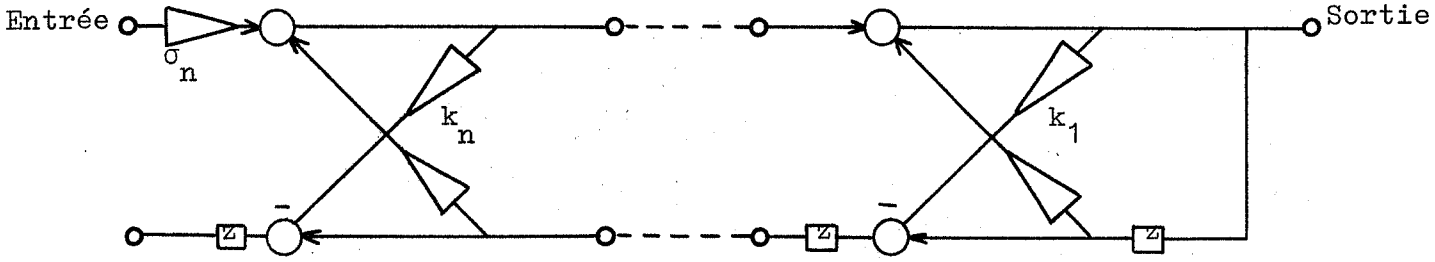
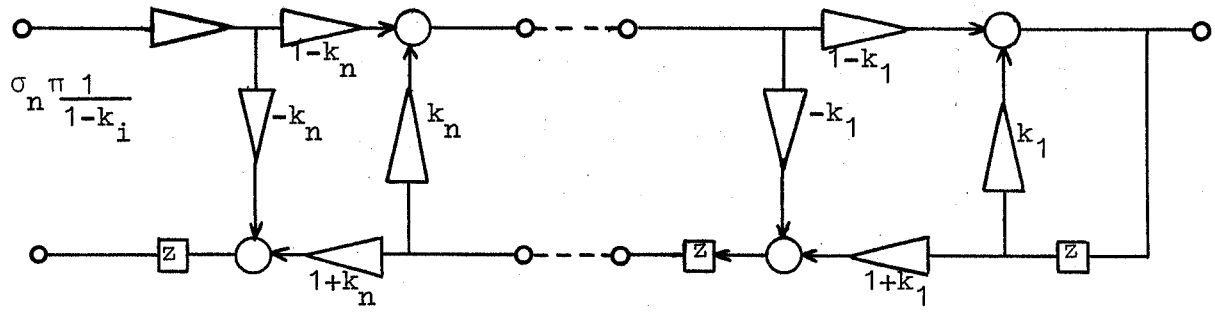


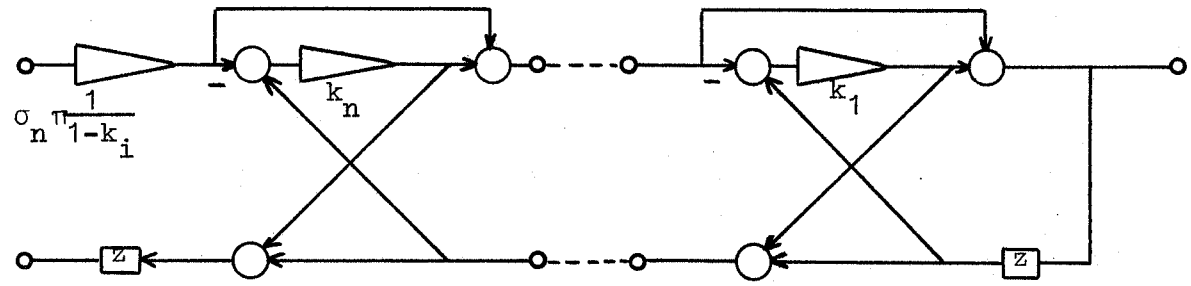
Figure 5 - Filtre numérique adapté pour le calcul des paramètres $\{k_i\}$



a) Filtre numérique à structure en échelle



b) Modèle équivalent [semblable à celui de Kelly]



c) Autre modèle équivalent avec moins de multiplications

Figure 6 - Filtres numériques à structure en échelle pour la synthèse [les filtres sont équivalents]

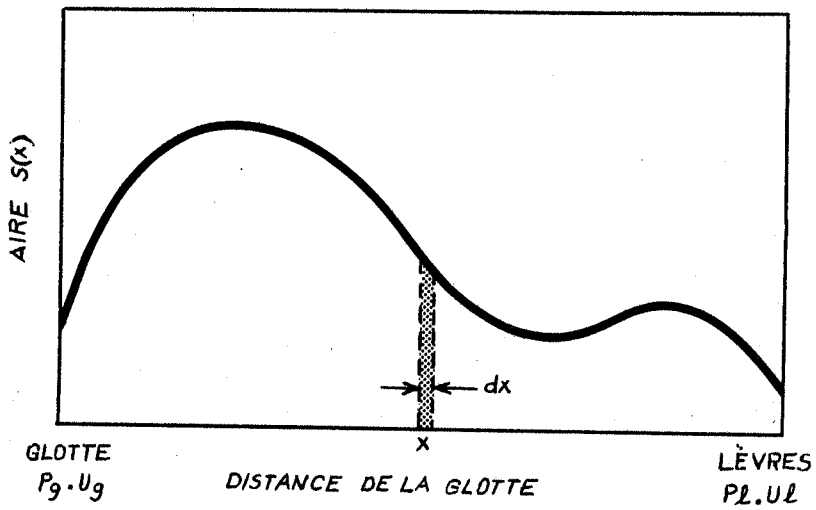


Figure 7 - Tube Acoustique à Aire de Section Variable

T A B L E R O N D E

MESURE DE FREQUENCE FONDAMENTALE

Animateurs : M. J.P. PECKELS (I.B.M. - La Gaude)
M. A. LANDERCY (Institut de Phonétique - Université Libre de BRUXELLES - Belgique)
M. L.J. BOE (Institut de Phonétique - ST MARTIN D'HERES)
M. ZURCHER (C.N.E.T. - LANNION)
M. MAISSIS (E.N.S.T. - PARIS)
M. DEMAN (Thomson C.S.F. - GENNEVILLIERS)
M. DECHAUX (Thomson C.S.F. - GENNEVILLIERS)

METHODE D'EXTRACTION DU PITCH

(Exposé présenté devant la table ronde sur la mesure de la fréquence fondamentale)

par

A.H. MAISSIS - E.N.S.T. PARIS

METHODE D'EXTRACTION DU PITCH

A.H. MAISSIS
E.N.S.T. PARIS

Résumé .

Les sons sonores sont produits par une excitation de nature impulsionelle du canal vocal . Ce dernier étant assimilable à un ensemble de résonateurs du second ordre , l'apparition d'une impulsion excitatrice est caractérisée par une brusque augmentation de l'énergie du signal . Une fonction spécialement introduite permet d'émettre des impulsions chaque fois que la variation de l'energie est importante . Enfin un algorithme testant la régularité du pitch obtenu , élimine les impulsions parasites .

Abstract .

The voiced sound signal is looked as the response of a system composed of a number of resonators and excited by a pulse series . So , every excitation pulse is caracterised by an abrupt increase of signal's energy .
A specially introduced function mesures signal's energy variations and emits a weighted pulsion when this variation takes a peak value . Then a regularity test algorithm selects useful pulses.

INTRODUCTION.

Le problème d'extraction éfficace du pitch n'a pas encore trouvé de solution définitive . La méthode souvent employée de filtrage à très basses fréquences a l'inconvénient de ne pas pouvoir toujours discerner la fréquence fondamentale de celle du premier formant . La méthode de cepstre necessite trop de calculs.

De plus ces méthodes ne peuvent pas prelever le début de chaque période c'est à dire le moment d'apparition de l'impulsion excitatrice .

La méthode de J.N Maksym[1] parait plus spécialement interessante. Elle est basée sur une identification adaptative du canal vocal, une prédiction du signal de sortie et le calcul de l'erreur entre les deux signaux.Cette erreur est particulièrement importante aux instants d'apparition des impulsions excitatrices de sorte qu'en prélevant et reconnaissant ce signal on peut alors disposer des repères du début du début de chaque période . La méthode est entièrement cablée en hardware , mais elle nécessite un algorithme supplémentaire qui d'une part decidera si le signal d'erreur est relativement important pour en tenir compte , et d'autre part testera la régularité du pitch obtenu.

La présente méthode tout en partant des principes differents, ressemble à celle de J.N. Maksym par le fait qu'elle fournit par hardware des repères possibles des instants d'excitation du canal vocal . De plus elle a l'avantage d'émettre une seule impulsion au début de chaque période et non pas un signal d'une certaine durée comme c'est le cas pour la méthode Maksym, la durée du signal posant toujours le problème du choix de l'instant cherché.

Présentation de la méthode.

Les sons sonores, comme il est bien connu , sont produits par une excitation de nature impulsionelle du canal vocal . Ce dernier peut être assimilé à une série de résonateurs du second ordre connectés en cascade (et en parallele dans le cas des sons nasals). On sait aussi que la réponse impulsionnelle d'un système du second ordre présente l'allure d'une sinusoïde amortie et par conséquent , l'apparition d'une impulsion excitatrice est marquée par une brusque augmentation de l'energie du signal.

Considerons le cas de la fig 1 . L'instant t est celui d'apparition de la nouvelle impulsion excitatrice. On mesure l'energie du signal pendant T sec à droite du point t et pendant le même temps à sa gauche ; On prend la différence des deux energies,et on obtient ainsi un indice quantitatif

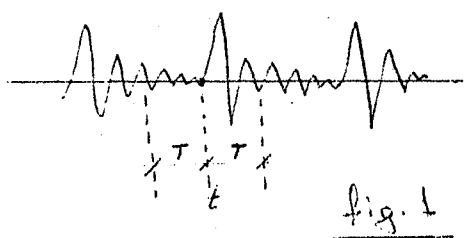


fig. 1

qui prend des valeurs relativement très grandes aux instants t. De façon plus analytique pour chaque point t on calcule la fonction:

$$m(t) = \int_{t}^{t+T} x^2(\tau)d\tau - \int_{t-T}^{t} x^2(\tau)d\tau \qquad (1)$$

La signification physique de la fonction m(t) peut être obtenue en considérant la courbe : energie du signal en fonction du temps. La fonction m(t) donne une mesure de l'angle entre deux tangentes sur le même point ; l'une à sa gauche et l'autre à sa droite . Ensuite à tout instant τ_i où la fonction m(t) passe par un maximum positif , une impulsion, munie d'un poids spécifique $m(\tau_i)$ est émise . De ce fait nous disposons d'un ensemble de repères du début des périodes fondamentales. Ajoutons ici que toutes les opérations déjà décrites peuvent être entièrement cablées en hardware . Ceci assure le temps "réel" de l'extraction du pitch.

La figure 2 donne l'exemple d'un signal analysé . Les impulsions qui apparaissent sur la figure marquent les instants τ_i prélevés par la méthode tandis que leur longeur est proportionnelle au poids $m(\tau_i)$.Il est à remarquer la netteté et la régularité des impulsions dans la partie des sons sonores tandis que le son sourd "p" est marqué par des impulsions irrégulières et d'un poids insignifiant . La figure 3 présente un second exemple dans lequel la fréquence fondamentale est presque le double de celle du premier exemple.

Discussion:

Il y a deux remarques à faire :

- la formule (1) peut être appliquée sur le signal brut . Il y a cependant une amélioration possible en préfiltrant le signal pour atténuer l'importance d des formants supérieurs et accentuer ainsi la valeur de m(t) aux instants désirés . En pratique un filtrage aux environ de 900 à 1000 Hz est suffisant .L'inconvénient de ce filtrage est l'introduction d'un retard (de l'ordre de 1 ms).
- La deuxième remarque concerne la durée T d'intégration. Sur les deux exemples présentés la valeur de T était égale à 2 ms . Pour le premier exemple , T étant relativement petite par rapport à la période fondamentale , des impulsions parasites apparaissent (mais d'un poids très faible , donc facilement éliminables). Au contraire des impulsions de ce type n'existent pratiquement pas dans le second exemple où la période est de l'ordre de 5 ms . Le problème du réglage de la durée T se pose donc comme suit : une durée trop petite introduit des impulsions parasites tandis que pour une durée trop g

grande on aura fatalement des "trous" . Il existe donc une durée optimale pour chaque fondamental ..

La meilleure solution pratique est de choisir T assez petite (correspondant par exemple à une valeur optimale pour $f_0 = 200 Hz$ et corriger ensuite les résultats par software . Cet algorithme aura comme tâche d'éliminer les impulsions parasites par simple comparaison du poids de l'impulsion qui se présente à celui de l'impulsion précedente , ou d'introduire eventuellement des impulsions supplémentaires pour combler des"trous" et ceci par un test de régularité du pitch obtenu . L'algorithme en question peut être très simple et peut fonctionner en temps réel " .

Un autre point qui va attirer notre attention est le cas de certains sons de faible puissance tel que le fricatif " V " . Le signal étant très bruité et sa puissance très faible , la fonction m(t) prend des valeurs très petites. Ceci entraîne une confusion possible entre les sons sonores et les sonssourds (ces derniers étant en fait distingués des premier par les faibles valeurs de la fonction m(t)). Une très bonne solution du problème est de normaliser la fonction m(t) par rapport à l'amplitude du signa

Considérons une segmentation du signal en segments de 10 ms . Soit A l'amplitude maximale du signal dans le segment qui contient l'instant t . Nous définissons alors la fonction normalisée par: $m_N(t) = m(t)/A^2$

la figure 3 donne un exemple de cette nature . Les fonctions $m(t)$ et $m_N(t)$ y sont présentées en même temps .Les repères des impulsions excitatrices y apparaissent clairemen

Conclusion.

L'avantage de la méthode présentée est la bonne précision avec laquelle les instants de début des périodes fondamentales sont marqués. Sa rapidité peut être assurée par le fait que les impulsions de base (instants t_i) peuvent être fournis par hardware . La combinaison d'une méthode qui sépare les sons sonores des sons sourds et la normalisation du poids des impulsions émises donne une solution efficace du problème d'extraction du fondamental dans la plupart des cas.

Références:

J.N Maksym : " Real time pitch period extraction by adaptive prediction of the speech waveform ".
1972 Conference on speech Communication and processing.
Boston April 1972 .

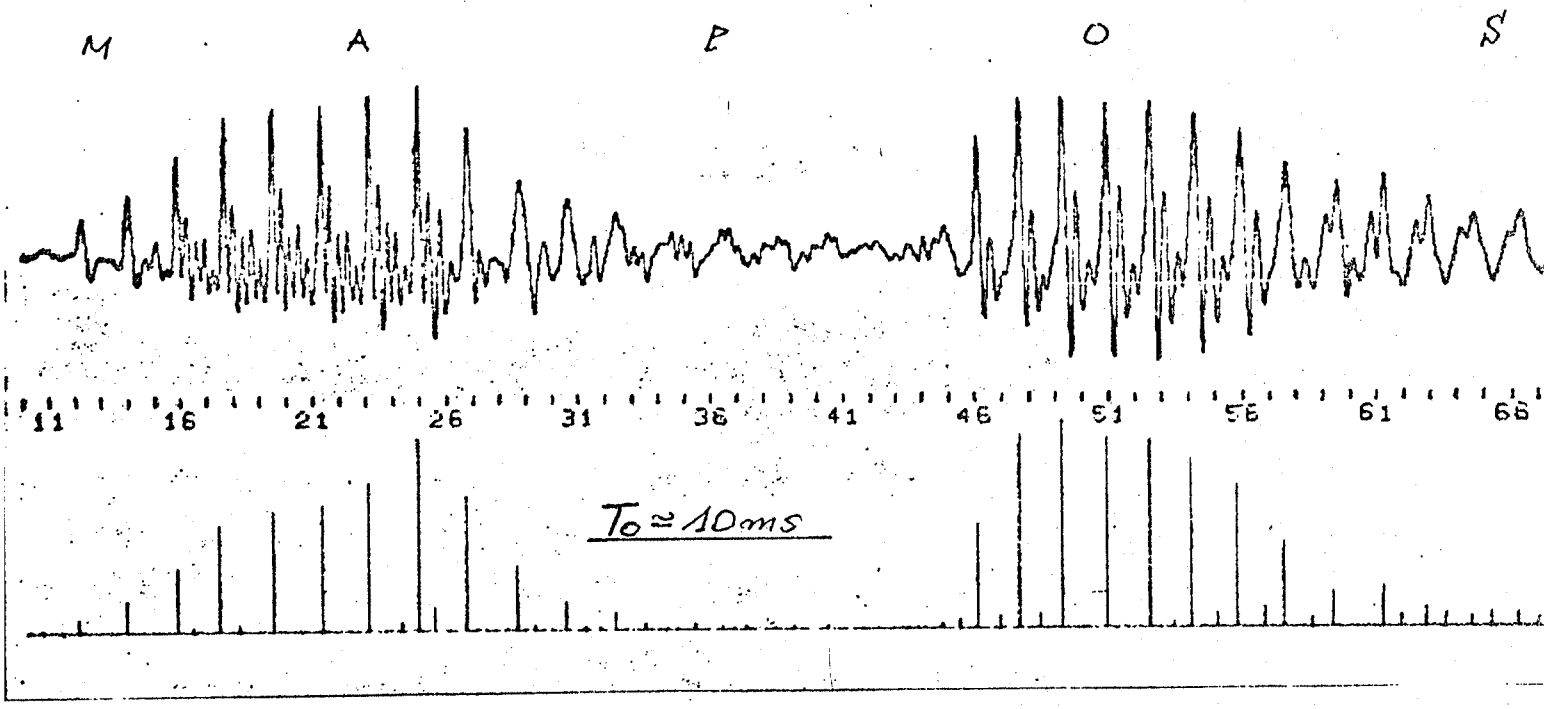


Fig 2 : Analyse de la phrase " ma position..."

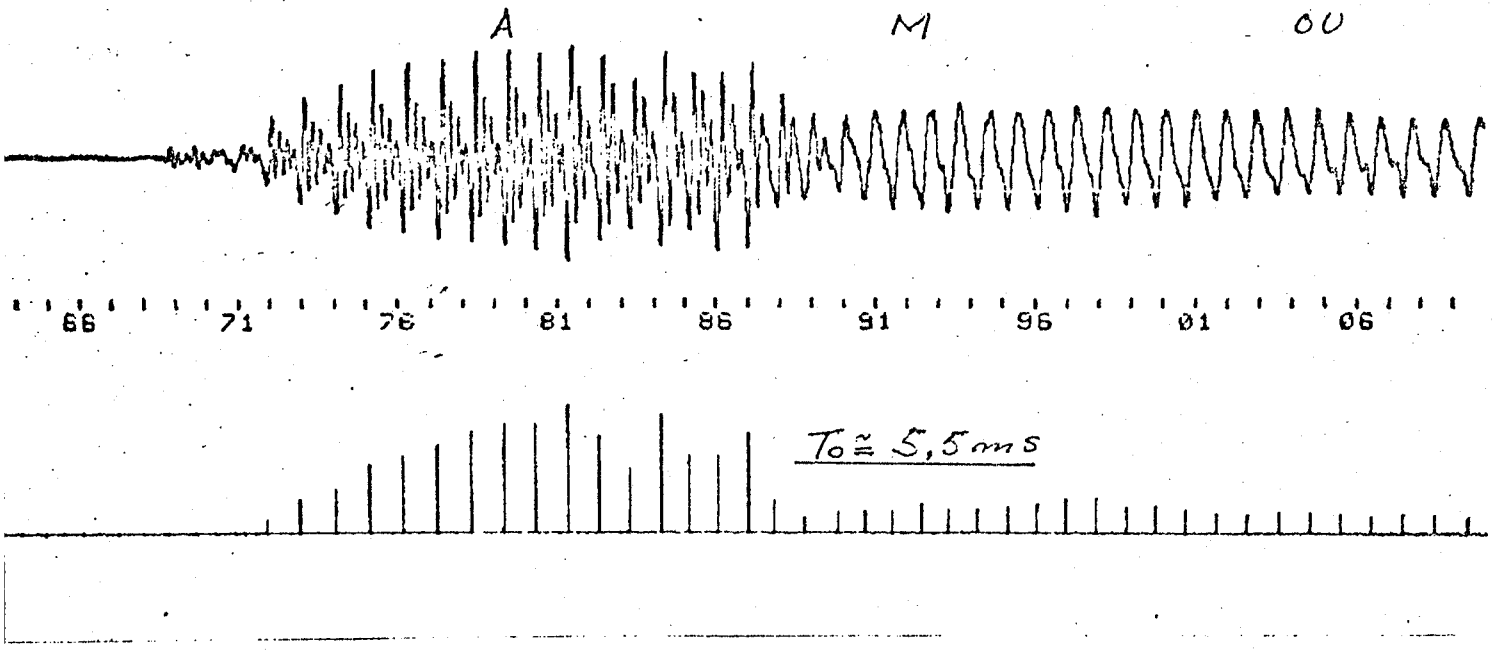


Fig 3 : Analyse du mot " amour"

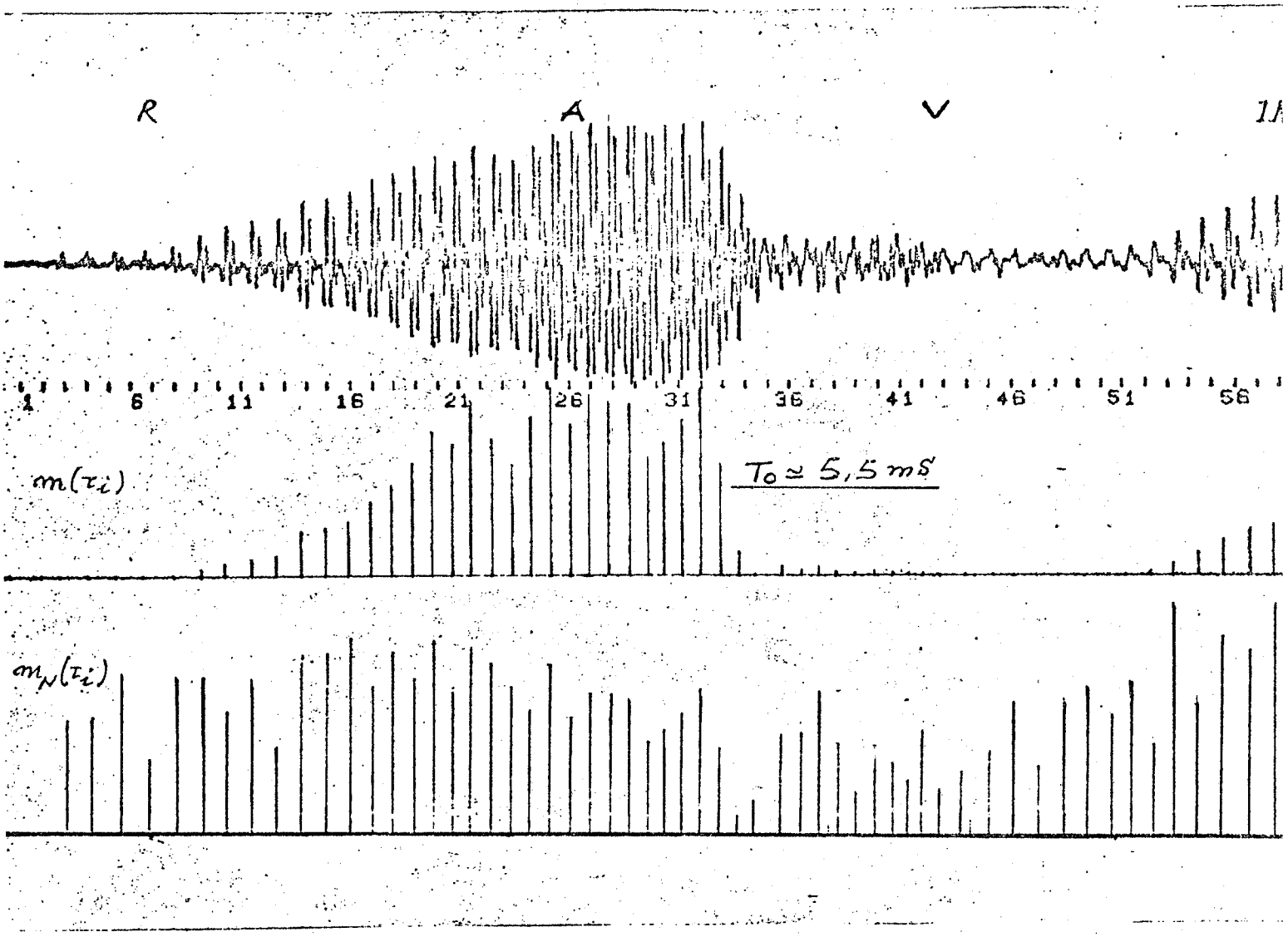


Fig 4 : Analyse du mot " ravin "

RECONNAISSANCE AUTOMATIQUE

ET SEGMENTATION DE LA PAROLE

J.P. QUANCARD

DIRECTION DES RECHERCHES ET MOYENS D'ESSAIS

SERVICE DES RECHERCHES

NOTE SUR LA RECONNAISSANCE DE LA PAROLE ET SES PERSPECTIVES

--

1 - Le congrès tenu à BOSTON en avril 72 sur le traitement automatique du signal vocal a été en particulier marqué par le changement de la politique de recherche menée aux U.S.A. (cf. avant propos des pré-prints - position de l'US Air Force Laboratoire de Cambridge) - notamment en reconnaissance de la parole - qui sera soutenue par des aides gouvernementales dans le cadre de plans d'ensemble.

En France, huit équipes (à ma connaissance) * effectuent des recherches sur le thème "reconnaissance de la parole" et la dizaine de projets étudiés sont simultanément :

- proches, quant à leur état actuel d'avancement.
- différents quant à leurs objectifs, leur conception et leur réalisation.

Rapprocher ces deux éléments conduit naturellement à se poser des questions sur ce domaine de recherches.

-doit-on faire (ou continuer) ces recherches ? Pourquoi ?

-où en est l'avancement des travaux et que peut-on en attendre ?

- que peut-on fixer comme objectifs à 5 ans, 10 ans, ... et quelles voies suivre pour les atteindre ?

Une réponse à ces questions peut sans doute venir en partie d'une synthèse détaillée et prospective des travaux connus ce jour. Cependant, les projets, ne serait-ce qu'en France, sont trop différents au niveau des options de réalisation et assez proches au niveau des résultats partiels enregistrés actuellement pour qu'il soit facile d'en tirer des conclusions définitives (par ex., quant aux options de réalisation) et surtout de les justifier. Plus intéressant (et certainement plus facile) parait donc de philosopher et il ne faut voir dans les lignes qui suivent que l'impression et l'opinion d'un non spécialiste, issues d'un vaste (mais hélas, superficiel!) tour d'horizon pour des projets actuellement conduits en France.

---

* L'équipe travaillant dans le cadre du CEA et dont les travaux ont été exposés au cours des journées d'étude, ne m'était pas jusqu'alors connue. Elle n'est donc pas, qu'elle veuille bien m'en excuser, mentionnée ici.

.../...

2- L'ambition générale des travaux menés en reconnaissance automatique de la parole peut s'exprimer comme la réalisation d'une machine ou d'un système capable de reconnaitre la parole, c'est-à-dire capable de traduire le support "signal vocal" en un autre support informationnel (par ex : suite de phonèmes, mots imprimés, ...) sans modifier le contenu sémantique du message vocal (et conservant éventuellement certaines caractéristiques de la forme de ce message).

Dans ce cadre général, il n'est pas inutile, pour juger de l'intérêt de ces recherches et des caractéristiques des différentes machines à réaliser, de détailler quelques applications :

2.1. - Communication à faible bande de transmission.

La réalisation de synthétiseurs de parole (vocoder, synthétiseur à formants, ...) permet de disposer d'appareils restituant le signal vocal (qui nécessite une bande de transmission d'environ 3000 à 10000 Hz), grâce à une commande nécessitant un débit faible d'information (1000 bits/s environ pour un synthétiseur à formant, moins de 100 bits/s pour une synthèse par règles - à partir des phonèmes).

Une (des) machine (s) réalisant l'extraction des paramètres de commande de ces synthétiseurs permettra (permettront) la transmission de messages vocaux sur des lignes de transmission à bande très étroite, ce qui est

. souvent utile (diminution des brouillages et des repérages, amélioration de la sécurité de transmission par codages, ...)

. ou indispensable (par ex. communication sous-marine).

2.2. - Communication avec des machines.

L'interface homme-machine est actuellement le plus souvent assuré par des supports informationnels ergonomiquement peu satisfaisants pour l'homme (cartes perforées, boutons-poussoirs, ...) nécessitant souvent la présence d'un personnel spécialisé.

L'apparition de moyens de communication graphiques, puis vocaux répondent aux besoins d'amélioration de ces rapports (en particulier par diminution ou suppression de l'apprentissage préalable de l'utilisateur).

a) La commande des machines non informatiques (machines-outils, systèmes de manutention, robots de manipulation, calculateurs de bureau, ...) nécessite des langages de faible importance (quelques centaines de mots), à syntaxe et vocabulaire pouvant dans la plupart des cas être définis arbitrairement (ce qui autorise en particulier une meilleure adaptation au système de reconnaissance) et peut se satisfaire de contraintes sur la variété des locuteurs.

.../...

b) Les système-interface avec les ordinateurs peuvent s'utiliser pour toutes sortes de réalisations :

- programmation(FORTRAN, COBOL, ... ou autre langage à définir)
- interrogation de banque de données (gestion de stock, réservation de place, ...)
- utilisation ou/et commande de systèmes complexes (systèmes de management, de commandement militaire, ...)

Ces applications mettent en jeu des langages très divers, depuis des langages simples pour un nombre élevé de locuteurs (réservation de place, programmation, ...) jusqu'à des langages complexes, quasi-naturels, mais pour éventuellement peu de locuteurs - ou du moins, une adaptation, même délicate, aux locuteurs peut être envisagée - comme par exemple mise en oeuvre de grands systèmes de commandement .

2.3. - machine à écrire phonétique :

Cela est probablement la forme la plus évoluée de système reconnaissant la parole, puisqu'il s'agit de transcrire un message parlé en un message écrit (par des symboles phonétiques).

Ce "dictaphone pour PDG de l'an 2000" ne se fera pas sans difficultés, puisqu'il est soumis aux contraintes maximales : langue naturelle parlée naturellement par un ensemble quelconque de locuteurs.

Cependant, des étapes de réalisation sont probables pour sérier les difficultés, étapes marquées par des restrictions sur la généralité du langage, sur le nombre et le type des locuteurs, des contraintes sur la prononciation.

2.4. - identification de locuteurs.

L'identification du locuteur est un moyen d'authentification qui peut s'avérer indispensable dans certains cas (par ex : communication radiotéléphonique entre chefs d'armée, ...). Notons que la "signature verbale" fait surtout appel à un aspect du message vocal plus "esthétique" (c'est à dire lié à sa forme) que sémantique. D'ailleurs, un système d'identification de locuteur ne pourra qu'être plus efficace si la même séquence de mots, spécialement choisie en fonction de l'application, est utilisée.

2.5. - aide aux handicapés.

2.6. - Justification de théorie d'intelligence artificielle :

La multiplicité des applications envisageables et les progrès qu'elles amèneront de par leur utilisation permettent d'estimer probable l'existence d'un marché pour les moyens de communication homme-machine vocaux et que donc ceux-ci se développeront comme le font actuellement les moyens graphiques. Dans cette perspective, il parait plus raisonnable d'envisager de petites réalisations, peu coûteuses et efficaces dans un certain domaine (par ex : commandes de machines à la voix, interrogation élémentaire de banque de données, ...) pour sensibiliser et développer le marché, que de viser dès maintenant des systèmes de reconnaissance très sophistiqués, mettant en jeu un hardware et un software très volumineux. Le parallèle avec le domaine des moyens de communication graphique le confirme, domaine en plein essor depuis l'apparition des matériels peu coûteux : (tablettes d'entrée graphique, displays, consoles alphanumériques), bien qu'il existe (chez IBM en particulier) des consoles avec crayon lumineux depuis longtemps.

3. - L'existence potentielle d'un marché n'est rien sans l'espoir de pouvoir développer des matériels convenables.

L'état actuel des recherches françaises permet un certain optimisme, et ce d'autant que les études entreprises, diverses dans leurs objectifs, laissent entrevoir des développements variés.

3.1. - Pour pouvoir utilement discuter les différents travaux (ou même effectuer des comparaisons), je pense intéressant de se référer à un modèle de machine à reconnaître la parole - bien que l'on puisse contester la notion de modèle pour des réalisations qui seront très différentes en raison de leurs objectifs. Des modèles ont été proposés d'abord par KOZHEVNIKOV et CHISTOVITCH (1968), puis par FANT (1970).

Le modèle de FANT est très complet et une version simplifiée peut être la suivante :

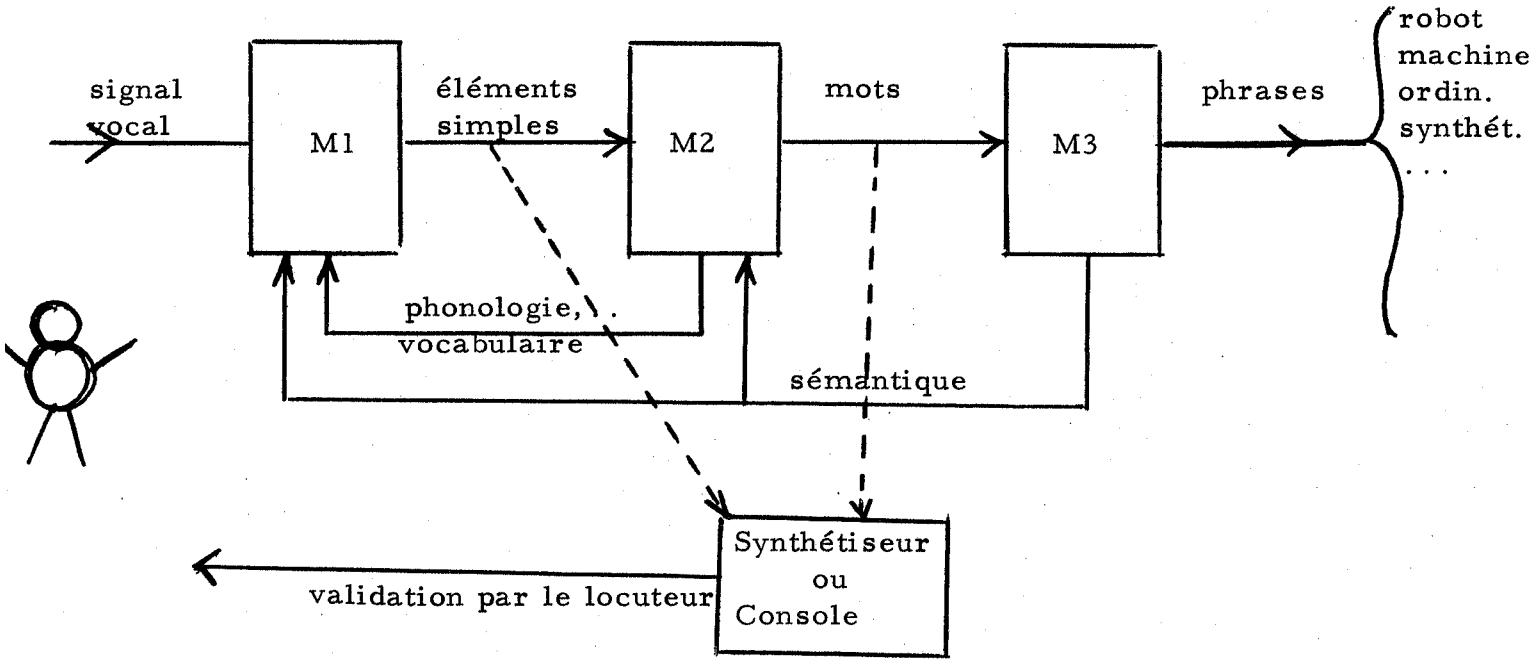
- M, machine (ou système) à réaliser est un ensemble des machines M1, M2, M3 agissant successivement :

  . M1 interprète le message vocal en micro-segments (phonèmes, phonatomes, ou autres...)

  . M2 assure, en utilisant ces micro-segments, l'identification de groupes de micro-segments (mots)

. M3 assure le passage des mots aux phrases (ou commandes, ou ...)

- chacune de ces machines peut être décomposée selon le schéma (classique en reconnaissance des formes) = capteur - prétraitement - classification.
- une machine de niveau inférieur est le "capteur" de la machine de niveau supérieur.
- un feed-back peut exister entre les machines de niveau supérieur et inférieur (ainsi par exemple, entre M2 et M1, probabilité d'existence de certaines associations de microsegments, ...). Ce feed-back autorise une certaine imperfection aux machines non terminales (M1 et M2).
- un feed-back par synthétiseur (ou affichage sur console) vers le locuteur autorise éventuellement une validation de l'information interprétée.



La conception d'un système selon ce modèle peut être qualifiée d'analytique.

Cependant, toutes les réalisations ne peuvent se réduire à cette conception analytique, à laquelle s'oppose la notion de réalisation globale, c'est celle d'une machine M non décomposable en machines plus élémentaires (voir P2. Les symboles Pi renvoient au tableau annexé décrivant succinctement les projets en étude par les différentes équipes).

.../...

L'examen de l'ensemble des travaux achevés ou en cours de réalisation sur le plan national, montre , exception faite de quelques études (P2 par ex.) qu'ils se basent sur une conception analytique du système à réaliser. De plus, même pour ceux d'entre eux que l'objectif ne limite pas à réaliser une machine de type M1 (P9 par exemple), l'état actuel d'avancement se borne à la réalisation de machines de ce type. (et dans un cas, P6, à leur évaluation). C'est donc seulement pour ce type (identification de "phonèmes" - ou autres éléments de base - dans le message vocal d'entrée) que se dégagent quelques philosophies de conception et de réalisation.

3. 2. - La première impression d'une revue des projets (voir tableau annexé) est l'incroyable diversité des conceptions, qui se traduit par l'absence apparente de points communs entre les réalisations (si ce n'est - et non pour tous ! - l'échantillonnage à 100 Hz - toutes les 10 millisecondes - du signal vocal - mais ne pouvait-on attribuer cela à la valeur agréable à l'esprit du nombre 100 !).

Cette diversité découle en grande partie des objectifs qui diffèrent, mais également des options à fixer a priori dans la conception et l'étude :

a) nature du langage et type de parole :

un vocabulaire limité, comprenant un nombre éventuellement faible de phonèmes différents (30 phonèmes environ dans la langue française) d'une part, une parole composée de mots isolés très distinctement prononcés (comme par exemple dans la "machine à calculer de bureau" réalisée par J. Y. GRESSER au CNET LANNION) pose des problèmes moins difficiles que du français naturel dans une conversation.

b) nature des locuteurs et faculté d'adaptation de la machine :

- les voix d'homme et de femme présentent des caractéristiques sensiblement différentes, encore que les résultats encore fragmentaires obtenus par LIENARD (P1 et P2) et ALINAT (P9) corroborent l'hypothèse que la différence se résout aisément (du moins pour la reconnaissance) par normalisation (calibrage) en fréquence obtenue par affinité, d'une constante variant globalement de 0, 7 pour les voix très graves à 1, 3 pour les voix "féminines").

De plus, la dispersion des qualités des voix est grande (accents régionaux, défauts de prononciation, . . .)

- la possibilité pour le système de l'adaptation à un (ou une classe de) locuteur (s) constitue une option intéressante soit pour la simplification de la machine, soit pour l'amélioration de ses performances tout en atténuant en partie les inconvénients dus à la dispersion des caractéristiques des locuteurs .

Cette adaptation peut se concevoir comme :

. pré-adaptation par intervention extérieure au système de reconnaissance, soit par améliorations successives des résultats de reconnaissance obtenus, soit a priori à partir de mesures sur les caractéristiques du locuteur .

. auto-adaptation, généralement obtenue grâce à un groupe de mot-clés, solution dont la souplesse et l'utilité doit susciter de plus amples efforts de mise au point.

L'adaptation s'effectue pratiquement par des modifications de caractéristiques de fonctionnement de la machine, qui peuvent intervenir à tous les niveaux, essentiellement :

- ajustement des caractéristiques du capteur (banc de filtre)
- utilisation de normalisation à paramètres variables au cours du prétraitement.
- modification de paramètres des algorithmes d'identification.

c) temps réel.

Le fonctionnement en temps réel ou différé n'a de signification que par rapport à l'application envisagée et non par rapport au signal analysé. Autrement dit, travailler en temps réel pour une machine à reconnaitre n'est pas d'identifier les phonèmes au fur et à mesure de leur apparition dans le message vocal, mais à élaborer les informations de sortie (par exemple, commandes pour un robot, paramètres de commande d'un synthétiseur, . . .) à une vitesse telle que les actions à entreprendre ne soient pas retardées. Compte tenu de l'utilisation générale des ordinateurs pour l'étape de reconnaissance, de leur vitesse de travail et des algorithmes de reconnaissance à effectuer, il n'est pas d'exemple où l'on ne puisse satisfaire l'impératif temps réel (à tout le moins, tant que les algorithmes ne sont pas exagérément complexes) pour les applications actuellement concevables.

Cependant, une sortie en temps différé, si le système possède d'importants moyens de mémorisation, peut autoriser des traitements poussés, donc une reconnaissance améliorée.

3. 3. - Les capteurs utilisés ont en commun de réaliser, non seulement l'acquisition (évidemment !) du signal vocal, mais aussi la réduction de l'énorme redondance de ce signal (du moins vis à vis de son contenu sémantique). Cette réduction s'obtient par échantillonnage et analyse du signal.

Les paramètres essentiels de réalisation paraissent donc être :

- utilisation de compression dynamique.

Il peut être nécessaire, selon les traitements ultérieurs, de calibrer le signal reçu en amplitude (en particulier lorsque les traitements sont plus axés sur les phénomènes stationnaires que transitoires).

- cadence de délivrement des évènements

Un évènement se définit comme l'ensemble des valeurs émis à chaque pas d'analyse par le capteur. Généralement, il est produit un évènement toutes les 10 milisecondes. Ce chiffre parait raisonnable, rapporté à la durée des éléments de base que l'on cherchera par la suite à identifier (par exemple, la durée d'un phonème voyelle est de 2 à 5, sinon plus, évènements).

Certains préfèrent une cadence plus rapide (mais jamais inférieure à la milliseconde) pour tenter de saisir des phénomènes transitoires fins, comme il peut en apparaitre dans l'émission de certaines consonnes ou dans l'interaction entre différents phonèmes.

- type de l'analyse et nature du banc de filtres.

La théorie du traitement du signal suggère plusieurs types d'analyse. Ces analyses sont pratiquement réalisées par des bancs de filtre.

Essentiellement :

- analyse fréquentielle (type vocoder)
- analyse temporelle (comptage des passages à zéro du signal dans une certaine bande de fréquence)
- extraction de cepstrum

Les bancs de filtre ont tendance à croitre en fréquence dans les travaux récents (pour en particulier mieux saisir les consonnes sifflantes) et couvrir une bande 100 à 8000 Hz semble un bon objectif. Le nombre de filtres, leur répartition dans cette bande ne répondent pas à des critères bien précis, mais souvent à ce dont on dispose. Par contre, les filtres sont des filtres très large-bande.

Les meilleurs résultats sont probablement à obtenir avec une trentaine de filtres très large-bande, plus concentrés entre 1000 et 2000 Hz, et se recouvrant partiellement.

.../...

Ces différents paramètres sont généralement fixés :

- soit en fonction des traitements ultérieurs que l'on envisage d'effectuer
- soit par référence à des traitements simples en théorie du signal ou des appareils existants (ou facilement réalisables) pour ces traitements simples (ex = vocoder : P6, P10, P3 - Icophone = P1 - P2)
- soit par référence à des modèles du système auditif (ex : imitation de cochlée, de membrane basilaire - en particulier d'après les travaux de Flanagan, de Bekesy - voir P5 et P9)

Les hypothèses adoptées donnent une grande diversité à la nature des évènements sortant du capteur. L'état actuel des réalisations ne permet pas de privilégier un point de vue : si le vocoder parait malgré tout assez imparfait, rien n'autorise encore à penser qu'un modèle bionique de l'oreille humaine est la solution (après tout, les avions ne battent pas des ailes ! ... et, par ailleurs, il convient de noter que jusqu'à présent, la bionique promet plus qu'elle ne tient dans le domaine du traitement du signal). Il semble plus raisonnable de croire que la qualité primordiale d'un capteur est de réduire de façon importante la redondance du signal vocal, et il est probable, dans un premier temps du moins (c'est-à-dire si l'on s'attache à réaliser une machine de type M1 non entièrement parfaite) que tout capteur est satisfaisant dès lors qu'il réduit la quantité d'information de façon pertinente.

3.4. - Le prétraitement doit prendre en compte la suite des évènements et les transformer de façon à présenter aux algorithmes d'identification des données satisfaisantes.

Plusieurs options sont là encore à prendre dont :

- élaboration d'autres évènements plus significatifs ou plus informants
- partition des suites d'évènements en segments.
- normalisation des évènements ou des segments.

Il n'existe pas d'accord général sur ces points. Là encore, les points de vue sont imposés par les algorithmes d'identification et surtout par le sens que l'on cherche à attribuer aux segments et aux évènements les composants, sens qui se réfère bien évidemment à une certaine interprétation de la structure du message vocal, car c'est au niveau de ce prétraitement qu'apparaît, dans la plupart des cas, la spécificité vocale de la machine à reconnaitre. A défaut d'accord, quelques tendances se discernent :

a) l'élaboration d'évènements peut prendre deux formes :

- réduction ou augmentation du nombre de paramètres d'un évènement (en fonction uniquement de cet évènement). Ceci a pour but de faire apparaitre des paramètres liés aux éléments stationnaires du message vocal : paramètres formantiques essentiellement .

.../...

- création de paramètres ou d'évènements à partir de plusieurs évènements ou dérivation entre évènements. Le but est de faire apparaitre des phénomènes transitoires par exemple interactions consonne - voyelle.

b) partition des séquences en segments.

Le découpage d'une suite d'évènements en segments est lié à l'idée que l'on se fait du signal vocal comme structuré par des "éléments simples"(phonèmes, phonatomes, ...) et des caractéristiques de ces éléments simples.

- une conception analytique des éléments simples conduit à ne pas découper la suite des évènements, mais à essayer d'attacher à chacun des évènements (ou groupe d'évènements constitué d'une fenêtre glissant sur la suite des évènements de façon continue) le nom d'un élément (voir les travaux de ROCHE P3 et CASTAN ET PERRENOU P10).

- une conception globale consiste à penser qu'un élément n'est pas caractérisé par un évènement, mais par un ensemble d'évènements (segment) soumis à certaines règles. Ces segments sont consécutifs et disjoints. Les règles, issues des hypothèses faites sur le sens physique des éléments, déterminent le (s) critère (s) de segmentation. Ce critère est généralement de segmenter la suite des évènements à ses périodes de stabilité, ou à des périodes déduites de celles des paramètres composant ces évènements (ce qui ne signifie pas que ces périodes de stabilité des paramètres sont les mêmes que celles du signal vocal); LIENART (P1) et GUEGUEN (P5) illustrent ces deux points de vue.

- comme bien sûr, la fréquence (100 Hz) d'apparition des évènements est telle qu'il existe généralement plusieurs évènements pour un élément simple, une conception intermédiaire revient à segmenter, non la suite des évènements, mais celle des noms d'éléments qu'on lui attache afin de préparer la décision. (par ex. P10)

c) normalisation des évènements ou des segments.

Le calibrage des données présente beaucoup d'intérêt dans une conception globale où la suite des évènements est découpée en segments consécutifs disjoints. En effet, le débit de parole, la façon de prononcer, la qualité du locuteur peuvent conduire, pour le même élément, à des segments au nombre d'évènements variable, à des évènements semblables mais non identiques, ...

Cette normalisation est une transformation portant, soit sur les évènements, soit sur les segments. Sur les évènements, il s'agit essentiellement de la modification des valeurs des paramètres le composant par un facteur correctif.

.../...

.../...

Ce facteur peut tenir compte (sous forme d'une fonction parfois complexe) des diverses caractéristiques présentant certain caractère de stabilité dans leur existence (par ex : fréquence des formants, pitch, ...) - voir GUEGUEN P5. Sur les segments, il s'agit plutôt d'affinité en temps, éventuellement d'un centrage du segment, et en fréquence (cf P1 et P2, travaux de J. S LIENART).

Notons :

. toute l'importance de cette normalisation qui tente de ramener à un standard commun les données de la reconnaissance et ce par des ajustements de paramètres que l'on peut avec profit utiliser pour l'auto-adaptation du système.

. l'aspect philosophiquement réaliste de la normalisation qui revient à dénier l'existence de valeurs absolues immuables des paramètres et à supposer leur relativité, donc à tenter de se ramener à des "échelles" standard.

3.5. - A l'issue de ce prétraitement, on se trouve en présence de suite d'évènements ou de segments qu'il s'agit d'identifier. La connaissance a priori des évènements ou segments correspondant aux éléments simples à identifier(qui sont assimilés aux phonèmes dans presque tous les projets), n'étant guère possible, l'utilisation de l'apprentissage comme moyen de constitution d'un dictionnaire d'éléments simples de référence s'impose. Ce dictionnaire sert de base (ou d'ensemble de constantes)aux algorithmes d'identification qui utilisent généralement l'ordinateur.

Selon le degré de confiance que l'on accorde à la représentativité des données à identifier (ou bien selon que l'on a plus ou moins de moyens!), les algorithmes peuvent être :

- des traitements très simples (et que l'on peut alors éventuellement câbler et donc se passer d'ordinateur) - par exemple : identification ou corrélation très simple sur les éléments du dictionnaire (LAMOTHE P8 a choisi cette voie).
- des traitements complexes utilisant des théories de reconnaissance des formes (CASTAN & PERRENOU - P10).
- l'utilisation de tout un ensemble d'algorithmes hiérarchisés (GUEGUEN P5) ou autogénérés (ROCHE - P3 et P4)

Dans les traitements issus de théories de reconnaissance des formes, les données (évènements ou segments) sont représentées généralement par des points dans un espace à n dimensions(n étant le nombre de paramètres distincts constituant l'évènement ou le segment à identifier). Si n est grand, des méthodes de réduction (analyse factorielle, régression multiple, ...) sont utilisées. Le problème est ensuite d'identifier le point nouveau en utilisant les points déjà connus par l'apprentissage (voir par ex : P10).

.../...

.../...

Ce problème n'a pas une solution unique et on dispose donc d'un ensemble d'algorithmes aux performances diverses (performance en temps de calcul et justesse de l'identification). Une amélioration peut provenir de l'utilisation d'un ensemble hiérarchisé de ces algorithmes (GUEGUEN P5) :

- utilisation préférentielle et successive d'algorithmes rapides, même si peu efficaces jusqu'à obtention d'une identification avec une probabilité de justesse estimée comme suffisante.
- utilisation de plusieurs algorithmes pour obtenir une convergence de l'identification.

La forme la plus sophistiquée de reconnaissance est l'obtention automatique des algorithmes d'identification (ROCHE - P3).

L'algorithme de reconnaissance est conçu comme la mise en oeuvre successive d'opérateurs qui élaborent à partir des données (évènements ou segments) des résultats permettant l'identification, c'est à dire le rattachement à une classe d'élément simple. Ces opérateurs sont fabriqués aléatoirement (par exemple, combinaison quelconque de multiplications et d'additions) et seuls sont retenus les meilleurs, ceux qui ont le plus grand pouvoir informant, c'est-à-dire ceux qui donnent les résultats les plus probants sur les données d'apprentissage. (Remarquons que par ce biais se réalise l'adaptation du système à un groupe de locuteurs, celui qui a fourni les données d'apprentissage). Cette méthodologie est satisfaisante en ce qu'elle semble analogue au fonctionnement de notre cerveau (malgré ce que j'ai écrit plus haut de la bionique!) et également pour les résultats obtenus dans d'autres problèmes de reconnaissance des formes.

4. - Un système est réalisé après avoir fait choix d'un certain nombre d'hypothèses. Des mesures vont permettre de juger son fonctionnement. De ces mesures, on doit pouvoir conclure sur deux points :

- valeur des concepts mis en oeuvre.
- évolution du système et prospective sur les recherches à effectuer.

4. 1. - Avant de tirer ces conclusions, il n'est pas inutile de s'interroger sur la valeur des mesures. En effet, il est très facile de justifier quelque hypothèse que ce soit par des mesures grâce à des chiffres "bien interprétés". Il importe donc de bien situer les mesures effectuées, en fixant un certain nombre de paramètres, liés au système et à son utilisation (ou son objectif). Ces précisions utiles sont fréquemment escamotées dans les publications et cela doit faire douter au plus haut point de la généralité des hypothèses que ces mesures justifient.

.../...

.../...

Les mesures effectuées sont généralement des taux = nombre de phonèmes bien reconnus/nombre de phonèmes entrés dans le système, et peuvent avoir plusieurs objectifs:

- validation des concepts mis en oeuvre
- détection d'imperfections et améliorations
- comparaison de différentes réalisations
- connaissance des performances d'une réalisation

Les paramètres fixant le contexte des mesures peuvent varier selon les objectifs, mais sont essentiellement :

- en ce qui concerne les phonèmes d'apprentissage
  - . nombre et type des locuteurs prononçant les phonèmes d'entrée
  - . nature de l'apparition des phonèmes : phonèmes isolés, mots, parole continue, ...
  - . types des phonèmes : représentativité de l'échantillon des phonèmes (par rapport à l'ensemble des phonèmes du français)
  - . nombre d'occurrence des divers phonèmes : nombre total, nombre pour chaque type de locuteurs, ...
- en ce qui concerne les phonèmes de test
  - . d'une part, des paramètres identiques à ceux mentionnés ci-dessus.
  - . d'autre part : -différences entre l'ensemble des phonèmes d'apprentissage et de tests.
    -différences entre l'ensemble des locuteurs d'apprentissage et de tests.

L'importance de ces précisions apparait comme évidente; on ne peut comparer des taux identiques obtenus par deux mesures dont

. une porte sur quelques dizaines de phonèmes utilisés pour l'apprentissage et le test et prononcés par un seul locuteur,

. l'autre met en jeu quelques centaines d'occurrences de chacun des phonèmes du français, et ceci pour un nombre important de locuteurs représentatifs, des phonèmes différents servant à l'apprentissage et aux tests.

4.2. - L'obstacle essentiel à la multiplication des mesures provient de la difficulté à disposer facilement de données sous la forme nécessaire à leur utilisation.

.../...

.../...

En effet, beaucoup de réalisations envisagées sont étudiées par des méthodes de simulation en ordinateur.

Les données d'apprentissage et de test doivent donc se présenter sous une forme numérique. Elles sont longues à créer, et cette création nécessite l'emploi de matériels(analogiques-digitaux par ex.) dont on ne dispose souvent pas. Aussi les équipes disposent-elles souvent d'ensemble de données peu important et rarement renouvelé. Cependant, grâce aux travaux du groupe "Reconnaissance de la parole", des échanges de données doivent susciter des progrès, chacun disposant de données plus nombreuses (donc des tests plus probants) et identiques (donc des mesures permettant les comparaisons).

5. -

5. 1. - Les mesures existantes (voir tableau en annexe) sont encore très fragmentaires et il n'est pas raisonnable d'en tirer des conclusions définitives quant à la validité des différents concepts mis en oeuvre. (Il se dégage cependant l'impression qu'atteindre des taux compris entre 60 et 80 % ne présente pas des difficultés insurmontables. Cela incline donc à penser que les options à prendre sont assez indifférentes au niveau des machines de type M1 (et en particulier leur partie capteur) (pourvu que ces options soient réalistes, eû égard au problème à traiter), si on ne désire pas atteindre des taux trop sévères ($>$ 95 %).

5. 2. - Le but d'un système n'est généralement pas de reconnaître des phonèmes, mais des mots ou des phrases.

Par conséquent la machine M1 n'existe pas de façon autonome, et sert d'entrée à une machine M2, qui interprète la séquence de phonèmes. Dans ce cadre se pose la question de la nécessité d'un taux de 100 % au sortir de M1 pour l'obtention d'une reconnaissance sans erreurs en sortie de M2. Il est bien évident que non, car les éléments que doit reconnaître M2 ont un sens (du moins dans le plus grand nombre de cas d'applications) et ne sont pas aléatoires. Par conséquent, ils sont soumis à un ensemble de règles (ou contraintes) dont l'utilisation autorise l'imperfection de la séquence de phonèmes entrante. (imperfection = phonèmes non identifiés, phonèmes erronés, phonèmes non présents dans le message original). Ainsi par exemple, soit un système reconnaissant uniquement les 10 chiffres. L'apparition de la séquence de phonèmes "E. O", où pourtant sont absents les phonèmes correspondants à Z et R (soit 50% d'erreurs), suffit à décider que le chiffre prononcé est "zéro".

.../...

.../...

5. 3. - Reconnaître qu'un taux de 100 ne constitue pas une absolue nécessité conduit à se poser des questions ouvrant probablement de fructueuses voies de recherche :

- des recherches, difficiles (et sans doute longues et coûteuses), en vue d'atteindre des taux approchant 100 se justifient-elles ?
- peut-on déterminer un ou plusieurs seuils minimaux du taux de reconnaissance, en fonction des contraintes exploitables dans les niveaux M2 et M3 ?
- comment utiliser au mieux une machine donnée de type M1 existante, en déterminant expérimentalement les contraintes de niveau M2 et M3 qu'elle satisfera le mieux ?

Ces deux dernières voies sont actuellement prometteuses. En effet, les efforts importants déjà fournis pour réaliser des machines de type M1 ayant des qualités satisfaisantes laissent penser que, pour des améliorations lentes et faibles du taux de reconnaissance, les recherches seront longues et coûteuses et les traitements à effectuer lourds et complexes. Par contre, le domaine des machines M2 et M3 est actuellement quasi-inexploré (tout au moins dans une perspective reconnaissance de la parole) et des études sur ces sujets peuvent être source de progrès rapides, notamment :

a) - détermination et étude des contraintes dues à l'usage de langages non arbitraires pour une application donnée.

- étude de l'utilisation de ces contraintes.

Ces contraintes interviendront à tous les niveaux. Cela sera par exemple :

- contraintes phonologiques : probabilité d'existence d'un phonème, d'une association de phonèmes, ...
- contraintes syntaxiques : vocabulaire du langage de l'application, règles sur les associations des mots de ce langage, ...
- contraintes sémantiques : signification des phrases obtenues par rapport à l'application, ... (voir en particulier à ce sujet, les idées exposées et mises en oeuvre par T. VINOGRAD pour la réalisation d'un robot manipulant des polyèdres colorés-MIT BOSTON U.S.A.).

L'étude de ces contraintes devra se faire dans une perspective de reconnaissance de la parole, mais en faisant appel à des chercheurs des disciplines concernées (phonéticiens, théoriciens de l'informatique, théoriciens de l'étude des langages, ...).

.../...

b) - étude expérimentale systématique d'un système existant (cf. travaux du CNET LANNION - P6).

- détermination des contraintes satisfaites au mieux par ce système (par ex : phonèmes les mieux reconnus).

De ces études sera déductible un (ou une classe de) langage(s) bien adapté(s) à ce système. Ce système pourra être par suite utilisé dans toutes les applications où le choix du langage (vocabulaire et règles syntaxiques) peut comporter une grande part d'arbitraire : commande de machines, interrogation élémentaire de banque de données, . . .

5. 4. - L'exploration de ces sujets ne permettra pas de résoudre tous les problèmes et il faudra nécessairement, lorsqu'un palier sera atteint dans les réalisations, reprendre des recherches plus fondamentales sur le signal vocal, de façon à se rapprocher d'un taux 100 au niveau de la reconnaissance des phonèmes. En effet, reconnaître la parole revient au fond à identifier les invariants composant cette parole. D'où l'utilité de rechercher des invariants dans le signal vocal, indépendants des mots où interviennent les phonèmes, des locuteurs, . . . Des études de phonétique, d'acoustique, . . . ont donc encore leur nécessité. Mais, dans l'immédiat, je crois que l'effort doit surtout se porter sur les niveaux M2 et M3.

6. - Dès maintenant, il apparait une foule d'applications pour des machines reconnaissant la parole. Les recherches actuelles et des réalisations comme par exemple la "machine de bureau" de J. Y. GRESSER (CNET-LANNION) permettent un raisonnable espoir de développement de systèmes qui couvriront peu à peu les différents créneaux d'utilisation. Le champ des recherches futures est très vaste, et s'ouvre sur des disciplines nouvelles (théorie des langages, par ex.), ce qui demandera la participation de chercheurs de tous horizons. La diversité des équipes intéressées par la reconnaissance de la parole et l'existence du groupe "Reconnaissance de la parole" placé sous l'égide du G. A. L. F. sont un gage de saine émulation et un espoir de résultats concluants.

Ainsi peut-on espérer que la "machine à écrire phonétique" ne soit plus en l'an 2000 le serpent de mer de l'intelligence artificielle.

ANNEXE - TABLEAU DES PROJETS EN COURS D'ETUDE.

Laboratoire

L.I.M.S.I. (Laboratoire d'Informatique pour la mécanique et les sciences de l'ingénieur) : LIENARD, TEIL, MLOUKA (Pr. MALAVARD).

Laboratoire d'acoustique PARIS VI° : M CASTELLONGO, LEIPP, SAPALY (Pr. SIESTRUNCK)

Objectifs

-Transcription d'un message vocal en un message écrit avec des signes phonétiques.

-10 ou 15 locuteurs représentatifs au point de vue acoustique / 90 % de reconnaissance quelque soit le texte / 1974

Options

-Isoler les phonatomes (association de deux phonèmes)
-Normaliser en temps et en fréquence les segments de reconnaissance.
-Reconnaître avec un dictionnaire de phonatomes.
-Adapter le système au locuteur avec un mot-clé.

Résultats actuels

-Dictionnaire de 100 phonatomes obtenus par apprentissage.
-Test avec 150 phonatomes (dont 100 différents) provenant de 7 phrases prononcées par un locuteur.
-70% de phonatomes et 85% de signes phonétiques reconnus correctement.

Réalisation :

. Capteur. Banc de 64 filtres (32 ultérieurement) linéairement répartis de 0 à 5000 Hz

. 1 évènement de 32 valeurs (niveaux d'énergie) toutes les 10 millisecondes.

Prétraitement

. Segmentation de la suite d'évènements par les périodes de stabilité de cette suite (déterminés par les maxima de la fonction de stabilité obtenue par corrélation des évènements avec ceux les précèdant de 3 pas - 30 millisecondes).

. Normalisation en temps des segments obtenus en 5 évènements, avec centrage sur les minima de la fonction définie ci-dessus. (2 évènements - extrémité, évènement du centre, 2 évènements médiateurs entre centre et extrémités font 5 évènements).

Traitement

. Corrélation avec tous les éléments du dictionnaire (ultérieurement, une meilleure organisation du dictionnaire permettra d'éviter de faire toutes les corrélations) et choix du meilleur corrélat.

.../...

Laboratoire idem P 1.

Objectifs -Reconnaissance de mots isolés (100 environ) avec utilisation d'un petit calculateur.
-15 locuteurs représentatifs /50 mots reconnus à 95% /fin 1973.

Options -Compression maximum du signal en amplitude, temps et fréquence.
-Reconnaissance globale par une méthode de correspondance.

Résultats actuels - 8 locuteurs différents prononçant 14 mots.
- 1 locuteur est utilisé comme référence (dictionnaire utilisé dans les correspondances)
7 autres locuteurs donnent 70 % de reconnaissance sans adaptation.

Réalisation

. Capteur 16 voies d'analyse réparties de 0 à 5000 Hz (avec une plus grande concentration entre 1000 à 2000 Hz).

. prétraitement : compression de séquence des évènements (correspondant à 1 mot) par conservation unique des évènements correspondant aux extrémas de la fonction de stabilité (cf P1). (Facteur de compression de 5 environ).

. Traitement : reconnaissance par la méthode des correspondances.

La méthode des correspondances consiste à associer à chaque évènement E (t) de la séquence à identifier un évènement R (t') de la séquence de référence. Ce dernier est choisi par programmation dynamique à l'intérieur d'une fenêtre temporelle (t', t' + θ) de manière à obtenir un taux de coïncidence égal ou supérieur à un certain seuil Co. La moyenne des taux obtenus sur une séquence est une mesure de sa ressemblance en temps normalisé avec la référence.

Laboratoire : Institut de Programmation (PARIS 6°) J.C. SIMON - ROCHE.

Objectifs . Reconnaissance de "phonèmes" pour utilisation dans une machine reconnaissant les mots.
. Nombre quelconque de phonèmes et de locuteurs / 60% de reconnaissance /1974.

Options : Capteur indifférent.
Génération automatique des opérateurs traitant les données et de l'algorithme utilisant ces opérateurs.
Travail en continu sur les évènements.
Traitement très rapide.

Résultats

. 300 phonèmes utilisés pour le test et l'apprentissage
. 80% de bonne reconnaissance.

Réalisation

-Capteur . Vocoder (données fournies par la C.G.E. sur un vocoder CNET)
. Echantillonnage à 100 Hz

-Retraitement/traitement

. Identification continue de segments de 5 évènements consécutifs, se déplaçant d'un évènement à chaque identification.
. Création des opérateurs par apprentissage en maximisant leur pouvoir informant.
. Création de la structure gérant ces opérateurs.

Objectif

- Reconnaissance de 1 mot parmi 500 en utilisant en entrée une suite (erronée) de phonèmes.
- Reconnaissance bien inférieure au temps réel /90% d'exactitude / 1974.

Options

. Suite des phonèmes d'entrée peut comporter jusqu'à 40% d'erreurs (mauvais phonèmes, phonèmes parasites ou omis).

. Opérateurs de traitement travaillant sur la position des symboles phonétiques dans la suite.

. Utilisation d'une methodologie de traitement identique à celle du projet P3.

Résultats

. Pas encore d'expériences.

Réalisation

. Capteur : constitué par la réalisation P 3

. Traitement : Méthodologie identique à celle de P 3.

Laboratoire: E . N . S . des Télécommunications.

GUEGUEN, MAISSIS, PAU (CARAYANNIS, GAAFAR)

Objectif: .Reconnaitre en temps réel un mot dans un dictionnaire de 200

.Reconnaitre des phrases de structure simple dans un contexte restreint.

Options: .i -Analyse pertinente du signal vocal par propagation dans un système régi par équations aux dérivées partielles (simple spectre insuffisant)

.ii -Procédures statistiques de reconnaissance modulaires et hiérarchisées

.iii -Ces options sont destinées à minimiser le temps de calcul. (i) est orientée vers 1' auto-adaptation par identification du locuteur.

Résultats:

.Base de données: 748 occurrences de 23 phonèmes extraits de 238 mots isolés 2 locuteurs de pitch tres différent (110 et 180 Hz)

.Reconnaissance phonèmique 1 locuteur sur phonèmes d' apprentissage 83% en 200 ms. sans normalisation non-linéaire.

.Reconnaissance phonèmique 1 locuteur sur phonèmes d' apprentissage 84% en 160 ms. avec normalisation non-linéaire.

.Reconnaissance des mots : Concaténation avec probabilités de transition d'ordre 1 : 54% (dictionnaire libre). Avec utilisation du dictionnaire (distance entre trajectoire de mots) supérieure à 88%.

Réalisation:

Capteur: Simulation approximative de la membrane basilaire

. 7 filtres larges à flanc doux $\Delta f/f$ constant de 300 à 5000 Hz
. 7 filtres de mème type légérement decalés pour créer la derivée spatiale
. 3 filtres rectangulaires pour calcul approximatif des formants(zéro-crossing
. Segments minimaux de 10 ms. (100 Hz)
. Codeur analyseur permettant une réalisation hardware.

Prétraitement:

. Segmentation quasi-phonèmique par zones de stabilité sur tous les paramètres
. Normalisation non-linéaire des occurrences et locuteurs (résultat expérimental confirmé par modèle théorique).

Traitement:

. Réduction des données par analyse factorielle à l'apprentissage(3axes 86%)
. Procèdures de reconnaissance prédéterminées ordonnées par pouvoir discriminant croissant (et donc temps calcul croissant)gérées hiérarchiquement.
. Utilisation séquentielle ou parallèle des procèdures pour obtenir un taux satisfaisantou motiver un rejet vers les niveaux inférieurs.

Laboratoire : CNET LANNION
GRESSER, MERCIER (Centre de Calcul)
Participation de CARTIER (Département d'acoustique)

Objectifs

. Reconnaissance d'un vocabulaire limité (quelques dizaines de mots isolés) mesure de systèmes existants pour application éventuelle.
Reconnaissance à l'aide du phonétographe de tous les chiffres de 0 à 9 prononcés dans un appareil téléphonique par n'importe quel locuteur sans adaptation .
. Etude du comportement de l'utilisateur (notamment pour rédaction de manuel d'utilisation)

Options

. Prendre système existant
. Choisir des types de vocabulaire
. Adapter éventuellement
. Reconnaître en temps réel
. Essais extensifs notamment sur un très grand nombre d'occurrences de chaque mot.

Résultats

. Vocabulaire de 20 mots (machine de bureau sur Ramsés 1 L avec Vocoder) : 80 à 85 % selon locuteur avec adaptation (essais en temps réel)
L'adaptation au niveau des mots(référence à un dictionnaire personnalisé) permet d'atteindre 90% au moins pour chaque locuteur (essais simulés)
. Vocabulaire de 8 chiffres (0 à 7) sur un vocoder : 100 % quel que soit le locuteur.

Réalisations :

Appareils testés

. Vocoder + algorithme de "base"
. phonétographe de Dreyfus-Graf (détection d'éléments de type phonémique)
. Vocoder + algorithme segmentant sur les pointes d'énergie (segmentation en "syllabes ouvertes").
. Vocoder + méthode des correspondances (cf. P2)

Laboratoire - Idem P6

Objectif . Reconnaissance d'un langage de communication parlée à vocabulaire limité (quelques centaines de mots), à syntaxe et sémantique définies a posteriori

Options . Adaptation au locuteur et organisation de la chaîne de reconnaissance pour obtenir 100 % de reconnaissance sémantique.
. Feed-back sur opérateur pour validation
. Capteur indifférent
. Analyse linguistique au sens classique
. Temps de traitement sans importance
. Mots isolés, puis passage progressif à la parole continue.

Résultats

Pas de résultats actuellement.

Réalisation

. Capteur Vocoder, phonetographe, codeur digital avec extraction de paramètres articulatoires, ...

. Prétraitement/Traitement : utilisation de méthodes déjà existantes au niveau segmentation élémentaire (machine de type M1)

. Implantation système en cours sur un ensemble 10020-10070

-Note : MERCIER poursuit parallèlement des études sur la recherche d'éléments simples optimaux (phonèmes, syllabes ouvertes,...) pour la reconnaissance.

Laboratoire L.E.A. (Laboratoire d'Electricité et d'Automatisme) NANCY I LAMOTTE, HATON.

Objectif . Reconnaissance d'un vocabulaire limité (100 à 200 mots) Utilisation pour la commande numérique de machine-outils.

. 1 ou 2 locuteurs / 95 % mots exacts / 1973

Options . reconnaissance acoustique au niveau phonème par hardware
. " " mot par un petit calculateur

Résultats . phonèmes : 50 % pour 1 locuteur sur 200 phonèmes (dont 1/2 pour apprentissage).

Réalisation

. Capteur . échantillonnage à 100 Hz
. banc de 25 filtres répartis exponentiellement de 150 à 7000 Hz donnant des niveaux d'énergie.

-Prétraitement

. 24 paramètres p : comparaison des sorties de 2 filtres consécutifs (p = 0 si amplitude filtre j + 1 amplitude filtre j et = 1 sinon)
. 6 paramètres élaborés en sortie des comparateurs (dont 1 sommateur pour détection de silence)
. Ces paramètres commandent des intégrateurs à cte de temps durée d'un phonème donnant signaux analogiques - 5, + 5 v.

. Traitement

. Identification par matrice d'apprentissage 30 x 30
. En entrée des colonnes : 30 tensions issues du prétraitement.
. En sortie - des 30 lignes (correspondant à 30 phonèmes), un détecteur de maximum choisissant la ligne à V max.
. Les poids de la matrice d'apprentissage (résistances) sont déterminés par apprentissage.
. Utilisation de contraintes linguistiques pour optimiser les séquences de phonèmes.
. Recherche des mots avec un dictionnaire.

Perspectives :

- branchement de ce système sur un ordinateur pour travail en temps réel.
- utilisation des contraintes syntaxiques pour l'entrée orale d'un programme en ordinateur.

Laboratoire THOMSON-CSF DASM CAGNES SUR MER
TOURNOIS - ALINAT

Objectif

- Transcription phonétique du message vocal en une suite de phonèmes (dans la mesure où ils sont prononcés).
- 95% des locuteurs males - articulation soignée.
  80% de réussite pour les phonèmes soutenus (2/3 des phonèmes)
  60% de réussite pour les consonnes explosives (1/3 des phonèmes)
  pour 1972 - 1973
- Par adjonction d'une partie programmée reconnaissance de mots appartenant à un vocabulaire limité pour une articulation moins soignée.
- Application : entrée de donnée dans un ordinateur : par exemple liste de cablage pour système logique représentant des milliers d'adresses.

Options

- capteur fortement inspiré de la nature (modèle de cochlée)
- phonèmes caractérisés par la présence simultanée ou séquentielle de critères dont les principaux sont relatifs à l'excitation et à la position des formants (zones formantiques).
- critères définis par rapport à l'organe de réception et non pas par rapport à l'organe d'émission de la parole.
- décisions prises par tout ou rien (pour simplification du système).
  Conséquence :
  la prononciation doit être soignée.
- Pour une parole mal articulée utilisation d'un vocabulaire limité ou prise de décisions indiquant des probabilités de présence et intervention des probabilités de succession des phonèmes et de la prosodie.

Résultats

- Une première maquette (temps réel) donnait pour les voyelles et (CH, J) prononcés lentement un taux de réussite de l'ordre de 70 % (1971)
- Cette maquette est en cours d'amélioration. Elle reconnaitra également les consonnes.

Réalisation

Capteur : -Emphases à l'entrée.
-Banc de 96 filtres (100 à 10.000 Hz) à fonction de transfert approximant celle de l'oreille (cf. travaux de Bekesy)
-Détection-Intégration des sorties et échantillonnage à 250 Hz.

Prétraitement - Recherche des formants par un ensemble de filtres numériques passe bande non récursifs.
- Extraction des autres critères : excitation, nasalisation des voyelles,...

Traitement - Utilisation de zones formantiques dans lesquelles doivent se produire les formants.
- Vérification de la présence ou de l'absence des autres critères.
- Décision de présence prise par tout ou rien en fonction des critères présents (simultanés et parfois séquentiels).

Laboratoire Cybernétique des Entreprises et Reconnaissance des formes.
(U E R Informatique - Université TOULOUSE) CASTAN PERENNOU

1) Objectif

. Reconnaissance d'un vocabulaire limité pour un nombre limité de locuteurs en temps réel.

Option

. Utilisation d'un vocoder à canaux
. Segmentation parole/non parole
. Reconnaissance par séparation linéaire sur formes globales.

Résultats

. 10 chiffres - 80 à 85 % de reconnaissance

2) Objectif

. Reconnaissance d'un vocabulaire tenant compte des imperfections du vocoder.

Réalisations

1).Reconnaissance des voyelles par méthode globale
.Segmentation du mot en "syllabes" par la recherche des zones d'instabilité.
.Interprétation linguistique et correction tenant compte du vocabulaire.

2).Reconnaissance de chaque évènement par séparation linéaire.
.Traitement de la chaîne d'éléments reconnus pour interpréter ensuite la syllabe
.Interprétation linguistique et corrections tenant compte du vocabulaire.

3) Objectif

. Etude des modèles mathématiques des cochlées en vue de la définition du traitement du signal de parole.

Travaux

. Connexion d'un vocoder du CNET à 16 canaux au 7044 IBM.
. Programmes :
-analyse et synthèse de la parole en temps réel
-segmentation parole/non parole
-segmentation automatique
-normalisation des mots en temps
-reconnaissance par séparation linéaire.

LABORATOIRE

Département d'Etudes et de Recherches en Automatique du CERT à TOULOUSE.

LABARRERE, GIMONET, KRIEF.

OBJECTIFS

- Recherche et extraction des paramètres significatifs au sens de l'information contenue dans le signal et non au sens de la physique du phénomène.

OPTION

- non accès aux formants, ni aux formes du conduit vocal,
- caractérisation de l'information temporelle contenue dans le signal.
- une première approche basée sur une étude des largeurs de commutation (distance séparant deux passages consécutifs à zéro). Les "patterns" proposées à l'algorithme de reconnaissance sont des courbes de distributions du nombre de commutations en fonction de leur largeur.
- une deuxième approche consiste, à partir des zones à forte variation d'énergie du signal vocal, à définir un train d'impulsion.
- recherche par une méthode d'identification classique (déterministe ou statistique, méthode paramétrique ou estimation à la Kalman) de la fonction de transfert ou de l'équation d'état permettant de passer du train d'impulsions au signal vocal.
- utilisation de ces nouveaux paramètres pour la reconnaissance.
- inclusion dans le modèle de la forme d'onde des cordes vocales, le spectre des explosions (hypothèse de non gravité de ce mélange d'information), ceci sera surtout valable pour des sons générés à partir d'une excitation dont l'énergie est concentrée dans le temps.

RESULTATS

Périodiquement, vérification des progres par des manipulations partielles. En ce qui concerne la reconnaissance à partir des histogrammes, une expérience de reconnaissance de voyelles isolées (A, E, I, O, U, ET, IN, AN, ON, OU) a conduit à des pourcentages de reconnaissance de 85%, pour un seul locuteur.

Ce pourcentage est global, et est pénalisé par la difficulté réelle de distinguer certains sons (ON, AN, IN) lorsqu'ils sont isolés. En ce qui concerne la dernière approche, uniquement des expériences de synthèse ont été faites comme guide dans la réduction de l'information.

REALISATION

IBM 360-44, et de son codeur permettant de coder la parole en temps réel, mais par tranches de 2 s (suffisant pour des phonèmes ou des mots isolés).

L'Information est immédiatement convertie en histogrammes ou en coefficients de fonction de transfert. Pour la reconnaissance, utilisation un algorithme de classification linéaire très simple.

SEGMENTATION AUTOMATIQUE DE LA PAROLE EN PHONATOMES

J.S.LIENARD et M.MLOUKA

Laboratoire d'Acoustique
de l'Université PARIS VI
(département de Mécanique)

Laboratoire d'Informatique
pour la Mécanique et les
Sciences de l'Ingénieur
(CNRS - ORSAY)

Pages

I - INTRODUCTION

La reconnaissance vocale peut à notre sens adopter deux démarches bien distinctes. L'une, analytique, consiste à isoler dans le message parlé des éléments tels que traits distinctifs, phonèmes, phonatomes, syllabes, etc , et à les normaliser (ou les paramétrer) avant de les identifier. L' autre, globale, considère dès le départ des mots isolés, comparés en bloc aux mots-références. La segmentation est l'une des principales difficultés de la reconnaissance analytique. Nous présenterons ici un processus de segmentation en phonatomes reposant sur l'étude de la stabilité temporelle de la parole.

Cette étude a bénéficié de l'aide du Comité de Recherche en Informatique, et de la Direction des Recherches et Moyens d'Essais.

II - CADRE EXPERIMENTAL

Un important corpus de parole (30 minutes), composé de mots, de phrases et de phonatomes prononcés par huit locuteurs différents (masculins et féminins) a été enregistré, numérisé et analysé par FFT sur l'ordinateur IBM 360 50 75 du CIRCE (CNRS). Cette phase laborieuse de l'expérimentation avait pour but d'obtenir une représentation de la parole comportant toute l'information sémantique telle que nos études de synthèse l'avaient mise en évidence dans la plan temps-fréquence (bib 1 et 2). Cette première étape a été atteinte moyennant des traitements décrits par ailleurs (bib 1 et 3), et notamment une pondération fréquentielle, une régulation de niveau et divers lissages.

L'ensemble du corpus est actuellement disponible en machine sous forme de sonagramme numérique couvrant la bande 0 - 5 kHz avec une résolution fréquentielle de 78 Hz (64 voies d'analyse) et une résolution temporelle de 10 ms.

III - STABILITE TEMPORELLE DU SQUELETTE SEMANTIQUE

Nous appelons squelette sémantique de la parole la Forme (Gestalt, ou totalité perceptive), décrite dans le plan temps-fréquence par un graphisme simple, qui reflète les mouvements du conduit vocal et convoie l'information sémantique du message. Ce graphisme, dont la pertinence peut être vérifiée par synthèse, obéit aux lois des Formes: il peut, en particulier, subir diverses anamorphoses et se prêter à une décomposition en formes élémentaires (phonatomes) de plus grande cohésion.

Sur la figure 1 se trouve représenté le squelette sémantique de la phrase "As-tu vu ce fameux lapin ?" et sa décomposition en phonatomes. La notation phonétique n'est utilisée ici que pour repérer commodément les frontières des phonatomes et n'implique nullement l'existence objective des phonèmes.

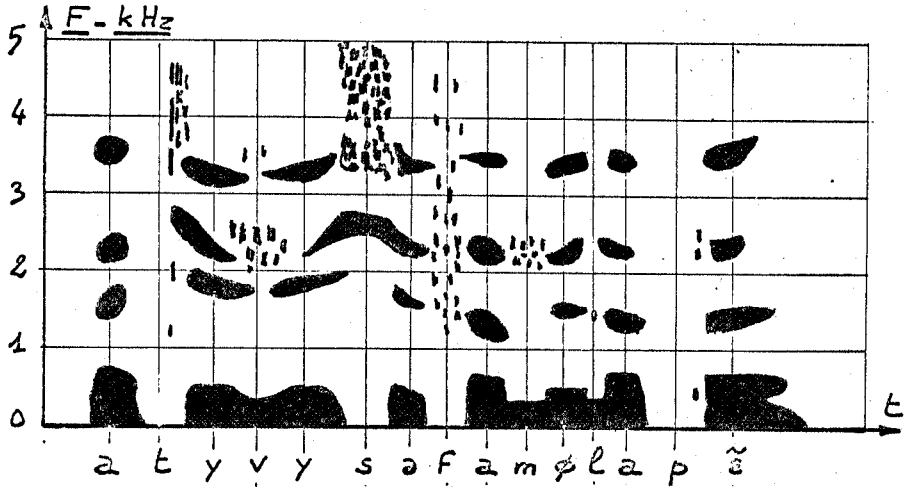


Fig 1 - Squelette sémantique de la phrase "As-tu vu ce fameux lapin ?" (schématisation manuelle)

Les phonatomes sont délimités par des états quasi-stationnaires qui par eux-mêmes ne convoient aucune information, car celle-ci est liée à l'imprévisibilité et au changement. On peut donc trouver les frontières des phonatomes en étudiant la stationnarité (ou stabilité temporelle) du squelette sémantique.

Le traitement s'apparente à l'autocorrélation : on compare le squelette sémantique à l'instant $(t+\theta/2)$ à ce qu'il était à l'instant $(t-\theta/2)$, le paramètre θ (distance d'ordre) étant choisi aux environs de 30 ms pour une vitesse normale d'élocution. Le résultat de cette comparaison est porté selon une courbe $S(t,\theta)$ évoluant entre 0 et 1, dont les maxima relatifs indiquent les instants de quasi-stabilité, donc les frontières des phonatomes. Les minima relatifs sont également intéressants, car ils marquent les instants de moindre stabilité, et correspondent ainsi aux événements les plus informatifs du message, que nous appelons les pivots des phonatomes.

IV - MISE EN OEUVRE DE LA SEGMENTATION-NORMALISATION

Les déterminations du critère de comparaison, de la distance d'ordre θ et du processus de normalisation méritent quelques précisions.

Nous avons développé dans les problèmes de reconnaissance la notion de taux de coïncidence entre deux configurations, c'est à dire le rapport de la partie commune au tout (bib 3).Entre deux ensembles A et B nous définissons le taux de coïncidence par le rapport

$$\tau_{AB} = \frac{K + A \cap B}{K + A \cup B}$$

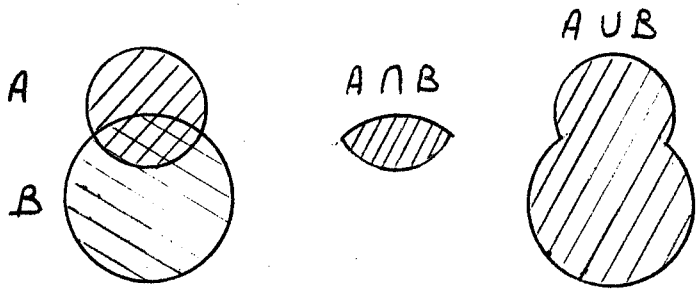


Fig 2 - Eléments de définition du taux de coïncidence

La constante K est nécessaire si l'on considère que deux ensembles nuls donnent une coïncidence unité. Elle permet en outre de définir un seuil au dessous duquel les ensembles peuvent considérés comme nuls (bruit de fond).

Si les deux configurations sont des spectres discrétisés (ou événements) notés $E_1(F)$ et $E_2(F)$, le taux de coïncidence s'exprime par

$$\tau_{12} = \frac{K + \sum_F \min [E_1(F), E_2(F)]}{K + \sum_F \max [E_1(F), E_2(F)]}$$

Ce type de comparaison s'applique parfaitement à l'étude de stabilité du squelette sémantique ; les exemples présentés ici ont été obtenus selon cette méthode. Mais d'autres définitions de la distance ou de la coïncidence peuvent être adoptées avec succès, sans modifier le principe de la segmentation.

La valeur de la distance d'ordre θ n'est pas critique. Des courbes $S_1(t)$ et $S_2(t)$ établies avec $\theta_1 = 20$ ms et $\theta_2 = 40$ ms sont pratiquement parallèles. Cependant, si θ devient trop faible (quelques millisecondes), $S(t)$ tend vers 1 et ne permet plus de détecter que les variations très petites (consonnes plosives, p.ex.). Inversement, si θ devient trop grand (de l'ordre de 100 ms), les seuls maxima notables sont obtenus sur les sons quasi-permanents (voyelles ou fricatives prolongées). La valeur $\theta = 30$ ms convient à peu près pour tous les échantillons de parole de notre corpus.

Pour détecter les extrema de la courbe $S(t)$ nous avons appliqué un traitement voisin d'une double dérivation, avec le souci de mettre en évidence les segmentations "difficiles", celles que l'on observe par exemple sur les liquides et les nasales, pour lesquelles le niveau sonore et les fréquences formantiques varient peu.

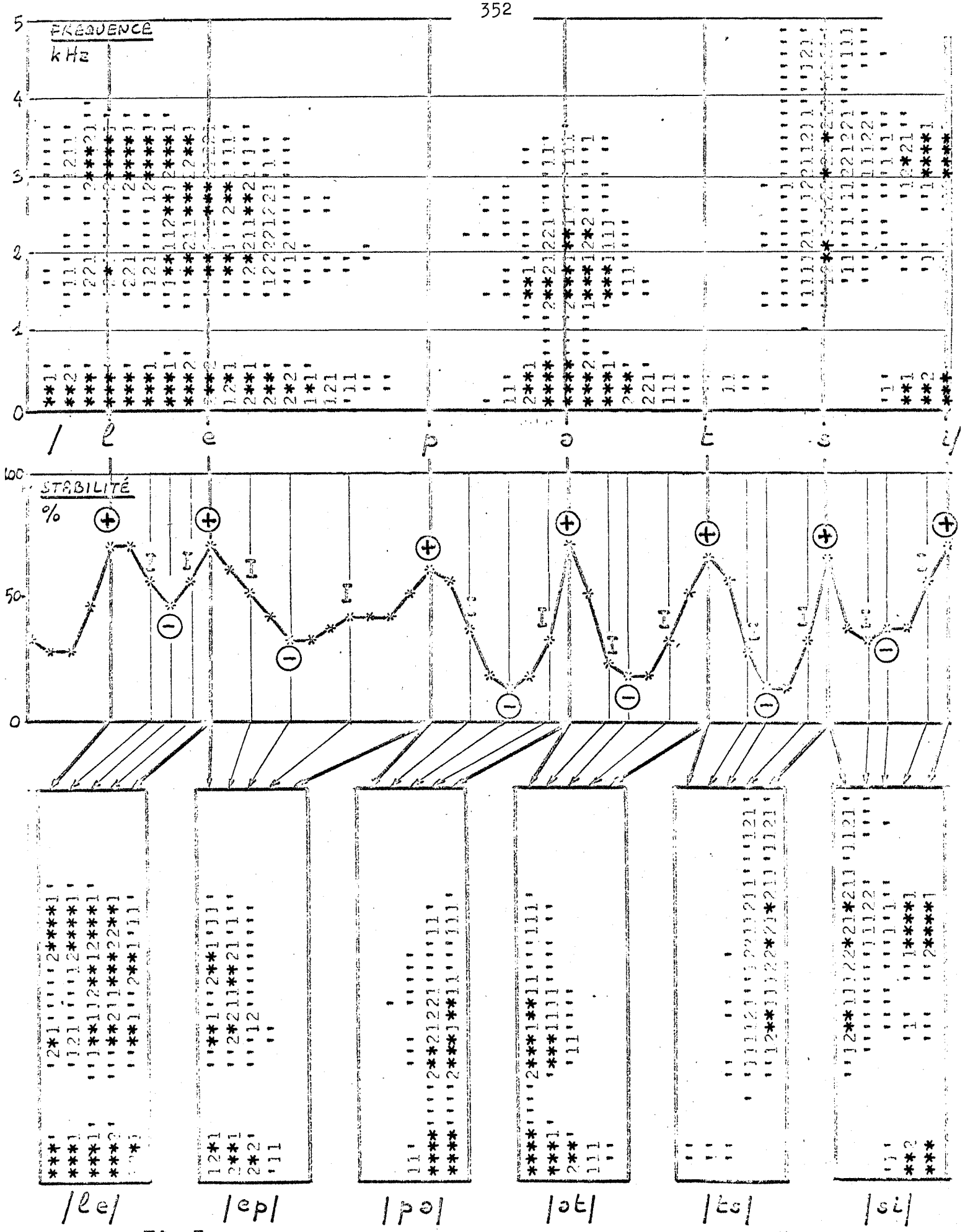


Fig 3

Principe de la segmentation du message en phonatomes, et de leur normalisation temporelle. Séquence "Les petits", locuteur JSL.

Le processus de segmentation-normalisation est illustré par la figure 3, qui représente la séquence "les petits" , prononcée par un locuteur masculin. Cinq événements significatifs sont prélevés dans le message : les deux frontières (deux segmentations positives consécutives), le pivot (segmentation négative), et deux événements intermédiaires. Les phonatomes ainsi reconstitués sont comparables à ceux que nous utilisons en synthèse.

La figure 3 montre une segmentation excédentaire sur la partie stable du /t/. Cette "erreur", d'ordre phonétique et non acoustique, ne prête pas à conséquence, dans la mesure où le spectre de bruit du /t/ est proche de celui du /s/. Le groupe /lepəti/ est alors assimilé à /lepətsi/, et le /s/ excédentaire peut être éliminé ultérieurement par des considérations de durée.

V - RESULTATS

L'efficacité de la segmentation a été estimée de la manière suivante. Nous avons repéré sur les sonagrammes numériques de 7 phrases prononcées sans contraintes particulières par un locuteur masculin les positions "idéales" des événements-frontières. Cette opération a fourni 145 repères, ceci sans savoir a priori quels événements seraient choisis par le programme. Nous avons ensuite comparé les segmentations trouvées par programme à ces segmentations idéales, et nous avons classé les erreurs en trois catégories :

a) Fautes "graves"

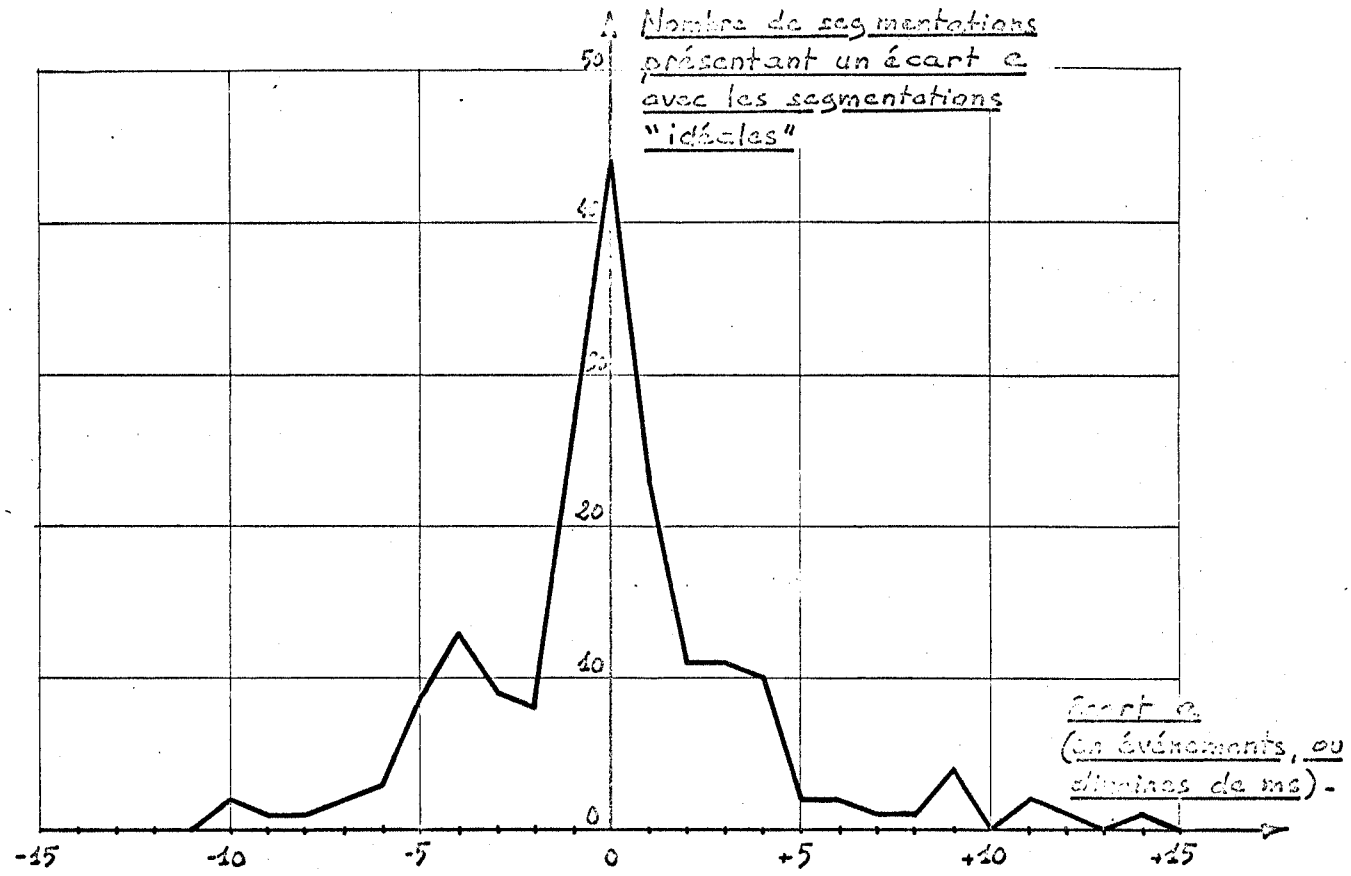
Ces fautes sont constituées par la non-détection d'une frontière. Nous n'en avons trouvé que trois, soit un taux de 2 %.L'une concernait une suite de deux liquides (/rl/), et les autres la partie quasi-stable de la voyelle /y/ dans le mot /syr/.

b) Fautes "rattrapables"

La présence d'une segmentation excédentaire dans une partie transitoire, pour être grave, n'est pas irrémédiable. Nous en avons cité un exemple (détection d'une frontière sur l'explosion du /t/). Nous avons compté 16 fautes de ce type, soit un taux de 11 %.

c) Fautes "bénignes"

La logique utilisée pour détecter les maxima de S(t) est telle qu'un son prolongé donne souvent lieu à des segmentations multiples. Nous avons compté 41 frontières excédentaires de ce type, dont la seule conséquence est de multiplier par un facteur d'environ 1,3 le nombre de phonatomes à reconnaître. Ceci n'est d'ailleurs pas nécessairement un inconvénient, dans la mesure où l'on souhaite conserver une information sur le rythme de la phrase étudiée.



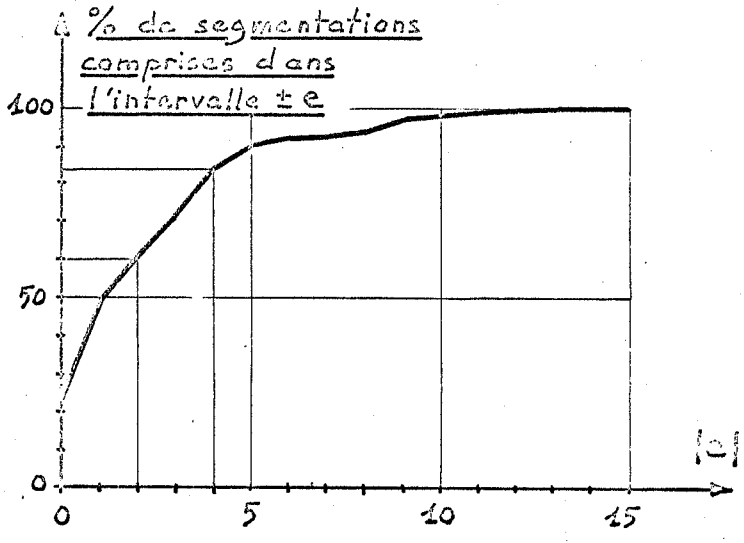


Fig 4

Efficacité de la segmentation. Au lieu de 145 segmentations "idéales" (frontières des phonatomes déterminées à l'examen du sonagramme, donc à ± 10 ms près), le programme a trouvé 186 segmentations. Trois segmentations idéales seulement ont été manquées. Les 44 segmentations excédentaires sont pour la plupart des segmentations multiples trouvées sur des sons stables et pourraient facilement être réduites. Tel qu'il est, le programme fournit environ 85 % de frontières "excellentes" ($|e| \leq 20$ ms) ou "correctes" ($|e| \leq 40$ ms).

Après ces considérations de quantité, il reste à caractériser la qualité de la segmentation obtenue. Pour cela, nous avons mesuré le décalage de toutes les frontières trouvées (sauf celles signalées en b) par rapport aux positions idéales, et construit un histogramme des écarts; ceux-ci représentent des nombres d'événements, c'est à dire des dizaines de millisecondes (fig 4). Pour l'interprétation de ce diagramme, on tiendra compte de ce que les positions dites idéales sont à au moins ± 10 ms près.

On peut admettre que la segmentation est excellente quand une frontière se trouve à moins de 20 ms de sa position idéale, et qu'elle est correcte jusqu'à 40 ms. Selon cette estimation, on obtient environ 60% de frontières "excellentes", et 85% de frontières "excellentes ou correctes".

VI - CONCLUSION

Le processus de segmentation-normalisation que nous avons décrit est intrinsèque. Il ignore la phonétique, et pourrait s'appliquer sans grand changement à la voix chuchotée ou chantée, à la musique, à certains bruits, en bref à tous les messages composés d'éléments variables dans leur durée et leur interpénétration. Le résultat en est satisfaisant, eu égard à l'aspect purement acoustique du traitement, et permet d'aborder dans de bonnes conditions la phase d'identification des phonatomes normalisés.

VII - BIBLIOGRAPHIE

1) - E.LEIPP, J.S.LIENARD, M.CASTELLENGO, J.SAPALY, D.TEIL, A.CALINET, M.MLOUKA - Colloque sur la parole, bulletin n° 53 du Groupe d'Acoustique Musicale de l'Université PARIS VI, janvier 1971.

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) - J.S.LIENARD - Analyse, synthèse et reconnaissance automatique de la parole. Thèse de doctorat d'état ès-sciences appliquées. Université Paris VI, avril 1972.

Questions de M. DEMAN à J. S. LIENARD

1°) Comment conciliez-vous les résultats de coarticulation entre deux voyelles successives séparées par une consonne, alors qu'après une segmentation qui se veut limitée à deux phonèmes (selon la définition de phonétique graphique) vous faites une corrélation topographique ?

2°) L'exemple que vous citez comme erreur (apparition d'une segmentation supplémentaire entre T et I sous la forme $\overline{TS}\ \overline{SI}$) doit-il être considéré comme une erreur ou au contraire comme la constatation d'une segmentation en deux atomes phonétiques de la transition TI ?

Réponse

Je répondrai d'abord à la seconde question, qui pose un problème de définition.

Nous avons toujours défini le phonatome comme l'élément de parole correspondant à l'évolution du conduit vocal entre deux positions quasi-stables. Cet élément de parole n'a qu'une relation très lâche avec les phonèmes et diphonèmes de la phonétique. C'est une forme élémentaire (au sens de Gestalt), composante d'une superforme qui est le mot ou la phrase. Généralement, un phonatome peut être repéré au moyen de deux symboles de la phonétique classique, mais ce n'est pas une règle absolue, surtout lorsque la parole est trop lente, trop rapide, ou mal articulée (cas général de la parole dite normale). En particulier, il est tout-à-fait normal, acoustiquement parlant, de détecter un instant de stabilité relative au beau milieu d'un son réputé transitoire, lorsque celui-ci est accentué en durée. Ce n'est une erreur que si l'on se refère à la phonétique, et si l'on confond phonatome et diphonème.

L'influence de la coarticulation sur la forme des phonatomes est beaucoup plus faible que vous ne semblez le penser, car c'est une notion associée à l'idée de phonème : si les états quasi-stables sont affectés par la coarticulation, les transitions le sont beaucoup moins. Cet effet ne s'avère sensible que dans les phonatomes incluant des liquides ou des nasales. En synthèse nous arrivons facilement à définir des phonatomes "moyens", c'est-à-dire valables dans tous les entourages phonétiques.

Par exemple le même phonatome /ra/ du dictionnaire de synthèse sera utilisé indifféremment dans les mots "ira" /ira/ et "hourrah" /ura/ sans aucune perte d'intelligibilité. En reconnaissance nous n'avons pas encore été gênés par ce problème ; mais en cas de besoin il serait aisé de définir deux phonatomes /ra/ de référence, l'un convenant après les voyelles aiguës, l'autre après les voyelles graves.

GENERATION AUTOMATIQUE D'OPERATEURS DE SEGMENTATION DE LA PAROLE

par

C. ROCHE

(Institut de Programmation - Université PARIS VI)

Pages

GENERATION AUTOMATIQUE D'OPERATEURS DE SEGMENTATION DE LA PAROLE

C. ROCHE

(Institut de Programmation - Université PARIS VI)

I. INTRODUCTION

La recherche d'un opérateur automatique de segmentation le plus près de l'idéal possible a été menée en collaboration avec LIENARD et MLOUKA.

Nous avons travaillé sur les mêmes données. Sur celles-ci, LIENARD nous a indiqué la position des "Top" de segmentation définissant les segments caractéristiques des diphonèmes à reconnaître.

Notre but est donc de trouver un opérateur travaillant sur les données vocales et donnant les mêmes positions pour ces "Top" de segmentation.

En se ramenant à un problème de reconnaissance de formes nons avons à créer sur ces données un opérateur qui reconnaît les formes "Top" et "Non-top", les résultats de l'opérateur idéal à approcher ayant été donnés par un spécialiste de l'étude des sons vocaux.

II. LE PROGRAMME DE GENERATION AUTOMATIQUE UTILISE

Un programme général a été mis au point pour la génération automatique d'opérateurs de reconnaissance des formes. Cet opérateur est engendré dans le but d'approcher un opérateur idéal inconnu dont on ne connaît que les réponses sur un ensemble de données d'apprentissage suffisant (apprentissage avec professeur).

On considère dans cette méthode un opérateur de reconnaissance comme une structure d'opérations élémentaires qui ne sont plus des opérations cablées d'un ordinateur comme +, -, x, mais comme des opérations définies de manière générale par leur table de correspondance (de même que l'addition est définie par la table d'addition).

Le programme choisit les variables A_i et A_j sur lesquelles il va construire un opérateur O_{m+1}. Ensuite, la table de l'opération $O_{m+1} : (A_i, A_j) \rightarrow A_{m+1}$ est engendrée par une méthode de regroupement inspirée de celle de MAC QUEEN (voir WATANABE [1]). La fonction à optimiser est dans ce cas l'information utile contenue dans A_{m+1} (information utile au sens de BONGARD [6] : voir SIMON et ROCHE [2] et ROCHE et SABAH [3]). Le choix des variables est fait de manière similaire à l'algorithme de HUFFMANN en choisissant le couple A_i et A_j le moins informant (voir ROCHE [4] SIMON, ROCHE et SABAH [5]).

Ce programme a engendré des algorithmes de reconnaissance ayant de bonnes performances dans des problèmes de reconnaissance visuelle (photo aérienne), d'approximation de notes de ressemblance intuitive et reconnaissance acoustique, dans des cas où les partages par hyperplans donnaient des résultats peu encourageants.

III. RESULTATS EN SEGMENTATION DE LA PAROLE

L'algorithme engendré travaille sur les données utilisées par LIENARD et MLOUKA. Les opérateurs de base sont au nombre de 14. Chacun d'eux est la somme de données situées dans un rectangle tel que sur la figure 1.

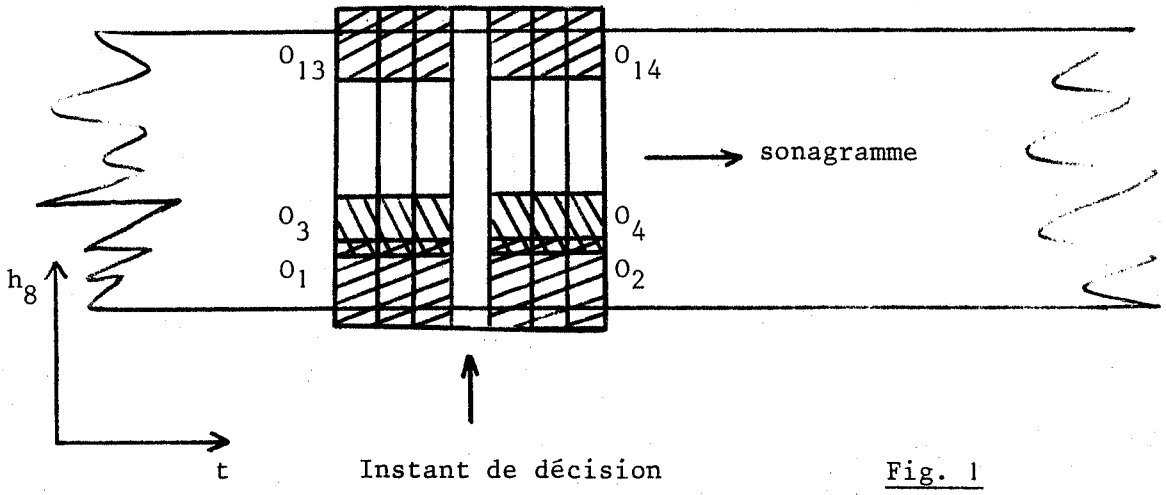


Fig. 1

Une 15ème variable est ajoutée pour tenir compte du contexte grâce à une programmation dynamique : elle représente la durée qui s'est écoulée depuis le dernier top de segmentation détecté.

Un coût a été introduit pour tenir compte du fait qu'il est plus grave de ne pas reconnaître un top de segmentation que de décider d'un top où il ne devrait pas en avoir.

Les performances de l'opérateur de segmentation ont été les suivantes. (Les paramètres utilisés sont :

coût relatif entre top non reconnu et top en trop = 6 ;
occupation mémoire de l'algorithme engendré = 3 500 bits ;

ces paramètres peuvent être transformés à volonté).

Sur les données d'apprentissage (82 top)

79 top reconnus
3 top non reconnus
52 top en trop.

Sur les données de test indépendantes des données d'apprentissage et introduites a posteriori : (57 top)

45 top reconnus
12 top non reconnus
18 top en trop

Le temps mis pour l'apprentissage a été de 8 mn de CII 10070, l'algorithme de segmentation engendré utilise 100 µs toutes les 10 ms (temps d'acquisition des données vocales).

La figure 2 montre un exemple de résultats (données de test) montrant les 2 sortes d'erreurs (top en trop, top non reconnu) : la première colonne de top montre la décision du spécialiste acousticien (Lienard), la 2ème colonne montre la décision de l'algorithme.

Remarque : Les méthodes d'apprentissage utilisant la séparation par hyperplans ont donné des résultats peu utilisables. (taux de réussite de 17%).

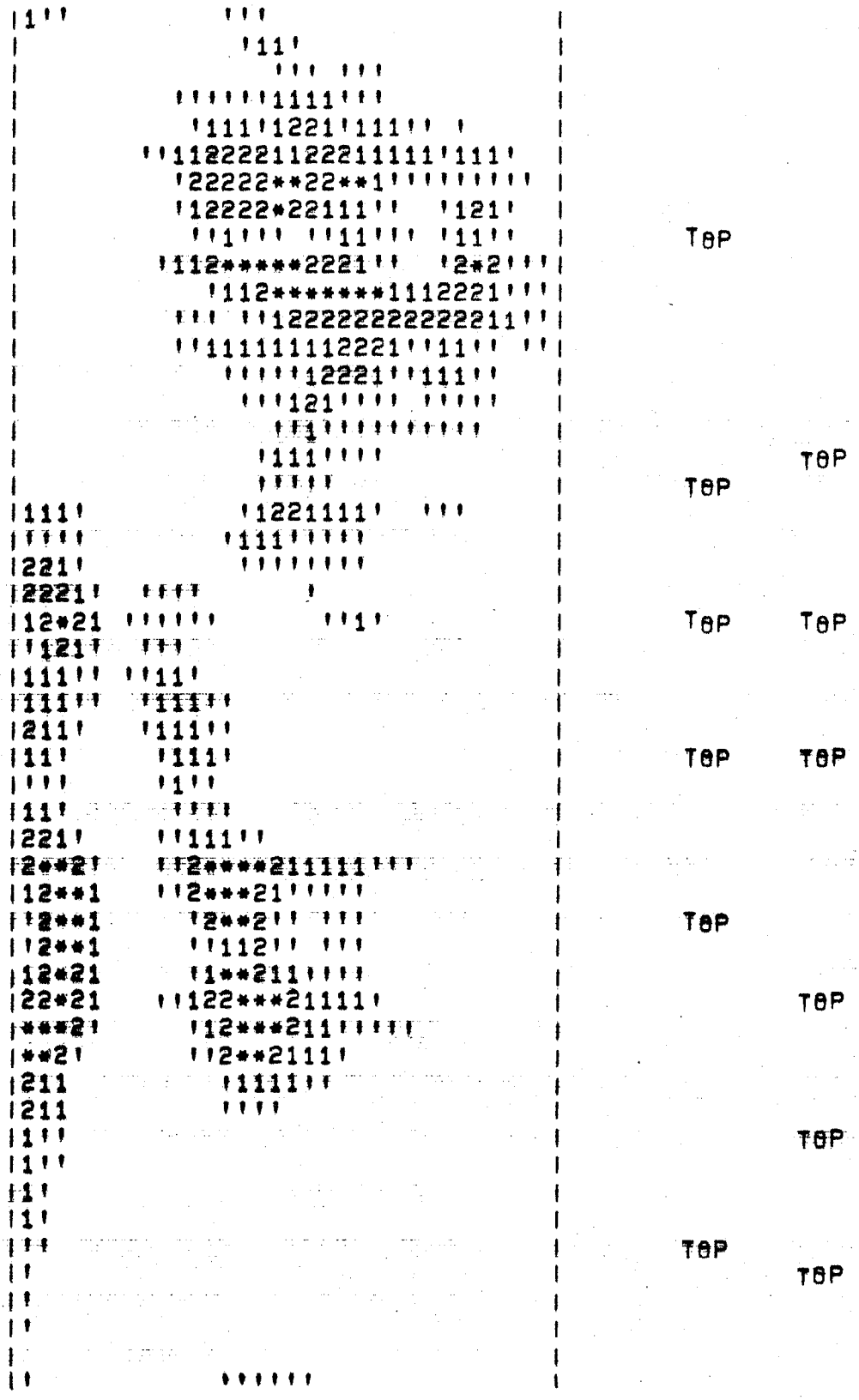


Fig. 2

BIBLIOGRAPHIE

[1] S. WATANABE : "Unified viem of clustering algorithms". IFIP Congress 1971, Ljubliana Booklet TA-2.

[2] J.C. SIMON et C. ROCHE : "Application of questionnaire theory to pattern recognition" International Joint Conference on Artificial Intellegence, Londres, sept 1-2-71.

[3] C. ROCHE et G. SABAH : "Définition d'une mesure de l'information utile d'un opérateur de reconnaissance, génération automatique d'opérateurs de reconnaissance" Compte rendu à l'Ac. des Sc. Février 72 - t. 274, p. 501 - 504.

[4] C. ROCHE : "Idées générales sur la reconnaissance des formes appliquées à la parole "Journée du Galf sur la parole - Aix en Provence, Mars 71. Publié dans Automatisme Mars 1972.

[5] J.C. SIMON, C. ROCHE, G. SABAH : "On automatic generation of pattern recognition operators" IEEE International Conference on Cybernetics, Washington, Octobre 6- 72.

[6] M. BONGARD : "Pattern Recognition" Sparton books 1970

[7] E. LEIPP, M. CASTELLENGO, J.S. LIENARD, J. SAPALY, A. CALINET, M. MLOUKA : Colloque sur la parole. GAM Janvier 71.

RECONNAISSANCE AUTOMATIQUE DE LA PAROLE
ETUDE DE LA SEGMENTATION

M. BAUDRY, M. B. DUPEYRAT, M. C. FRANK.
C.E.A. - GIF SUR YVETTE -

Pages

SES/SERA/SERF/R- 95

RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

- Etude de la Segmentation -

1 - INTRODUCTION.

Nous présentons une étude sur le prétraitement et la segmentation d'un signal acoustique en vue de la reconnaissance automatique de la parole. Nous exposerons deux méthodes différentes de segmentation qui sont en cours de comparaison.

Nous avons écarté l'analyse en fréquence à cause de la dispersion des caractéristiques obtenues pour un même son prononcé par des locuteurs différents. En effet on a alors une déformation en fréquence et en temps.

Nous avons préféré travailler sur le signal acoustique fonction du temps et étudier l'information contenue dans les intervalles de temps entre les passages par zéro du signal ou de ses dérivées. Par cette méthode nous avons réussi à reconnaître les voyelles du français dans 95% des cas sur 15 sons prononcés deux fois chacun par trois locuteurs masculins.

Parallèlement et à des fins de contrôle des différentes étapes nous avons testé la validité des méthodes utilisées en synthètisant la parole à l'aide des résultats du prétraitement. Ceci nous permet de vérifier si la segmentation est valable et nous montre les limites des expériences entreprises.

Le prétraitement n'est pas spécifique de la parole et il a été appliqué à d'autres signaux à une dimension : Electro-encéphalogramme (détection des signaux alpha et delta).

Les traitements ont été simulés sur ordinateur puis un équipement câblé qui effectue le prétraitement en temps réel a été réalisé et est en cours de mise au point. La figure 1 montre l'organisation générale du système de reconnaissance.

.../...

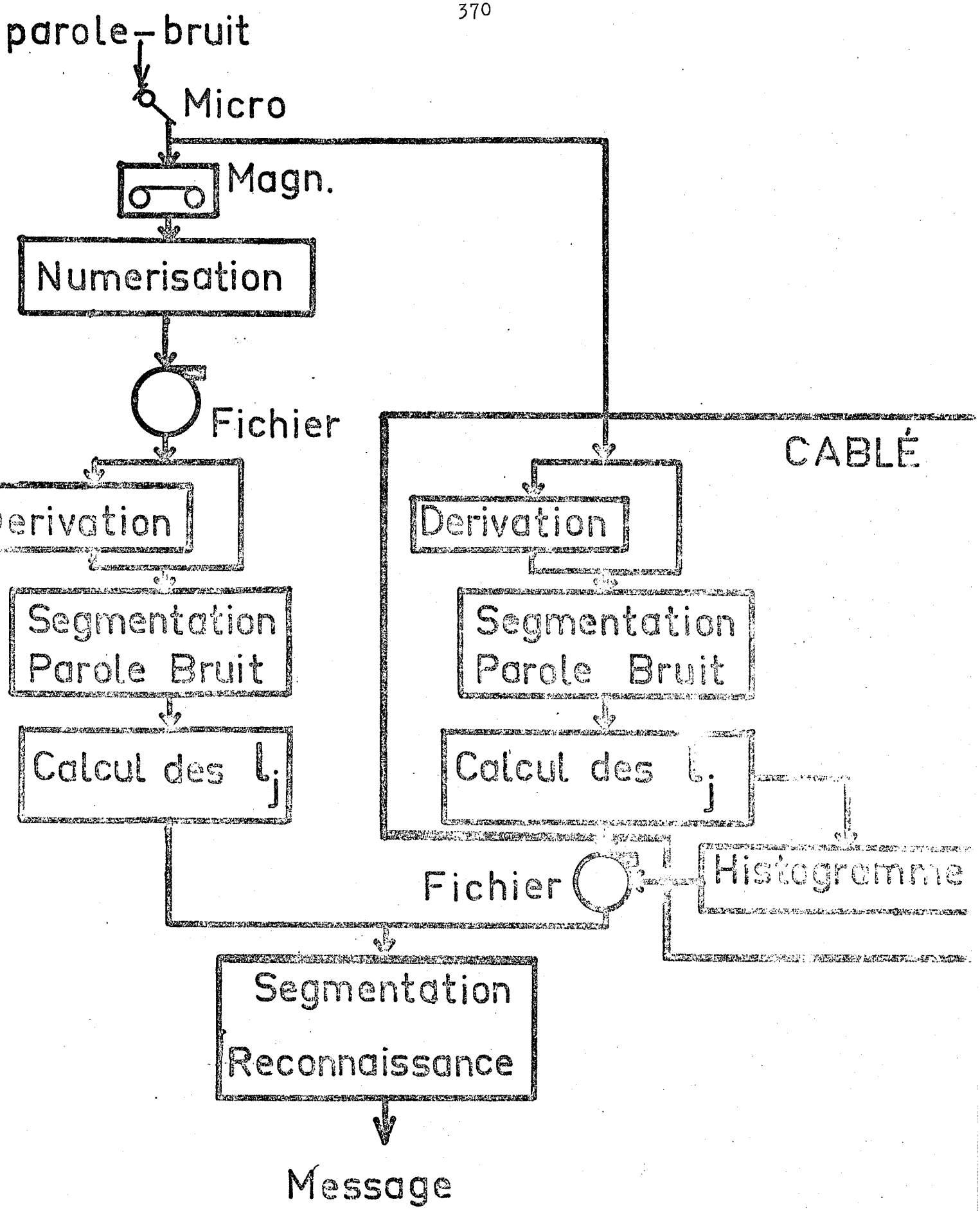

parole-bruit
Micro
Magn.
Numerisation
Fichier
CABLÉ
Derivation
Segmentation
Parole Bruit
Calcul des l_j
Derivation
Segmentation
Parole Bruit
Calcul des l_j
Fichier
Histogramme
Segmentation
Reconnaissance
Message

FIG. 1 -

2.- NUMERISATION DES DONNEES.

Nous avons enregistré deux bandes magnétiques sur un magnétophone de qualité ordinaire sans précaution particulière. Une bande contient un texte continu lu successivement par trois locuteurs. L'autre contient une série de sons élémentaires ainsi que des mots isolés obtenus en combinant ces sons élémentaires. [1]*

La numérisation est faite à une fréquence de 9.800 Hz et sur 16.384 niveaux en amplitude. On peut donc reconstituer le signal par interpolation avec une erreur connue jusqu'à 2.800 Hz [2] et avoir une bonne approximation jusqu'à 4.900 Hz (erreur non minimisable).

3.- PRETRAITEMENT ET RECONNAISSANCE.

3.1 - PRétraitement.

Des expériences anciennes [3] ont montré que l'intelligibilité de la parole est maintenue lorsqu'on applique au signal acoustique la transformation suivante : Amplification avec écrétage du signal ou de ses dérivées (l'intelligibilité est meilleure dans le cas de la dérivée 1ère fig. n° 2a à 2d).

* Les chiffres entre crochets renvoient à la bibliographie in fine.

.../...

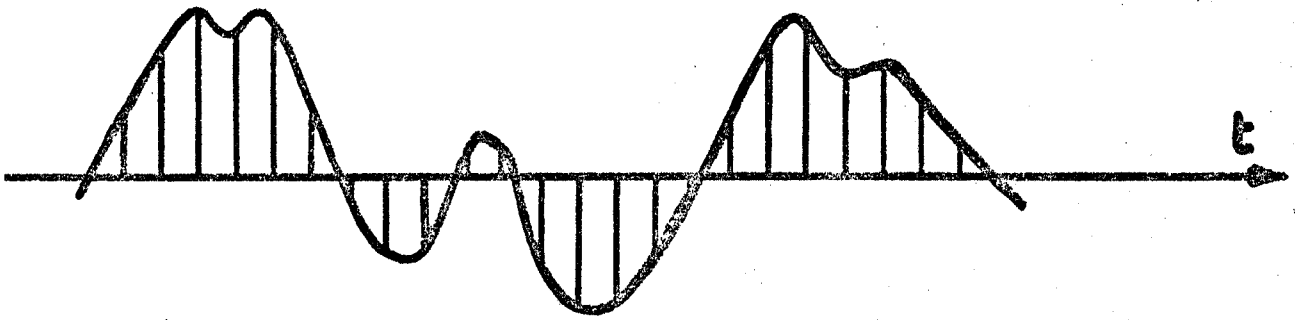


FIG. 2 a -

FIG. 2 b -

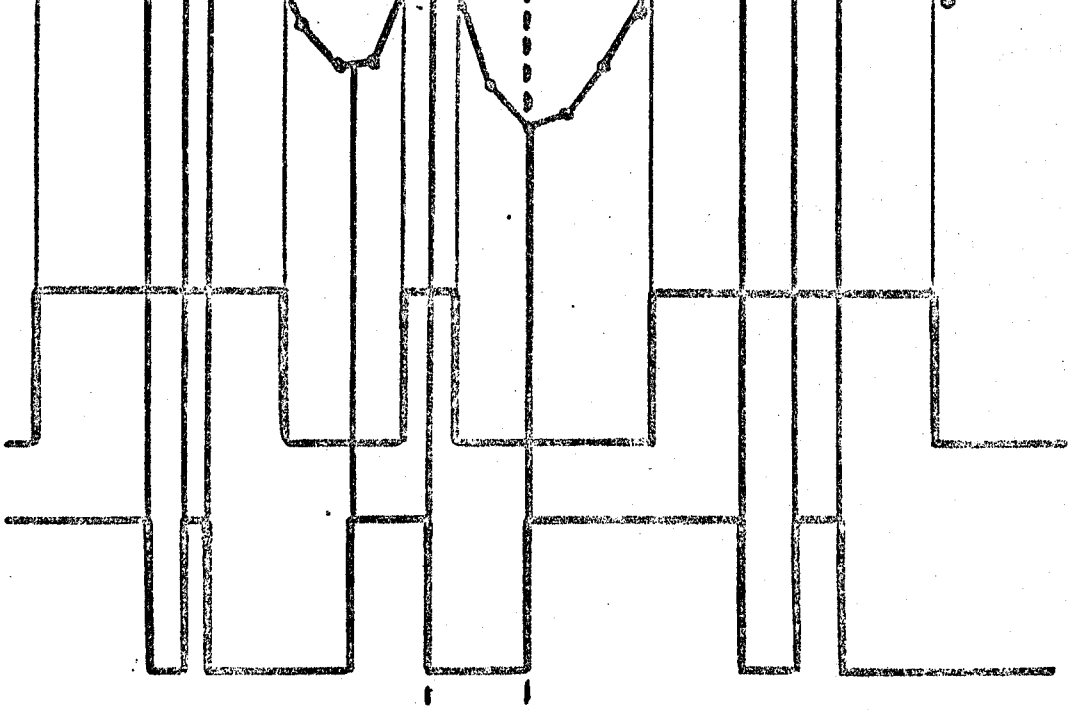


FIG. 2 c -

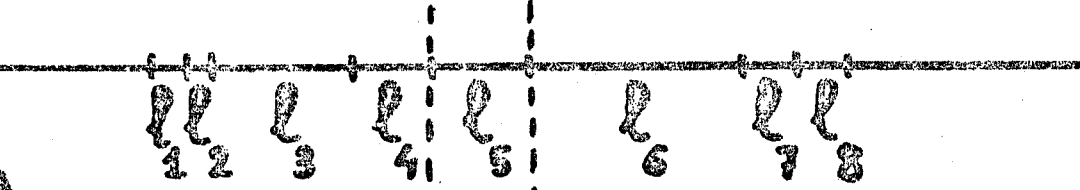


FIG. 2 d -

FIG. 2 e -

$$l_j = t_{j+1} - t_j$$

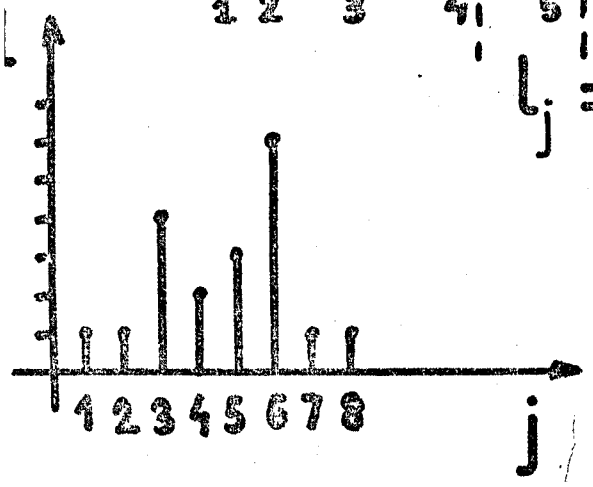


FIG. 2 f -

En première approximation, la seule information qui subsiste dans ce cas est l'intervalle de temps séparant les passages par zéro successifs du signal ou de ses dérivées. Nous avons donc étudié sur ordinateur certaines propriétés des intervalles de temps séparant les passages par zéro. Dans une première phase et pour ne pas avoir de problèmes de segmentation, nous avons fait cette étude sur des sons isolés : voyelles seules.

Nous avons mis au point un opérateur qui élimine les silences (absence de son) en ne conservant que leur longueur . Cette élimination se fait suivant plusieurs critères de seuil. Un recalage automatique du zéro est possible.

Pour nos enregistrements nous avons constaté que l'écart quadratique moyen du bruit est plus important dans les zones de silences longs que dans les zones de silences courts.

Soient : t_j, t_{j+1}, t_{j+2}..., les abscisses en temps des passages par zéro d'un signal, soient l_j, l_{j+1}, l_{j+2} les longueurs des intervalles de temps entre passages par zéro du signal (fig. 2 e). On a :

$$l_j = t_{j+1} - t_j$$

$$l_{j+1} = t_{j+2} - t_{j+1} \qquad \text{etc.}$$

Le programme calcule par interpolation linéaire dans les zônes de son les longueurs des intervalles de temps l_j séparant les passages par zéro du signal ou d'une de ses dérivées et il les enregistre sur bande.

Il est aisé de montrer que pour un son, la valeur de la somme des différents l_i nous donne le temps correspondant au j ième passage par zéro :

$$t_j = \sum_{i=1}^{j} l_i$$

.../...

Une observation peut être faite à propos de ces intervalles :

- soit s_j le rapport $\frac{l_{j+1}}{l_j}$, s_j est invariant à un changement linéaire de l'échelle du temps près. Cette observation permet d'étendre l'étude à une plus grande variété de locuteurs.

3.2 - Reconnaissance.

La reconnaissance a d'abord porté sur les sons isolés. La segmentation est faite alors par l'opérateur qui élimine les silences.

Nous avons construit l'histogramme des intervalles de temps entre les passages par zéro pour chaque son élémentaire avec une largeur de classe égale au pas d'échantillonnage. Nous pouvons remarquer que l'histogramme ne conserve pas l'ordre d'apparition des passages par zéro.

Un premier essai avec les intervalles provenant du signal non dérivé donne des résultats peu interessants. Par contre on remarque qu'il y a ressemblance des histogrammes obtenus sur le signal dérivé pour des sons identiques prononcés par des locuteurs différents et dissemblance pour des sons différents. Ceci permet de construire des modèles correspondants aux sons élémentaires. (fig. 3a, 3b et 4)

Pour comparer les histogrammes nous sommes amenés à les normaliser et nous définissons un indice de dissemblance entre l'histogramme d'un son élémentaire inconnu et les histogrammes modèles.

L'indice de dissemblance est obtenu en faisant le rapport entre l'aire des portions d'histogrammes non communes et la somme des deux aires des histogrammes.

.../...

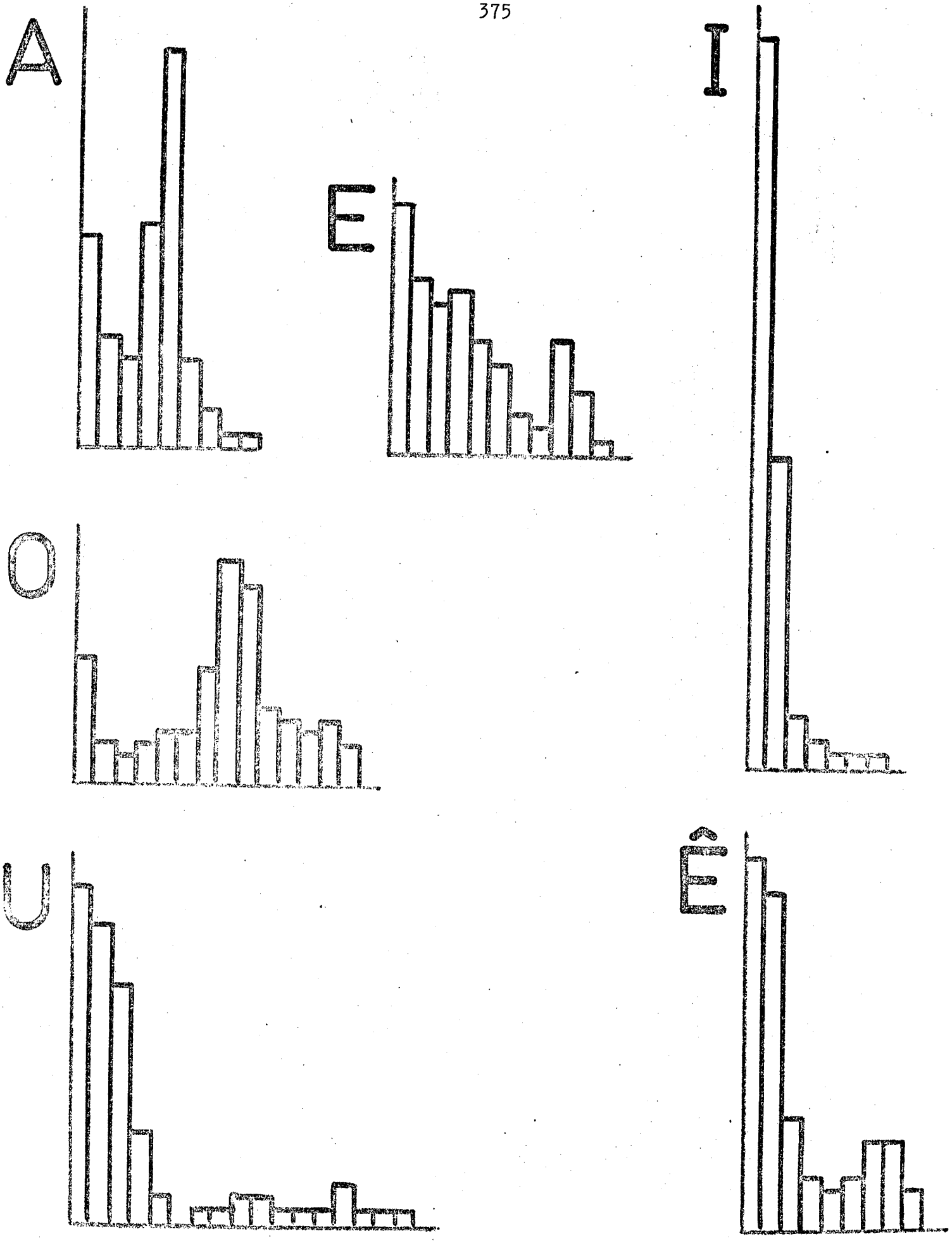


FIG. 3 a -

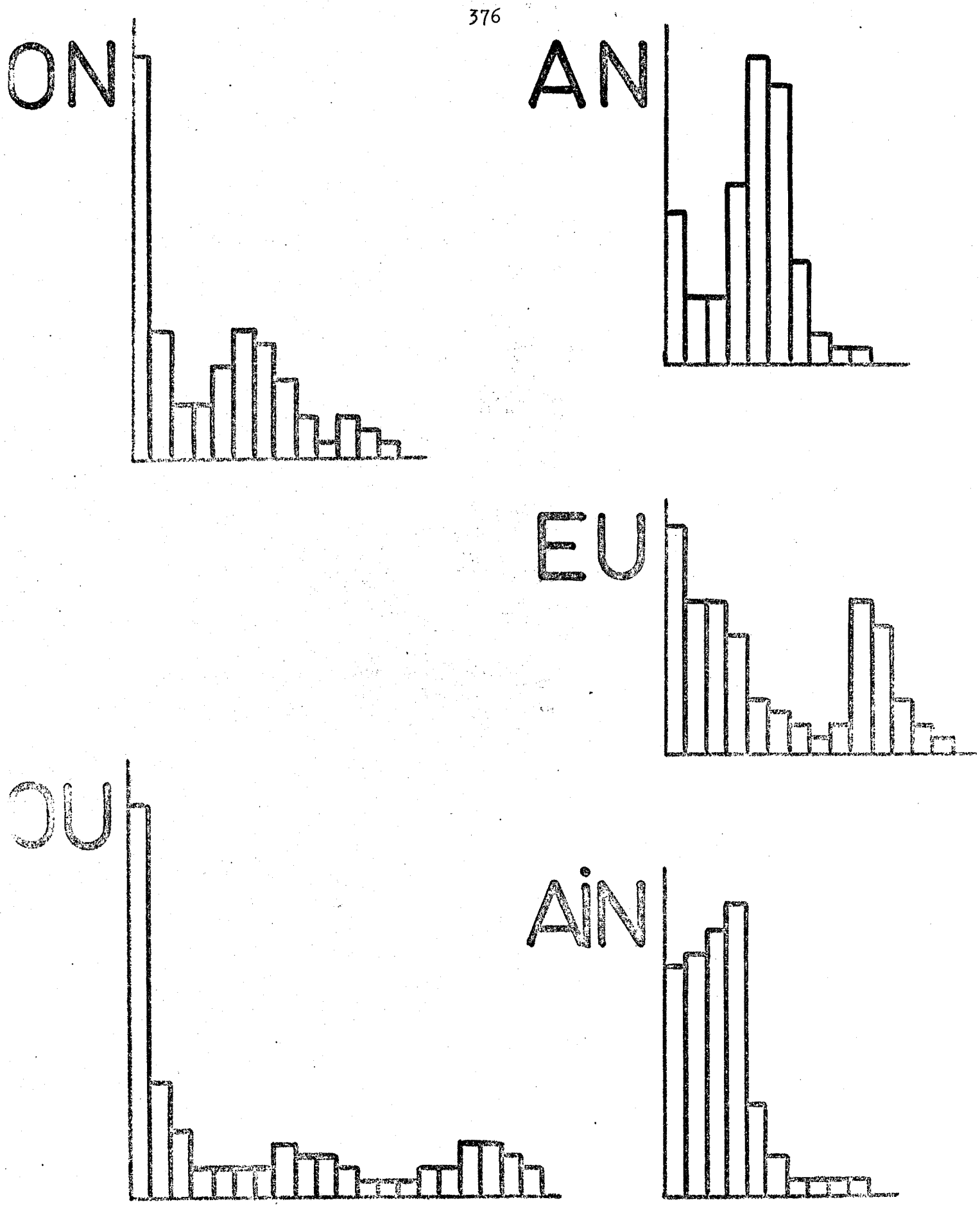


FIG. 3 b -

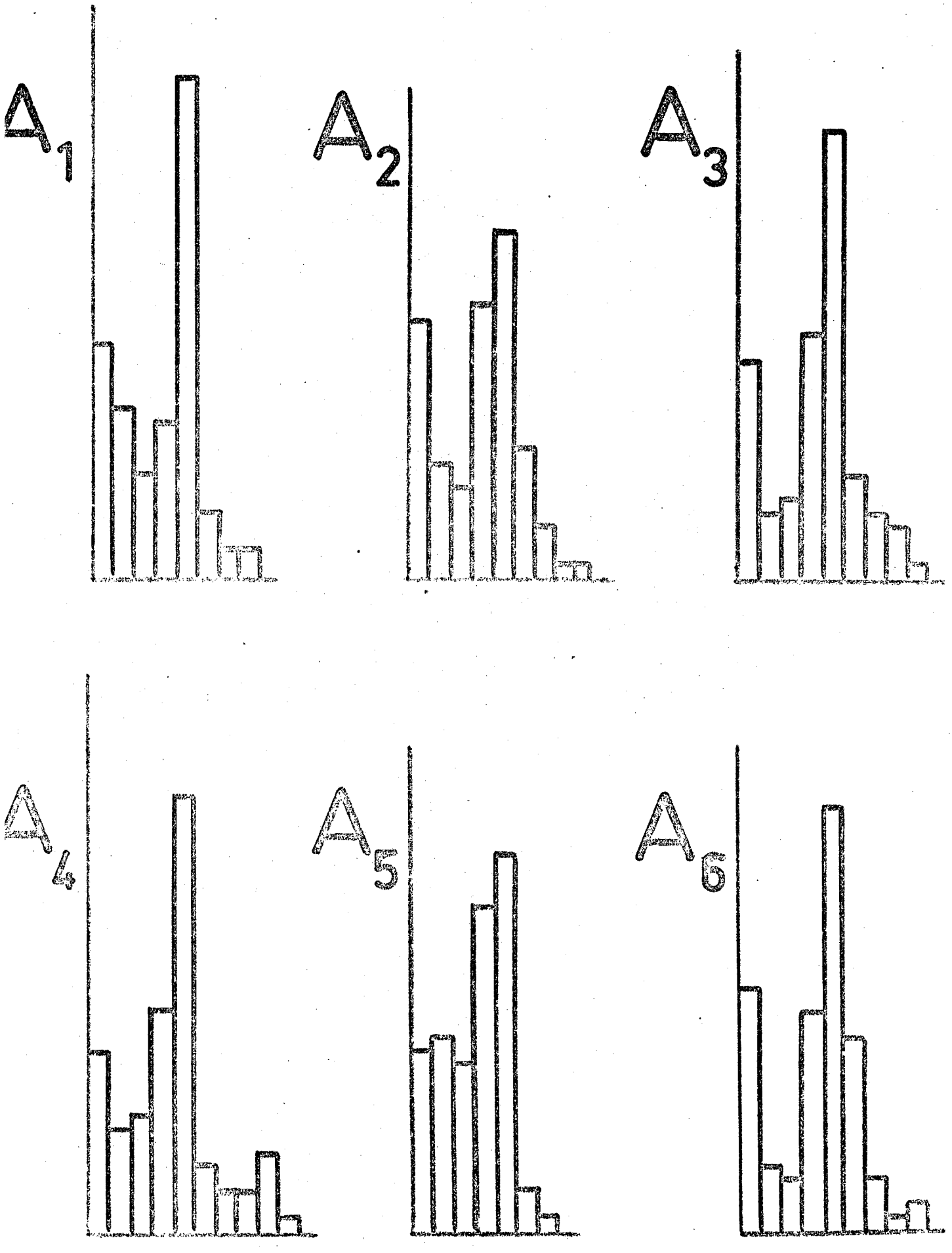


FIG. 4 -

Soit h_m (i) la valeur de la ième classe de l'histogramme modèle normalisé, soit h_g (i) la valeur correspondante pour l'histogramme inconnu. L'indice de dissemblance D est obtenu en calculant :

$$D = \frac{\sum_{i=1}^{20} |h_m(i) - h_g(i)|}{\sum_{i=1}^{20} (h_m(i) + h_g(i))}$$

Pour un son inconnu, nous calculons les indices de dissemblance avec tous les modèles. Nous classons ensuite les indices par ordre croissant. Le plus petit indice correspond au modèle le plus probable.

Pour les sons élémentaires nous arrivons à reconnaître 95% de nos 90 essais. (le son classé premier est celui qui a été prononcé).

Une machine câblée a été réalisé pour effectuer ce prétraitement, elle nous donne soit les longueurs entre passages par zéro de la dérivée du signal, soit directement les histogrammes avec possibilité à tout moment de faire une remise à zéro de l'histogramme (cf. figure n° 1).

Il n'est pas possible à ce niveau de réaliser l'analyse d'un son composé, étant donné que l'histogramme produit alors un mélange de tous les sons élémentaires. D'où la nécessité de segmenter les mots.

4. - SEGMENTATION.

Remarquons tout d'abord que le traitement utilisé pour éliminer les silences entre les sons est une première segmentation.

Ce traitement nous sert également à segmenter les mots en unités plus petites.

Une segmentation plus fine nous permet de trouver les consonnes et dans le cas de certaines consonnes longues d'exhiber des modèles d'histogrammes pour les reconnaître.

Deux méthodes de segmentation sont en test :

4.1- L'une est basée sur le comptage par unité de temps du nombre de passages par zéro du signal ou de la dérivée. [4] [5]

soit n (t) le nombre de passages par zéro du signal compris entre les instants t et t + 5 Δt , Δt étant un paramètre ajustable de 3 à 6 ms.

On construit la courbe

n (t) , n (t + Δt), n (t + 2 Δ t), ...

Le recouvrement des intervalles de temps considérés permet de moyenner la courbe.

Il y a segmentation quand la pente de la courbe est supérieure en valeur absolue à un seuil.

Cette méthode, très simple et très rapide ne donne pas des résultats complets. Nous cherchons des critères supplémentaires pour la perfectionner.

4.2 - L'autre méthode consiste à étudier l'image obtenue quand on porte les différents intervalles de temps, dans l'ordre d'arrivée sur la table D ZERO (fig. 2f). Il apparait à l'oeil des segmentations qui sont caractérisées par la variation de densité des points (fig. 5)

.../...

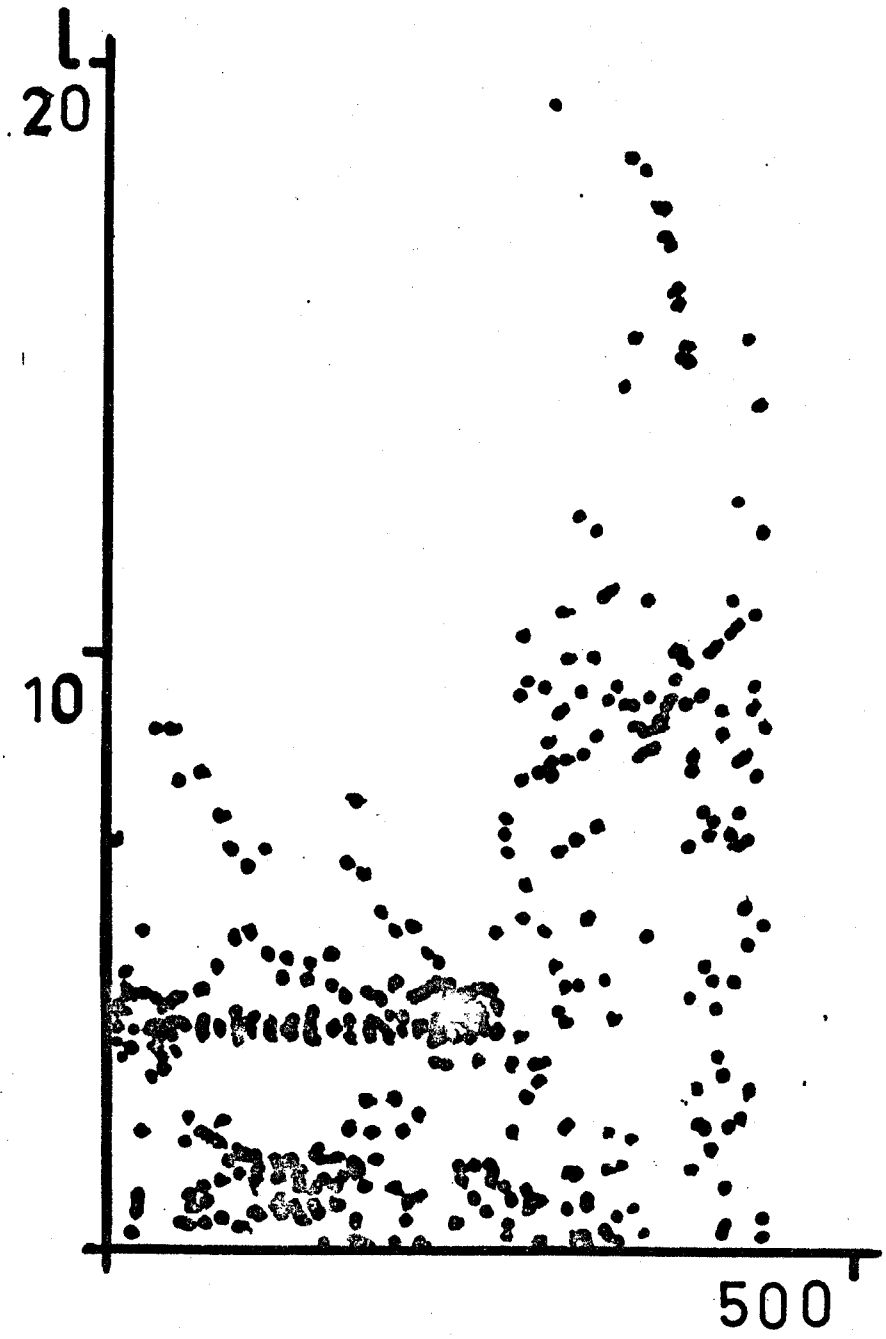


FIG. 5 -

4.2.1 - Segmentation basée sur l'analyse globale de l'image d'un son.

Le principe du programme de traitement de l'image est de retrouver les mêmes segmentations que celles qui apparaissent à l'oeil. Les l_j sont discrétisés sur 20 valeurs.

Le programme détermine les vides : Ce sont les intervalles de temps au cours desquels il n'y a aucun point dans une certaine classe. Dans les intervalles de temps non vides on calcule le rapport R du nombre de points dans la classe au nombre de point total compris dans l'intervalle de temps correspondant.

Un seuil sur R pondéré suivant la classe, détermine ce qu'on appelle les pleins. Les intervalles ayant un R trop faible ne sont pas pris en compte (ni plein ni vide).

Les pleins et les vides sont limités localement par des coupures.

On segmente s'il y a un nombre suffisant de coupures dans l'ensemble des classes pour un intervalle donné (cf. fig n° 6a et 6 b). Il faut ajuster les paramètres de pondération, la longueur d'intervalle de temps et le nombre de coupures.

Les résultats sont satisfaisants mais la méthode n'a été testé que sur un nombre limité de sons (15 mots isolés).

La segmentation des consonnes courtes est envisageable dans le cas où elles sont entourées de voyelles préalablement reconnues.

.../...

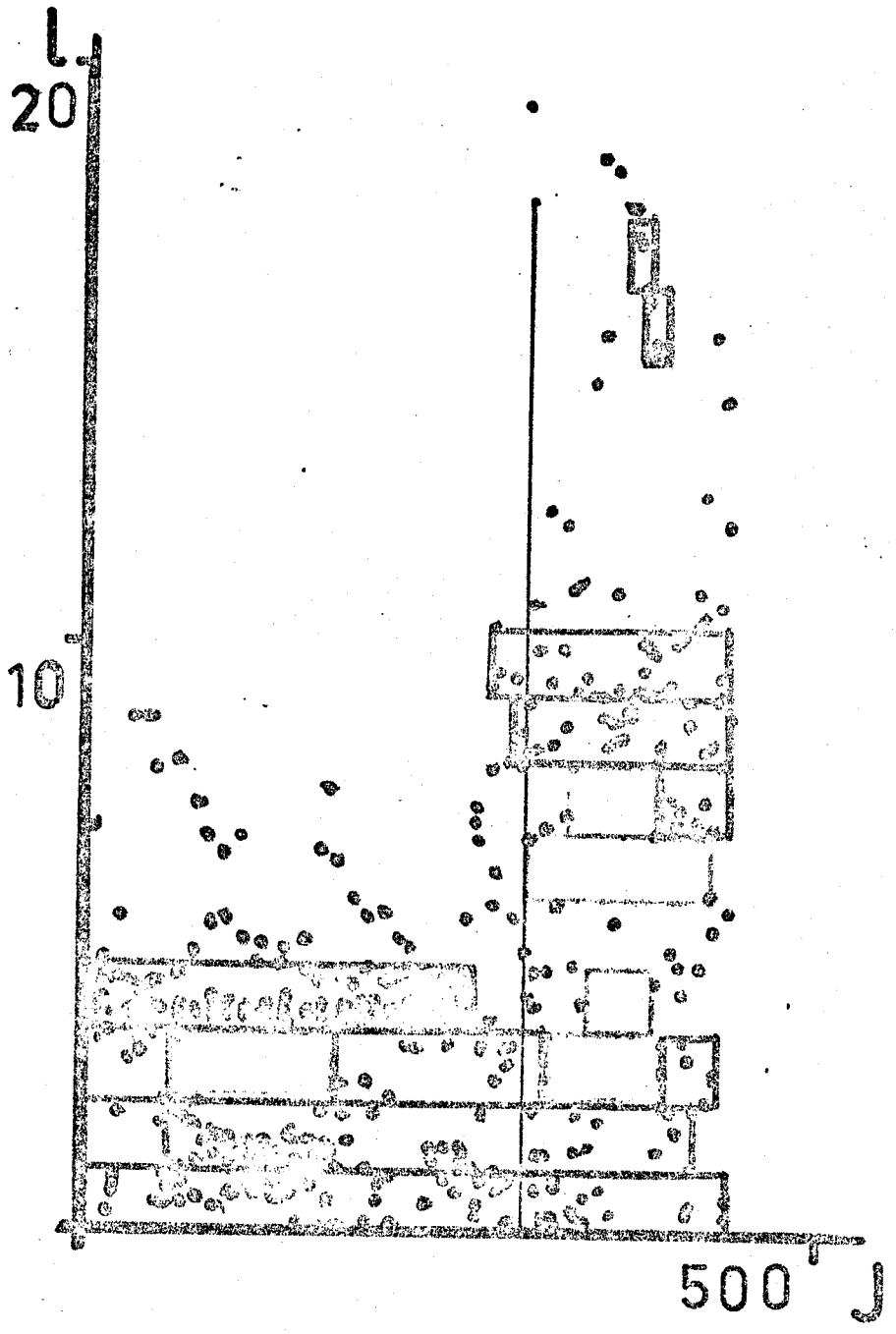


FIG. 6 a -

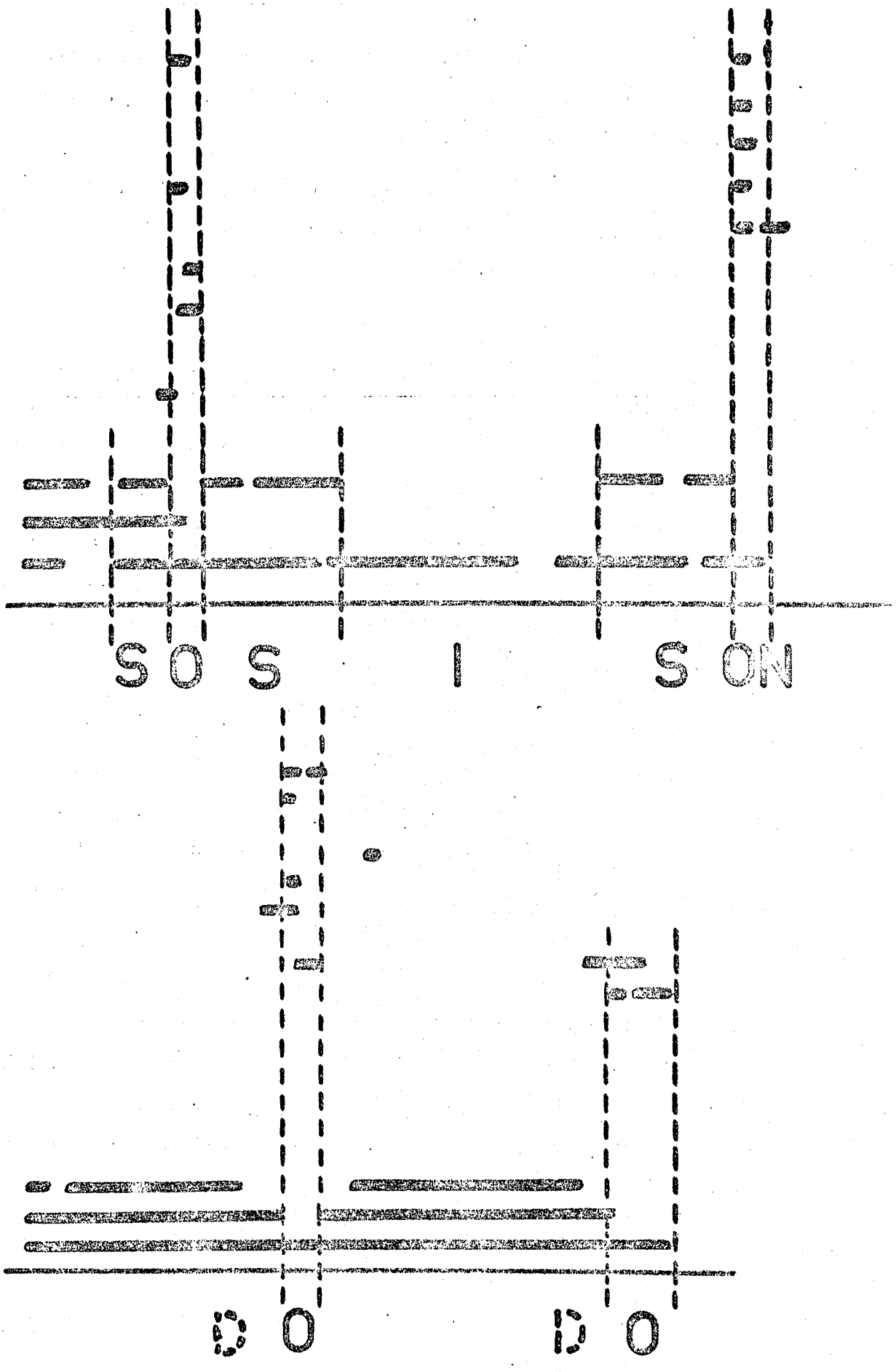


FIG. 6 b -

4.2.2. Utilisation d'histogrammes partiels.

On se fixe un pas en temps T et on construit des histogrammes successifs pendant les intervalles T : c'est-à-dire de O à T puis de T à 2 T etc...

Pour chaque histogramme on calcule l'indice de dissemblance avec le précédent et on construit une courbe donnant cet indice en fonction du numéro d'ordre de l'histogramme partiel.

Nous avons essayé de segmenter en utilisant simplement un seuil sur cette courbe. Ce seuil est difficile à déterminer. Pour améliorer la segmentation nous pensons plutôt considérer les extremas de la courbe en question.

Une transition rapide entre deux voyelles est bien détectée (sons OA, OIN). Mais nous ne réussissons pas encore à localiser correctement les phénomènes transitoires comme les plosives.

5. - SYNTHESE.

A des fins de diagnostic et de mise au point des programmes de segmentation, nous avons essayé de réaliser la synthèse de sons à partir des l_j (dérivée première) à l'aide d'un calculateur CII-90-10. Il faut noter que nous ne cherchons pas dans les expériences de synthèse la qualité sonore, mais plutôt l'intelligibilité.

Pour cela, nous construisons un signal électrique fonction du temps qui bascule chaque fois qu'un intervalle de temps l_j s'est écoulé. Ce signal est écouté par l'intermédiaire d'un haut parleur. La qualité sonore est mauvaise mais le son reste intelligible. Le traitement numérique appliqué que nous avons exposé plus haut n'apporte pas une perturbation trop importante.

La mauvaise qualité du son est due au fait que l'on donne une égale importance à tous les sons, consonnes, voyelles, bruit de fond. Elle est due aussi à la perte de précision sur les longueurs lors des traitements numériques. (échantillonnage, arrondis).

.../...

Nous pouvons également modifier la suite des l_j avant de synthètiser ce qui nous permet de mettre en évidence certaines propriétés

a) synthèse par juxtaposition.

Ainsi, nous avons pu synthètiser de nouveaux mots en juxtaposant les l_j de certains sons ou fragments de sons. Cette juxtaposition n'entraîne aucune discontinuité audible si les sons composants sont extraits à partir d'une bonne segmentation et moyennant l'ajustement des paramètres.

Néanmoins notre segmentation des plosives n'est pas actuellement suffisamment précise pour pouvoir faire de bonnes synthèses

b) longueur minimale des sons.

Nous prenons au hasard, à l'intérieur d'un son élémentaire (voyelles) plusieurs séquences disjointes d'une vingtaine de l_j consécutifs (chaque séquence a une durée d'environ 5 à 10 ms). Nous juxtaposons ces séquences dans un ordre arbitraire, le son reproduit est toujours intelligible.

Cette dernière expérience nous conduit à chercher s'il n'est pas possible de caractériser un son par une vingtaine de l_j consécutifs.

6.- CONCLUSIONS.

Tous ces résultats nous semblent intéressant dans la voie de la reconnaissance automatique de la parole utilisant l'information contenue dans les longueurs d'intervalles de temps compris entre les passages par zéro du signal acoustique ou de sa dérivée.

La synthèse nous a montré que cette approche était valable.

Nous poursuivons nos travaux dans les directions suivantes

a) L'équipement câblé qui effectue le prétraitement est en cours de raccordement sur le calculateur CII 90-10. Nous allons pouvoir disposer d'un plus grand nombre de données. Nous pourrons tester nos opérateurs sur une plus grande variété de voix (voix de femmes y comprises).

b) L'opérateur de reconnaissance sera optimisé en utilisant des méthodes de classification. Nous extrairons de nouveaux modèles pour les consonnes longues. Actuellement nous n'avons pas de caractéristiques permettant d'identifier les consonnes courtes.

c) Perfectionnement des méthodes de segmentation dont l'importance est fondamentale pour résoudre les problèmes de synthèse et de reconnaissance automatique de la parole.

REMERCIEMENTS

Nous remercions vivement Messieurs B. MEYER et G. CHOLET qui ont réalisé respectivement les programmes décrits aux paragraphes 4.1 et 4.2.1. au cours de stages de D.E.A. (Université de PARIS 6, Institut de Programmation, D.E.A. de Reconnaissance des formes, Professeur : M. J.C. SIMON).

BIBLIOGRAPHIE -

[1]- Bulletin du Groupe Reconnaissance de la Parole (AFCET) n° 4
Supplément au bulletin du groupe Communication Parlée (GALF)

- Description de la bande de FRANK.

[2]- HOUDAS, HORVAT
Thèse 3ème cycle, Juin 1970
Institut de Programmation Université de PARIS 6

[3]- PETERSON, E (1951), "Frequency Detection and Speech Formants", J. Acoust. Am., 23, 668- 674.

[4]- VICENS, P.J. (1969), "Aspects of Speech Recognition by computer", Ph. D. Thesis, Stanford University, A.I. Memo N° 85, Stanford Artificial Intelligence Project, Stanford.

[5]- DE MORI R., GILLI L., MEO A.R., (1970) "A flexible réal time recognizer of spoken words for man-machine communication". Int. Journal on Man-Machine Studies, 2, 317-326.

-CHANG, S. al. (1951), "Representation of speech Sounds and some of their Statistical Properties". Proc. IRE, Feb 1951, 147-153

-SCARR, R. W. A. (1968), "Zero-Crossing as a means of obtaining spectral information in speech analysis", IEE Trans., Vol AU-16, 2, 247-253

-EWING G.D., TAYLOR J.F. (1969), " Computer Recognition of Speech Using zero-Crossing Information". IEEE Trans., vol AU-17, 1, 37-40.

BIBLIOGRAPHIE

[1]- ITO M.R., DONALDSON R.W. ," Zero - crossing measurements for analysis and recognition of speech sounds". I EEE Trans. AU, Vol. AU-19, n° 3, Sept 1971, pp 235- 241.

[2]- BOND F.E., CAHN A.R., " On sampling the zeros of Bandwidth Limited Signals". IRE Trans. IT, Sept 1958, pp 110 - 114.

QUESTIONS POSEES A LA SUITE DE L'EXPOSE

M. GRESSER :

Quel est le but du C.E.A. en commençant cette étude ?
M. De MORI a t-il des commentaires à faire au sujet de ce travail ?

REPONSE :

- le but poursuivi par le C.E.A. est lié à la diversification des activités de recherches. Nous faisons un travail sur la parole avec en vue les applications habituelles : commande de machine.... Par ailleurs, le traitement n'est pas spécifique de la parole et a été appliqué à d'autres signaux à une dimension : Electro-encephalo-grammes : séparation des signaux alpha et delta.

INTERVENTION DE M. De MORI :

M. De MORI nous a encouragé à poursuivre ce travail sur les passages par zéro. En effet, les résultats qu'il a obtenus en utilisant un système basé sur les passages par zéro lui permet de reconnaître un vocabulaire programmable de mots.

Il a confronté les résultats obtenus par son système avec ceux qu'on obtient par l'analyse en fréquence : Ils sont complémentaires. Pour l'Italien, l'analyse en fréquence fait une grande confusion entre les deux chiffres deux et neuf (duo et nove), alors que les passages par zéro les séparent très bien. Le m et le v se distinguent très bien à l'aide des passages par zéro.

M.De MORI sépare le signal en deux bandes de fréquences, il doit ajuster les filtres pour certains locuteurs. La distance au microphone influe également sur les résultats. Si on respecte une certaine distance au microphone, les résultats sont très bons pour un vocabulaire limité.

M.De MORI signale quelques confusions entre i et é et entre a et o. Il a observé que ses résultats concordent assez bien avec la théorie de DONALDSON [1] Les distributions de passages par zéro sont les mêmes.

M. De MORI nous signale pour la synthèse à partir des distances entre passages par zéro les travaux de BOND [2].

QUESTION :

-M. ROVE :

Qu'est-ce qui se passe pour les signaux faibles, par exemple quand le locuteur se tait ? Est-ce que vous ne comptez rien ou tous les passages par zéro dus au bruit thermique.

REPONSE :

- Nous avons un programme de segmentation qui isole les zones de parole et le traitement n'est fait que sur celles-ci. Les critères portent sur un nombre d'échantillons consécutifs dont l'amplitude est supérieure à un seuil fixé.

QUESTION :

-M. ROVE :

Même si le son comporte un signal très petit ?

REPONSE :

- Les conditions d'entrée dans un son et de sortie d'un son ne sont pas les mêmes, ce qui nous permet de régler ce cas.

M. De MORI :

Il y a d'autres techniques possibles citées dans la littérature : par exemple celle qui est d'additionner au signal origine un signal à fréquence plus haute dont les intervalles sont trop petis pour être considérés.

.../...

SEGMENTATION DE LA PAROLE
BASEE SUR LA RECHERCHE DE CRITERES *

P. ALINAT

THOMSON-C.S.F. D. ASM - 06 - CAGNES-SUR-MER

* Etude financée par la D.R.M.E.

Pages

1. - DEFINITION ET BUT DE LA SEGMENTATION

La segmentation de la parole consiste à la découper en une suite d'informations successives. Cela peut être fait à différents niveaux : phonèmes, diphones, syllabes, mots, groupes de mots, phrases. Il existe naturellement une analogie avec notre écriture phonétique décomposable en lettres, mots et phrases, mais comme nous le verrons par la suite, il ne faut pas la pousser trop loin.

Le but principal de la segmentation est de permettre la reconnaissance des informations ainsi isolées. Cette opération sera rentable dans la mesure où la classification des informations porte sur un choix plus restreint d'informations qui sont elles-mêmes combinées par la suite. Ainsi, il est préférable de reconnaître des phonèmes, plutôt que des mots : il n'y a qu'une trentaine de phonèmes dans la langue française mais des milliers de mots. A court terme, la reconnaissance directe des mots appartenant à un vocabulaire limité est peut-être plus simple, mais à long terme la reconnaissance des phonèmes constituant ces mots est plus rentable. C'est pour cela que le but de l'étude décrite ci-après est la décomposition en phonèmes.

2. - LE PROBLEME DE LA SEGMENTATION EN UNE SUITE DE PHONEMES

On pourrait être tenté de considérer que le signal parole n'est qu'une suite de phonèmes, chaque phonème étant un signal bien déterminé, relié à ses voisins par des transitions. On pourrait alors découper la parole en une suite de portions dont chacune représenterait un phonème déterminé. C'est ce qui se produit pour les lettres dans le cas de l'écriture imprimée. Il a été montré (Haskins - Laboratories [2]) que, bien que cette vue des choses soit presque vraie pour les phonèmes soutenus (voyelles, consonnes fricatives), elle est fausse dans le cas des phonèmes faisant intervenir le temps (diphtongues, consonnes explosives).

L'analogie avec l'écriture manuscrite est plus intéressante que celle avec l'écriture imprimée. En effet, le degré d'application de celui qui écrit prend une grande importance : dans une écriture ordinaire, la plupart des lettres subissent des déformations graves et certaines même sont omises. Ces défauts dépendent en général de la position des lettres dans les mots. Il est alors impossible de reconnaître certaines lettres et certains mots si on n'est point aidé par la connaissance à priori du vocabulaire utilisé et du contexte. De façon analogue, le degré d'application des locuteurs a une énorme importance : dans une conversation courante, les phonèmes subissent des déformations graves et peuvent être omis.

Etant donné que nous nous sommes placés au niveau phonème, il faut que les phonèmes soient prononcés correctement pour être reconnus. Cela impose que les locuteurs parlent lentement et articulent avec soin. L'intervention du vocabulaire représente l'étape suivante dont la complexité est d'ailleurs très variable selon l'étendue et la nature du dit vocabulaire.

Mais même avec une parole bien articulée, nous nous trouvons devant l'impossibilité de segmenter la parole en une suite de phonèmes en découpant simplement le signal en tranches de durée variable de façon à ce que chaque tranche corresponde à un phonème. En fait la solution consiste à faire correspondre aux tranches des informations, appelées critères, qui se combineront elles-mêmes pour former les phonèmes. Avant de rechercher ces critères, nous allons faire subir au signal une transformation en vue de nous faciliter la suite des opérations.

3. - LA COCHLEE ARTIFICIELLE

Dans les problèmes de reconnaissance des formes, le choix de la base dans laquelle on décrit le signal est très important : la complexité du système en dépend pour une bonne part. La parole est un signal présentant certaines stationnarités et la phase a peu d'importance (tout au moins au point de vue reconnaissance des phonèmes). Un pré-traitement adapté peut être une transformation du type défini par la relation ci-dessous :

$$F(t,\omega) = \int_{-\infty}^{+\infty} f(\lambda)\ r_\omega(t-\lambda)\,d\lambda \qquad (1)$$

suivie d'une détection intégration (pour éliminer la phase).
$f(\lambda)$ est le signal fonction du temps et $r_\omega(\lambda)$ est la réponse impulsionnelle d'un filtre passe bande. Il faut déterminer le $r_\omega(\lambda)$ optimum.
Or, il se trouve que l'oreille humaine ou plus exactement la cochlée fait subir au signal reçu une transformation du type défini par la relation (1) [1] .
On a ainsi un moyen de déterminer les fonctions $r_\omega(\lambda)$. Dans notre cas, l'oreille est simulée par une batterie de 96 filtres passe-bande.

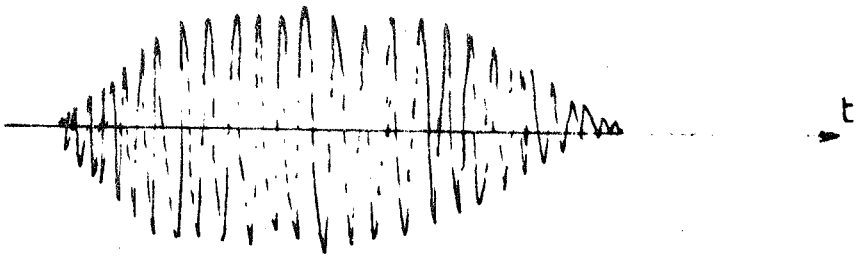
4. - LES CRITERES

Un critère est un fait particulier qui caractérise un groupe de phonèmes [3]. Ainsi, par exemple, certains phonèmes sont sonores et d'autres sourds. Le nombre decritères est bien inférieur à celui des phonèmes et chacun de ces derniers sera caractérisé par un ensemble de 2 ou 3 critères se produisant simultanément ou séquentiellement. Les principaux critères ont trait à la nature de l'excitation, à l'appartenance des 2 premiers formants à des zones déterminées et à des conditions de durée. La liste des critères qui semblent utiles pour la langue française est donnée en Annexe 1.
Les critères ne dépendent que de la langue ; ils sont indépendants du locuteur et de l'auditeur, bien que certains "accents" puissent les déplacer légèrement. Nous effectuons la segmentation au niveau des critères, c'est-à-dire que l'on recherche en permanence quels sont les critères vérifiés et on en déduit les phonèmes en cours, ou qui viennent de se produire. La figure 1 donne des exemples de cette façon de faire. Expliquons le cas c): la consonne explosive [B]. Il y a 3 critères qui doivent se produire successivement : une phase dite de "Silence vocalisé" (le conduit vocal est fermé mais les cordes vocales sont en mouvement) une explosion (variation brusque de l'énergie ou des positions des formants s'il y en avait avant) une transition vers la voyelle suivante (ici un A) au cours de laquelle les basses fréquences sont particulièrement importantes (par opposition à D et G).

Pour pouvoir employer cette méthode, il faut observer pour un bon nombre d'individus (hommes uniquement jusqu'à présent) les différents critères et chiffrer le mieux possible leurs paramètres. A ce niveau la cochlée artificielle s'est avérée très utile, surtout pour la détermination des plages dans lesquelles doivent se trouver les formants des voyelles et des consonnes fricatives. On s'est en effet aperçu que pour chacune des voyelles les 2 premiers formants se produisent dans des zones de fréquences spécifiques à la voyelle. La largeur de ces zones est de l'ordre de la moitié de la fréquence centrale de la zone (figure 2). Ce fait n'a été pour le moment vérifié que pour des voix d'homme, mais il est remarquablement indépendant du locuteur. Ajoutons que la position du premier formant doit être fixée par la valeur d'un pole de la fonction de transfert des cavités vocales et non pas par le plus grand pic du spectre.

Figure 1

a) CH U



2ème formant dans la zone U

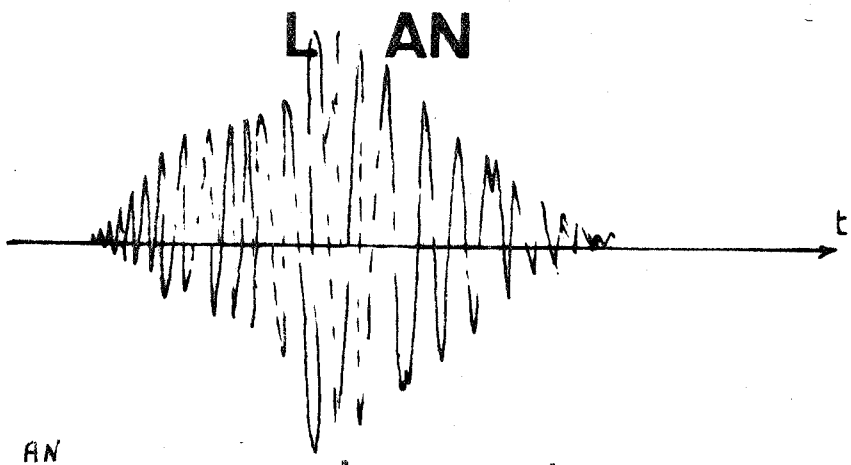
1er formant dans la zone U

Excitation → fondamental pur

Durée > Durée minimale

→ Présence d'un U

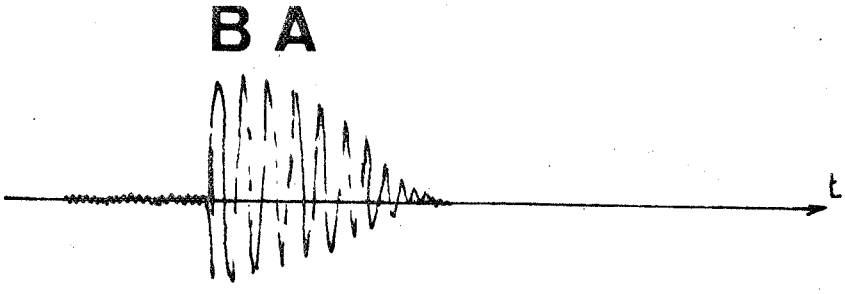
b) L AN



2ème formant dans la zone AN
Excitation → fondamental pur
Amplitude F2 > Amplitude F1
Durée > Durée minimale

→ Présence de AN

c) B A



Explosion

Silence vocalisé

certaines caractéristiques de transition

→ Présence du B

Zones formantiques (francais)

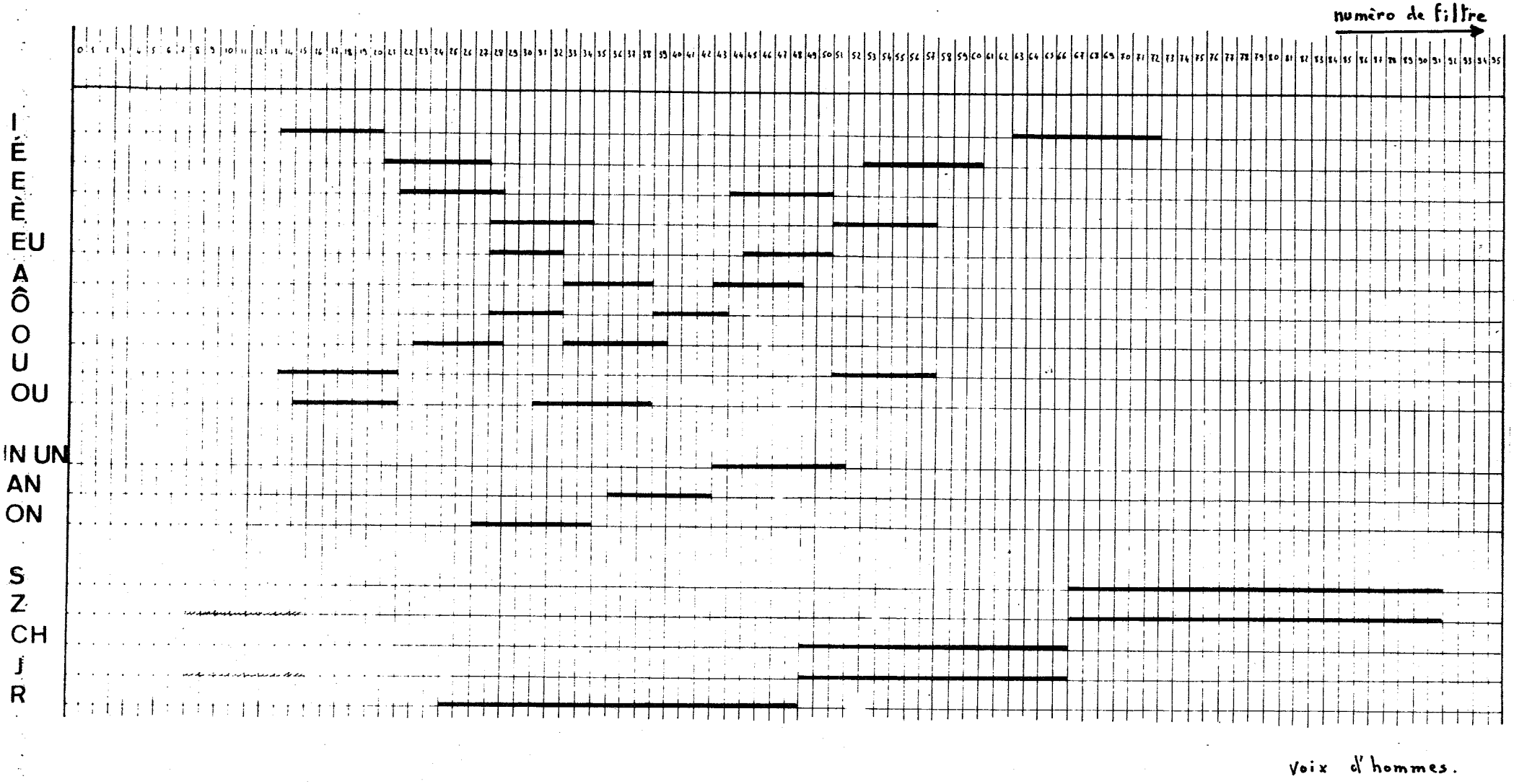


FIGURE 2

Cette façon d'agir est utile lorsque le premier formant est bas (I, U, OU) pour les voix d'homme (fondamental vers 120 Hz) et pour presque toutes les voyelles pour les voix de femmes.

Les 2 formants n'ont pas forcément la même valeur au point de vue information. Ainsi par exemple on trouve, rarement d'ailleurs, des [É] dont le second formant se trouve dans la zone correspondant au [I] , le premier formant par contre restant bien dans la zone correspondant au [É] . Dans ce cas là, on entend parfaitement [É] . On peut donc supposer que quoique normalement [I] et [É] se distinguent par les positions de leur 2 formants, c'est la position du premier formant qui entraine la décision.

A une voyelle donnée, on peut en général associer une autre voyelle qui ne se distingue de la première que par la position d'un seul formant : par exemple le premier formant pour [O] et [OU] . La frontière est alors bien marquée.

Pour les consonnes fricatives (sourdes et sonores), il existe également des zones dans lesquelles se situe l'unique formant (figure 2). Seules les consonnes [F] , [V] semblent ne pas obéir à cette règle : elles seraient caractérisées par un bruit de faible intensité et de bande large.

On a vérifié pour les voyelles nasales [IN, UN] [AN] et [ON] que le second formant était plus grand que le premier, alors que pour les voyelles pures correspondantes il est plus faible [4] . Toutefois, ce critère est bien moins rigoureux que celui des zones formantiques.

L'étude des consonnes explosives est encours. Un premier critère consiste en une explosion, c'est-à-dire une variation brusque de l'énergie ou une déformation rapide du spectre. La nature du signal précédant l'explosion intervient également. Ces 2 critères se mettent aisément en relief [3] . Le dernier critère est relatif à la transition vers le phonème suivant. Il n'est pas encore entièrement défini.

Les diphtongues n'ont point été étudiées pour l'instant.
On donne en Annexe 2 les associations de critères caractérisant les différents phonèmes.

5. - LA DECISION

Dans tout ce qui précède et au stade actuel de notre étude, on considère que les critères sont ou ne sont pas vérifiés et que les phonèmes sont ou ne sont pas reconnus. Cette façon de penser n'est pas utopique. Les études de perception faites au Haskins Laboratory [3] montrent en effet que pour les consonnes explosives, il existe, lorsque l'on fait varier de façon continue les paramètres, des frontières nettes quant à la perception. De même, les résultats de la figure 2 montrent pour les voyelles et les consonnes fricatives que, bien que les locuteurs aient la possibilité physique de faire varier de façon continue la position des formants, ils se limitent à les répartir dans des zones dont les frontières sont bien délimitées.

Cependant, toutes ces expériences se produisent dans des conditions idéales : bon rapport signal à bruit, articulation soignée, attention soutenue, etc.. Au cours d'une conversation, ces conditions sont rarement réunies, et on peut penser qu'il vaut mieux alors chiffrer des probabilités de présence, plutôt que de prendre des décisions par tout ou rien.

Les résultats précédents n'en restent pas moins vrais et doivent être utilisés. Les notions de probabilité de présence interviendraient plutôt au sujet des critères de durée, de stabilité des formants, c'est-à-dire en fait pour tout ce qui fait intervenir le temps.

Il faudra également faire intervenir à ce niveau la prosodie (intensité, mélodie et durée) et les probabilités de successions (par exemple [R] suit souvent [Ô] et [EU]).

Dans ces conditions, on pourra obtenir un faible taux d'erreur pour un vocabulaire restreint, les locuteurs prononçant naturellement, sans application excessive. Bien entendu, le système sera valable pour toutes les personnes parlant la même langue dans la mesure où ils en respectent presque toutes les règles : quelques erreurs de prononciations, même systématiques, seront parfaitement tolérées, Aucun apprentissage ne sera nécessaire.

En revanche, l'utilisation d'un vocabulaire très étendu, structuré par une grammaire complexe, n'est guère envisageable pour l'instant.

6. - SYSTEMES CONSTRUITS DANS LE CADRE DE L'ETUDE

Nous avons commencé par la réalisation d'une cochlée artificielle de 66 filtres. La fonction de transfert du filtre élémentaire est directement inspirée des résultats de Bekesy. Toutefois, elle est un peu plus surtendue. Ayant mis en évidence les zones formantiques des voyelles et des consonnes fricatives prononcées par des hommes, on a construit un système de reconnaissance simple et sommaire pour reconnaître ces phonèmes en temps réel.

Ce système donnait des résultats encourageants. Cependant de nombreuses erreurs se produisaient soit à cause des voyelles nasales, du [R] et des consonnes explosives que l'on ne cherchait pas à reconnaître, soit du fait que le système, trop simple, créait par erreur des formants là où il n'y en avait pas. De plus, la nature de l'excitation et la nasalisation n'était pas prise en compte et la batterie de filtre ne montait pas assez haut en fréquence pour mettre en évidence le formant de [S] , [Z] . L'annexe 3 montre des résultats obtenus avec ce premier système.

Dans un second temps, on a complété la batterie de filtre (96 filtres) et on est en train de perfectionner le système de reconnaissance. La détection des formants a été améliorée, les critères relatifs à l'excitation, la nasalisation des voyelles, la détection des explosions et la nature de la phase soutenue des consonnes explosives sont mis en évidence. En revanche la diphtongaison n'est pas prise en compte pour le moment. On peut raisonnablement espérer que, toujours pour une parole bien articulée par un locuteur male quelconque, les résultats pour les voyelles et les consonnes fricatives seront bons (de l'ordre de 80 % de réussite). Pour ce qui est des consonnes explosives, le critère permettant de distinguer entre les 3 classes [P, B, M] , [T, D, N] et [K, G] n'est pas encore défini avec assez de précision. A cause de cela, son extraction est réalisée de façon très sommaire et les erreurs dues aux confusions entre les 3 classes atteindront certainement 30 à 40 % des cas.

Comme nous l'avons indiqué au paragraphe 4, les décisions sont prises par tout ou rien. Les résultats seront inscrits sur ruban perforé en utilisant un code à 5 moments. Il est à remarquer que le système est resté relativement simple.

7. - APPLICATION D'UN TEL SYSTEME

De tels systèmes peuvent être utilisés comme périphériques d'ordinateurs. Ils permettront pour une prononciation bien articulée, mais quel que soit le locuteur, de rentrer des informations nécessitant un vocabulaire réduit, par exemple : les listes de cablages des systèmes logiques en cours de construction (qui représentent très vite des milliers d'adresses à perforer), ou bien des informations bancaires. Cette catégorie d'applications est, économiquement parlant, très importante.

Le même système sert pour tous les vocabulaires différents, seule la programmation du vocabulaire est à modifier (soft). Par contre, il ne sert que pour une langue. On peut toutefois aisément lui faire tenir compte de variations locales ("accent"). L'adaptation aux voix féminines semble très possible.

Comme cela a été dit plus haut, l'étape suivante consistera à modifier le système de décision pour y incorporer des probabilités de présence : le degré d'application demandé aux locuteurs sera alors bien diminué.

ANNEXE 1 - LISTE DES CRITERES INTERVENANT POUR LES PHONEMES FRANCAIS

Certains critères tels que sourd-sonore, voyelle nasalisée ou non etc.. sont connus depuis longtemps, mais souvent, ils sont définis en se référant à l'organe d'émission, tandis que ceux employés ici le sont par rapport à l'organe de réception (oreille).

- Excitation : voisée pure — voyelles
  voisée + friction — consonnes fricatives sonores
  friction pure — consonnes fricatives sourdes
  explosion, c'est-à-dire variation brusque de l'énergie ou déformation rapide du spectre

- Formants : pour les voyelles pures, appartenance des 2 premiers formants à 2 zones données,

  pour les voyelles nasales, appartenance du 2e formant (parfois unique) à 1 zone donnée,

  pour les consonnes fricatives, appartenance du formant à 1 zone donnée.

- Nasalisation des voyelles : l'amplitude du second formant est supérieure à celle du premier formant

- Nature de la phase soutenue des consonnes explosives :

silence pur	[P, T, K]
voisé	[B, D, G]
nasalisé	[M, N]
pour L	[L]

- Critères pour distinguer les groupes de consonnes explosives :

  [P, B, M]
  [T, D, N]
  [K, G]

- Diphtongaison (vitesse de variation des formants)

- Critères de durée et de stabilité pour les voyelles

ANNEXE 2 - CRITERES CARACTERISANT LES DIFFERENTS PHONEMES

1. Voyelles non nasales : [I] [E] [E] [E] [EU] [A] [O] [O] [U] [OU]
   - excitation voisée pure (la voix chuchotée n'est pas considérée),
   - appartenance des 2 premiers formants à 2 zones données,
   - durée minimale.

2. Voyelles nasales : [IN, UN] [AN] [ON]
   - excitation voisée pure,
   - appartenance du 2e formant à 1 zone donnée,
   - 2e formant supérieur au 1er,
   - durée minimale.

3. Consonnes fricatives sourdes (sauf F) : [S] [CH]
   - excitation friction pure,
   - appartenance du formant à 1 zone donnée
   - durée minimale.

4. Consonnes fricatives sonores (sauf V) : [S] [J]
   - excitation friction + voisée
   - appartenance du formant à 1 zone donnée
   - durée minimale.

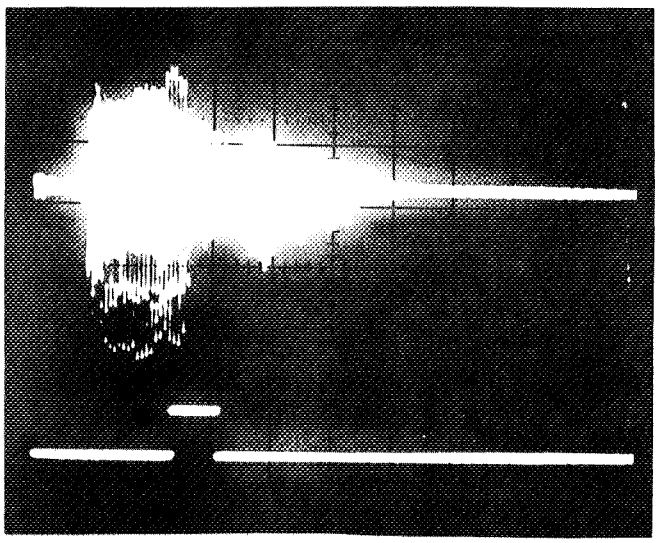
   Le [R] grasseyé est une consonne fricative indifféremment sourde ou sonore, donc appartenant à la fois aux 2 catégories 3 et 4.

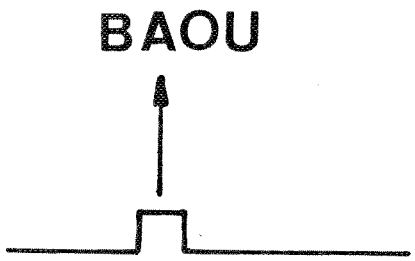
   Pour [F] et [V], il semble qu'il faille remplacer la condition sur le formant par le fait que le spectre doit être large et de faible intensité.

5. Consonnes explosives : [P] [B] [M] [T] [D] [N] [K] [G] [L]
   - explosion (avec obligatoirement modification spectrale dans le cas des explosives sourdes)
   - phase soutenue :
     silence pur [P] [T] [K]
     voisé [B] [D] [G]
     nasalisé [M] [N]
     pour L (2 formants) [L]
   - caractéristique de la transition pour distinguer entre [P B M] [T D N] [K G]

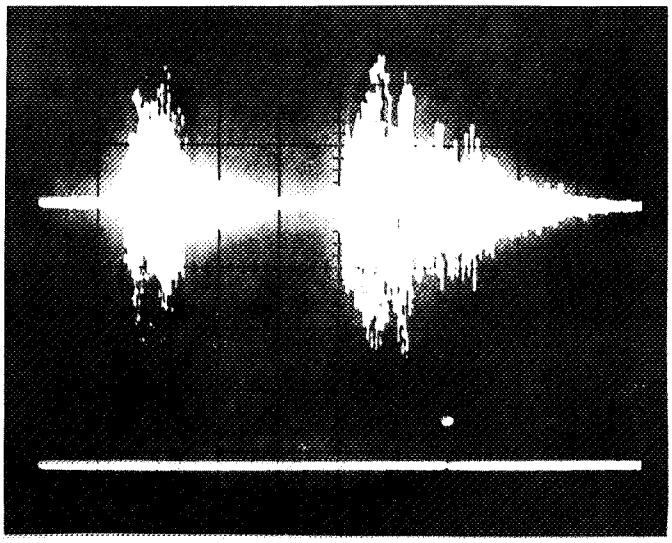
6. Diphtongues
   - excitation voisée pure
   - points de départ des 2 premiers formants
   - évolution suffisamment lente des formants vers ceux de la voyelle jointe (pour ne pas déclencher le critère d'explosion).

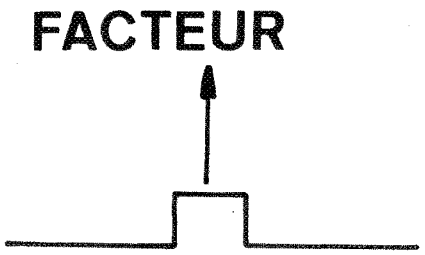
ANNEXE 3



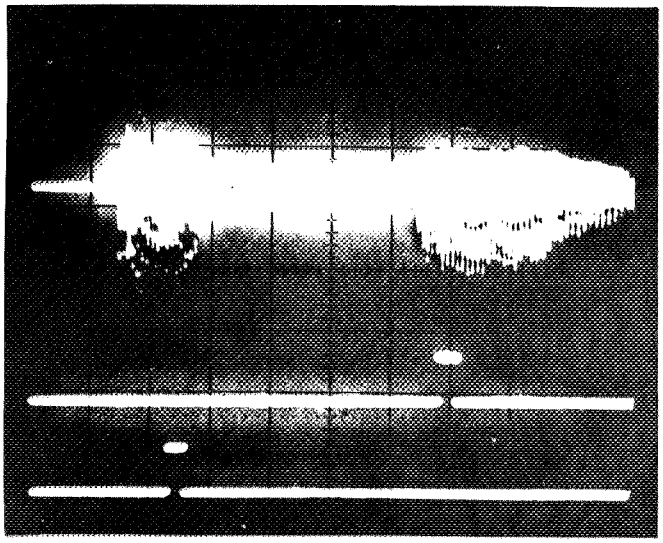


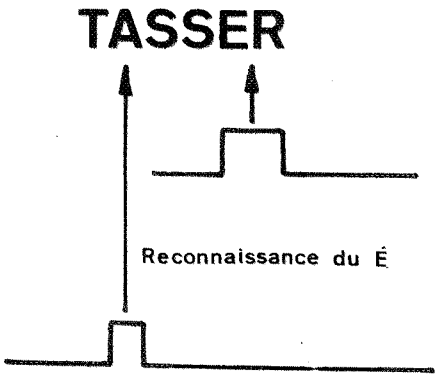
Reconnaissance de la voyelle A





Reconnaissance du EU





Reconnaissance du A

BIBLIOGRAPHIE

1 Von Bekesy
"Experiment in hearing"
Mc Graw Hill Book

2 A.M. Liberman, F.S. Cooper, D.P. Shankweiler et M. Studdert Kennedy
"Perception of the speech code"
Psychological Review, vol 74 n° 6, Nov 1967, p. 431 à 461

3 A.M. Liberman
"Some results of research on speech perception"
JASA, vol 29 n° 1, Jan 1957, p. 117 à 123

4 P. Delattre
"Les indices acoustiques de la parole : Premier Rapport"
Phonetica, vol 2, p. 108 à 118, 1958

TABLE RONDE

LINGUISTIQUE ET RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

ORATEURS : MM. HERAULT (Faculté des Sciences - Paris)
ROCHE (Institut de Programmation - Paris)
QUANCARD (D.R.M.E.)

ANIMATEUR : M. GRESSER (C.N.E.T.)

Ont participé à la discussion :

MM. CONTENSOU (C.R.I.)
DEMAN (Thomson-C.S.F.)
J.S. LIENARD (Fac. Sciences PARIS)
PAU (E.N.S.T.)
ROSSI (Fac. Lettres AIX EN PROVENCE)
GUEGUEN (E.N.S.T.)
RISSET (Conservatoire de Musique MARSEILLE)
VINCENT-CARREFOUR (C.N.E.T.)

Le compte rendu sera publié dans le bulletin du groupe de travail AFCET Reconnaissance Automatique de la Parole.

T A B L E R O N D E

INSTRUMENTATION

Modérateur : M. M. CARTIER (C.N.E.T. - Lannion)

TABLE RONDE SUR L'INTELLIGIBILITE DE LA PAROLE ET LA NORMALISATION DES TESTS

DERNIERS RESULTATS

TABLE RONDE SUR L'INTELLIGIBILITE DE LA PAROLE ET LA NORMALISATION DES TESTS

DERNIERS RESULTATS

-=-=-=-=-=-=-=-

M. ROSSI

Nous ne reviendrons par sur les conclusions qui ont été tirées à la suite du dernier colloque qui s'est tenu à Aix-en-Provence en avril ; nous essayerons plutôt de dégager certains des problèmes qui se posent pour la normalisation des Tests d'intelligibilité.

Nous avons travaillé en collaboration avec M. PECKELS pour voir si on pouvait obtenir des résultats constants avec de nouveaux sujets et sur un même type de Vocoder. Les résultats ont été dans l'ensemble encourageants. En effet avec trois groupes (chacun d'entre eux était composé de 14 sujets) - (deux groupes l'année dernière et un groupe cette année) nous obtenons des résultats homogènes dans la direction suivante :
Le % d'erreurs et de bonnes réponses peut évidemment varier selon le degré d'entraînement des sujets : l'année dernière nous avions constaté une différence importante entre un groupe de sujets phonétiquement naïfs et un groupe de non-naïfs, cette année le groupe que nous avons utilisé était relativement naïf. De ces deux séries d'expérience il se confirme que la constance des réponses des auditeurs ne repose pas sur le pourcentage de fautes, mais plutôt sur le type de fautes commises et ce pour les trois groupes en question.
EX: Dans le cadre de notre test passé avec un locuteur et un vocoder I.B.M. les consonnes aigües étaient confondues dans une forte proportion avec les consonnes graves contrairement à ce que trouvait VOIERS. (par ex : dette était confondu avec bette), mais par contre l'erreur inverse (c'est à dire les consonnes graves identifiées comme des consonnes aïgues) était beaucoup plus rare.

Nous retrouvons le même type d'erreurs dans les trois groupes comme l'indiquent clairement les pourcentages :

1) Consonnes aiguës → consonnes graves
- 1° groupe : 20 % d'erreurs
- 2° groupe : 15 % "
- 3° groupe : 20 % "

2) consonnes graves → consonnes aiguës
- 1° groupe : 12 % "
- 2° groupe : 5 % "
- 3° groupe : 10 % "

Nous pouvons dire également la même chose à propos de l'opposition compact-non-compact. ex : / t ∿ K/ .

Il semble s'agir dans ce cas d'un type d'erreurs qui dépend non pas des auditeurs, mais du locuteur. Nous avons testé sur le Vocoder I.B.M. un locuteur I.B.M. et un locuteur de Lannion.
- Avec le locuteur I.B.M. on a le même type d'erreurs dans les trois groupes. C'est-à-dire que ce sont les consonnes comme /t/ qui sont mises pour /k/. Donc c'est l'erreur diffus → compact qui est caractéristique.
- Avec les mêmes auditeurs et le même Vocoder, mais avec un locuteur différent le type d'erreurs est inversé, ce qui veut dire que le locuteur lui-même peut être responsable d'une somme de perturbations très importante.

En d'autres termes et ceci peut être très important pour la normalisation ultérieure des tests, certains locuteurs appartenant à des groupes linguistiques différents peuvent utiliser des indices acoustiques différents pour la détermination de telle ou telle opposition.

Nous en avons ici un exemple très significatif. Le 2° locuteur semble utiliser un indice acoustique différent de celui qu'utilise le 1° locuteur. Le locuteur 2 utilise un bruit plus fort d'explosion pour /k/ alors que l'autre utilisait comme indice acoustique la convergence des transitions de F_2 et de F_3.

Un autre problème qui s'est posé est celui du choix binaire. Pour certain types de fautes ex : entre /f, s, ʃ/ les auditeurs placés devant le choix : <u>furent</u> ou <u>sûrent</u> répondent <u>sûrent</u> dans 60 à 70 % des cas.

Cela veut-il dire qu'ils entendaient <u>sûrent</u> ?

Ce n'est pas évident car si dans ce cas bien précis on lui avait proposé un choix ternaire avec /ʃ/ par exemple, des résultats différents et peut-être plus valables auraient été obtenus.

Faut-il donc méler dans un même test le choix binaire et le choix ternaire ? Nous ne le pensons pas car cela compliquerait peut-être inutilement l'exploitation du test ; mais dans le cas où des problèmes comme celui-ci se présenteraient, c'est-à-dire dans le cas de réponses distribuées au hasard un test complémentaire à choix ternaire pourrait certainement fournir des renseignements fort utiles.

PROPOSITIONS DE TECHNIQUES DE PRESENTATION ET DE DEPOUILLEMENT

PECKELS

Certains problèmes sont présentés et notamment celui de la création de la bande maîtresse et de la voix du locuteur professionnel.
Prenons par exemple les oppositions suivantes :

- voisée - non-voisée ex : bile-pile
- grave - aigu ex : fils-six

Au cours du test chaque mot d'une paire est présenté deux fois, donc il est permis de penser que les réalisations peuvent ne pas être exactement identiques et que le meilleur locuteur professionnel ne prononcera pas deux fois le mot (ex. bile) de la même façon et que cette différence dans les réalisations peut être une source d'erreur.

C'est là un point important qu'il convient de souligner dans l'élaboration du test.

Plusieurs moyens sont à notre disposition pour remédier à cet inconvénient. D'abord la technique du doublage.

Le mot est enregistré une seule fois, des copies de cet enregistrement sont faites et serviront au montage ultérieur du Test. Mais ce travail est long et fastidieux et il est évident que les copies seront toujours moins bonnes que les originaux. Il faut donc avoir recours à d'autres techniques et notamment à celle de la digitalisation.

Voici un bref résumé de cette technique :

"Lorsqu'on a une onde temporelle, on peut l'échantillonner à certains moments (To + ΔT) (To + 2 ΔT) etc...
Un théorème démontre que l'on peut échantillonner toute onde analogique un certain nombre de fois en conservant toutes les finesses de cette onde à condition que l'échantillonnage se fasse suivant un certain ordre et plus particulièrement si l'on veut échantillonner un signal qui aurait des composantes jusqu'à 5 - 6000 hz ; il faut l'échantillonner à des intervalles Δ T qui sont distants $\frac{1}{12000hz}$. On fait prononcer la liste au locuteur une seule fois, on digitalise et ensuite à l'aide d'un programme très simple sur ordinateur, on peut faire un assemblage et également faire l'inverse de l'opération décrite c'est-à-dire reproduire sur bande magnétique une onde analogique rigoureusement identique à la première. Par cette méthode, nous avons ainsi la certitude que le locuteur a présenté exactement le même mot à l'audition des sujets".

Au cours du dépouillement des Tests, nous avons relevé certaines anomalies dans les erreurs faites par les auditeurs. Les erreurs peuvent avoir deux raisons diverses : elles sont à chercher soit dans le principe même du Vocoder, soit dans les différences de réalisation du locuteur.

Dans le premier cas l'erreur reste. S'il y a 12 erreurs pendant le premier passage du mot, il y en aura à peu près le même nombre lors du deuxième passage. Dans le deuxième cas les chiffres sont moins égaux. La première fois il y a sept erreurs, la deuxième fois deux erreurs seulement. Certains diront que c'est l'effet de l'apprentissage ; or VOIERS démontre qu'il n'y a pas d'apprentissage possible dans ce test.

Pour notre part nous pensons que l'explication se trouve dans le fait que les mots présentés n'étaient pas rigoureusement les mêmes dans les deux cas et à plus forte raison à la sortie du Vocoder qui par la nature même des choses distort les sons.

CARTIER — Quel est l'intérêt de figer le mot prononcé alors que dans la réalité le locuteur n'émet pas toujours les sons de la même façon ?

DECHAUX Cette technique présente un double avantage.
- d'abord sur un plan pratique elle permet d'éviter des différences quelque fois considérables qui sont dues à l'effet de fatigue du locuteur.
Ex : le mot Vosne prononcé la première fois[von], la deuxième[von].

Il est possible aussi de générer rapidement une bande mágnétique et de réaliser un très grand nombre de Tests dans un ordre aléatoire différen

QUESTION

En multipliant les matériaux, n'aura-t-on pas des problèmes ultérieurs de dépouillement ?

REPONSE DE PECKELS :

Nous pensons qu'on a intérêt à rendre la génération des tests le plus aléatoire possible car dans le cas d'un nombre restreint de tests un problème peut se poser :
Si l'auditeur a suffisamment de mémoire, il est capable de s'apercevoir que s'il a déjà entendu deux fois "bile" et une fois "pile", la quatrième fois, ça ne peut-être que"pile". Un autre facteur intervient au niveau digital : les intervalles et la cadence du test sont rigoureusement identiques.

CARTIER On peut effectivement employer cette technique pour mêler les tests, mais peut-être pas de façon totalement aléatoire. Ce qu'il est possible de faire en ce qui concerne les derniers tests, c'est au moins d'intervertir les passages essentiels , car si le phénomène de mémoire ne joue pas en général, un certain effet d'accoutumance peut quand même se produire. C'est pourquoi malgré les risques d'apprentissage on a fait passer plusieurs fois le même test aux mêmes auditeurs en le faisant commencer aux trois-quarts, au quart ou à la moitié. Les auditeurs se rappellent peut-être le début d'un certain passage du test, mais l'effet d'ordre est ensuite très rapidement neutralisé.

DECHAUX Le problème du dépouillement peut être résolu assez facilement sous réserve qu'on puisse perforer sur carte les résultats (80 cartes perforées par sujet)

PECKELS Nous avons fait l'expérience d'une autre solution pour le dépouillement des tests. Après le test et pendant que les auditeurs sont encore présents, on fait lire à haute voix et de la façon la plus intelligible possible les mots réels qui ont été prononcés par le locuteur. Les auditeurs corrigent donc leurs propres fautes et ceci de la manière suivante : à l'aide d'un stylet, ils cochent le mot où ils ont fait une erreur en faisant sauter la perforation d'une carte du type pré-perforée. Les cartes passent dans l'ordinateur et le dépouillement est pour ainsi dire immédiat.

DECHAUX — A propos de voix de femmes : n'y-a-t-il pas intérêt, vu la nature différente des signaux que l'on fait passer dans le Vocoder, à opposer éventuellement deux listes dont l'une serait enregistrée par une voix de femme ?

ROSSI — Ces listes existent, et nous pensons les utiliser d'ici peu.

PECKELS — On peut remarquer que VOIERS n'a jamais utilisé de voix de femmes dans ses propres tests.

DEMAN — Au sujet de la présentation des tests : il est plus commode d'avoir des nombres avec leurs indices pour le dépouillement sur calculateur. Nous proposerons donc de composer les nombres de la façon suivante :
- il existe 9 phonèmes pour différencier les paires ; on pourrait donc leur donner un premier chiffre de 1 à 9.

- Ensuite il y a six traits : ces différents traits pourraient recevoir un deuxième chiffre allant de 1 à 6.

- Enfin le troisième chiffre pourrait être 0-1-2 ou 3 puisque le trait voisé est testé deux fois. Le caractère + serait la parité, donc 0 ou 2 et le caractère - serait 1 ou 3.

Ce système présenterait l'avantage que dans le cas de tests complémentaires, on aurait ainsi des numéros faciles à exploiter et à manipuler.

Les causes d'erreurs

CARTIER — A propos des conditions d'écoute des auditeurs nous avons rencontré un certain nombre de problèmes et notamment celui-ci :
L'an dernier nous avions constaté une différence significative entre les résultats des tests faits à Lannion et à Aix-en-Provence et ceux-ci avaient été faits avec la même bande. En principe les niveaux d'écoute étaient relativement semblables.

Cette différence a été d'abord attribuée au fait que les auditeurs d'Aix-en-Provence étaient munis de deux écouteurs avec le casque Socapex enveloppant tandis que ceux de Lannion avaient un écouteur téléphonique. Mais en fait des vérifications ultérieures ont montré que les courbes de réponses n'étaient

pas fondamentalement différentes.

La deuxième explication proposée était relative à la plus grande expérience des auditeurs d'Aix qui étaient peut-être moins "phonétiquement naïfs" que ceux de Lannion. Pour juger de l'importance de ce facteur nous avons fait l'expérience suivante:
Nous avons réuni deux groupes de huit auditeurs suivant leur expérience aux tests, leur origine, leur lieu de travail (département de langues, centre de calcul etc...). Nous les avons répartis de façon aléatoire (par ex : 4 auditeurs travaillant au centre de calcul dont deux étaient munis d'écouteurs et les autres d'un casque etc...). L'origine des auditeurs était ainsi neutralisée et nous avons organisé six sessions à raison d'une par jour.

Les auditeurs ont dans l'ensemble reconnu que la vitesse du test et le nombre de mots réunis (432) ne leur permettaient pas d'apprendre par coeur une partie du test.

Cette expérience a été instructive sur deux plans :

d'une part elle nous a montré que s'il y avait un problème d'accoutumance, ce problème était le même pour tout le monde. Je pense qu'il faut trois ou quatre essais pour stabiliser l'auditeur (VOIERS considère qu'il en faut quatre). D'autre part nous avons remarqué que les pourcentages d'erreur varient de façon considérable selon que l'écoute est binaurale ou monaurale.

DECHAUX N'y-a-t-il pas un deuxième phénomène qui interviendrait dans l'accoutumance et qui serait dû aux mots rares qui peuvent surprendre le locuteur ?

ROSSI A ce propos, nous avions conduit l'an dernier une expérience pour voir la fréquence d'apparition des sons (voyelles et consonnes) dans la langue parlée. Les résultats n'ont pas permis d'établir de corrélation entre la fréquence d'apparition des voyelles et des consonnes dans la langue et la fréquence dans le test. La fréquence des mots, est difficile à tester, mais VOIERS s'appuyant sur les recherches des psychologues affirme que dans un test à choix forcé, l'effet dû à la fréquence d'apparition des mots dans la langue est négligeable ou nul. En revanche, si le sujet n'est pas accoutumé à un mot rare, il peut être surpris et donner une réponse au hasard. Ce cas-là est toujours possible mais on peut éventuellement y remédier en lisant auparavant l'ensemble des mots et en donnant des explications sur les mots difficiles à comprendre. De toutes façons une lecture préalable est à conseiller.

Nous avons cherché à savoir si les sujets ne faisaient pas la faute toujours dans le même sens (par exemple dans la paire "saint-zain", le nombre de fautes est-il plus grand sur "zain" ?); nous avons eu rapidement la preuve que la rareté du mot (à condition que les auditeurs aient lu au préalable la liste) n'avait aucune incidence sur le sens des fautes.

CARTIER Pour revenir au problème de l'écoute, nous avons fait des mesures subjectives sur une oreille pour obtenir la même force sonore avec l'écouteur Audio 15 et les écouteurs du casque. S'il y a un problème de force entre une écoute monaurale et une écoute binaurale elle se situe entre 3 et 6 db du point de vue du gain. Afin de comparer les deux méthodes, nous avons retranché 6db pour avoir en principe la même force subjective. Malgré cette correction on observe une très grande différence entre les deux modes d'écoute. On a deux fois plus de fautes avec un écouteur qu'avec un casque.

Quelle est l'origine de cette différence ? On peut avancer plusieurs hypothèses : chaque écouteur ne donne pas tout à fait la même réponse; le fait d'avoir les deux écouteurs permet au sujet de mieux se concentrer; autre explication pratique : l'auditeur qui a son casque sur les deux oreilles est plus à l'aise que l'auditeur qui tient son écouteur. D'autres expérience seront nécessaires (écoute monaurale avec casque).

DEMAN <u>L'écouteur était-il sur l'oreille gauche ou droite</u> ?

CARTIER Sur l'oreille gauche, tous les auditeurs écrivent de la main droite.
Il y a deux sortes d'écoute : l'écoute monaurale et l'écoute binaurale.
Si on veut tester un système, comme le font les phonéticiens, on utilisera l'écoute binaurale avec un casque enveloppant. Mais l'écoute monaurale présente un intérêt capital puisque dans le cas d'une conversation téléphonique on n'utilise qu'une oreille.

DEMAN Quelle est l'incidence du type d'écoute au niveau des traits ?

CARTIER Nous avons pris en considération les résultats avec les écouteurs et avec les casques, en prenant non pas le % de fautes par rapport au nombre de fois où les paires ont été prononcées, mais le nombre de fautes par rapport au nombre total de fautes.

D'après les tests réalisés sur 5 écoutes et 8 auditeurs dans chaque groupe, soit en tout 40 écoutes avec casque et autant avec écouteur, les chiffres sont extrêmement proches. Tous les traits semblent être affectés de la même façon.

QUESTION Avez-vous essayé de faire varier la structure de l'organe de transmission, haute fidélité, Vocoder, etc... ?

PECKELS Les tests ont seulement porté sur un Vocoder ; en haute fidélité, les erreurs sont par définition extrêmement minimes.

QUESTION Quand vous avez fait l'essai en direct, disposiez-vous d'une bande téléphonique ou aviez-vous toute la bande dont étaient munis les casques.

ROSSI Les casques sont munis d'une bande passante qui va jusqu'à 6000 hz. (courbe de réponse : 100 - 4000 hz. à $\pm$ 6 db., fréquence de référence 400 hz.). En haute fidélité nous obtenons 99 % de bonnes réponses. Le nombre d'erreurs est de l'ordre de 1 ou 2 par sujet, ce qui est négligeable.

Nous avons également étudié le rapport des fautes par trait en comparant un Vocoder analogique, un Vocoder à 5- 6000 bits/sec et un Vocoder à 2400 bits/sec

Les résultats montrent que les traits Nasal, Vocalique, Interrompu sont pratiquement perturbés au même degré par les 3 Vocoders. En revanche, il existe une différence importante entre le Vocoder analogique et les autres types de Vocoder pour les Traits Compact/Non-Compact et Voisé/Non-Voisé. Nous en attribuons la raison à l'échantillonnage. La différence semble être significative entre les trois types de Vocoders en ce qui concerne le trait Grave/Aigu. Alors que la reconnaissance du trait Grave sur le Vocoder à 2400 bits/sec est très perturbée, elle l'est beaucoup moins sur celui à 6000 bits/sec et encore moins sur le Vocoder analogique.

La raison de ces différences dans la reconnaissance du trait Aigu/Grave peut être recherchée dans la largeur de la bande passante. Nous avons effectué un test en champ libre dans une chambre anéchoïque à partir de la bande originale filtrée (filtrage à 24 db/octave ; fréquence de coupure 1000 hz.) : dans ces conditions, c'est encore le trait Aigu/Grave qui a été le plus perturbé.

QUESTION — Quel est le nombre limite d'auditeurs pour passer les tests ?

CARTIER — Nous pensons qu'une dizaine d'auditeurs devrait suffire pour ce genre de tests.

DECHAUX — Avez-vous établi une corrélation entre les erreurs relevées sur chacun des traits et des défauts physiques du Vocoder ?

PECKELS — Pour certains traits, c'est absolument caractéristique. Nous laisserons de côté le trait Voisé/Non-Voisé où l'on voit tout de suite si le détecteur de pitch est en cause. Pour l'instant nous savons seulement que les plosives sont affectées par les filtres passe-bas du Vocoder. Pour les autres traits une corrélation nette et définie n'a pas pu être encore établie (cf. VOIERS, SMITH).

QUESTION — Des essais ont-ils été faits avec d'autres systèmes de compression de la parole, la téléphonie ou des systèmes analogiques ?

de MORI — Des chercheurs de l'Université de Turin ont entrepris un travail où les échantillons du signal sont groupés en trois groupes. Les segments quasi-stationnaires, les segments non-sationnaires et les segments de silence. Quand il y a un segment de silence, on représente ce segment seulement par sa durée. Quand il y a un segment quasi-stationnaire, le pitch est répété. Certains échantillons du pitch sont seulement transmis et un compteur indique leur nombre. Une période fondamentale est représentée seulement par un maximum et un minimum relatif et en réception on fait une reconstruction du signal avec des fonctions d'interpolation du type $\frac{5x}{x}$.
Il y a des déformations dans le domaine du temps car les maximums et les minimums n'ont pas de différences de temps constantes ; en fait, c'est une forme mathématique un peu plus complexe qui a été développée pour interpoler entre un maximum et un minimum.
La même technique a été employée pour les segments non-stationnaires. Nous sommes arrivés à 4000 bits/sec et des essais d'intelligibilité ont été faits, mais pour l'instant des expériences aussi poussées qu'en France n'ont pas été réalisées sur la langue italienne. Nous avons utilisé des mots qui diffèrent seulement pour la première consonne ou le premier phonème. Les mots ont été reconstruits au moyen de la technique exposée ci-dessus et des essais sont en cours. La qualité est en général assez bonne, mais le défaut princi-

pal est que la longueur des segments quasi-stationnaires n'est pas parfaite. Ainsi le système des segments quasi-stationnaires est un peu plus long que dans la réalité. En d'autres termes le système accentue un peu trop la longueur des voyelles. Il y a d'autres problèmes rythmiques qui sont soulevés pour l'identification des segments, mais nous espérons les résoudre en réduisant le nombre de bits perdus, c'est-à-dire en employant une sorte de modulation.

CONCLUSION

M. ROSSI Nous remercions toutes les personnes qui sont intervenues dans ce débat et qui ont par leurs questions et leurs réponses permis de clarifier certains points importants concernant les tests d'intelligibilité. Des tests vont être effectués par différentes équipes. Ce qui a été dit ici contribuera à leur normalisation ; des rencontres ultérieures pourront apporter des éléments de réponse à des points encore obscurs dans la mesure où, grâce à cette normalisation, les résultats seront comparables.

pal est que la longueur des segments quasi-stationnaires n'est pas parfaite. Ainsi le système des segments quasi-stationnaires est un peu plus long que dans la réalité. En d'autres termes le système accentue un peu trop la longueur des voyelles. Il y a d'autres problèmes rythmiques qui sont soulevés pour l'identification des segments, mais nous espérons les résoudre en réduisant le nombre de bits perdus, c'est-à-dire en employant une sorte de modulation.

CONCLUSION

M. ROSSI — Nous remercions toutes les personnes qui sont intervenues dans ce débat et qui ont par leurs questions et leurs réponses permis de clarifier certains points importants concernant les tests d'intelligibilité. Des tests vont être effectués par différentes équipes. Ce qui a été dit ici contribuera à leur normalisation ; des rencontres ultérieures pourront apporter des éléments de réponse à des points encore obscurs dans la mesure où, grâce à cette normalisation, les résultats seront comparables.

A1

LISTE DES PARTICIPANTS

M. ALINAT	C.S.F.	06800 - CAGNES SUR MER -
M. AUTESSERRE Denis	Institut de Phonétique - Université de Provence -	13100 - AIX EN PROVENCE -
M. AUZIMOUR	Compagnie Industrielle des Télécommunications - Centre de Villarceaux -	NOZAY - 91310 - MONTLHERY -
M. BAUDRY	C.E.A. S.E.A.S. G.I.A.C. BP. N° 2	91190 - GIF SUR YVETTE -
M. BENOIT J.	E.N.S.E.R.G. 23 Rue des Martyrs	38000 - GRENOBLE -
M. BERGER-VACHON	Fac. des Sciences - Université C. Bernard - Lab. de Physique Electronique - 43 bd du 11 Novembre - 69100 - VILLEURBANNE -	
M. BOE L.J.	Institut de Phonétique - Domaine Universitaire -	38400 - ST MARTIN D' HERES -
Melle BOUARD D.	Département E.T.A. - C.N.E.T. - Route de Trégastel	22301 - LANNION -
M. BOULOGNE	E.N.S.E.R.G. 23 rue des Martyrs	38000 - GRENOBLE -
M. BOURGENOT J.S.	Thomson C.S.F. G.1.	92230 - GENNEVILLIERS -
M. BRETECHER Yves	S.L.E. - Route de Perros -	22300 - LANNION -
M. BUISSON L.	Département C.E.I. - C.N.E.T. - Route de Trégastel	22301 - LANNION -
M. CAELEN	I.U.T. Ave de Rangueil	31400 - TOULOUSE -
Melle CAMUS Eliane	Laboratoire R.P.B. - T.S. C.I.T.	91460 - MARCOUSSIS -
M. CARAYANIS Georges	D.S.E.	Rue Barrault 75013 - PARIS -
M. CARRE R.	E.N.S.E.R.G. 23 Rue des Martyrs	38000 - GRENOBLE -
M. CARRIL Ramon	Consejo Sup. de Invest. Cient. Lab. de Fonetica Duque de Medinaceli, 4	MADRID (14) - ESPAGNE
M. CARTIER M.	Département E.T.A. - C.N.E.T. - Route de Trégastel	22301 - LANNION -
M. CASTAN S.	I.U.T. Avenue de Rangueil	31400 - TOULOUSE -

Melle CASTELLENGO	Laboratoire d'Acoustique - Fac. des Sciences - Tcur 66 - Quai Saint-Bernard	75005 - PARIS -
M. CHABERT	C.N.E.G. - Lab. L.E.T.I./MCTE/TS -	38000 - GRENOBLE-GARE - CEDEX
Mme CHAUVALLON	Secrétariat Général - Défense Nationale - 51 Bd de la Tour Maubourg	75007 - PARIS -
Mme CHEVILLARD	Laboratoire du Pr. Scherrer - Hôpital Salpétrière	75013 - PARIS -
Mme CHEVRIE C.	Laboratoire du Pr. Scherrer - Hôpital Salpétrière	75013 - PARIS -
M. CHOCHOLLE	Collège de FRANCE 11 Place M. Berthelot	75005 - PARIS -
M. CONTENSOU	IRIA Domaine de Voluceau	ROCQUENCOURT 78150 - LE CHESNAY -
M. CONTINI M.	Institut de Phonétique Domaine Universitaire	38400 - ST MARTIN D'HERES -
M. CSILLAG Pierre	Laboratoire GAPSE E.N.S.E.E.I.H.T.	31000 - TOULOUSE -
M. DECANTE Patrick	17 Ave St Exupéry	92160 - ANTONY -
M. DECHAUX	THOMSON C.S.F.	92230 - GENNEVILLIERS -
M. DEMAN	THOMSON C.S.F.	92230 - GENNEVILLIERS -
M. DE MORI-R.	CENS-IENGF POLITECNICO CORSO DUCA DEGLI ABRUZZI 24	10129 - TORINO (ITALIE)
M. DESCOUT	Département E.T.A. - C.N.E.T. - Route de Trégastel	22301 - LANNION -
M. DISSOUBRAY	Institut National des Jeunes Sourds 254 Rue St Jacques	75005 - PARIS -
M. DUMAS M.	247 - LAGACÉ - DORVAL 780, P.Q. - CANADA -	
M. DUPEYRAT	C.E.A. S.E.A.S. G.I.A.C. BP N° 2	91190 - GIF SUR YVETTE -
M. EL MALLAWANY	Département C.E.I. - C.N.E.T. - Route de Trégastel	22301 - LANNION -
M. GAGNEPAIN	Département de Linguistique Fac. des Lettres de Villejean	35000 - RENNES -
Melle GALVAGNY	35, rue de la République	59164 - MARPENT -
M. GENIN	Département E.T.A. - C.N.E.T. Route de Trégastel	22301 - LANNION -

M. GENUIST	I.U.T. - Route de Perros	22300 - LANNION -
M. GIMONET	C.E.R.T. 32 ave ET. BILLERES	31300 - TOULOUSE -
M. GODIN	222 Av. J. Wanters	7230 - FRAMERIES (BELGIQUE)
M. GRANGE M.F.	Lab. de PHONETIQUE - FACULTE DES LETTRES Rue MEGEVAND	25000 - BESANCON -
M. GRESSER J.Y.	Département C.E.I. - C.N.E.T. - Route de Trégastel	22301 - LANNION -
M. GUEDJ	THOMSON - C.S.F. L.C.R. Domaine de Corbeville	91400 - ORSAY -
M. GUEGUEN	E.N.S.T. - Lab. d'Automatisme 46 rue Barrault	75013 - PARIS -
M. GUERIN	E.N.S.E.R.G. 23 rue des Martyrs	38000 - GRENOBLE -
M. GUIBERT	R.P.B. - T.S. C.G.E.	91460 - MARCOUSSIS -
M. GUIGLIO	Fac. des Lettres et Sc. Hum. 98 Boulevard Carlone	06200 - NICE -
M. HATON	Faculté des Sciences Bd des Aiguillettes	54000 - NANCY
M. HECAEN H.	Centre Neurochir. Sainte Anne 1 Rue Cabanis	75014 - PARIS -
M. HERAULT D.	Centre de Linguistique Quantitative Université PARIS 6	91910 - ST SULPICE DE FAVIERES -
M. JACOB J.B.	S.L.E. Route de Perros	22300 - LANNION -
M. JACQUEMIN	Institut de Phonétique Domaine Universitaire	38400 - ST MARTIN D'HERES
M. JAFFRES	S.L.E. Route de Perros	22300 - LANNION -
M. JOSPA P.	Institut de Phonétique - Université Libre Ave F.D. Roosevelt	BRUXELLES (Belgique)
M. KAMMINGA	Delf University of Technologyee Dept. Mekelweg 4	DELF (Hollande)
Mme KONOPCZINSKI	Maitre assistante de Phonétique - Fac. des Lettres et Sciences Hum.	- 25000 - BESANCON -
M. LAMOTTE M.	Fac. des Sciences Bd des Aiguillettes	54000 - NANCY -
M. LANDERCY A.	Institut de Phonétique - Université Libre Ave F.D. Roosevelt	BRUXELLES (Belgique)

M. LAURENT Gérard	S.L.E. Route de Perros	22300 - LANNION -
M. LAVANANT P.	S.L.E. Route de Perros	22300 - LANNION -
M. LE CORNEC	Département C.E.I. - C.N.E.T. - Route de Trégastel	22301 - LANNION -
MMe LHOTE	Faculté des Lettres Rue Mégevand	25000 - BESANCON -
M. LIENARD J.S.	Lab. d'Acoustique - Faculté des Sciences - Tour 66 - 9 Quai St Bernard	75005 - PARIS -
M. LINDBLOM	Institute of Linguistics - University of Stockholm - Box 23144	STOCKHOLM 23 (Suède)
M. LOOSE PETER	HENDISNALAND 64,	DEN HAAG 2020 (Pays-Bas)
M. LORAND P.	Département E.T.A. - C.N.E.T. - Route de Trégastel	22301 - LANNION -
M. MAISSIS	E.N.S.T. - Lab. d'Automatisme - 46 Rue Barrault	75013 - PARIS -
M. MARCIE P.	Pathologie du Langage - V.111 de l'INSERM 2 Ter rue d'Alésia	75014 - PARIS -
M. MERCIER G.	Département C.E.I. - C.N.E.T. - Route de Trégastel	22301 - LANNION -
Mme METTAS O.	42 Ave R. Coty	75014 - PARIS -
M. MLOUKA M.	C.C.A. Bat. 508 BP N° 30	91400 - ORSAY - PLATEAU -
M. MRAYATI	E.N.S.E.R.G. 23 Rue des Martyrs	38000 - GRENOBLE -
M. PAILLE J.	E.N.S.E.R.G. 23 Rue des Martyrs	38000 - GRENOBLE -
M. L.F. PAU	E.N.S.T. Lab. D'Automatisme 46 Rue Barrault	75013 - PARIS -
M. PECKELS	I.B.M.	06610 - LA GAUDE -
M. PERENNOU G.	I.U.T. Ave de Rangueil	31400 - TOULOUSE -
M. PERSON J.M.	Département T.M.A. - C.N.E.T. Route de Trégastel	22301 - LANNION -
M. PHAM VAN VUI	Chargé de Recherche au CNRS - Lab. GAPSE E.N.S.E.E.I.H.	31300 - TOULOUSE -
M. PIMONOW	Président du GALF - C.N.E.T. -	92130 - ISSY LES MOULINEAUX -

M. PINEL J.	THOMSON CSF - L.C.R. Domaine de Corbeville	91400 - ORSAY -
M. PYNN	THE PLESSEY CO. LTD - TAPLOW COURT - TAPLOW, NR.	MAIDENHEAD - BERKSHIRE (Ang.)
M. QUANCARD	D.R.M.E. - Ministère de l'Air 5 Bis Ave de la Porte de Sevres	75015 - PARIS -
M. QUERRE M.	Département C.E.I. - C.N.E.T. - Route de Trégastel	22301 - LANNION -
M. QUILIS	Consejo Sup. de Inv. Cientif. Lab. de Fonetica Duque de Medinaceli, 4	MADRID (14) - ESPAGNE -
M. RAMAKRISHNAN KOLL	LAB. GAPSE - F.N.S.E.E.I.H.T.	31300 - TOULOUSE -
M. RISSET J.C.	4 Bd Auguste Blanqui	75013 - PARIS -
M. ROCHE C.	Institut de Programmation - Lab. de Reconnaissances des Formes et d'Intelligence Art. - 9 Q. St Bernard -	75005 - PARIS -
M. ROGER	C.G.E.	91460 - MARCOUSSIS
M. ROINSOL	Institut National des Jeunes Sourds 254 Rue St Jacques	75005 - PARIS -
Mlle RONAT	18 Rue St Lazare	75009 - PARIS -
M. ROSSI M.	Institut de Phonétique Fac. des Lettres	13100 - AIX EN PROVENCE -
M. ROSTOLLAND	Physiologie du Travail du CNAM 41 Rue Gay-Lussac	75005 - PARIS -
M. SALLE Y.	Dept Transmission C.I.T. - Cent. de Villarceau - Rte de Villejust	NOZAY - 91310 MONTLHERY -
M. SANTERRE L.	Dep. de Linguistique C.P. 6128 Université de Montreal	MONTREAL (Canada)
M. SAP	C.I.T.	91460 - MARCOUSSIS -
M. SERRA ANGELO	CENS - IENGE - POLITECNICO-TORINO	TORINO (Italie)
Mme SIMON P.	Institut de Phonétique 25 Rue du Soleil	67000 - STRASBOURG -
M. TESTON	Institut de Phonétique - Univ. de Provence 29 Ave R. SCHUMAN	13100 - AIX EN PROVENCE -
Mlle THIEBERGER	Fac. des Lettres et Sciences Humaines 98 Bd Carlone	06200 - NICE -
M. TILKOV DIMITAR	Institut de Phonetique - Université III	38000 - GRENOBLE-CARE CEDEX -
M. TUAUDEN Jean	S.L.E. Route de Perros	22300 - LANNION -

M. VASSEUR J.C.	THOMSON CSF LCR - DR 5 Domaine de Corbeville	91400 - ORSAY -
M. VINCENT-CARREFOUR J.	Département C.E.I. - C.N.E.T. - Route de Trégastel	22301 - LANNION -
M. VIVES R.	Département C.E.I. - C.N.E.T. - Route de Trégastel	22301 - LANNION -
M. VOMSCHEID C.	Fac. des Sciences Boulevard des Aiguillettes	54000 - NANCY -
M. WAJSKOP	Institut de Phonetique - Université Libre Ave F.D. Roosevelt	BRUXELLES (Belgique)
M. WIOLAND	Institut de Phonetique 25 rue du Soleil	67000 - STRASBOURG -
M. ZURCHER	Département E.T.A. - C.N.E.T. - Route de Trégastel	22301 - LANNION -