Groupement des Acousticiens de Langue Française

n° 1172

4èmes JOURNEES D'ETUDES du GROUPE de la " COMMUNICATION PARLEE "

SOUS LE PATRONAGE DE
L'ASSOCIATION BELGE DES ACOUSTICIENS (A.B.A.V.)
ET AVEC
LA PARTICIPATION DE L'A.F.C.E.T.

Organisées par
l'INSTITUT de PHONETIQUE
de l'UNIVERSITE LIBRE de
BRUXELLES

Bruxelles
23 - 25 mai 1973.

n° 1172


UNIVERSITE DE PROVENCE
CENTRE D'AIX
Bibliothèque
de l'Institut
de Phonétique
1172

Nous tenons à exprimer nos plus vifs remerciements à Monsieur le Ministre de la Culture Française, représenté à ces Journées par Monsieur V. FEAUX, Chef de Cabinet Adjoint, au FONDS NATIONAL DE LA RECHERCHE SCIENTIFIQUE et au MINISTERE DE L'EDUCATION NATIONALE pour l'aide matérielle qu'ils nous ont fournie ainsi qu'à Monsieur A. JAUMOTTE, Recteur de l'UNIVERSITE LIBRE DE BRUXELLES.

Pour le Groupe de la Communication Parlée

Le Président, R. CARRE

Comité Organisateur

MM. P. RIJLANT
M. LEROY
P. BERTELSON
M. WAJSKOP

1973
EDITIONS DE L'UNIVERSITE DE BRUXELLES
Parc Léopold, 1040 Bruxelles

TABLE DES MATIERES

Tables Rondes.

Allocution prononcée par M. le Recteur A. JAUMOTTE, lors des 4èmes Journées d'Etudes de la Communication parlée .

Monsieur le Représentant du Ministre,
Monsieur le Délégué du Gouvernement,
Mesdames et Messieurs,

Depuis février 1970, le Groupement des Acousticiens de Langue Française a vu se spécialiser l'une de ses structures de recherche : le "Groupe de la Communication parlée".

Les études entreprises alors à Aix, Bruxelles, Grenoble, La Gaude, Lannion, Marcoussis, Nancy, Paris, Saint-Martin d'Hères et Toulouse portent sur
- l'analyse,
- la synthèse,
- la perception,
- la reconnaissance,

et l'intelligibilité de la parole.

Les progrès rapides de ces recherches interdisciplinaires et multidisciplinaires ont requis des rencontres annuelles. Celles-ci ont eu pour cadre Grenoble en 1970, Aix en 1971, Lannion en 1972.

Voici qu'en 1973 Bruxelles accueille à son tour cette rencontre. Les thèmes inscrits à l'ordre du jour sont : la perception de la parole, la psycholinguistique et la reconnaissance automatique de la parole. Ils touchent plus particulièrement, dans ce dernier cas, à la normalisation selon le locuteur, à la recherche des paramètres phénoménologiques et à l'utilisation des règles linguistiques.

On le voit, ceci nécessite plus qu'une collaboration entre ingénieurs, linguistes, phonéticiens et psychologues. En fait chacun d'eux a dû pénétrer et maîtriser les connaissances, les langages et les modes de pensée de tous les autres. Il s'agit d'une structuration dynamique réciproque d'univers épistémologiques particuliers. Faut-il souligner la difficulté de ce travail ? Sinon se réjouir de voir qu'elle a suscité un groupe jeune, informel, enthousiaste et souple. Ce sont bien les qualités requises d'un collectif de recherches confronté avec un domaine en voie d'explosion, plus encore que d'expansion.

Qu'il s'agisse de sciences humaines, expérimentales ou spéculatives, que leurs aspects soient fondamentaux ou appliqués, les résultats acquis par le groupe constituent un nouveau corps de réflexion essentiellement collective. Son ambition est d'en étendre l'espace à la neurophysiologie, dont les développements récents permettent d'espérer prochainement le repérage et l'explication en termes objectifs de ce phénomène capital qu'est "un-homme-comprenant-ce-qu'il-écoute".

Les efforts consentis et les résultats obtenus permettent d'envisager favorablement l'incorporation, à la fois structurante et enrichissante, de la neurophysiologie aux approches fructueuses du "Groupe de la Communication parlée". Ce nouvel objectif ambitieux et hardi, élargira encore le rayonnement dont ses recherches bénéficient en Europe et en Amérique. Dès 1971, l'Angleterre et la Suède, par le truchement de leurs grandes écoles de phonéticiens et de linguistes manifestaient leur vif intérêt pour les démarches entreprises. Cela soulignait l'ouverture d'esprit du Groupement des Acousticiens de Langue Française qui n'a cessé de proclamer (et de prouver) sa volonté de ne pas demeurer enfermé dans les limites rigides de la francophonie.

Aujourd'hui, la richesse de cette attitude et de ses conséquences assumées marque les nécessités qui s'imposent à la poursuite des activités entreprises. Ces nécessités sont celles de la coopération, intensifiée, sur le plan européen, par la mise en place d'une action thématique programmée.

Cette coopération lierait les laboratoires de reconnaissance de la parole aux laboratoires de phonétique d'une part, et, d'autre part aux laboratoires de psychologie expérimentale.

Elle manifesterait la nécessaire ouverture sur le monde des recherches fondamentales et appliquées. Plus encore la féconde ouverture de l'une sur l'autre, lorsqu'une méthodologie appropriée aux problèmes en cause permet de surmonter les préjugés- ou même les méfiances éventuelles ! - que certaines fractions du monde académique pourraient nourrir pour le monde industriel ou réciproquement.

Dans un univers, où "l'Etranger" d'Albert CAMUS est depuis longtemps le symbole de l'incommunicabilité entre les êtres, le seul moyen d'éprouver celle-ci est de faire communiquer entre eux ceux qui s'attachent aux aspects complexes de la communication parlée. C'est chose faite. Leur dialogue est noué. Souhaitons-lui de se poursuivre fructueusement à l'occasion de ces Journées d'Etudes.

Allocution prononcée par M. V. FEAUX, Chef de Cabinet Adjoint, représentant le Ministre de la Culture Française, lors des 4èmes Journées d'Etudes de la Communication Parlée.

Monsieur le Recteur,
Monsieur le Président,
Monsieur le Délégué du Gouvernement,
Mesdames, Messieurs,

Je voudrais tout d'abord excuser l'absence de Monsieur le Ministre de la Culture Française. En effet , le Ministre Falize aurait beaucoup aimé pouvoir assister à ces travaux notamment parce qu ' aujourd'hui la Culture ne se limite plus aux seuls domaines des Arts, des Lettres et des Sciences mais s'étend à tous les aspects de la vie en société, et il est bien certain dès lors que la Communication Parlée est un des éléments essentiels d'une politique culturelle.
Il aurait aimé être des vôtres également, parce que, par sa formation professionnelle de base, il a toujours été intéressé par les problèmes de phonétique. Malheureusement ce matin, il doit devant une Commission du Conseil Culturel défendre son budget 1973, c'est évidemment une obligation impérieuse à laquelle il ne peut se soustraire.
Il m'a demandé de le représenter et j'en suis vraiment très honoré parce que d'une part cela me permet pour quelques instants de me replonger dans le cadre de vie universitaire qui a été le mien pendant une bonne dizaine d'années puisque j'ai fait de la recherche scientifique dans cet Institut et d'autre part parce que cela me permet d'apprécier combien est sereine et reposante cette atmosphère de l'Institut de Sociologie comparée à la trépidante quotidienneté des problèmes que je dois traiter. Mais j'arrête ici ces considérations personnelles et un tantinet nostalgiques.

Je voudrais simplement dire, combien je suis heureux que l'on ait choisi pour ces 4èmes Journées de la Communication Parlée, Bruxelles, et plus particulièrement l'Université Libre de Bruxelles et je voudrais remercier tout spécialement l'organisateur de cette rencontre : l'Institut de Phonétique qui a mis sur pied ces Journées à Bruxelles. Je voudrais également me réjouir des progrès réalisés dans le domaine de la Phonétique et de l'Acoustique qu'elle soit physique, technique ou physiologique et souligner en tout cas le rôle joué dans ce développement par le Groupement des Acousticiens de langue française et plus particulièrement pour ce groupe qui s'est créé à Grenoble : le Groupe de la Communication Parlée.

Cela me permet de dire que la Francophonie, à laquelle inévitablement un Ministre de la Culture Française est attaché, n'est pas une attitude de repli, mais au contraire, une attitude d'esprit largement ouverte et dynamique sur le monde.

J'en trouve la preuve dans l'ensemble des chercheurs qui sont ici, réunis pour ces Journées d'Etude et où se retrouvent des chercheurs allemands, italiens, suisses, hollandais, français et bien entendu belges.

Les différents thèmes qui vont être évoqués : "perception", "psycholinguistique", "reconnaissance automatique de la parole" apparaissent un peu divers, dans leur approche ainsi d'ailleurs que les différents horizons scientifiques qui sont représentés dans cette salle. Mais cela démontre aussi combien aujourd'hui la recherche en Sciences Humaines a elle aussi cessé d'être cantonnée dans les limites étroites des Belles Lettres pour s'étendre à tous les aspects des plus théoriques aux plus appliquées du comportement humain dans ce qu'il a de plus essentiel, à savoir la communication linguistique. Je ne suis ni un spécialiste des problèmes de phonétique, ni un spécialiste des problèmes d'acoustique, par conséquent je ne me lancerai pas dans de grands développements à ce sujet, mais je ne peux quand même pas m'empêcher de souligner, en tout cas, l'énorme chemin parcouru ces dernières années dans ces domaines de recherche.

Il est incontestable que l'analyse acoustique, l'apport des théories physico-mathématiques, tous les concepts qui ont été générés par la linguistique contemporaine et je crois aussi l'aide appréciable qu'ont apportée les ordinateurs, tout a, fondamentalement révolutionné la phonétique expérimentale.

Celle-ci a cessé d'être une espèce d'auxiliaire modeste de la philologie pour devenir aujourd'hui un confluent scientifique, une science-carrefour dont les résultats s'intègrent immédiatement avec un impact d'ailleurs grandissant, dans la plupart des disciplines connexes. J'ai lu les différents titres des communications qui vous seront soumises et quelques résumés également : tout démontre un regroupement qui s'opère entre les différentes disciplines en même temps que s'effacent les cloisonnements et les hiérarchies qui sont des cicatrices de l'histoire des Sciences.
Les recherches qui ont été menées au sein de cette Université entre l'Institut de Phonétique et le Laboratoire de Psychologie Expérimentale, enre le laboratoire de l'Institut de Phonétique et le Laboratoire de Synthèse de la Parole de Grenoble, tout cela sont des exemples éloquents de ces fusions fructueuses qui sont garantes de l'avenir et on ne peut vraiment que se réjouir de cette collaboration et de cette rencontre qui sont appelées à avoir des lendemains prometteurs pour les Sciences Humaines. Je voudrais terminer en réitérant aux organisateurs de ce Colloque mes plus vives félicitations et formuler au nom du Ministre de la Culture Française, en mon nom personnel, les voeux sincères pour que ces travaux apportent une contribution nouvelle aux Sciences Phonétiques et Acoustiques.

Allocution prononcée par M. René CARRE, Président du Groupe de la Communication Parlée, représentant le G.A.L.F.

Monsieur le Recteur,
Monsieur le Représentant du Ministre,
Monsieur le Délégué du Gouvernement,
Mesdames et Messieurs,

Je suis vraiment très heureux de participer à ces Journées d'Etudes, très heureux et aussi très honoré de me retrouver à cette place.

En effet, Monsieur Pimonow, Président du G.A.L.F. n'a pu, comme il en avait l'habitude, inaugurer ces Journées. Il m'a chargé de vous transmettre le salut du G.A.L.F. et ses meilleurs voeux de réussite.

Je suis très heureux de participer à ces Journées pour différentes raisons mais qui peuvent, en fait, se résumer en une seule.

C'est que nous sommes à Bruxelles et que, il y a 5 ans, c'était en février 1968, notre ami Wajskop avait pris l'initiative d'organiser des Journées d'Etudes sur la Parole.

Il était convaincu, et il a eu raison, de l'importance d'une collaboration étroite entre chercheurs d'origine diverse.

A cette rencontre, participaient les équipes de l'Institut de Phonétique et du Laboratoire de Psychologie Expérimentale de l'Université Libre de Bruxelles; il y avait aussi le directeur du Département de Phonétique de Londres D.B. Fry, Madame Mettas (de Paris), Mario Rossi (d'Aix) et de Grenoble : Lancia, Paillé, Beauviala et moi-même.

Nous avions été accueillis royalement. Les Journées étaient très chargées, à cause du travail, et les nuits très courtes, pour d'autres raisons.

Je vous conseille par exemple, une dégustation de gueuze ou bien une soupe à l'oignon vers 4 heures du matin, ou bien un bon repas de moules.

Bref, à la suite de cette rencontre, tout nous paraissait favorable pour envisager le renouvellement de l'expérience.

Et, en effet, la réunion suivante eut bien lieu en 1970, à Grenoble cette fois. Nous étions une quarantaine et c'est à l'issue de cette rencontre que fut créé, à l'intérieur du G.A.L.F., le goupe "Communication Parlée".

En 1971, M. Rossi organisait les Journées à Aix. Nous étions environ 80.

Puis en 1972, 100 personnes se retrouvèrent à Lannion dans les locaux du C.N.E.T. sous la houlette de Mercier.

Notre groupe représente aujourd'hui en Europe le premier groupement inter-disciplinaire organisé autour de la Communication Parlée.

Si l'on considère le chemin parcouru depuis sa création, nous pouvons admettre qu'il a dépassé le stade de sa formation. Peut-être faudra-t'il repenser ses structures, revoir ses orientations, coordonner davantage ses activités ?

Le nombre élevé de participants rendra plus difficile la formule de travail adoptée laquelle, de par sa souplesse et le caractère informel des rencontres, a donné de bons résultats.

Les problèmes seront plus difficiles encore si nous ouvrons plus largement le groupe aux chercheurs étrangers, mais, malgré tout, ceci paraît très souhaitable.

Le chemin qui a été parcouru depuis 1968, grâce à l'initiative de Max Wajskop et à partir de projets élaborés autour de verrres de gueuze, est donc considérable.

Vous comprenez maintenant pourquoi j'éprouve un vif plaisir à me retrouver ici.

Cette réunion a été possible grâce à l'obligeance et au dévouement de nombreuses personnes.

Il m'est particulièrement agréable de remercier, au nom du bureau de notre groupe, Monsieur le Professeur Jaumotte, Recteur de l'Université de Bruxelles qui a bien voulu accueillir notre réunion et Monsieur le Ministre de la Culture Française qui a accepté de nous accorder son patronage.

Mes remerciements s'adressent également à Monsieur le Professeur Bosquet, Président de l'Association soeur, l'Association Belge des Acousticiens, à Monsieur le Professeur Doucy qui a mis les locaux de l'Institut de Sociologie à notre disposition.

Je remercie aussi le Fonds National de la Recherche Scientifique dont l'aide matérielle nous a été des plus précieuses.

Enfin, nous tenons à exprimer notre gratitude aux membres du Comité organisateur et, en particulier, à notre ami Wajskop qui a tout fait pour que ces Journées soient une réussite.

Celles-ci sont consacrées, cette année, à la psycholinguistique, à la perception de la parole et à sa reconnaissance.

Malheureusement, le programme a été quelque peu perturbé par la défaillance de dernière minute de Madame Chistovich de Léningrad.

Jeudi dernier, un télégramme nous apprenait qu'il lui était impossible de venir à Bruxelles.

Nous prions donc tous les participants de bien vouloir nous excuser pour les changements intervenus au programme.

Après tous ces rappels qui permettent de mieux situer nos Journées, il nous reste à souhaiter un franc succès à cette rencontre.

LANGUAGE AND PERCEPTION, A FEW OBSERVATIONS.

Summary.

Two experiments, one with audithory the other with visual presentations, are presented. From both it becomes likely that Ss strategies for grasping perceptually sentences involves usage of the surface structure, or some homologous structure. Another experiment on the location of cliks superposed on sentences suggests that it is base structure parameters that guide Ss strategies. Experiments as well as theoretical formulations are suggested to gather a greater understanding of language perception at the sentential level.

Résumé.

Deux expériences, la première à présentation visuelle et la seconde avec présentation auditive, sont décrites dans cet exposé. Leurs résultats indiquent que les stratégies des sujets pour appréhender perceptivement les phrases impliquent l'usage des structures de surface ou de structures d'un type homologue.

Une autre expérience sur la localisation de clics superposés sur des énoncés semble, par contre, donner priorité aux paramètres de la structure profonde. Des expériences et des formulations théoriques sont suggérées afin d'aboutir à une meilleure compréhension de la perception de la parole au niveau de la phrase.

Jacques MEHLER
C.N.R.S. - PARIS

Following the publication of Chomsky's "Syntactic Structures", psycholinguists became initially more interested in discovering linguistically based parameters that had "psychological reality", than in understanding the psychological mechanisms that make language possible. Although this approach may now seem somewhat arbitrary, its effects were not entirely negative since it led to a better understanding of psychological mechanisms than had been acquired in the preceding years.

Up to and including the early sixties, most students of linguistic behavior were concerned with areas such as the informational, probabilistic and associational structure of messages. However, some of the better known researchers, after having acquired a deeper understanding of the subject, demonstrated that informational content, probabilistic structure and associational values do not tell the whole story. In a series of experiments, G. A. Miller attempted to prove that articulation index scores were a function of the information contained in a message. Indeed, many of his experiments seemed to confirm that assertion. Nonetheless, Miller went on to prove, a few years later, that in language perception, syntax was at least as important as the parameters that had been observed in the past. He also demonstrated that although the latter cannot account for the former, notions such as grammaticality and acceptability are much more vast in that they include many of the aspects conveyed by information and probability structure.

Miller also proved that a five word sentence was perceived better when transmitted through a noisy channel if it was a grammatically correct English sentence like 'John has eaten the soup' than when it broke the rules of English as in 'soup the eaten has John'. Furthermore, the superiority of the good sentence persisted even when the Sŝ had learned all the words presented during an extensive pre-testing session. This was an excellent demonstration of the organizational value of syntax. However, what remained to be further clarified was the manner in which syntax organizes perceptual performance in the language user.

Mehler and Carey then tried to ascertain whether perceptual processes are organized in a way that relates to the surface structure of a sentence, or whether the more abstract underlying sentence structure is not also involved. However, if surface structure plays a role, and if the Ss are prepared to expect it, then any change in surface structure should lead to impoverished performance.

The same can be said for the base structure of sentences. In consequence, the first step was to establish that Ss can be led to expect a structure, either base or surface, under different lexical embodiments. This experiment having been performed without any major difficulty, it was possible to proceed with the intended experiment.

Surface structure is similar in many respects to the parsing of a sentence. Consider the following sentences:

(1) They are forecasting cyclones
(2) They are conflicting desires

While in (1), forecasting is a unit and forecasting cyclones is not, in (2) are conflicting is clearly not a unit, while conflicting desires clearly is. These differences in bracketing are often correlated with pronunciation and intonation cues. In many cases, these kind of surface structure features do not exhaust the linguistic knowledge we have of sentences. For instance, consider sentences (3) and (4):

(3) They are delightful to embrace
(4) They are hesitant to travel,

whose bracketings are essentially identical. However, while in (3) they is the direct object of embrace , in (4) they is the subject of travel. Since many of the differences between (3) and (4) are not represented in the bracketing structures, it is said that such differences are to be described at a deeper level, namely at the deep structure level.

In our experiment, we used forty sentences: ten with surface structures like that in (1), ten like that in (2). Another ten were equivalent in base structure to that in (3), and ten like that in (4). The syntactic uniformity of the four groups of sentences formed a set for the common structure. A syntactically different sentence followed each group of ten homogeneous sentences. For the group of sentences sharing the surface structure of 'They are forecasting cyclones', the test sentence was 'They are recurring mistakes'. 'They are describing events' was the test sentence of the ten sentences sharing the surface structure of 'They are conflicting desires'. The same procedure was followed in the case of sentences where the base structure was not the same. Ten sentences

like 'They are reluctant to consent', were followed by the test sentence 'They are delightful to embrace'. The sentences corresponding to the structure 'They are troublesome to employ' were followed by a test sentence like 'They are hesitant to travel'. In all cases, the test sentence also appeared as the tenth sentence in the set, inducing a sample corresponding to its structure. A sentence is thus a test sentence if it is preceded by ten structurally homogeneous sentences and is different in structure from them. The same sentence becomes a control sentence, when it is preceded by nine sentences all sharing the same structure as itself.

Sentences were recorded in a monotone fashion and calibrated for maximum deflection on the VU meter. Subjects had been informed that they would hear sentences mixed with noise. Their task was to write down as much of the sentences as they could recall. The results are presented in Table I.

The scores show that sentences that differ in surface structure are significantly easier to perceive in control position than in test position. These results demonstrate that a sentence which differs in surface structure from an expected structure is perceived significantly less accurately than when the expected and obtained surface structures match. These results are thus compatible with the view that surface structure plays an important role in the perception of simple sentences. On the other hand, with the base structure sentences, the scores show that the subjects responded to the sentences about equally well in both the control and the test positions. This indicates that the mismatch in the expected and received base structure is of little importance in perceptual performance. The results of these experiments thus tend to justify the existence of a perceptual performance model that refers to, and only to, sentence surface structure. Otherwise, we would postulate that a structure plays a role but that its change does not make any difference. It seems hard to conceive of such a possibility.

In another entirely unrelated experiment, a similar result was obtained. Using a visual mode of presentation of simple sentences, and eye movements as the dependent variable, the results of the experiment carried out by Mehler, Bever and Carey seemed to be compatible with those presented above.

Subjects were tested by recording their eye movements when reading simple English sentences displayed in their visual field. The sentences, typed in large

print, were presented at a distance of sixty centimeters. Eye movements were recorded with the Mackworth eye movement device. This device photographs a beam of light reflected from the subject's left eye, and as the eye moves, the angle of reflection changes and the reflected beam sweeps across a photographic field. Whenever the eye movement stops briefly for a fixation (of at least 1 second), the light beam stays on the same point in the photographic field and thus overexposes that part. The visual stimulus seen by the subject is superimposed on the film at the same time. Thus a response protocol consisted of a photograph of a sentence coupled with bright dots indicating where in sentence the subject had fixated for a short time.

Each subject was shown five sequentially presented sentences making up a short story. The fourth sentence was ambiguous but highly predictable as to its interpretation given the context of the short story. Four kinds of sentences were considered as experimental sentences. Sentences like (5) and (6)

(5) They gave her dog candies
(6) They told her cat stories

having two surface structure readings, depending on the interpretation given and sentences (7) and (8)

(7) The shooting of the hunters
(8) The punching of the sisters

varying according to interpretation in underlying structures. On the basis of these four types of sentences, eight contexts were prepared with an ambiguous sentence inserted as the fourth sentence in the story. In consequence, the question was whether eye movements were affected by the syntactic structure of the material that was being read, and, if the answer was affirmative, whether surface structure and deep structure played equally important roles in determining an eye fixation patters. Finally, depending on whether the preceding points could be clarified, a calculus predicting eye scans was attempted.

Forty Ss provided 123 records which were satisfactory. These allowed for a scoring to the nearest letter since the recording device did not allow for greater accuracy. The results of the experiment are presented in Table II.

In figure 1, the same information is presented graphically in such a way

that the significance is apparent immediately.

As can be seen, eye fixation patterns for sentences which imply two different bracketings for the same words, differ substantially when the reader changes from one type of parsing to another. However, in a sequence of words that can be interpreted two different ways, but has only one surface parsing, eye fixations are similar for readings of one and the other interpretation. These results would hence seem further to indicate that the perceptual processing of sentences is organized in relation to surface structure and that base structure parameters would hardly appear to be involved. However, a word of caution is warranted here as this conclusion may well be proved wrong in the long run. Nonetheless, empirical evidence still favors the above statement.

Our results and the analysis of them were made more relevant to the subject under disucssion thanks to the discovery of a simple calculus that is fairly accurate in predicting eye-fixations: assign to the first half of each surface structure constituent one fixation. If a sequence of letters initiates simultaneously 'n' constituents, assign 'n' fixations to it. Repeat this simple rule for all levels of the surface phrase structure including the rough morphological structure of the lexical items in the sentence. The fact that this simple rule predicts the eye fixation patterns gathered in our experiment reasonable well, indicates that the assumption that reading is organized according to surface structure constituents cannot be dismissed lightly.

Many other experiments in language perception also seem to indicate that surface structure constituents are the major organizational parameter in perceptual performance. There is, however, an area in which such a result may seem something of a paradox since it somehow implies the existence of a passive component of perceptual organization independent of sentence comprehension. However, it is fairly hard to conceive of an encoding stage entirely cut off from the comprehension stage. Thus, it may well be that perception and comprehension of sentences could be demonstrated to proceed in highly connected and dependent ways, if other techniques were used.

Bever, Fodor and Garret have carried out a number of studies related to this question. In a first experiment, Bever and Fodor showed that sentences with a click superimposed on them are heard with the click perceptually in a

different position from the one it objectively occupies. The experiment uses a simple technique. A sentence is recorded on one channel, while the click is placed on another channel on a stereo system tape. If it can be demonstrated that the subjective displacement of clicks is not due to some accoustical or phonetic phenomenon, then one plausible hypothesis may be that the displacements are due to the uninterruptibility of mental operations necessary to the perception of sentences. Thus, the clicks should be subjectively placed at some point in the sentence after the point where they really occurred, i.e., postponed. However, such is generally not the case. Reports indicate that there are as many postpositions as there are prepositions and there are some researchers who have found mainly prepositions. Alternative proposals must be formulated to account for this phenomena. One such study could be based on the mnemonic treatment of sentences. Even so, before raising any further questions, the Bever, Fodor and Garrett findings should be summarized.

These findings indicate that clicks are generally attracted by the major syntactic seizures of a sentence. When a click is objectively positioned within a seizure, it is subjectively heard in the seizure. Clicks on a list of nonsense words or an agrammatical succession of words, suffer fewer and less substantial subjective displacements, than clicks in sentences.

In an experiment, Bever et al., re-evaluated former interpretations in terms of some new data. Bever et al., now claim that the reported experiments demonstrate perceptual processing determined by base structure variables, even when these are poorly embodied in the surface representation. However, it is har to understand how a listener can infer the base structure of a sentence without also inferring the phonological representation, the immediate constituents, etc. Since the structure of the sentence must be constructed out of cumulative information, it is difficult to describe a mechanism that could directly extract say, the underlying sentences, without previously going through many other operations dealing with more superficial aspects of sentences.

There is always one major possibility, which has been more or less explored by Bever, Kimbal and others, and that is that the subject is constantly making hypotheses about the sentences he is going to hear. Hence on the basis of very little information, the subject makes a tentative analysis of the rest of the sentences he will hear and than compares the input with them. Of course, given the vast productivity of language, it is difficult to predict how the subject can do this so that the result of the experiment can be reliably accounted for.

Given all these complications, Mehler and O'Regan thought of carrying out an experiment that could account for the level at which click displacements occur. If the subject listens to a sentence that has a click superimposed on it, and if the subject knows either the base or the surface structure he is about to hear, then the displacement of the click should, in fact, be very reliable indeed. On the other hand, if the structural expectancies are not confirmed, greater shifts, or at least larger dispersions, should be observed.

In this experiment, we used the same materials as the ones used by Carey and Mehler and the sentences were only minimally changed to make the set more balanced. The groups of set-induced sentences were reduced from ten to six as we had learned from prior experience that the set induction was established with that number of sentences. Clicks were superimposed on the last sentence of the set induction group and the experimental sentence (which followed the set induction sentences). The subject's task was to locate the clicks as accurately as possible while writing down the sentence heard.

The results of this experiment were quite interestingly negative. In fact, as can be seen in the figures presented by sentence as well as by type of syntactic structure investigated, there were few, if any, differential effects. In fact, one may assume as we have, that the clicks are not displaced by any of the major syntactic aspects of the sentence structure. Thus, many other rigorous indices must be considered before one can accept the conclusion that the perceptual errors of sentence and interruption processing are due to the syntactic characteristics of the sentences.

Before concluding, it would be interesting to sketch at least one of the possible models of sentence perception that could be profitably tested in the coming years.

It is generally assumed that the listener is in an active state of search and inspection of the environment. If this is not so, then all models that assume that a stimulus is a stimulus under any circumstances, are surely going to miss the essential aspects of the problem.

The level of performance and the structural constraints that determine the subject's performance are both set by the structure of the language and by the contextual determinants of task adaptation. Hence, if the subject is in a

situation in which he has to respond to sentences, the kinds of heuristical procedures that he employs are suited to sentences, while if the subject is tuned to respond to phonemes, the former heuristics are less than ideally suited and another set of heuristics will take over. Whenever the subjects respond to the highest level of structure with their heuristics, the lower level heuristics are very difficult to measure since they are not used at all, or only partially.

TABLE I

NUMBER OF SENTENCES RIGHT AND WRONG

Sentence	Control	Test
They are recurring mistakes.		
Right	15	1
Wrong	7	22
They are describing events.		
Right	21	9
Wrong	2	13
They are reluctant to consent.		
Right	7	5
Wrong	16	17
They are troublesome to employ.		
Right	12	8
Wrong	10	15

Table 2. Number of Eye Fixations Summarized by Sentence Structure, Normalized to 10 Ss for Each Individual Sentence.

											Actual No. Ss Contributing Data	Actual No. Eye Fixations
		gave		(her		dog)		cand	ies			
		told		(her		cat)		stor	ies			
SS	(a)	14.3	.35	13.7	2.6	8.7	2.1.	10.3	9.8		24	73
		gave		her		(dog		cand	ies)		21	70
		told		her		(cat		stor	ies)			
SS	(b)	13.5	3.6	7.9	3.6	14.3	2.8	10.4	10.7			
		(ore		int	erest	ing)		farm	ers		19	54
		(are		sur	pris	ing)		auth	ors			
SS-DS	(c)	10.7	1.6	12.6	7.3	3.2	4.2	12.6	8.2			
		are		(int	erest	ing		farm	ers)		18	61
		are		(sur	pris	ing		auth	ors)			
SS-DS	(d)	9.8	2.8	8.3	13.9	5.3	2.7	15.6	9.1			
		(Did shooting)										
		(Did punching)										
		shoot	ing		of		the		hunt	ers	19	63
	(e)	punch	ing		of		the		sis	ers		
DS		14.7	5.8	2.2	6.9	4.1	9.5	2.2	14.3	6.4		
		(Got Shot)										
		(Got Punched)										
		shoot	ing		of		the		hunt	ers	22	64
	(f)	punch	ing		of		the		sis	ers		
DS		12.7	7.2	2.5	4.6	2.0	8.1	3.9	12.7	5.8		

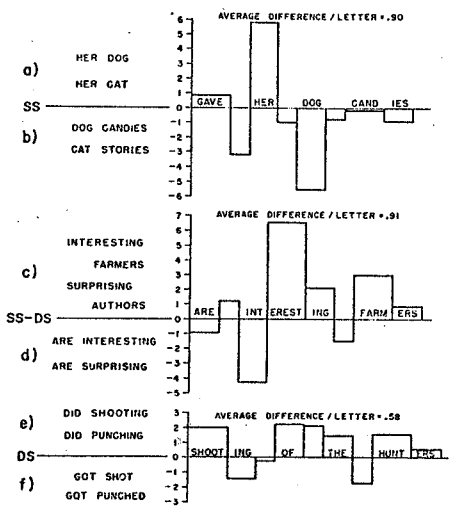


Fig. 1. Differences in eye-movement patterns correlated with differences in symbolic structure. See text.

OPTIMISATION EN CLASSIFICATION AUTOMATIQUE ET RECONNAISSANCE DES FORMES

Résumé

Les algorithmes actuellement opérationnels, consistant à fournir une "bonne partition" d'un ensemble fini, produisent des solutions dont rien ne permet d'affirmer qu'elles soient optimales. Le principal but de ce texte est une étude synthétique de propriétés d'optimalité dans des espaces formés de partitions d'un ensemble fini. On formalise et on prend pour modèle de cette étude, une famille de techniques particulièrement efficaces du type "nuées dynamiques". Après avoir développé l'aspect programmation, on illustre les différents résultats par un exemple artificiel et surtout par deux applications concrètes; l'une, en géologie minière pour la recherche de familles géographiques de sondages verticaux, l'autre en médecine pour le dépistage de profils biologiques permettant une aide au diagnostic.

Summary

Algorithms which are operationnally efficient and which give a good partition of a finite set, produce solutions that are not necessarily optimum. The main aim of this paper is a synthetical study of properties of optimality in spaces formed by partitions of a finite set. We formalize and take for a model of that study a family of particularily efficient technics of "clusters centers" type. The proposed algorithm operates on groups of points or "cores" (called "noyau"); these cores adapt and evolve into interesting clusters. Compared with other clustering algorithms, this algorithm requires less machine time and storage. After having developed the notion of "strong" and "weak" patterns, and the computer aspects, we illustrate the differents results by an artificial example and by two applications; one in mineral geology, the other in medicine to determine biological profiles.

E. DIDAY
I.R.I.A.

introduction

1.1 – LE PROBLEME

Dans différents domaines scientifiques (médecine, biologie, archéologie, économie etc.), il apparaît fréquemment de vastes ensembles d'objets représentés par un nombre fini de paramètres. Pour le spécialiste, l'obtention des groupements " naturels et homogènes ", ainsi que des éléments les " plus représentatifs " d'un tel ensemble constitue une étape importante dans la compréhension de ses données.

Une bonne approche de la solution de ce problème est fournie par les techniques de classification automatique qui consistent à trouver une partition d'un ensemble fini E telle que chaque objet ressemble plus aux objets intérieurs à son groupe, qu'aux objets extérieurs. En termes mathématiques, le problème peut s'énoncer sous l'une des deux formes suivantes ; étant donné un certain critère W :

A – Trouver la partition de E qui optimise W.

B – Trouver la partition de E qui optimise W parmi toutes les partitions en K classes.

La famille de méthodes dont il sera question concerne surtout le problème B, mais elle pourra également aider le praticien dans la résolution du problème C suivant :

C – **Chercher parmi toutes les partitions en K classes celles dont chaque classe aura le noyau le plus représentatif.** (Un noyau est un groupe de points de la population à classifier)*.

Dans le paragraphe 1.2 nous donnerons sommairement les principales propriétés des méthodes des nuées dynamiques**. Cette famille de méthodes servira de modèle au but véritable de cette étude qui sera développé dans le paragraphe 1.3.

1.2 – LES MÉTHODES DES NUÉES DYNAMIQUES

On se donne une fonction g permettant de transformer une partition de E en un ensemble fini de noyaux et une fonction f permettant le passage de plusieurs noyaux à une partition. Le principe de ces méthodes est simple, il consiste à appliquer de manière alternative les fonctions f et g à partir d'un choix initial de noyaux. Moyennant certaines hypothèses qui seront données, la décroissance du critère W est assurée jusqu'à la convergence. Le formalisme que nous donnons permet d'obtenir de nombreuses variantes de cette technique et notamment, comme cas particuliers, la méthode de HALL et BALL (1965) de FREEMAN (1969) et de DIDAY (1970). Nous avons pris cette famille de méthodes comme modèle de notre étude pour de multiples raisons.

a) Elles permettent d'éviter la mise en mémoire du tableau $\frac{N.(N-1)}{2}$ (où N = card(E)) des similarités des objets deux à deux. Cela permet le traitement de populations beaucoup plus importantes que par d'autres techniques plus classiques (SOKAL et SNEATH (1963), JOHNSON (1967), ROUX (1968), LERMAN (1970)).

---

* Le problème C est formalisé et un exemple simple est donné en 2.1.

** Voir également DIDAY E. (1970), DIDAY E. (1971).

b) Ces techniques sont très rapides, par exemple la variante étudiée dans DIDAY (1970), permet le traitement sur IBM 360/91 d'une population de 900 objets caractérisés chacun par 35 paramètres en 3 $\frac{1}{2}$ minutes.

c) Ces techniques ne souffrent pas de l'effet de chaîne (voir JOHNSON (1967), ZAHN (1971)). Autrement dit, elles n'ont pas tendance à rapprocher deux points éloignés si ces deux points sont liés par une file serrée de points.

d) Il n'est pas nécessaire de définir des seuils arbitraires pour la détermination des classes ni pour l'arrêt du processus, (voir SEBESTIEN (1966), BONNER (1964), HILL (1967) etc.).

L'utilisation des noyaux introduits de plus les avantages suivants :

a) La possibilité de prendre des noyaux de **plusieurs** éléments de la population permet une reconnaissance plus efficace des formes. On pourra voir, par exemple, dans le chapitre VI de DIDAY (1971), comment un bon choix de f et g revient à construire des noyaux épousant la structure des formes de E, ce qui ne serait évidemment pas possible si ces noyaux étaient réduits à un seul point. L'utilisation de noyaux permet également un vaste choix des fonctions f et g, par exemple l'utilisation de la distance de Mahalanobis (voir ROMEDER (1969)) qui n'aurait pas de sens si chaque noyau était réduit à un point unique.

b) En prenant pour noyaux des éléments de la population elle-même, plutôt que des centres de gravités, on évite l'effet artificiel que ces centres peuvent créer (cf. fig. 11). D'un autre côté, pour certains types de données, la notion de centre de gravité peut ne pas avoir de sens.

c) L'utilisation de noyaux permet la réalisation de partitions autour des agglomérations à forte densité en atténuant l'effet des points marginaux (cf. fig. 14 et fig. 15).

d) Signalons enfin, que l'utilisation des noyaux permet de donner des " optimum locaux " au problème C, permettant ainsi d'aider l'utilisateur intéressé par de bons échantillonnages.

1.3 – ETUDE SYNTHÉTIQUE DES SOLUTIONS OBTENUES

Toutes les techniques réalisables dont le but est de minimiser le critère W, fournissent des solutions dont rien ne prouve qu'elles soient optimales. Or, les différentes études faites récemment sur l'état actuel des recherches en " clustering " (voir BOLSHEV (1969), FISHER et VAN NESS (1971), BALL (1970), WATTANABÉ (1971), CORMACK (1971)) **font ressortir l'inexistence d'étude synthétique des solutions obtenues pour un algorithme donné. C'est à cette étude que ce texte est consacré.** Nous nous sommes restreint à un type particulier d'algorithme mais, évidemment, cette analyse pourrait s'étendre à d'autres techniques.

On appellera V_k l'ensemble des solutions possibles. Chaque solution* obtenue par un algorithme des nuées dynamiques est optimale vis-à-vis d'une certaine partie de V_k qui est une arborescence particulière. Cela conduit à donner une structuration à l'espace V_k. On montre en particulier que, sous certaines hypothèses, cet espace peut être partitionné en un nombre fini d'arborescences qui ont pour racine une solution stable dite "non-biaisée" et pour sommets pendants des éléments d'un certain type, appelés " éléments impasses ". On applique les différents résultats obtenus de la manière suivante :

a) On construit une variable aléatoire permettant de se faire une idée réelle de la structure de V_k . On obtient ainsi un invariant intéressant pour de multiples raisons, notamment pour les données évoluant dans le temps et pour comparer l'efficacité des différentes techniques.

b) On définit différents types de " fuzzy-sets "* dans W_k : les formes " fortes " et " faibles " ainsi que les points " charnières ". **Bien mieux que l'optimum global, ce sont à notre avis ces " fuzzy-sets " et les optimums locaux obtenus qui fourniront véritablement à l'utilisateur les différentes facettes de la réalité qu'il désire saisir.**

c) Nous donnons un nouveau type de techniques permettant, par passage d'une arborescence à l'autre, une approche de l'optimum global.

Les exemples d'application qui seront donnés font notamment ressortir l'intérêt des " formes fortes " qui sont un outil d'une grande utilité pour le praticien, en lui permettant d'extraire de sa population les groupes de points les plus significatifs.

Signalons enfin, que nous avons évité les développements théoriques, en nous restreignant aux résultats intéressants pour la compréhension et l'utilisation **informatique** des méthodes.

---

*. Voir ZADEH (1965) et RUSPINI (1970).

quelques notations et définitions

E : l'ensemble des objets à classifier, il sera supposé fini.

$\mathbb{P}(E)$: l'ensemble des parties de E.

$\mathbb{P}_k$: l'ensemble des partitions de E en un nombre $n \leqslant k$ de parties.

$\mathbb{L}_k \subset \{L = (A_1, \ldots, A_k) \ / \ A_i \subset A\}$ où, selon les cas A représentera E ou $\mathbb{R}^n$ par exemple.

$V_k = \mathbb{L}_k \times \mathbb{P}_k$.

W une application injective : $V_k \to \mathbb{R}^+$.

Un optimum local sur $C \subset V_k$ sera un élément v^*:

$W(v^*) = \underset{v \in C}{\text{Min}} \ W(v)$.

Si $C = V_k$ on a un optimum global.

Exemple 1 :

Soit $E = \{a, b, c, d, e, h\}$ 6 points du plan (*voir figure 1*).

W est défini comme suit : soit $v = (L, P)$ où $L = (x_1, x_2) \in \mathbb{L}_2 \equiv E^2$

et $P = (P_1, P_2) \in \mathbb{P}_2$ alors $W(v) = \sum_{i=1}^{2} \sum_{y \in P_i} d(x_i, y)$ où d est la distance Euclidienne.

On voit que dans ce cas, l'optimum global est donné par $v^* = (L^*, P^*)$ où $L^* = (b, d)$ et

$P^* = \{\{a, b, c\}, \{e, d, h\}\}$.

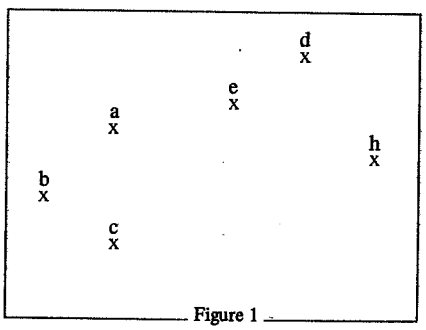


Figure 1

Exemple 2 :

Soit E l'ensemble des 17 points indiqués figure 2. Prenons

$\mathbb{L}_2 = \{ L = (A_1, A_2) / A_i \subset E, \text{card}(A_1) = 3, \text{card}(A_2) = 2 \}$ et $V_2 = \mathbb{L}_2 \times \mathbb{P}_2$

Choisissons $W(v) = \sum_{i=1}^{2} \sum_{x \in A_i} \sum_{y \in P_i} d(x, y)$ où d est encore la distance Euclidienne.

L'optimum global $v^* = (L^*, P^*)$ où $L^* = (A_1^*, A_2^*)$, est donné figure 3. Les traits pointillés indiquent les points de E qui constituent P_1^* et P_2^* ; les trois points indiqués par le signe $*$ forment A_1, le signe $\otimes$ sert à représenter les deux points qui constituent A_2.

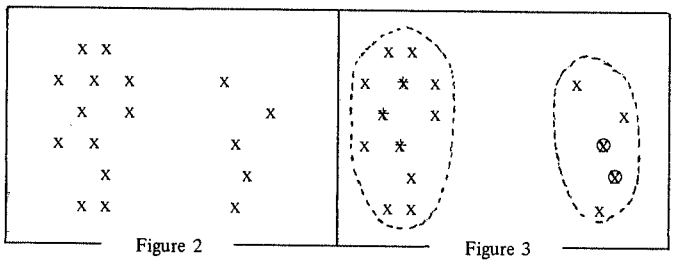


Figure 2

Figure 3

construction de triplets (f,g,w) par les nuées dynamiques

3.1 – FORMULATION GÉNÉRALE :

Nous noterons $v = (L, P) \in V_k$ où $L \in \mathbb{L}_k$: $L = (A_1, \ldots, A_k)$ avec $A_i \subset A$ et $P \in \mathbb{P}_k$: $P = (P_1, \ldots, P_k)$ où les P_i sont les classes de la partition P de E. On se donne également les quatre applications suivantes :

$D : E \times \mathbb{P}(E) \rightarrow \mathbb{R}^+$ qui dans la pratique exprimera la similarité d'un élément de E avec une partie de E.

$R : E \times T \times \mathbb{P}_k \rightarrow \mathbb{R}^+$ (où T est l'ensemble des entiers compris entre 1 et k). Cette application servira à agréger et à écarter les classes entre elles. On peut prendre par exemple $R(x, i, P) = D(x, P_i)$. On aurait pu également définir $R : E \times T \times V_k \rightarrow \mathbb{R}^+$ (cf. [9]) ; comme le montre l'exemple ci-dessous, cette définition de R peut rendre des services, cependant, afin de simplifier nous nous restreindrons dans tout ce texte à la première définition.

Exemple : $R(x, i, v) = \dfrac{D(x, P_i)}{(D(x, A_i))^{\frac{1}{n}}}$

Plus n est grand, moins les éléments des noyaux seront dispersés dans chacune des classes obtenues et plus ils exprimeront le squelette des formes qu'ils déterminent (cf. [10] chap. VI, un exemple de cas limite avec $n \rightarrow \infty$). Par un choix adéquat de n l'utilisateur pourra ainsi tenter d'obtenir des noyaux dont la distribution soit une bonne image de la distribution des classes qui leur correspond.

Le triplet (f, g, W) est construit comme suit :

$$W : V_k \rightarrow \mathbb{R}^+ : v = (L, P) \rightarrow W(v) = \sum_{i=1}^{k} \sum_{x \in A_i} R(x, i, P)$$

$f : \mathbb{L}_k \rightarrow \mathbb{P}_k : f(L) = P$ avec

$P_i = \{x \in E / D(x, A_i) \leqslant D(x, A_j) \text{ pour } j \neq i\}$, en cas d'égalité on affecte x à la partie de plus petit indice.

$g : P_k \rightarrow L_k$: $g(P) = L$ avec

A_i = les n_i éléments $a \in A$ qui minimisent $R(a, i, P)$. La valeur des n_i dépendra de la variante choisie (cf. 3.2).

Dans [9] nous avons pris l'habitude d'appeler A_i " noyau " de la $i^{ième}$ classe, et " étalons " les éléments qui le constituent.

Remarque : si $R : E \times T \times V_k \to \mathbb{L}_k$, il faut choisir $g : V_k \to \mathbb{L}_k$.

L'algorithme des nuées dynamiques consiste à appliquer alternativement la fonction f puis la fonction g sur le résultat obtenu et cela à partir de $L^{(o)} \in \mathbb{L}_k$ estimé ou tiré au hasard.

3.2 – LES DIFFÉRENTES VARIANTES ET INTÉRET COMPARÉ :

Nous ne prétendons pas exposer ici toutes les variantes possibles ; nous exposerons celles qui ont paru les plus intéressantes, en faisant simplement varier le choix de g et de R. (laissant au lecteur le loisir d'en imaginer d'autres !).

a) Pour cette variante on a : $A \equiv \mathbb{R}^n$, $n_i = 1 \; \forall i$;
si on prend de plus $R(x, i, P) = D(x, P_i)$, $g(P) = L$ est tel que A_i soit le centre de gravité* de P_i au sens de D.
HALL et BALL proposent une méthode de ce type dans [13].

b) $A \equiv E$ et $n_i = \text{card}\,(F_i)$ où :
$F_i = \left\{ x \in E \;/\; R(x, i, P) \leqslant R(x, j, P) \forall j \neq i \right\}$, si $i < j$ et $R(x, i, P) = R(x, j, P)$ on affecte x à F_i. On voit alors que les A_i sont identiques aux F_i et constituent une partition de E.
On trouvera une étude approfondie de ce cas dans [9] (qui est une généralisation de la méthode proposée par FREEMAN dans [12] où $\mathbb{L}_k \equiv \mathbb{P}$ et g est remplacé par f. Notons qu'une variante intéressante de cette méthode consisterait à choisir $\forall i \in \{1, 2, \ldots, k\}$ $n_i = \alpha \; \text{card}\,(F_i)$ avec $\alpha = \frac{1}{3}$ par exemple.

c) $A \equiv E$ et n_i fixé une fois pour toute $\forall i \in \{1, \ldots, k\}$; n_i sera choisi par l'utilisateur s'il a quelque idée du contenu de ses données, sinon il pourra prendre $n_i = \frac{\alpha \; \text{card}\, E}{k}$ pour tout i. (Voir [9] et [10]).

d) $A \equiv E$, n_i fixé ou égal à $\alpha \; \text{card}\, P_i$ avec $0 < \alpha < 1$; on définit A_i comme étant les n_i éléments de P_i qui minimisent $R(x, i, P)$. Quand n_i est fixé et dans le cas où le nombre d'étalons d'un noyau devient supérieur au nombre d'éléments de la classe correspondante, on prendra par exemple, $n_i = \text{card}\, P_i$ s'il s'agit de la classe P_i.

Remarque :
Dans le cas où l'utilisateur désire obtenir des classes empiétantes, il lui suffit de prendre $\alpha > 1$ dans les variantes b) et c).

---

* x est appelé centre de gravité de P_i au sens de D si $D(x, P_i) = \inf_{x \in \mathbb{R}^n} D(x, P_i)$.

On peut construire des méthodes mélangeant les variantes :
On pourrait ainsi commencer par une variante du type c) pour localiser les formes, puis terminer par une variante du type d) pour que les A_i donnent assurément les éléments les plus représentatifs de la classe P_i. **Dans tous les cas où l'utilisateur a besoin de définir des contraintes sur les noyaux, on choisira A de manière à ce que les éléments des noyaux satisfassent ces contraintes.**

Exemples :

1) faire une typologie d'un ensemble d'entreprises mais en imposant aux noyaux de n'être formé que d'entreprises modèles.

2) faire une typologie sur un ensemble de formes en imposant aux noyaux d'être pris parmi un ensemble de formes types.

Dans le cas où K = 1 on peut utiliser la variante suivante qui permet d'obtenir des étalons aux endroits à forte densité :

$L^{(o)} = A_1^{(o)}$ = n points tirés au hasard dans E.

$P^{(q)}$ = les m points de E les plus proches de $A_1^{(q-1)}$ au sens de D avec m > n (par exemple m = n + 1).

$A_1^{(q)}$ = les n points de E les plus proches de $P^{(q)}$ au sens de R.

Cette technique peut donner à l'utilisateur une idée à priori du nombre de classes de E.

Comparaison des différentes variantes : par rapport à la variante a), les variantes b), c) et d) ont l'avantage d'atténuer l'effet artificiel créé par des centres de gravité en utilisant des noyaux d'éléments de la population elle-même.

La variante b) a l'avantage par rapport à c) de ne pas nécessiter l'introduction des paramètres n_i, cependant elle a une forte tendance à osciller au lieu de converger, elle donne plus d'importance aux éléments marginaux puisque ses noyaux recouvrent E, alors que les noyaux de la variante c) ne tiennent compte que des éléments les plus représentatifs ; de plus, elle nécessite beaucoup plus de calculs et de place mémoire qu'une utilisation de c) avec $\Sigma\, n_i << \text{card } E$.

La variante d) permet d'assurer la représentativité des noyaux vis-à-vis de leur classe et réduit les calculs et la place mémoire ; cependant le choix des noyaux étant moins vaste à chaque itération (puisqu'on astreint les étalons à n'appartenir qu'à l'une des classes précédentes) elle peut donner des classes moins pertinentes que pour la variante c).

3.3 – CONSTRUCTION DE TRIPLETS RENDANT LA SUITE u_n DÉCROISSANTE

Définition des suites u_n et v_n :

Soit h l'application $V_k \to V_k$ telle que $v = (L, P) \in V_k$
$\Rightarrow h(v) = (g(P), f(g(P)))$.

Une suite $\{v_n\}$ est définie par v_o et $v_{n+1} = h(v_n)$.
Une suite $\{u_n\}$ est définie à partir d'une suite $\{v_n\}$ par $u_n = W(v_n)$.

Définition de S :

Soit $S : \mathbb{L}_k \times \mathbb{L}_k \to \mathbb{R}^+$:

$S(L, M) = \sum_{i=1}^{k} \sum_{x \in A_i} R(x, i, Q)$ où $Q = f(M)$.

Définition d'une fonction carrée*

On dira que R est carrée si :
$S(L, M) \leqslant S(M, M) \Rightarrow S(L, L) \leqslant S(L, M)$.

Théorème 1 :

Si R est carrée le triplet (f, g, W) rend la suite u_n décroissante pour les variantes où le nombre d'étalons par noyau est fixé.

Démonstration :

Etant donnée la suite $v_n = (L^{(n)}, P^{(n)})$ on a :

$W(v_n) = \sum_{i=1}^{k} \sum_{x \in A_i^{(n)}} R(x, i, P^{(n)})$ où $L^{(n)} = (A_1^{(n)}, \ldots, A_k^{(n)})$

d'où : $u_n = W(v_n) = S(L^{(n)}, L^{(n)})$.

Posons $z_n = S(L^{(n+1)}, L^{(n)})$.

Si R est carrée et si $z_n \leqslant u_n$ on a nécessairement :

$S(L^{(n+1)}, L^{(n)}) \leqslant S(L^{(n)}, L^{(n)}) \Rightarrow S(L^{(n+1)}, L^{(n+1)}) \leqslant S(L^{(n+1)}, L^{(n)}) \Rightarrow u_{n+1} \leqslant z_n$.

Montrons que $z_n \leqslant u_n$; en effet :

$u_n = \sum_{i=1}^{k} \sum_{x \in A_i^{(n)}} R(x, i, P^{(n)}) \geqslant \sum_{i=1}^{k} \sum_{x \in A_i^{(n+1)}} R(x, i, P^{(n)}) = z_n$ par construction même de $A_i^{(n+1)}$.

On voit ici l'intérêt de fixer le nombre d'étalons par noyau, car cette dernière inégalité n'est pas nécessairement vérifiée dans le cas de la variante b).

On a finalement montré que : $u_{n+1} \leqslant z_n$ et $z_n \leqslant u_n$ d'où $u_{n+1} \leqslant u_n$.

c. q. f. d.

N. B. Dans toute la suite on se restreindra au cas où le nombre d'étalons par noyau est fixé.

---

* Nous avons exhibé dans [9] un exemple de fonction R carrée.

structuration de l_K , p_K , v_K et propriétés d'optimalité

Considérons le graphe $\Gamma = (V_k, h)$. Il apparaît alors, des éléments particuliers dans V_k :

1) **Les éléments non biaisés**

Les propriétés suivantes sont équivalentes et caractérisent un élément non biaisé *

$v = (L, P) \in V_k$.

a) v est racine d'une arborescence bouclée de Γ.

b) v est un point fixe de h.

c) $L = g(P)$, $f(L) = P$.

Les propriétés d) (resp. e)), permettent de caractériser les éléments non biaisés de $\mathbb{L}_k$ (resp. $\mathbb{P}_k$).

d) $g(f(L)) = L$.

e) $f(g(P)) = P$.

2) **Les éléments impasses**

Les propriétés a) et b) suivantes sont équivalentes et caractérisent un élément impasse

$v = (L, P) \in V_k$.

a) v est un sommet pendant de Γ.

b) $P \neq f(L)$ ou $f^{-1}(g^{-1}(L)) = \Phi$

Signalons que les propriétés c (resp. d) suivantes, permettent de caractériser les éléments impasses de $\mathbb{L}_k$ (resp. $\mathbb{P}_k$).

c) $g^{-1}(L) = \Phi$ ou $f^{-1}(g^{-1}(L)) = \Phi$.

d) $f^{-1}(P) = \Phi$ ou $g^{-1}(f^{-1}(P)) = \Phi$ ou $f^{-1}(g^{-1}(f^{-1}(P))) = \Phi$.

Le théorème suivant se déduit immédiatement des définitions de la proposition 2. (cf. annexe 1) et du théorème 1.

---

* Cette appelation vient du fait que les noyaux correspondants à un tel élément sont au centre (au sens de g) de la classe qu'ils déterminent au sens de f.

Théorème 2 :

Si R est carrée alors :

a) Chaque composante connexe de $\Gamma = (V_k, h)$ est une arborescence bouclée.

b) Il existe dans V_k au moins un élément non-biaisé.

c) Si un élément non-biaisé $v \in V_k$ est sommet d'une arborescence C, alors v est un optimum local* vis-à-vis de l'ensemble des sommets de C.

d) Si $w \in V_k$ n'est pas un élément non-biaisé alors w appartient à une arborescence bouclée de racine w^*, et $W(w) > W(w^*)$. Et l'optimum global est un élément non-biaisé.

Remarque :

On a deux énoncés équivalents de ce théorème, en remplaçant partout V_k par $\mathbb{L}_k$, puis par $\mathbb{P}_k$. Il suffit pour cela d'utiliser les fonctions $\varphi_1 : V_k \to \mathbb{L}_k$ et $\varphi_2 : V_k \to \mathbb{P}_k$ telles que :

si $v = (L, P)$ alors $\varphi_1(v) = L$ et $\varphi_2(v) = P$.

On voit d'après ce théorème que dans le cas où R est carrée, il existe trois types d'éléments dans V_k (et de même dans $\mathbb{L}_k$ et $\mathbb{P}_k$) : les éléments impasses, les éléments non-biaisés et les éléments restants qui seront dit " biaisés ".

Si R n'est pas carrée, on est dans le cas de la proposition 1** ; il y a des circuits dans V_k (et de même dans $\mathbb{L}_k$ et $\mathbb{P}_k$) ; autrement dit on peut trouver des suites $\{v_n\}$ pour lesquelles il existe $N > 1$ tel que $v_o = v_N$. Les éléments de ces circuits constituent donc un quatrième type d'éléments de V_k, $\mathbb{L}_k$ ou $\mathbb{P}_k$.

Exemple de différents types d'éléments de $\mathbb{L}_k$:

Reportons nous au cas de la figure 1 et prenons le triplet (f, g, W) des nuées dynamiques (cf. 3.1), avec $k = 2$, $n_1 = n_2 = 1$, $A \equiv E \equiv \{a, b, c, d, e, h\}$, $\mathbb{L}_2 \equiv A \times A$, $D(x, Y) = \sum_{y \in Y} d(x, y)$ où d est la distance euclidienne et $R(x, i, P) = D(x, P_i)$.

---

* On peut montrer que v,est un optimum local pour une certaine topologie basée sur la différence symétrique. On peut montrer également que l'algorithme de Mac Queen [28] ne converge pas nécessairement vers une solution non-biaisée.

** Cf. Annexe 1.

Eléments impasses :

Il n'existe pas deux points de E permettant d'engendrer à l'aide de f la partition $P = (P_1, P_2)$ où :
$P_1 = \{a, d, h\}$ et $P_2 = \{b, c, e\}$. Ainsi $f^{-1}(P) = \Phi$, comme $L = (e, c) = g(P)$ on peut dire que L est un élément impasse. Un autre exemple d'élément impasse est le point $L = (b, h)$ car $g^{-1}(L) = \Phi$.

Elément non-biaisé :

$L = (b, d) \in \mathbb{L}_2$ est un élément non-biaisé car on voit simplement d'une part que $f(L) = P = (P_1, P_2)$ avec $P_1 = \{a, b, c\}$ et $P_2 = \{d, e, h\}$ et d'autre part que $g(P) = L = (b, d)$.

Elément biaisé :

$L = (c, e)$ est biaisé car $g(f(L)) = (b, d) \neq L$.
Nous avons utilisé l'appelation " élément biaisé " car un tel élément (c, e) n'est pas le plus proche de la partition qu'il engendre, contrairement à (b, d) qui est non-biaisé car il vérifie cette propriété.

recherche d'invariants

5.1 – MESURE DES ARBORESCENCES :

On supposera d'abord que le triplet (f, g, W) rend u_n décroissant $\forall u_o$, (autrement dit que $\forall x \in V_k$, W (h (x)) < W (x)). L'espace probabilisé (Ω, $\mathcal{A}$, P) des familles d'arborescences bouclées est définit comme suit :

$\Omega = V_k$. $\mathcal{A}$ = l'algèbre engendrée par la partition de Ω en arborescences bouclées (c.à.d. ensemble des parties de Ω qui sont réunions d'arborescences bouclées).
P : $\mathcal{A} \rightarrow$ [0, 1] est telle que si $C \in \mathcal{A}$ est réunion de n arborescences bouclées $C_1, \ldots, C_n$ alors $P(C) = \frac{1}{\text{card } (\Omega)} \sum_{i=1}^{n} \text{card } C_i$.

La variable aléatoire X (dite des familles d'arborescences) de (Ω, $\mathcal{A}$ P) dans (IR, B) où B est la tribu borélienne, est l'application $\Omega \rightarrow$ IR telle que $X(\vec{v}) = W(w)$ où w est l'élément non biaisé de l'arborescence bouclée contenant v. X est bien une variable aléatoire car si $I \in B$, X^{-1} (I) est la réunion d'arborescences de V_k ayant pour sommets les éléments v tels que $X(v) \in I$. La fonction de répartition $F(x) = pr [X < x]$ exprime la probabilité d'obtenir un élément $v \in V_k$ dans une arborescence bouclée ou une boucle contenant un élément non-biaisé w tel que $W(w) < x$. On donnera en 7.1 un exemple de fonction de répartition empirique correspondant à un n-échantillon de V_k. Dans le cas où $\exists x : W(f(x)) > W(x)$ (autrement dit, si on ne suppose plus la suite u_n décroissante $\forall u_o$), on peut également définir une variable aléatoire des composantes connexes de V_k. La variable aléatoire de (V_k, $\mathcal{A}$, P) dans (IR, B) est telle que $X(v) = \inf_{y \in C} W(y)$ où C est la partie connexe de V_k à laquelle appartient v.

L'introduction de ces variables aléatoires permet de se faire une idée du nombre de composantes connexes et de leur taille respective, grâce aux fonctions de répartitions empiriques. Cela donne également un outil de comparaison des différentes techniques, la meilleure étant celle pour laquelle les racines des arborescences de plus grande taille correspondent aux plus petites valeurs prises par W, (voir 7.1).

5.2 – FORMES FORTES, FUZZY-SETS ET INFORMATION :

5.2.1 Caractérisation des différents types de formes :

Soient $C_1, \ldots, C_n$ parties connexes du graphe (V_k, h) et

$C = C_1 \times C_2 \times \ldots \times C_n$; on définit l'application $Z : C \to \mathbb{R}^+$ par
$Z(V) = W(v_1) + \ldots + W(v_n)$ où : $V = (v_1, \ldots, v_n) \in C$ et $v_i \in C_i$.
Soit $V^\star$: $Z(V^\star) = \underset{V \in C}{\text{Min}}\ Z(V)$. Soit $V^\star = (v_1^\star, \ldots, v_n^\star)$ et $v_i^\star = (L^{i\star}, P^{i\star})$.

(Si R est carrée, C_i est une arborescence bouclée ou une boucle et $v_i^\star$ est l'élément non-biaisé de C_i). Notons P_j^i la $j^{\text{ième}}$ classe de **la partition P^i**.

Soit H l'application $E \to \mathbb{N}^n$ qui, à chaque élément $x \in E$ fait correspondre le vecteur $(\alpha_1, \ldots, \alpha_n)$ où α_i est le numéro de la classe ou apparaît l'élément α dans $P_i^\star$. Soit $H(y) = (\beta_1, \ldots, \beta_n)$ et $\delta\ (x,, y)$ le nombre d'indices i pour $i = 1, 2, \ldots, n$ tels que $x_i - y_i = 0$.
Soient F_n et F_1 deux applications multivoques définies sur E telles que
$F_n(x) = \{y \in E / \delta\ (x, y) = n\}$ et $F_1(x) = \{y \in E / \delta\ (x, y) \geqslant 1\}$.

Définition des formes fortes * :

Les propriétés suivantes sont équivalentes et caractérisent la partition $P^\star$ de E dont chaque classe est une forme forte.

1) $P^\star = P^{1\star} \cap P^{2\star} \cap \ldots \cap P^{n\star}$
2) $P^\star$ est la moins fine** des partitions qui sont plus fines que $P^{1\star}, \ldots, P^{n\star}$
3) $P^\star$ est la partition définie par l'espace quotient E/H.
4) $P^\star$ est la partition définie par les parties connexes du graphe $\Gamma_n = (E, F_n)$.

Définition des formes faibles :

Les propriétés suivantes sont équivalentes*** et caractérisent la partition $Q^\star$ de E dont chaque classe est une forme faible.

1) $Q^\star$ est la plus fine des partitions qui sont moins fines que $P^{1\star}, \ldots, P^{n\star}$.
2) $Q^\star$ est la partition définie par l'ensemble des parties connexes du graphe $\Gamma_1 = (E, F_1)$.

Plus généralement, si on pose $F_p\ (x) = \{y \in E / \delta\ (x, y) \geqslant p\}$ et

$\Gamma_p = (E, F_p)$, l'ensemble des parties connexes de Γ_p pour $p = 0, 1, 2, \ldots, n$ constitue une hiérarchie. Cette hiérarchie induit l'ultramétrique sous dominante de la différence symétrique (cf. annexe 3)

Remarque :

On voit, d'après ces définitions que $P^\star$ est une partition plus fine que $Q^\star$.

Définitions des points charnières et des points isolés

On les caractérise par le fait que ce sont les formes fortes réduites à un seul point. Ils se distinguent par la propriété suivante :

---

* L'intersection de deux partitions est l'ensemble des parties obtenues en prenant l'intersection de chaque classe de l'une par toutes les classes de l'autre.
** Une partition P est dite plus fine qu'une partition P' de E si toute classe de P est union de classe de P'.
*** Pour la démonstration de cette équivalence cf. annexe 2.

un point $a \in E$ est isolé si $\delta(a, x) = o \ \forall x \in E$.
un point $a \in E$ est charnière si $\exists x \in E : o < \delta(a, x) < n$.

5.2.2. Fuzzy sets*

L'intérêt des "fuzzy sets" de Zadeh que nous introduisons ici est qu'ils permettent :

a) d'obtenir de nouvelles formes à la suite d'opérations ensemblistes sur les formes fortes (réunion, intersection etc.)

b) de caractériser ces nouvelles formes sans avoir besoin de définir des profils types (par des calculs de moyenne par exemple) ni même de connaître les éléments qui les constituent

c) d'utiliser au maximum l'information apportée par le tableau des formes fortes.

Chaque forme forte A peut être considérée comme un "fuzzy-set" caractérisé par l'application $h_A : E \to \{0, 1\}$ telle que $h_A(x) = \frac{\delta(x, a)}{n}$ où $a \in A$. On voit d'après la définition (3° propriété) que $h_A(a) = 1 \ \forall a \in A$. On peut utiliser h_A pour avoir une idée du degré de ressemblance avec A d'un point charnière ou d'une autre forme forte. On peut également utiliser l'application $F : \mathcal{F} \to \{0, 1\}$ où $\mathcal{F}$ est l'ensemble des formes faibles de E et $F(B) = \frac{1}{\text{card}(B)} \sum_{x \in B} \left(\frac{1}{m} \sum_{j=1}^{m} h_{Aj}(x) \right)$

où les Aj sont les m formes fortes qui constituent B. Cette application F exprime le degré de faiblesse de B car plus les formes fortes Aj sont dissemblables plus F(B) sera petit. La plus grande valeur de F est 1, c'est-à-dire quand B est une forme forte.

5.2.3 Stabilité des formes fortes et information

Si le nombre de classes demandées est K et si n est le nombre d'optimum locaux obtenus, il est clair que $\forall x \in E$, H(x) peut prendre $(k)^n$ valeurs, cela souligne la cohésion des éléments d'une forme forte A puisque si x et $y \in A$ on a $H(x) = H(y)$. Cependant l'utilisateur, désireux d'avoir une assurance supplémentaire en ce qui concerne la cohésion et la stabilité des formes fortes, peut utiliser l'information apportée par l'augmentation de n. Considérons les classes $P_1^{q\star}, \ldots, P_k^{q\star}$ de la partition $P_q^\star$, et soient $A_1, \ldots, A_m$ les formes fortes obtenues pour $n = q - 1$; soit $P(j/i) = \frac{1}{\text{card } A_i} \text{card}(A_i \cap P_j^{q\star})$, cette quantité exprime la probabilité pour un élément d'être dans $P_j^{q\star}$ sachant qu'il est dans A_i.

On peut maintenant mesurer l'information apportée par $P_q^\star$ connaissant $A_1, \ldots, A_m$:

$$I(P^{q\star}/P^{1\star}, \ldots, P^{q-1\star}) = - \sum_{i=1}^{m} \sum_{j=1}^{k} P(j/i) \log_k P(j/i).$$

---

* Voir [20] et [26].

Si la partition des formes fortes $A_1, \ldots, A_m$ est plus fine que la partition $P^{q\star}$ l'information apportée par $P^{q\star}$ est nulle puisque $P(j/i) = 1$ ou 0. Si chacune des formes fortes A_i est répartie de manière égale dans chaque classe $P_j^{q\star}$ pour $j = 1, \ldots, k$ alors l'information apportée est maximum et vaut 1.

L'invariance des formes fortes est donc assurée pour $n = q$ si $\forall n > q$ on a $I(P^{n\star}/P^{1\star}, \ldots, P^{n-1\star}) = 0$. Ainsi, sur les données de RUSPINI (cf. 7.1) on s'aperçoit qu'au-delà de $n = 4$ l'information nouvelle reste généralement nulle.

Signalons, enfin, que le nombre $J(k) = \sum_{q=1}^{n} I(P^{q\star}/P^{1\star}, \ldots, P^{q-1\star})$ peut donner une idée sur la valeur du choix du nombre de classes k demandé ; la plus petite valeur de J(k) correspondant au meilleur choix de k.

5.3 – OPTIMUM GLOBAL DE V_k.

Théorème 3 :

Si les hypothèses suivantes sont vérifiées :

1) R est carrée.
2) $I(P^{n\star}/P^{1\star}, \ldots, P^{n-1\star}) = 0 \; \forall n : q < n < N$ où N est le nombre d'arborescences bouclées.

   Alors, la partition des formes fortes $P^{1\star} \cap \ldots \cap P^{q\star}$ est plus fine que la partition correspondant à l'optimum global $v^\star$ de V_k.

Démonstration :

L'optimum global $v^\star$ est racine d'une arborescence bouclée ou d'une boucle puisque d'après 1) on peut utiliser le théorème 2.
Soit $P^{i\star}$ la partition correspondant à $v^\star$ si $i \leqslant q$, $P^{i\star}$ est moins fine que la partition des formes fortes $P^\star$ car $P^\star = P^{1\star} \cap \ldots \cap P^{q\star}$. Si $i > q$ l'information apportée par $P^{i\star}$ est nulle et donc $P^{i\star}$ est moins fine que $P^\star$.

c. q. f. d.

Ainsi, sous les hypothèses de ce théorème, chaque classe de la partition correspondant à l'optimum global est une réunion de formes fortes qui doivent être proches. Pour représenter cette proximité on peut par exemple utiliser une analyse factorielle du triple ou un " minimum spaning tree " (cf. [23]), sur le tableau $T(i, j) = h_{A_i}(a_j)$ où A_i est la $i^{ième}$ forme forte et a_j un élément de la $j^{ième}$ forme forte ou encore la méthode des connexités descendantes (cf. annexe 3). Cette méthode permet (sous les hypothèses du théorème 3) une bonne approche de l'optimum global.
Remarquons que la 2ème hypothèse est vérifiée pour q d'autant plus petit qu'il existe effectivement k formes dans la population E.

5.4 – APPROCHE DE L'OPTIMUM GLOBAL PAR CHANGEMENT D'ARBORESCENCES

Il s'agit de construire à l'aide de deux éléments non biaisés v_1 et v_2 de V_k un troisième élément non biaisé v_3 qui améliore le critère. Nous supposerons R carrée.

Nous noterons $v^i = (L^i, P^i) \in V_k$ avec $L^i = (L^i_1, \ldots, L^i_k)$ et $P^i = (P^i_1, \ldots, P^i_k)$; $v^i_j = (L^i_j, P^i_j)$.

Nous supposons que W est additive (ce qui est souvent le cas dans la pratique) autrement dit, qu'il existe une application $z : \mathbb{P}(E) \times \mathbb{P}(E) \to \mathbb{R}^+$ telle que :
$W(v^i) = \sum_{j=1}^{k} z(v^i_j)$.

Supposons que v^1 et v^2 soient deux solutions non biaisées et soit $\{v^i_{j_1}, \ldots, v^i_{j_k}\}$, les k plus petites valeurs prises par $z(x)$ avec $x \in \{v^i_j / i = 1, 2 \text{ et } j = 1, 2, \ldots k\}$. Notons $P = (P^i_{j_1}, \ldots, P^i_{j_k})$ et $L = (L^i_{j_1}, \ldots, L^i_{j_k})$. On montre alors facilement la proposition suivante.

Proposition :

Si $L \neq L^1$, $L \neq L^2$ et $P \in \mathbb{P}_k$ alors l'arborescence bouclée contenant $v = (L, P)$ à pour racine un élément non biaisé v_3 : $W(v_3) < \inf (W(v_1), W(v_2))$.

programmation du tableau des formes fortes et interprétation heuristique

En ce qui concerne la programmation des méthodes des nuées dynamiques, de larges développements pourront être trouvés dans [8], [9] , [10]. Nous signalerons donc simplement l'apport qui a été fait depuis, par la sortie automatique des formes "fortes" et "faibles". Le programme donne en sortie un tableau (dit des formes fortes*) représentant les différents types de formes ; la première colonne de ce tableau (cf. tableau 1) donne le nom de chaque élément de E dans un ordre tel qu'à la suite de chaque élément x apparaît l'élément y rendant $\delta(x, y)$ minimum**. On trouve sur chaque ligne le nom de l'élément x suivi des valeurs $\alpha_1, \ldots, \alpha_n$ qui décrivent le vecteur H(x). La valeur $\Delta(x, y) = n-\delta(x, y)$ correspondant à deux éléments consécutifs x, y est donnée en dernière colonne et permet une détection aisée des formes fortes et faibles : si $\Delta(x, y) = n$ cela signifie que x est le dernier élément d'une forme faible (c.à.d. d'une partie connexe de Γ_1) ; si $\Delta(x, y) = 0$ cela signifie que x et y font partie de la même forme forte ; les files de 0 dans cette dernière colonne caractérisent donc les formes fortes.

En ce qui concerne l'interprétation du tableau des formes fortes , nous ferons les remarques suivantes :

a) soit m le nombre total de tirages effectués, m_i le nombre d'apparitions de la $i^{ième}$ solution v^i (d'où $\sum_i m_i = m$) et C_i l'arborescence bouclée ayant v^i pour racine ; si m est suffisament grand on peut considérer $\frac{m_i}{m}$ # Prob $(x = W(v_i)) = \frac{\text{card } C_i}{\text{card } V_k}$; si m_i est grand, on peut donc considérer que v_i est racine d'une arborescence de taille importante (card C_i grand) ; d'après le théorème 2, v^i est donc une solution particulièrement significative, puisque c'est un optimum local pour une grande partie de V_k.

b) Si q est le nombre de solutions obtenues, et si parmi ces solutions $v^* \in V_k$ est la solution qui minimise W, v^* est un optimum local vis-à-vis de l'ensemble des sommets des q arborescences bouclées obtenues.

c) D'après le théorème 2, les arborescences bouclées et boucles forment une partition de V_k, en conséquence plus le nombre de solutions obtenues est faible plus la taille de ces arborescences est grande et plus les solutions obtenues sont donc significatives. Dans le cas

---

* Par opposition au tableau des formes faibles qui aurait en général une structure différente. Signalons toutefois qu'il est possible de construire un tableau respectant simultanément la structure en formes fortes et en formes faibles.

** Nous revenons ici aux notations données en 5.2.1.

où il n'y a pas les formes le nombre d'arborescences bouclées est important et donc les solutions obtenues sont moins significatives.

d) Soit v^* la racine d'une arborescence bouclée C de $\Gamma = (V_k, h)$; **soit $h^{-q}(v^*)$ le $q^{ième}$ niveau de C.** Il s'avère dans la pratique que le nombre de niveaux pour une arborescence donnée est très faible : il oscille en général autour de 4 ou 5 et dépasse rarement 12, même pour des tableaux comportant 3.000 éléments à classifier. Dans le cas où il y a peu d'arborescences bouclées, le nombre de sommets d'un niveau donné est donc très grand.

Exemple :

Supposons que card(E) = 100, k = 15, $n_i = 3$, le nombre de niveaux = 5, le nombre de solutions non-biaisées = 6, R est carrée. Alors la taille d'un palier est supérieure à

$$\frac{2^{100}(C_{100}^3)^{15}}{5 \times 6} \ !$$

On a ainsi une explication de la rapidité et de l'efficacité de la méthode, qui à chaque itération permet de passer d'un niveau à l'autre en améliorant la solution.

exemples d'applications

7.1 – L'EXEMPLE ARTIFICIEL DE RUSPINI

Nous avons appliqué la variante c) sur les données de Ruspini (cf. Fig. 6). Cela a permis d'abord de constater la rapidité de la méthode, par rapport à celle de Ruspini ; ainsi en prenant $K = 4$, $n_1 = n_2 = n_3 = n_4 = 5$, $R(x, i, L) = D(x, C_i) = \sum_{y \in C_i} d(x, y)$ où d est la distance Euclidienne, nous avons réalisé 50 passages de la méthode (en changeant à chaque fois le tirage de $L^{(0)}$) en 2, 57 mn. sur CII 10 070. Ces 50 passages ont fait ressortir l'existence de 6 arborescences bouclées. Les fréquences d'apparition de chacune des 6 solutions correspondantes sont indiquées figure 12. Ce graphique est en fait l'histogramme de la variable aléatoire qui a été définie en 5.1. En abcisse est représenté $U = \text{Lim } U_n$ (cf. 3.3) ; la convergence est généralement atteinte au bout de 4 itérations. La solution qui apparaît le plus fréquement est celle correspondant aux quatre meilleurs classes ; la valeur de U pour cette solution est nettement meilleure que pour les autres solutions, ce qui montre qu'elle correspond bien à la meilleure partition. La meilleure solution correspond à la racine de l'arborescence bouclée de plus grande taille, ce qui est satisfaisant pour la méthode. Les solutions qui apparaissent le plus fréquemment sont indiquées figures 7, 8, 9, 10. On voit facilement que les solutions correspondant aux figures 9, 10 et 11 n'apportent aucune information (cf. 5.2) à la solution donnée figure 7. A partir de ces solutions on obtient 4 "formes fortes" correspondant exactement aux quatre classes de la meilleure solution.

Remarquons qu'en appliquant la proposition 3 aux solutions données figure 8 et 9, on fait apparaître l'arborescence bouclée dont la racine est la solution correspondant à la figure 7. On donne figure 11 une solution obtenue en utilisant la variante du centre de gravité (cf. 3.2.a)) ; cette solution n'apparaît jamais par les variantes utilisant des noyaux car elle doit correspondre à une position instable.

On donne (tabl. 1), le tableau des formes fortes, obtenu en prenant cette fois $K = 6$, $n_5 = n_6 = 5$, sans changer les autres paramètres et en réalisant 5 passages de la méthode ($n = 5$). Ce tableau fait ressortir l'existence de 6 formes fortes et 3 formes faibles. Notons, B_1, B_2, B_3 les formes faibles et A_i les formes fortes (cf. Fig. 13). On peut mesurer la "faiblesse" de B_i en utilisant la fonction F (cf. 5.2.2.). Comme

$\sum_{j=1}^{3} h_{A_j}(x) = 1 + \frac{4}{5} + \frac{4}{5}$ $\forall x \in B_1$, $\sum_{j=4}^{5} h_{A_j}(x) = 1 + \frac{1}{5}$ $\forall x \in B_2$ et $h_{A_6}(x) = 1$ $\forall x \in B_3$

on a : $F(B_1) = \frac{13}{15}$, $F(B_2) = \frac{3}{5}$ et $F(B_3) = 1$.

On voit que B_3 est une forme forte, que B_1 est presque une forme forte et que B_2 est une forme relativement faible. Ces formes fortes et faibles telles qu'elles apparaissent Fig. 13 expriment bien ces valeurs. Signalons enfin que dans 4 des 5 solutions il apparaît des classes vides ce qui signifie que le nombre de classes existant réellement doit être plus petit que 6.

7.2 – CLASSEMENT DE SONDAGES D'UN GISEMENT MINIER

L'étude dont il est question a été réalisée par J. Picard pour sa thèse de 3e cycle (cf. [17]). Elle porte sur 149 sondages géologiques. Chaque sondage est caractérisé par 24 teneurs métal mesurées de mètre en mètre sur une profondeur totale de 24 mètres. Les diverses méthodes d'analyse de données utilisées (analyse factorielle des correspondances et en composantes principales) n'ont pas permis de distinguer les différents types de courbes teneur métal/profondeur, contenues dans la population.

J. Picard a alors utilisé la méthode des nuées dynamiques d'une part pour une classification à l'aide d'étalons initiaux choisis, d'autre part pour une recherche de profils types à l'aide des formes fortes. L'enrichissement apporté aux méthodes classiques peut être résumé comme suit :

1) Le procédé de classification permet un partitionnement du plan 1-2 de l'analyse factorielle non décelable à priori, ce découpage en éléments disjoints est confirmé encore plus nettement dans l'espace tridimensionnel.

2) Le tableau des formes fortes fait apparaître 5 formes fortes dont les sondages moyens sont très significatifs d'un type de terrain.

3) Les 5 formes fortes réunies ne contiennent que 59 sondages. J. Picard a vérifié la représentativité de ces 5 formes en montrant que le nuage obtenu par les deux analyses factorielles* suivantes ne présentait pas de modification sensible :

- une analyse factorielle des correspondances des formes fortes avec en éléments supplémentaires les sondages restants.
- une analyse factorielle des correspondances de toute la population.

Une autre expérience a permis de confirmer ce résultat : la position des paramètres dans une analyse factorielle sur la population totale est la même que dans une analyse factorielle faite uniquement sur les 59 sondages des formes fortes.

4) Il a en outre été procédé à une classification hiérarchique qui ne contredisait pas les résultats obtenus mais ne faisait pas apparaître nettement les formes fortes.

On peut remarquer, que les deux méthodes peuvent être utilisées conjointement, la méthode hiérarchique permettant une évaluation du nombre de classes à priori pour la

* Ces analyses factorielles sont représentées dans la thèse de J. Picard.

méthode des nuées dynamiques ; cette dernière déterminant des classes très typées et les profils principaux de la population.

7.3 – RECHERCHE DE PROFILS BIOLOGIQUES*

Il s'agit de découvrir des groupements types dans une population de 990 sujets. Nous avons tenu compte de 16 paramètres chez chaque malade : des proportions des 5 fractions électrophorétiques : albumine, alpha-1, alpha-2, bêta, gamma, des trois paramètres de la fiche réticulo-endothéliale et du taux de 7 globulines individuelles, déterminées par la méthode immunochimique de diffusion radiale.

Ces 900 sujets ont été classés à priori en 33 groupes distincts représentant soit des entités nosologiques, soit des syndrômes.

Sont représentés en particulier, le cancer des tissus solides sans ou avec atteinte du foie, les leuco-réticuloses, des maladies infectueuses très variées, des collagénoses, la cirrhose du foie et l'hépatite virale, des dermatoses variées, des maladies atopiques, le diabète et autres troubles endocriniens, la macroglobulinémie des africains, des ulcères gastro-duodénaux. Un travail approfondi sur les mêmes données a été réalisé par le Pr. Lenoir et M. Kerbaol à l'aide de l'analyse factorielle des correspondances (cf. [21] ; cette analyse a été faite sur un nombre restreint de paramètres et a fait apparaître des nuages d'un grand intérêt pour les praticiens ; **cependant, une délimitation objective de ces nuages est difficile et de plus sur les 16 paramètres avec 990 sujets le nuage obtenu par l'analyse factorielle est d'interprétation difficile ; d'un autre côté, une classification donnant une hiérarchie n'est pas praticable vu la taille des données.**

Nous avons utilisé la variante c) de la méthode des nuées dynamiques avec $k = 10$, $n_i = 10$ et la distance du χ^2 (cf. [10]). Le tableau des formes fortes a été calculé avec $n = 15$; les formes fortes obtenues sont particulièrement significatives puisque chaque sujet a une chance sur 10^{15} d'appartenir à une forme forte donnée.

Le Pr Sandor a trouvé très commode la représentation en nombres entiers (donnée par le tableau des formes fortes) pour exprimer la position réelle des sujets dans $\mathbb{R}^{16}$. En effet, grâce à ce tableau on a un moyen de saisir les multiples aspects de la position des points dans $\mathbb{R}^{16}$, de manière bien plus proche de la réalité que toute classification rigide n'aurait pu le faire.

---

* **Ce travail a été réalisé en collaboration avec le M. le Professeur Sandor de l'Institut Pasteur, MM. Lechevalier et Barré de l'IRIA ; il a fait l'objet d'une communication à l'Académie des Sciences [22], et d'un rapport de stage IRIA, pour plus de détails le lecteur pourra se reporter à ces textes.**

Une étude détaillée du tableau des formes fortes et une analyse factorielle du triple (cf. [4]) sur les 51 formes fortes qui sont apparues ont permis de dégager nettement l'existence de huit formes, ces formes permettent de tracer 8 profils types.

Nous ne donnerons pas ici l'interprétation détaillée des profils types obtenus, le lecteur intéressé pourra se reporter au compte-rendu à l'Académie de Médecine (séance du 29.2.1972) à paraître prochainement. Nous nous bornerons à donner la conclusion de compte-rendu.

Après avoir signalé une anomalie en ce qui concerne le classement de l'ataxie télangiéctasique*, le Pr Sandor conclut ainsi : " Il reste non moins vrai que les résultats que nous apportons constituent une exellente base d'un diagnostic objectif. L'appartenance à un type de profil donnera, en effet, le plus souvent tous les renseignements que le praticien peut tirer sur le plan des diagnostics et des pronostics d'un protéinogramme et il n'aura pour cela besoin d'aucune connaissances concernant la nature et l'origine des diverses protéines sériques ".

---

* Nous pensons que cette anomalie vient du fait que la distance utilisée donne plus d'importance aux augmentations au-dessus de 1 qu'aux diminutions entre 0 et 1.

conclusion

Un large champ de recherche reste ouvert ; sur le plan pratique, il faudrait développer à l'aide des méthodes d'apprentissage par exemple, le choix de f et g , développer les techniques permettant le choix de k (le nombre de classes demandées à priori), réaliser une comparaison exhaustive des différentes variantes de la méthode des nuées dynamiques, approfondir et développer les techniques du passage d'une arborescence à l'autre, faire une étude statistique de la structure de l'espace V_k en liaison avec E, notamment en ce qui concerne le nombre relatif d'éléments impasses, d'éléments non biaisés, la taille des arborescences, des niveaux etc. . . Mettre au point des techniques permettant une vision plus nette du tableau des formes fortes (du type " minimum spaning tree " par exemple). Utiliser les formes faibles afin de détecter entre les formes fortes les zones à faible densité (l'obtention des " trous " débouchant sur de nombreuses applications pratiques).

Sur le plan théorique il faudrait caractériser les familles de fonctions carrées, développer des théorèmes de convergence pour les différentes variantes et dans le cas où le nombre d'objets à classifier tend vers l'infini, ce dernier point est d'un grand intérêt pratique car il devrait permettre de développer et justifier des techniques purement séquentielles.

REMERCIEMENTS.— Je tiens à témoigner ma reconnaissance à M. le Professeur J.C. Simon* pour ses conseils et ses encouragements ainsi qu'à M. Chavent** pour ses judicieuses remarques lors de la rédaction définitive de ce texte. Egalement MM. M. Roux*** pour ses conseils, Y. Lechevallier*** et J. Barré*** pour leurs remarques et leur aide à la programmation.

---

* Professeur à la Faculté des Sciences de Paris-6.

** Chef de projet à l'IRIA dans le département de M. le Professeur Lions.

*** Laboratoire de Statistique mathématique de la Faculté des Sciences de Paris (dirigé par M. le Professeur J.P. Benzecri).

bibliographie

[1] BALL G.H., 1970 – Classification Analysis – Technical Note, Stanford Research Institute. Menlo Park, California 94025 USA

[2] BARBU M., 1968 – Partitions d'un ensemble fini : leur treillis – M.S.H. n° 22

[3] BENZECRI J.P., 1971 – Algorithmes rapides d'agrégation – Sup. Class. n° 9, Laboratoire de Statistique Mathématique - Université de Paris-6

[4] BENZECRI J.P., 1970 – Représentation Euclidienne d'un ensemble muni de masses et de distances – Université de Paris-6

[5] BERGE C., 1967 – Théorie des graphes et ses applications – Dunod Editeur, Paris

BOLSHEV L.N., 1969 – Cluster Analysis – I.S.I.R.S.S.' 69

[6] BONNER R.E., 1964 – On some clustering technics – IBM Journal of Research and Development

[7] CORMACK R.M. 1971 – A review of Classification – The journal of the Royal Statistical Society, Serie A, vol. 134, Part 3

[8] DIDAY E., BERGONT M., BARRÉ J., 1970-71-72 – Différentes notes sur la programmation de la Méthode des nuées dynamiques – Note IRIA, Rocquencourt 78

[9] DIDAY E., 1970 – La méthode des nuées dynamiques et la reconnaissance des formes – Cahiers de l'IRIA, Rocquencourt 78

[10] DIDAY E., 1971 – Une nouvelle méthode en classification automatique et reconnaissance des formes – Revue de Statistique Appliquée, vol XIX, n° 2

[11] FISHER L., VAN NESS J.W., 1971 – Admissible Clustering Procedures – Biometrika, 58, 1, p.91

[12] FREEMAN N., 1969 – Experiments in discrimination and classification – Pattern Recognition J. vol. 1, n° 3

[13] HALL D.J., BALL G.H., 1965 – Isodata a Novel Method of Data Analysis and Pattern Classification – Technical Report, 5 R I Project 5533, Stanford Research Institute, Menlo Park, California U.S.A.

[14] HILL D.R., 1967 – Mechanized Information Storage, retrevial and dissemination – Proceedings of the F.I.D./I.F.I.P. Joint Conference Rome

[15] JOHNSON S.C., 1967 – Hierarchical clustering schemes – Psychometrica 32, 241-45

[16] LERMAN H., 1970 – Les bases de la classification automatique – Gauthiers-Villars, 1970

[17] PICARD J., 1972 – Utilisation des méthodes d'analyse de données dans l'étude de courbes expérimentales – Thèse de 3e cycle. Laboratoire de Statistique Mathématique, Université Paris 6

[18] ROMEDER J.M., 1969 – Méthodes de discrimination – Thèse de 3e cycle. Statistique Mathématique. Faculté des Sciences de Paris 6

[19] ROUX M., 1968 – Un algorithme pour construire une hiérarchie particulière – Thèse de 3e cycle. Laboratoire de Statistique Mathématique, Université de Paris 6

[20] RUSPINI H.R., 1970 – Numerical Methods for fuzzy clustering – Information Science 2, p. 319-350

[21] SANDOR G., LENOIR P., KERBAOL M., 1971 – Une étude en ordinateur des corrélations entre les modifications des protéines sériques en pathologie humaine – C.R. Acad. Sc. Paris, t.272, p. 331-334

[22] SANDOR G., DIDAY E., LECHEVALLIER Y., BARRÉ J., 1972 – Une étude informatique des corrélations entre les modifications des protéines sériques en pathologie humaine – C.R. Acad. Sc. Paris, t. 274, d.p. 464-467

[23] SEBESTIEN G.S., 1966 – Automatic off-line Multivariate Data Analysis – Proc. Fall Joint Computer Conference pp. 685-694

[24] SOKHAL R.R., SNEATH P.H.R., 1963 – Numerical Tasconomy – W.H. Freeman and Co., San Francisco and London

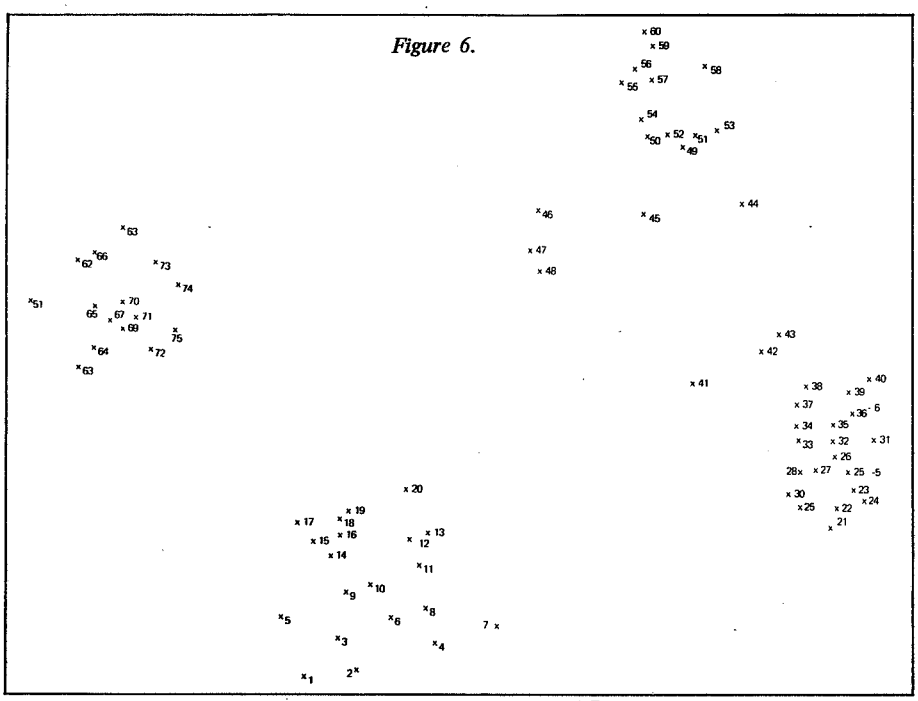
[25] WATANABÉ M.S., 1971 – A unified view of clustering algoritms – IFIP Congress 71, Ijubiana, Booklet TA-2

[26] ZADEH L.A., 1965 – Fuzzy sets – Inf. Control 8, pp. 338-353

[27] ZAHN C.I., 1971 – Graph theoretical methods for detecting and describing Gestalt Clusters – I.E.E.E. Trans. on Computers, vol. C-20, n° 1, January

[28] McQUEEN J., 1967 – Some Methods for Classification and Analysis of Multivariate Observations – 5th Berkeley Symposium on Mathematics, statistics and probability, vol. 1, n°1, pp. 281-297

Figure 6.



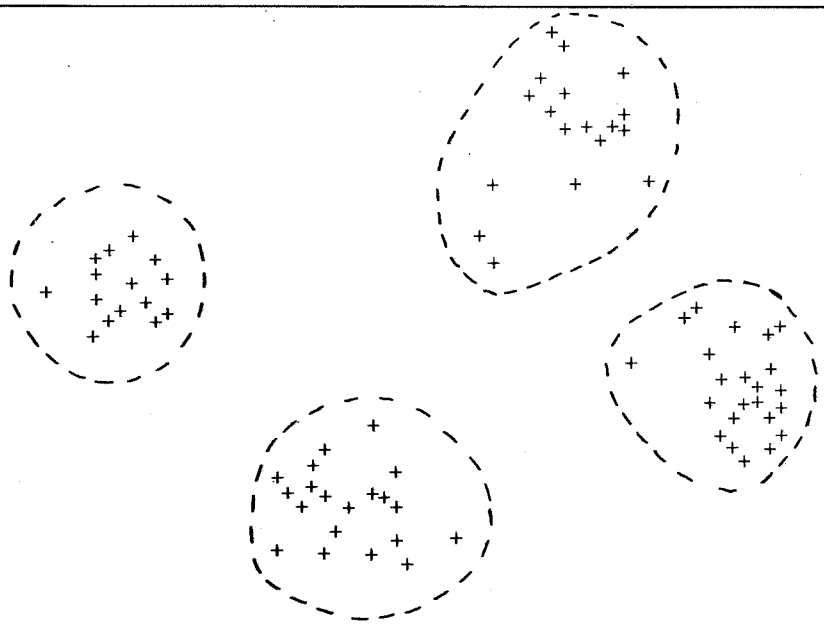


Figure 7

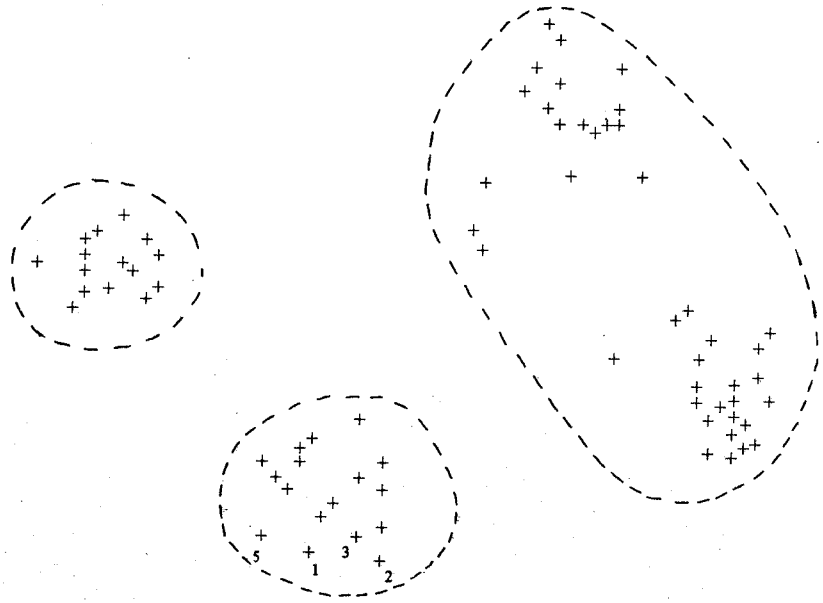


Figure 8

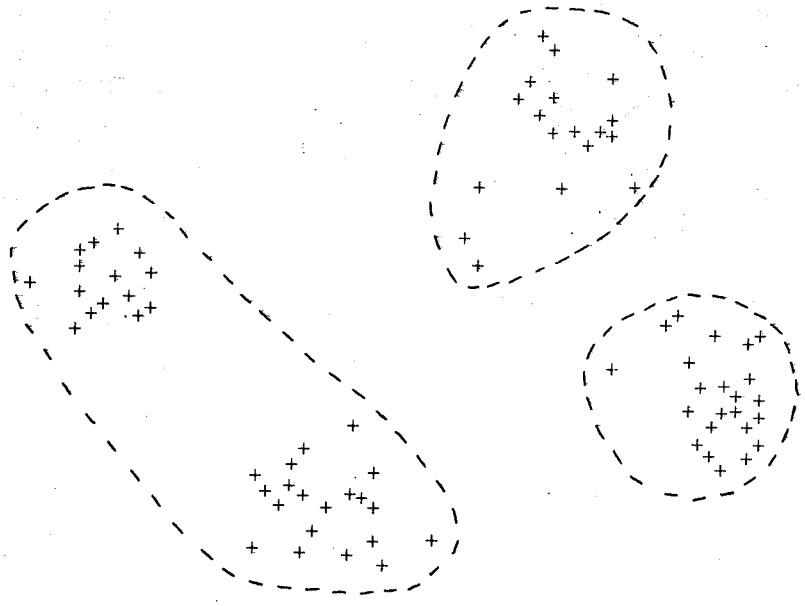


Figure 9

Figure 10

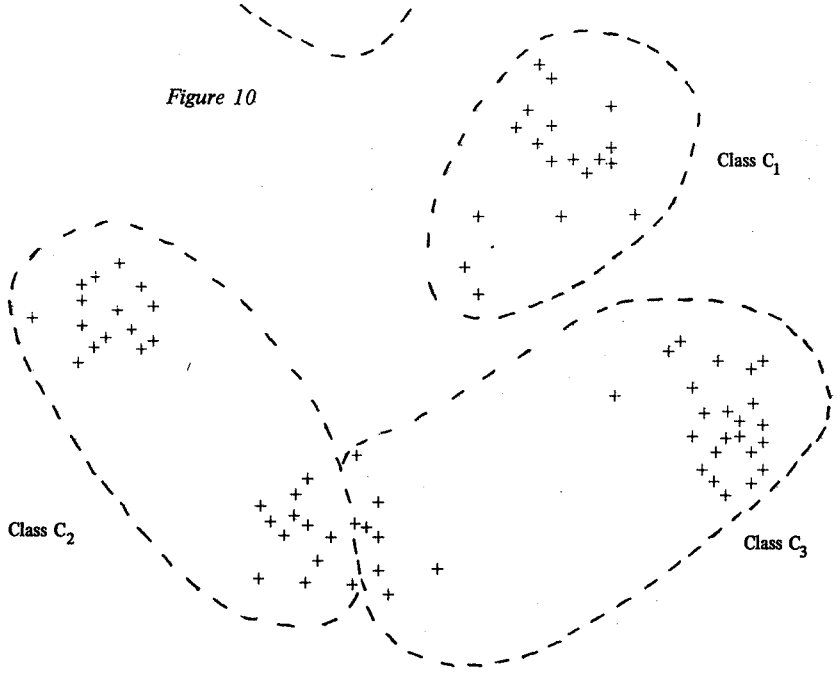


Figure 11

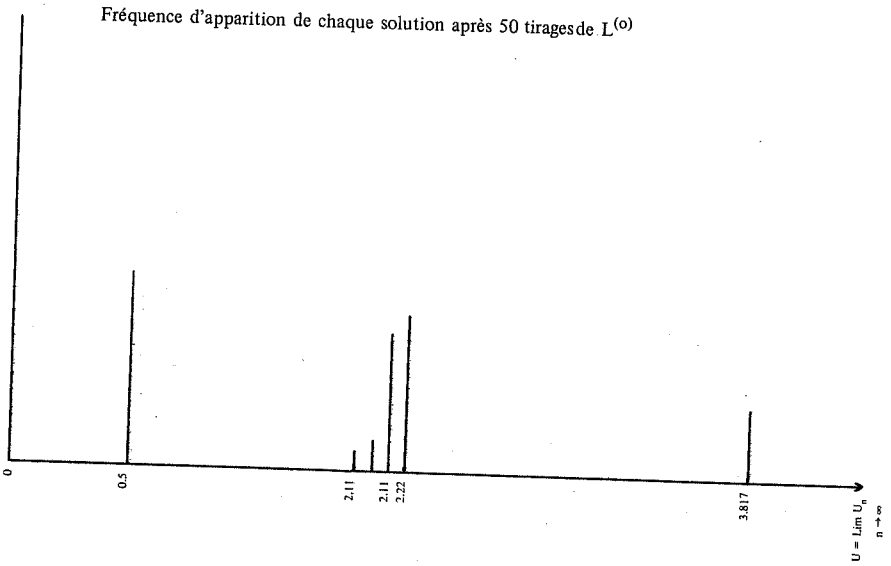


Figure 12

La valeur U = 0.5 correspond à la solution donnée figure 7
U = 2.11 correspond à la solution donnée figure 8
U = 2.22 correspond à la solution donnée figure 9
U = 3.817 correspond à la solution donnée figure 10

Numéro des points	Solutions obtenues					δ_{ij}
	1ère	2ème	3ème	4ème	5ème	
1	1	3	3	1	4	0
2	1	3	3	1	4	0
3	1	3	3	1	4	0
5	1	3	3	1	4	0
6	1	3	3	1	4	0
9	1	3	3	1	4	0
10	1	3	3	1	4	0
8	1	3	3	1	6	1
4	1	3	3	1	6	0
7	1	3	3	1	6	0
11	1	3	3	1	6	0
12	1	3	3	1	6	0
13	1	3	3	1	6	0
14	1	3	3	1	5	1
15	1	3	3	1	5	0
16	1	3	3	1	5	0
17	1	3	3	1	5	0
18	1	3	3	1	5	0
19	1	3	3	1	5	0
20	1	3	3	1	5	0
21	2	2	1	2	2	5
22	2	2	1	2	2	0
23	2	2	1	2	2	0
24	2	2	1	2	2	0
25	2	2	1	2	2	0
26	2	2	1	2	2	0
27	2	2	1	2	2	0
28	2	2	1	2	2	0
29	2	2	1	2	2	0
30	2	2	1	2	2	0
31	2	2	1	2	2	0
32	2	2	1	2	2	0
33	2	2	1	2	2	0
34	2	2	1	2	2	0
35	2	2	1	2	2	0
36	2	2	1	2	2	0
37	2	2	1	2	2	0
38	2	2	1	2	2	0
39	2	2	1	2	2	0
40	2	2	1	2	2	0
41	2	2	1	2	2	0
42	2	2	1	2	2	0
43	2	2	1	2	2	0
44	3	4	1	4	1	4
45	3	4	1	4	1	0
46	3	4	1	4	1	0
47	3	4	1	4	1	0
48	3	4	1	4	1	0
49	3	4	1	4	1	0
50	3	4	1	4	1	0
51	3	4	1	4	1	0
52	3	4	1	4	1	0
53	3	4	1	4	1	0
54	3	4	1	4	1	0
55	3	4	1	4	1	0
56	3	4	1	4	1	0
57	3	4	1	4	1	0
58	3	4	1	4	1	0
59	3	4	1	4	1	0
60	3	4	1	4	1	0
61	4	1	2	3	3	5
62	4	1	2	3	3	0
63	4	1	2	3	3	0
64	4	1	2	3	3	0
65	4	1	2	3	3	0
66	4	1	2	3	3	0
67	4	1	2	3	3	0
68	4	1	2	3	3	0
69	4	1	2	3	3	0
70	4	1	2	3	3	0
71	4	1	2	3	3	0
72	4	1	2	3	3	0
73	4	1	2	3	3	0
74	4	1	2	3	3	0
75	4	1	2	3	3	0
$\cup$ =	1.2135	1.2135	4.8910	1.2135	1.0000	

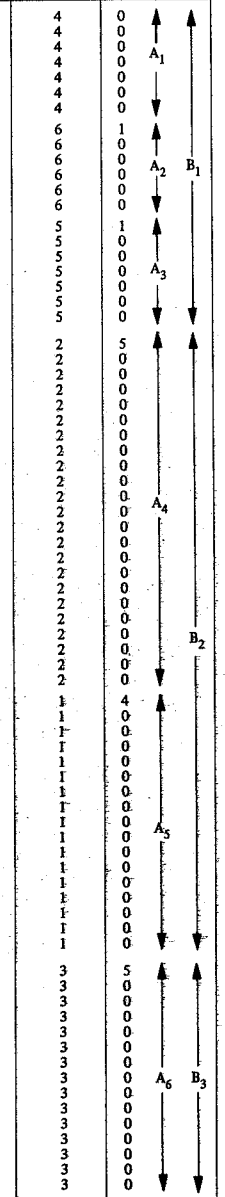


Tableau 1 : formes fortes pour les données de Ruspini.

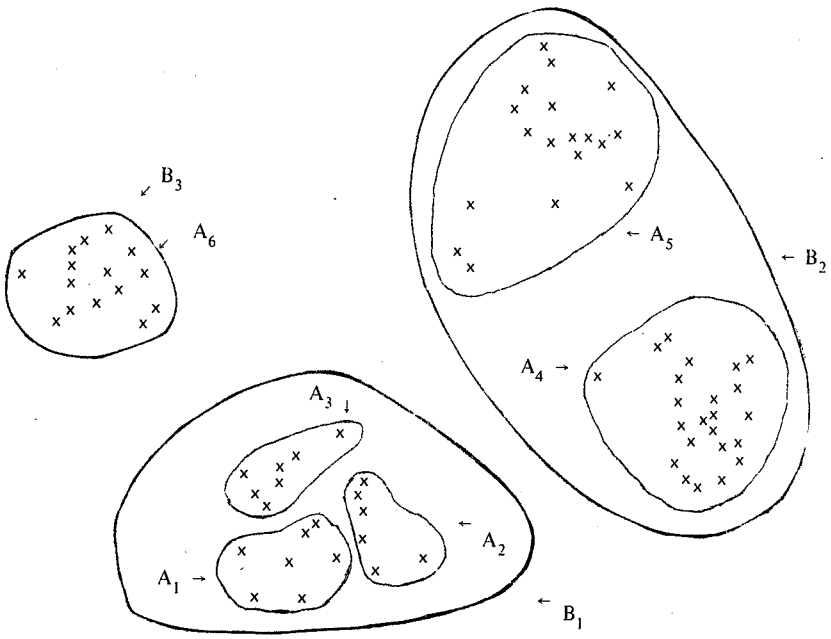


Figure 13

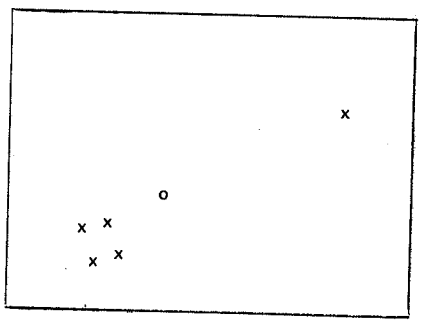


Figure 14

Les signes "x" représentent les éléments à classifier alors que le signe "o" représente le centre de gravité des 5 éléments.

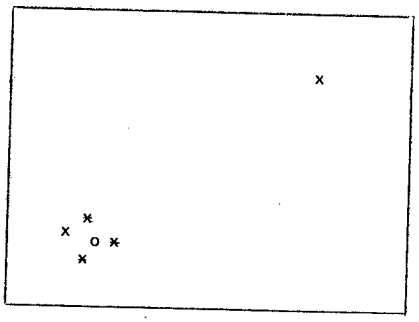


Figure 15

Les 3 éléments les plus proches de la population sont représentés par le signe "⁎" ; le centre de gravité de ces 3 éléments atténue l'effet du point marginal.

annexes

ANNEXE 1

Soit B un ensemble fini et une fonction h : B → B. Le graphe défini par B et l'ensemble des arcs (h (x), x) sera noté $\Gamma = (B, h)$. On sait que l'ensemble des composantes connexes de Γ constitue une partition de B ; chacune de ces composantes à une forme particulière :

Proposition 1 :

Chaque composante connexe de Γ contient un circuit au maximum.

Démonstration :

Si dans une composante connexe il existe un circuit, cela implique l'existence d'une suite finie de sommets $C = \{ x_0, \ldots, x_n \}$ telle que $h(x_i) = x_{i+1}$ et $h(x_n) = x_0$. L'existence de deux circuits dans une composante connexe implique la possibilité de sortir d'un circuit, c'est-à-dire l'existence de i et $y \notin C$ tels que $h(x_i) = y$. Cela n'est pas possible puisque $h(x_i) = x_{i+1}$ et que h est une fonction.

c. q. f. d.

On donne (fig. 4) un exemple de composante connexe de Γ. Nous dirons que x est un point fixe si $x = h(x)$. Une arborescence ayant pour racine un point fixe sera appelée arborescence bouclée (voir figure 5). Soit W une application $B \to \mathbb{R}^+$.

Proposition 2 :

Si W est injective sur toute suite v_n et vérifie la propriété $W(h(x)) \leqslant W(x)$ alors :

1) Chaque composante connexe de Γ contient une boucle et une seule et ne contient pas un autre circuit.
2) Chaque composante connexe de Γ est une arborescence bouclée ou une boucle.
3) Si $y \in B$ n'est pas un point fixe, il existe un point fixe x tel que $W(x) < W(y)$.

Démonstration :

1) Soit un circuit $C = \{ x_0, \ldots, x_n \}$ où $h(x_i) = x_{i-1}$, la condition $W(h(x)) \leqslant x$ implique $W(x_0) \leqslant W(x_1) \leqslant \ldots \leqslant W(x_n) \leqslant W(x_0)$ d'où $W(x_0) = W(x_i)$ $\forall i = 1, 2, \ldots, n$. Comme W est injective on a : $x_0 = x_i$. Donc tout circuit de Γ est une boucle.
Soit un chemin $C = \{x_0, \ldots, x_n\}$, la suite $u_n = W(x_n)$ est décroissante et minorée

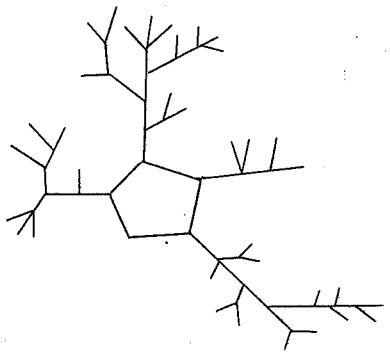
par 0 ; elle converge donc, de plus elle atteint sa limite (cf. [5]). Il existe donc $N : \forall n \geqslant N\ u_n = u_{n+1}$, d'où $x_n = x_{n+1}$ d'où $h(x_n) = x_n$. Donc toute partie connexe de Γ contient une boucle et une seule, elle ne contient pas un autre circuit d'après la proposition 1.

2) D'après ce qui vient d'être prouvé, toute composante connexe a un sommet et un seul qui est un point fixe. Soit x_1 ce sommet. Soit $X = \{x \in B / h(x) = x_1,\ x \neq x_1\}$. Si $X = \emptyset$, la composante connexe contenant x_1 est réduite à une boucle. Si $X \neq \emptyset$, la composante connexe contenant x_1 est une arborescence bouclée ; en effet, si on supprime la boucle $(h(x_1), x_1)$, les trois propriétés définissant une arborescence de racine x_1 sont vérifiée *:

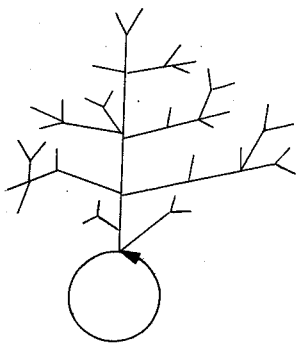
- Tout sommet $\neq x_1$ est l'extrêmité terminale d'un seul arc ; cela vient du fait que h est une fonction.
- x_1 n'est l'extrêmité terminale d'aucun arc ; puisqu'on a supprimé la boucle $(h(x_1), x_1)$.
- La composante connexe contenant x_1 n'a pas de circuit ; cela d'après la première assertion de cette proposition.

3) Les composantes connexes de Γ constituent une partition** de B ; si $y \in B$, y est un sommet d'une composante connexe de Γ ; si y n'est pas un point fixe, il appartient à une arborescence bouclée. Soit x_1 la racine de cette arborescence ; il existe un chemin de x_1 à y, soit $C = \{x_1, x_2, \ldots, x_n\}$ ce chemin, où $x_n = y$. On a : $W(x_1) \leqslant W(x_2) \leqslant \ldots \leqslant W(x_n)$ d'où $W(y) \geqslant W(x_1)$. Comme W est injective et y n'est pas un point fixe on en déduit $W(y) \geqslant W(x_1)$.

c. q. f. d.



Composante connexe de Γ
Figure 4



Arborescence bouclée
Figure 5

* Voir [5] page 154.

** Voir [5] page 9

ANNEXE 2

Il s'agit de montrer que les deux propriétés suivantes sont équivalentes pour caractériser une forme faible.

1) $Q^\star$ est la plus fine des partitions qui sont moins fines que $P^{1\star}, \ldots, P^{n\star}$.

2) $Q^\star$ est la partition définie par l'ensemble des parties connexes du graphe $\Gamma_1 = (E, F_1)$.

Démonstration :

Nous allons d'abord montrer que 1) ⇒ 2). Soit $x \in Q_j^\star$ alors $\forall z \in \complement_E Q_j^\star$ et si $W(z) = (\gamma_1, \ldots ; \gamma_n)$ on a $\alpha_j - \gamma_j = o$ car sinon $Q^\star$ ne serait pas moins fine que $P^{i\star}$ $\forall i$. Cela revient à dire que $\forall z \in \complement_E Q_j^\star$ et $\forall x \in Q_j^\star$ on a $z \notin F_1(x)$. Pour montrer que $Q_j^\star$ est bien une partie connexe de Γ_1, il reste à vérifier que : $\forall x \in Q_j^\star, \exists y \in Q_j^\star : y \in F_1(x)$. Prenons x quelconque dans $Q_j^\star$ et supposons qu'il n'existe pas y appartenant à $Q_j^\star$ tel que $y \in F_1(x)$; cela signifierait que $\delta(x, y) = o$ $\forall y \in Q_j^\star$ autrement dit $\alpha_i \neq \gamma_i$ $\forall i$; ainsi en remplaçant $Q_j^\star$ par les classes $Q_j - \{x\}$ et $\{x\}$ on définit une partition Q qui tout en étant moins fine que les $P^{i\star}$ serait plus fine que $Q^\star$ ce qui est contraire à l'hypothèse.

Montrons maintenant que 2) ⇒ 1). Par définition de Γ_1 une partie connexe Q_j est telle que $\forall x \in Q_j$ et $\forall z \in \complement_E Q_j$ on a : $z \notin F_1(x)$, c'est-à-dire $\alpha_i - \gamma_i = o$ $\forall i$; ainsi l'ensemble des parties connexes Q_j de Γ_1 constitue bien une partition moins fine que les partitions $P^{i\star}$ $\forall i$. Il reste à montrer que la partition Q des parties connexes de Γ_1 est bien la plus fine parmi celles qui sont moins fines que les $P^{i\star}$; soit x et y deux éléments appartenant à la même partie connexe de Γ_1 ; il existe une chaîne $z_1, \ldots, z_n$ tel que $x = z_1$, $y = z_n$ et $z_{q+1} \in F_1(z_q)$ $\forall q = 1, 2, \ldots, n - 1$; deux sommets consécutifs quelconques de cette chaîne appartiennent nécessairement à une même classe de l'une des partitions $P^{i\star}$, par définition même de F_1. Pour toute partition plus fine Q' que Q qui découpe par exemple Q_j en deux parties A et B il existe $x \in A$ et $y \in B$ et donc $q : z_q \in A$ et $z_{q+1} \in B$; Q' découpe donc l'une des classes de la partition $P^{i\star}$ et n'est donc pas moins fine que les partitions $P^{i\star}\forall i$. La plus fine de ces partitions est donc bien Q.

c. q. f. d.

ANNEXE 3

Théorème (des "connexités descendantes")

Soit Δ l'application * $E \times E \to \mathbb{N}$ telle que $\Delta(x, y) = n - \delta(x, y)$ et soit E' l'espace quotient** E/H. Si F_p est la multi-application $E' \to \mathbb{P}(E')$ telle que $F_p(x) = \{y \in E' / \delta(x, y) \geqslant P\}$ et Γ_p est le graphe (E', F_p), alors :

---

* Δ est en fait la métrique de la différence symétrique (voir définition de δ en 5.2.1).

** H est l'application qui a été définie en 5.2.1.

1) L'ensemble des parties connexes de Γ_p pour p = 0, 1, 2, . . . , n constitue une hiérarchie sur E'.

2) Cette hiérarchie induit l'ultramétrique sous-dominante de Δ.

Démonstration

1) Soit G_i^p la $i^{ième}$ classe de la partition définie par les parties connexes du graphe Γ_p. Soit $G = \{G_i^p / i = 1, \ldots, q_p ; p = 1, \ldots, n\}$ où q_p est le nombre de parties connexes de Γ_q ; nous allons montrer que G est une hiérarchie sur E' :

   – $E \in G$, car Γ_o est réduit à une seule partie connexe qui est identique à E.

   – $\forall x \in E'$ on a $x \in G$; en effet, chaque partie connexe de Γ_n est réduite à un seul élément et l'ensemble de ces parties constitue une partition de E'.

   – Quels que soient a et b, éléments de G, si $a \cap b \neq \phi$ alors on a soit $a \subset b$, soit $b \subset a$. En effet, posons $a = G_i^p$, $b = G_i^m$, deux cas peuvent se produire : p = m, alors $a \cap b = \phi$ puisque a est une partie connexe et b une autre partie connexe du même graphe Γ_p.

   $p > m$, soit $x \in G_i^p$ alors pour tout élément $y \in G_i^p$ il existe une chaîne $z = z_1, \ldots, z_q$ avec $z_1 = x$ et $z_q = y$ telle que $\min_i \delta(z_i, z_{i+1}) \geqslant p > m$; donc tous les éléments connexes à x dans Γ_p sont dans une même partie connexe de Γ_m, autrement dit G_i^p est contenu dans une des parties connexes de Γ_m, on a donc $G_i^p \subset G_j^m$ ou bien $G_i^p \cap G_j^m = \phi$.

   Ainsi G est une hiérarchie.

2) On peut indicer G par l'application $X : G \to [0, 1]$ telle que $X(a) = p$ si p est le plus grand entier tel que $a \equiv G_i^p$. Cette application définit bien une hiérarchie indicée puisque $x(a) = 1$ si a est réduit à un seul élément car alors il existe i : $a \equiv G_i^n$; d'autre part $a \subset b \Rightarrow X(a) > X(b)$ car si $a = G_i^p$ et $b = G_j^m$, on déduit de 1) que $a \subset b \Rightarrow p > m$ d'où $X(a) > X(b)$.

   La hiérarchie G ainsi indicée, permet de définir sur E' un indice de similarité de la manière suivante :

   $d(x, y) = n\,(1 - \max_a \{X(a) / x, y \in a\})$; autrement dit $d(x, y) = n - q$ où q est le plus grand entier tel que x et y appartiennent à la même partie connexe de Γ_q. Cela implique l'existence d'une chaîne $z = (z_1, \ldots, z_p)$ telle que $z_1 = x$, $z_p = y$ et $\inf_{z \in \varphi_{xy}} \max_i (n - \delta(z_i, z_{i+1})) = n - q$ où φ_{xy} est l'ensemble des chaînes de x à y.

   En effet, dire que x et y appartiennent à une même partie connexe de Γ_q signifie qu'il existe une chaîne z telle que $\min_i \delta(z_i, z_{i+1}) \geqslant q$ d'où :

$\underset{z \in \mathcal{C}_{xy}}{\text{Max}} \ \underset{i}{\text{Min}}\ \delta(z_i, z_{i+1}) \geqslant q$; si $\underset{z \in \mathcal{C}_{xy}}{\text{Max}} \ \underset{i}{\text{Min}}\ \delta(z_i, z_{i+1}) \neq q$ on a pour tout $z \in \mathcal{C}_{xy}$ $\underset{i}{\text{Min}}\ \delta(z_i, z_{i+1}) \geqslant q + 1 > q$ ce qui est contraire au choix de q qui est le plus grand des entiers tel que x et y appartiennent à une même partie connexe de Γ_q. On a donc $\underset{z \in \mathcal{C}_{xy}}{\text{Max}} \ \underset{i}{\text{Min}}\ \delta(z_i, z_{i+1}) = q$ d'où $\underset{z \in \mathcal{C}_{xy}}{\text{Min}} \ \underset{i}{\text{Max}}\ (n - \delta(z_i, z_{i+1}) = n - q$, ce qui implique : $d(x, y) = \underset{z \in \mathcal{C}_{xy}}{\text{Min}} \ \underset{i}{\text{Max}}\ \Delta(z_i, z_{i+1})$. Cette condition suffit à prouver que d est la sous-dominante de Δ (cf. M. Roux [19]).

c. q. f. d.

Remarque :

Ce théorème permet d'obtenir un bon rangement du tableau des formes fortes. Plus généralement, pour tous les problèmes où le tableau des données ne contient que des nombres entiers et pour lesquels la distance de la différence symétrique est significative, ce théorème donne une méthode commode pour la construction de la hiérarchie indicée induite par la sous-dominante. En effet, on économise du temps et de la place mémoire puisqu'on peut construire cette hiérarchie sans avoir besoin de calculer et de mettre en mémoire le tableau des distances deux à deux des éléments pour la différence symétrique et pour la distance associée à la sous-dominante.

METHODES D'ANALYSE FACTORIELLE APPLIQUEES A LA RECONNAISSANCE DE LA PAROLE.

Résumé.

Les méthodes d'analyses factorielles sont des instruments d'observation permettant de visualiser l'espace des paramètres issus du prétraitement de la voix. Les deux méthodes utilisées en pratique sont l'analyse en composantes principales de HOTELLING et l'analyse des correspondances de BENZECRI. Elles permettent également de réduire le nombre de paramètres en exhibant ceux des paramètres initiaux qui sont les plus indépendants, et qui ont la valeur informative la plus forte.

L. LEBART
C.R.E.D.O.C. (Paris).

METHODES D'ANALYSE FACTORIELLE APPLIQUEES A LA RECONNAISSANCE DE LA PAROLE.

Les méthodes d'analyse factorielle interviennent, de par leur caractère d'instrument d'observation, au moment de la description de l'espace des paramètres caractéristiques issus du prétraitement de la voix. Elles permettent une visualisation de cet espace, et, par voie de conséquence, une réduction du nombre de ces paramètres. Si un vocodeur fournit de façon automatique une caractérisation d'un phonème sous forme de vecteur à, par exemple, 20 composantes, il est clair que ces composantes constituent un prélèvement arbitraire, chargé de redondance et grevé par des lacunes. Il est nécessaire, dans une première phase exploratoire de "photographier" l'espace des paramètres, afin d'exercer un contrôle et une analyse raisonnés du fonctionnement aveugle de l'appareil de prétraitement.

Cette "photographie" est rendue possible grâce aux techniques d'analyse factorielle descriptive que sont l'analyse en composantes principales et l'analyse des correspondances. Ces techniques permettent, en effet, de mettre en évidence le nombre minimum de variables nécessaires pour reconstituer de façon satisfaisante l'information fournie par l'ensemble des paramètres. Ces deux techniques peuvent également donner lieu à des applications dans des domaines connexes : études des cooccurences des différents phonèmes dans des mots ou expressions (ou phrases ou vers) appartenant à un corpus donné, ou, à un autre niveau d'articulation, études des cooccurences de mots dans des unités statistiques (articles, messages, pièces de théâtre, poèmes, ...) appartenant à un corpus encore plus vaste.

Nous n'avons donc pas affaire à une technique permettant d'aider de façon directe à la résolution des problèmes de reconnaissance de la parole, mais à un instrument d'observation utile dans une phase préliminaire d'analyse, de contrôle, d'orientation des recherches.

Nous allons brièvement donner les principes de ces méthodes en mettant en évidence leur noyau théorique commun, après un rapide aperçu historique.

I. Origine des méthodes.

Les grands principes de réduction linéaire des données sont à attribuer aux psychologues du début de ce siècle, et en particulier à SPEARMAN. En fait, l'arsenal mathématique nécessaire à ces analyse est plus ancien. L'analyse factorielle des psychologues, ou analyse en facteurs communs et spécifiques, s'est développée durant la première moitié de ce siècle dans le champs relativement clos de la psychologie expérimentale, sans grande relation, ni avec la statistique mathématique, ni avec la physique ou les mathématiques appliquées. Cet isolement a eu pour conséquence la création d'un système de notation, d'un style de formulation des problèmes très particulier qui furent et sont encore une entrave à la communication interdisciplinaire. La méthode d'analyse factorielle des psychologues pose le modèle a priori suivant : chacun des paramètres initiaux dépend linéairement d'un petit nombre de nouveaux paramètres appelés facteurs, ou variables latentes. En fait, l'effet de ces facteurs communs aux paramètres initiaux est complété par celui de facteurs spécifiques à chacun des paramètres, qui reconstituent en quelque sorte le bruit du phénomène étudié. Un tel modèle pouvait être posé "a priori", moyennant quelques hypothèses supplémentaires, par les psychologues pour lesquels la notion de facteur avait déjà une existence conceptuelle : les facteurs cachés reconstituant les notes initiales (notes à des épreuves ou à des tests) représentent des aptitudes (intelligence, mémoire, etc...).

Le modèle préconisé par les psychologues est trop particulier pour que son emploi puisse être généralisé. On lui préfère généralement d'autres méthodes de réduction, également particulières, mais beaucoup plus simples dans leur principe. Ces méthodes nous donneront une représentation à partir du tableau de données initiales, et des règles d'interprétation permettant de remonter inductivement vers la structure réelle des données à partir des structures représentées.

Précisons que l'analyse factorielle de SPEARMAN a été perfectionnée par THURSTONE, puis rattachée à la statistique mathématique par LAWLEY et

et MAXWELL. Elle a donné naissance à l'analyse en composante principale (HOTELLING). Des grands principes de ces méthodes est issue l'analyse des correspondances (BENZECRI).

II. Exposé théorique général.

Nous allons expliciter le "noyau théorique" commun à la plupart des techniques d'analyse de données. Toutefois, nous ne démontrerons pas les résultats purement mathématiques.

Nous désignerons par D le tableau de données, supposé avoir n lignes et p colonnes. Nous allons essayer d'opérer une approximation de ce tableau de valeurs numériques sans faire d'hypothèse sur la structure initiale de D.

On peut considérer indifféremment que les colonnes de D constituent p points de l'espace vectoriel R^n (Espace des vecteurs à n composantes), ou au contraire que les lignes de D sont n point de R^p (Espace des vecteurs à p composantes).

Nous supposerons que les espaces vectoriel vectoriels R^n et R^p sont munis de la métrique euclidienne usuelle, somme des carrés des coordonnées, ce qui signifie par exemple que la distance de deux points x et y de R^p s'écrit :

$$d^2(x,y) = (x - y)'(x - y) = \sum_{i=1}^{p} (x_i - y_i)^2$$

(Le symbole " ' ", à la suite d'un vecteur ou d'une matrice, désigne le transposé de ce vecteur, ou de cette matrice).

Notons que cette supposition, bien qu'étant la plus naturelle et la plus simple possible, peut être considérée comme une hypothèse. On peut, sans aucune difficulté, supposer que les deux espaces sont munis de distances euclidiennes quelconques, mais l'exposé en serait alourdi (cf. paragraphe suivant).

Plaçons-nous donc dans l'espace R^p, où notre tableau D représente n points. Nous allons chercher un sous-espace à une dimension qui ajuste au mieux le nuage de n points. (Si cet ajustement est bon, à partir des p composantes de ce sous-espace, et des n abscisses des n points

sur ce sous-espace, on pourra reconstituer les positions des points dans l'espace R^p, et par conséquent, le tableau D initial).

Soit donc u le vecteur des cosinus directeurs d'une droite dans R^p. (On a la relation : u'u = 1, c'est-à-dire $\sum_{i=1}^{p} u_i^2 = 1$, puisque le vecteur u est unitaire). Le vecteur z = Du, à n composantes, n'est autre que le vecteur des projections des n points sur la droite joignant l'origine des axes au point u. Chaque composante de z est bien le produit scalaire du vecteur unitaire u par l'un des n vecteurs de R^p.

Comme critère d'ajustement de la droite au nuage de points, on peut retenir la somme des carrés des écarts, ce qui constitue une seconde hypothèse, bien que ce critère soit également l'un des plus simples et des plus naturels. Or, pour chaque point, le carré de la distance à l'origine des axes (quantité constante) se décompose en somme des carrés de la projection sur la droite et de la distance à la droite. Il revient donc au même de maximiser la somme des carrés des différents points sur la droite, ou de minimiser la somme des carrés des distances à la droite, puisque ces deux quantités ont une somme constante (somme des carrés des distances des points à l'origine).

Nous chercherons donc la droite joignant l'origine au point u , qui rendra maximum la quantité : S = u'D'D u

avec la contrainte u'u = 1 .

Il s'agit là d'un problème mathématique connu dont la solution est la suivante: le vecteur u est le vecteur propre de la matrice symétrique D'D correspondant à la plus grande valeur propre : D'D u = u

En prémultipliant les deux membres de cette relation par u', on voit que la valeur propre est précisemment la valeur du maximum de S .

On verrait d'une façon analogue que le sous-espace à deux dimensions qui ajuste au mieux le nuage des n points est engendré par les deux premiers vecteurs propres de la matrice D'D, puis que le sous-espace à k dimension ayant les mêmes propriétés est engendré par les k vecteurs propres de D'D correspondant aux k plus grandes valeurs propres.

En se limitant à deux dimensions, si u et v désignent les deux premiers vecteurs propres de D'D, les vecteurs Du et Dv représentent

respectivement les abscisses et les ordonnées des n points dans le système d'axe (u, v).

Il suffit donc de porter ces deux quantités sur deux axes orthogonaux pour obtenir une approximation à deux dimensions du nuage de n points, et donc pour représenter les proximités existant entre ces n points vis-à-vis des p variables.

Plaçons nous maintenant dans R^n, où les colonnes du tableau de données D représentent p points. Nous procéderons comme dans R^p. Soit s le vecteur unitaire dont les composantes sont les cosinus directeurs d'un sous-espace à une dimension de R^n.

Le vecteur ayant pour composantes les p projections des p points sur cet axe s'écrit :

$w = D's$ (D' désignant la transposée de D)

Un raisonnement analogue au précédent va nous conduire à chercher le maximum de la forme quadratique :

$T = s' D D' s$

Avec la contrainte $s's = 1$

Le vecteur s sera donc le vecteur propre de D D' correspondant à la plus grande valeur propre.

Liaison entre les représentations dans les deux espaces.

Considérons par exemple l'équation aux vecteurs propres que nous avons été conduits à résoudre dans R^p :

$D'D\, u = \lambda u$

En prémultipliant les deux membres de cette relation par la matrice D , on obtient :

$D' D'D\, u = \lambda D'u$

ou encore :

$D\ D'(D\, u) = \lambda (D'\, u)$

En posant : $s = D\, u$

On constate que si u est vecteur propre de D'D, relatif à la valeur propre non nulle λ , alors $s = D\, u$ est vecteur propre de la matrice D D' relatif à la même valeur propre.

Les matrices à diagonaliser dans les deux espaces ont donc les mêmes valeurs propres non-nulles. De plus, les vecteurs propres sont liés par une relation simple.

On doit cependant normer le vecteur s :

$s's = u'D'Du = \lambda$

On posera donc, pour que s soit unitaire :

$s = (1/\sqrt{\lambda}\ ) D u$

De façon entièrement symétrique, on aura la relation :

$u = (1/\sqrt{\lambda}\ ) D' s$

Notons que c'est le tableau de données initial qui sert de matrice de passage entre les axes factoriels dans les deux espaces.

Nous avons vu plus haut que le vecteur $z = D u$ avait pour composantes les coordonnées des n points de R^p sur l'axe u.

De même, le vecteur $w = D' s$ a pour composantes les coordonnées des p points de R^n sur l'axe s.

Les relations précédentes nous montrent que :

Les cosinus directeurs d'un axe factoriel dans un espace sont proportionnels aux coordonnées des points sur l'axe factoriel correspondant de l'autre espace.

En pratique, on choisira, parmi les matrices symétriques $D D'$ et $D' D$ celle qui a les plus petites dimensions, pour la diagonalisation.

Reconstitution approchée du tableau de données D.

Nous désignerons maintenant par s_q et u_q les $q^{\text{ièmes}}$ vecteurs propres des matrices DD' et $D'D$.

On a la relation :

$$\sqrt{\lambda_q}\ s_q = D\, u_q$$

Postmultipliant les deux membres de cette relation par u'_q

$$\sqrt{\lambda_q}\ s_q u'_q = D u_q u'_q$$

En sommant par rapport à l'indice q, on obtient

$$\sum_q \sqrt{\lambda_q}\ s'_q u'_q = D\left\{\sum_q u_q u'_q\right\}$$

Or, la matrice entre accolade du second membre de cette relation n'est autre que la matrice unité. (Puisque les vecteurs propres sont orthonormés).

On a donc la relation : $D = \sum_q \sqrt{\lambda_q}\, s_q u'_q$

En se limitant aux k premiers facteurs, si les valeurs propres correspondant à $q > k$ sont petites, on a la formule approchée :

$$D \approx \sum_{q=1}^{k} \sqrt{\lambda_q}\, s_q u'_q$$

Cette formule fondamentale de l'analyse des données nous montre à quel degré d'approximation on arrive en se limitant à k facteurs. Mathématiquement, cette formule exprime la façon dont on peut faire l'approximation d'un élément du produit tensoriel de R^n et de R^p (ici cet élément est D), à partir de sommes de tenseurs de rang 1.

Cette dernière formulation est celle qui permet d'exposer de façon la plus synthétique les techniques d'analyse de données, en respectant les rôles symétriques des deux dimensions du tableau à analyser. Nous avons préféré en donner un exposé plus élémentaire, pour atteindre un plus grand nombre d'utilisateurs.

III. Spécifications du modèle général.

Nous insisterons plus spécialement sur deux spécifications particulières : l'analyse en composantes principales et l'analyse des correspondances.

1. Analyse en composantes principales simultanée

Supposons que le tableau des données D contiennent n réalisations d'un même vecteur aléatoire à p composantes.

Pour décrire les associations existant entre les lignes et les colonnes du tableau, c'est-à-dire entre les observations et les variables, on pourrait procéder comme cela a été fait au paragraphe précédent. Cependant, il existe, de par la structure du tableau (existence de variables pouvant être hétérogènes, existence d'observations ayant un caractère répétitif) des informations exogènes dont nous devons tenir compte, sous peine de trouver des résultats assez triviaux.

Ainsi, les variables peuvent avoir des moyennes extrêmement différentes. Dans R^p, le point moyen du nuage peut être loin de l'origine, et par conséquent, le premier axe passant par l'origine ne représentera pas une caractéristique de la forme du nuage, mais une caractéristique de la position de ce nuage par rapport à l'origine. Il est donc plus raisonnable de se placer dans un système d'axes admettant pour origine le point moyen. De même, les échelles des différentes variables seront souvent différentes.

On sera finalement conduit à centrer et réduire les variables, c'est-à-dire à remplacer les données brutes d_{ij} par les données transformées :

$$x_{ij} = \frac{d_{ij} - \bar{d}_j}{s_j}$$

$\bar{d}_j$ et s_j désignant respectivement la moyenne et l'écart-type de la variable j.

Si l'on appelle X le tableau transformé, il ne reste plus qu'à appliquer les principes théoriques évoqués au paragraphe précédent en substituant X à D pour obtenir une analyse en composantes principales.

Principe de la représentation simultanée.

Les relations liant les représentations des deux ensembles, l'existence de valeurs propres communes, nous conduisent à représenter sur un même graphique les points observations et les points variables.

Ce qui n'est pour l'instant qu'un artifice graphique (facilitant beaucoup l'interprétation) sera beaucoup plus justifié dans le cas de l'analyse des correspondances. Ce sera même une façon d'introduire la méthode.

La représentation graphique nous permet d'interpréter les proximités entre variables en terme de corrélation, les proximités entre individus en terme de similitude de "comportement" vis-à-vis des p variables. La proximité entre un point représentant une variable et un point représentant un individu n'a aucun sens, par contre, la position relative d'un point d'un ensemble (par exemple l'ensemble des individus) par rapport à tous les points de l'autre ensemble (l'ensemble des variables pour notre exemple) pourra être interprétée. La cohérence des deux

représentations vis-à-vis de l'interprétation des facteurs est souvent frappante.

Si l'on appelle X le tableau de terme général $x_{ij} = \frac{d_{ij} - \bar{d}_j}{s_j}$

La matrice à diagonaliser est maintenant proportionnelle à :

$C = (1/n).\ X'X$

Ce n'est autre que la matrice des corrélations entre les p variables.

Calculons le coefficient de corrélation entre une variable et les valeurs d'un facteur (c'est-à-dire les coordonnées des individus sur un axe factoriel déterminé).

Comme nos nouvelles variables ont pour variance 1, et que la combinaison linéaire $u'x$ a pour variance λ, λ étant la valeur propre associée au vecteur propre u, on voit facilement que le coefficient de corrélation de la ième variable avec le facteur u vaut : $u_i \sqrt{\lambda}$ (u_i étant la $i^{ème}$ composante de u)

L'interprétation des facteurs sera ainsi grandement facilitée par le fait que les abscisses des points variables sur les axes factoriels correspondant sont proportionnels aux coefficients de corrélation entre ces variables et les facteurs. En pratique, on portera souvent les quantités $u_i \sqrt{\lambda}$ à la place des u_i, sur les axes, afin de lire directement le degré de corrélation.

2. Analyse des correspondances.

Cette théorie, contrairement à celle de l'analyse en composantes principales, n'est pas un simple cas particulier du modèle général exposé plus haut, mais relève d'un modèle encore plus général, que nous allons esquisser brièvement.

a) Analyse d'un tableau transformé avec deux métriques euclidiennes quelconques.

Nous avons rencontré deux conventions, lors de l'exposé du modèle général, qui, bien que faisant appel à des hypothèses naturelles, n'en étaient pas moins arbitraires.

D'une part, la distance dans R^p était la métrique usuelle, d'autre part, le critère d'ajustement retenu a consisté à rendre minimum la somme des carrés des écarts.

Il se peut que l'on ait des raisons de choisir des distances ou des critères plus compliqués, en raison de la nature des données et des hypothèses qu'elles suscitent.

Cela revient, par exemple, à choisir dans R^p une métrique euclidienne définie par une forme quadratique dont les coefficients forment une matrice symétrique Q_p. De même, cela revient à choisir dans R^n, une distance définie de la même façon par un tableau Q_n.

Pour l'ajustement des n points dans R^p, on peut également choisir une métrique-critère S_n (L'indice n nous rappelle la dimension du tableau). De même, pour l'ajustement des p points de R^n, on peut choisir une métrique-critère S_p.

Pour ajouter à la généralité du modèle, ajoutons que le tableau de données D peut être transformé initialement de façon à modifier la géométrie du nuage.

Nous supposerons que le tableau D donne lieu à la construction d'un nuage X_p dans R^p et X_n dans R^n. (Les n lignes de X_p sont les coordonnées des n points de R^p).

Formulée de cette façon, la théorie générale est évidemment trop complexe, et trop indéterminée. Reprenons la théorie précédente en nous plaçant à nouveau dans R_p.

Dans cet espace, muni de la métrique Q_p, les n projections des n points sont maintenant les composantes du vecteur $z = X_p Q_p u$, u désignant comme précédemment les composantes du premier axe factoriel.

La quantité à maximiser sera, en utilisant la métrique-critère $S_n = z'S_n z$, soit :

$$u'Q_p X'_p S_n X_p Q_p u$$

Avec la contrainte $u'Q_p u = 1$ (u est unitaire pour Q_p)

La solution consiste encore en la recherche des vecteurs propres d'une matrice, qui sera en général non symétrique :

$$X'_p S_n X_p Q_p u = \lambda u$$

u est appelé axe factoriel, alors que l'opérateur projection $Q_p u$ est appelé facteur. (Ces deux vecteurs coïncidaient précédemment, car les espaces vectoriels coïncidaient avec leur dual).

Le facteur $y = Q_p u$, vérifie l'équation matricielle, déduite de la précédente :

$Q_p X'_p S_n X_p y = \lambda y$

Si nous nous plaçons dans R^n, le même raisonnement nous conduira à résoudre l'équation matricielle symétrique de la précédente.

Pour l'axe factoriel s :

$X_n S_p X'_n Q_n s = \lambda' s$ (avec $s' Q_n s = 1$)

Pour le facteur $t = Q_n s$

$Q_n X_n S_p X'_n t = \lambda' t$

Dans le cas général, il n'y a évidemment aucune relation de correspondance simple entre ces deux espaces, sauf si les matrices Q_p, Q_n, S_p, S_n sont liées par des relations simples, et si les matrices transformées X_p et X_n se déduisent du tableau initial par des transformations simples.

b) Cas de l'analyse des correspondances.

Les matrices définissant les métriques et les critères sont toutes diagonales, et les tableaux transformés se déduisent des tableaux de départ D par une transformation également diagonale.

Avant de justifier statistiquement le choix des métriques, des critères, des transformations initiales, indiquons brièvement en quoi elles consistent.

Si D désigne le tableau de données, constitué dans le cas de tableaux de contingence par les effectifs d_{ij} des individus appartenant à la fois à la modalité i d'un ensemble et à la modalité j de l'autre ensemble, nous désignerons par $d_{i.}$ et $d_{.j}$ les sommes marginales :

$$d_{i.} = \sum_{j=1}^{p} d_{ij} \qquad \text{et} \qquad d_{.j} = \sum_{i=1}^{n} d_{ij}$$

Nous supposerons que le tableau d'effectifs D a, en fait, été divisé par la somme de tous ses éléments $k = \sum_{ij} d_{ij}$, de façon à l'interpréter comme un tableau de fréquence . Les quantités $d_{i.}$ et $d_{.j}$ peuvent alors s'interpréter en terme de fréquences marginales. On a en particulier les relations $\sum_i d_{i.} = 1$ et $\sum_j d_{.j} = 1$.

La métrique Q_p de R^p sera définie de la façon suivante :

$$(Q_p)_{ij} = \delta_{ij} / d_{.j} \quad (\delta_{ij} = 0 \text{ si } i \neq j ; \delta_{ii} = 1)$$

Autrement dit, la matrice Q_p est diagonale, et ses p éléments diagonaux valent $1/d_{.j}$. De même, la matrice Q_n sera diagonale, et ses n éléments diagonaux vaudront $1/d_{i.}$.

La matrice critère S_n vaut : $S_n = Q_n^{-1}$ soit : $(S_n)_{ij} = \delta_{ij} d_{i.}$

La matrice critère S_p vaut : $S_p = Q_p^{-1}$ soit : $(S_p)_{ij} = \delta_{ij} d_{.j}$

Enfin, ce sont toujours ces mêmes matrices diagonales qui nous donnent les tableaux transformés X_p et X_n à partir du tableau de données initial.

Pour le nuage dans R^p : $X_p = S_n^{-1} D = Q_n D$

Pour le nuage dans R^n : $X_n = D S_p^{-1} = D Q_p$

Justification des transformations, des métriques, des critères.

Les transformations initiales visent à éliminer les effets de taille et d'échelle, en construisant des nuages de profils (chaque ligne, ou chaque colonne est divisée par sa somme).

Les métriques choisies (Distance dites "du Chi-deux") vérifient la propriété d'équivalence distributionnelle, qui assure une stabilité des résultats obtenus lors de l'agrégation éventuelle de lignes ou de colonnes homothétiques.

Enfin, le critère choisi revient à donner à chaque point-profil une masse proportionnelle à la somme des éléments de la ligne ou de la colonne correspondante.

INTRODUCTION A LA CLASSIFICATION AUTOMATIQUE

Maurice ROUX
Université de Paris.

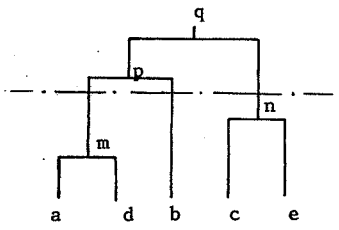
Introduction à la classification automatique.

Afin de préciser rapidement nos idées décrivons succintement le rapport entre classification automatique et reconnaissance des formes. Le but de cette dernière discipline est de pouvoir faire dire par une machine qu'une image ou un son (ou des assemblages d'images ou de sons) entre dans telle ou telle classe d'images ou de sons, classe qui appartient à un ensemble fini fixé à l'avance appelé ensemble des formes possibles. Ainsi la lecture automatique est du domaine de la reconnaissance des formes : étant donné un caractère imprimé - image en noir et blanc - dans laquelle des 26 classes possibles va-t-on le ranger? Bien que celà soit assez simple pour un cerveau humain, c'est un problème assez compliqué pour une machine, compte tenu de la diversité des caractères typographiques usuels.

Nous dirons que la reconnaissance des formes est un problème de classement. Nous réserverons en effet le terme de classification, ou taxinomie, pour désigner le travail préliminaire au classement, savoir la recherche et la définition de l'ensemble des classes "naturelles" dans le domaine étudié, à partir d'un échantillon dont on demande (ou plutôt on suppose) qu'il comprenne au moins un exemplaire de chacune des formes possibles.

Historiquement la classification s'est développée d'abord chez les naturalistes pour qui cette discipline reste la base indispensable à toute recherche approfondie. Ce qui explique, qu'en réalité la classification a un but un peu plus général que celui que nous avons dit ; on désire en effet non seulement reconnaître les types élémentaires (les espèces pour les naturalistes) mais aussi les classes en lesquelles certains de ces types peuvent se regrouper (les genres) puis les ensembles de classes (ordres), etc...

En bref, on ne se contente pas, en général, d'une partition de l'échantillon, on désire en fait obtenir une hiérarchie, c'est-à-dire une suite de partitions "emboîtées" telles que deux parties quelconques (appartenant ou non à la même partition) soient incluses l'une dans l'autre ou bien soient d'intersection vide.



Une telle hiérarchie peut être aisément résumée par un arbre hiérarchique (cf figure) dont les noeuds symbolisent les parties, constituées par les objets qui leurs sont reliés.

Le niveau des noeuds, qui est souvent chiffré, est censé indiquer un degré de ressemblance entre les objets correspondants. Ainsi, sur notre figure, les objets a et d se ressemblent plus que les objets c et e. Remarquons, en passant, que si l'on coupe cet arbre à un niveau intermédiaire entre n et p on obtient une partition en 3 de l'ensemble étudié, savoir les parties {a,d}, {b}, {c,e} ; et en faisant varier ce niveau de "troncature" on obtient les diverses partitions constituant la hiérarchie.

De ce qui précède on voit que la taxinomie (ou encore typologie) est assez éloignée de votre préoccupation : la reconnaissance de la parole. Elle nous paraît pourtant pouvoir être un auxiliaire utile ; tout d'abord dans la recherche directe des formes : à un ensemble de sons inconnus on pourra adjoindre des sons connus et soumettre le tout à une procédure de classification, la présence d'un son connu au sein de l'une des classes obtenues permettra d'identifier cette classe ; d'autre part il pourrait être utile de faire une classification préalable des locuteurs, c'est-à-dire chercher s'il existe des types de voix et lesquels ; il serait alors peut-être plus facile de faire des machines spécialisées dans le décodage d'un seul type de voix.

Quoi qu'il en soit la classification automatique se rattache plutôt à l'analyse des données dans la mesure où elle en suit le schéma habituel : on part d'un tableau rectangulaire de description - ce peut être, par exemple, une série de sons analysés par un "Vocoder", donnant l'énergie dissipée dans une dizaine de fréquences - et l'on cherche à en donner une image simplifiée, qui permette d'en découvrir la structure... si elle existe. Mais, tandis que cette simplification est obtenue en analyse factorielle par projection d'un nuage de points, représentant l'échantillon étudié, sur des plans privilégiés, en classification on cherche à rendre compte des niveaux de liaison entre les différents groupes qui sont supposés constituer le nuage, par une arborescence où la hauteur des noeuds indique ces niveaux.

Pour que le problème soit "bien posé" il faudrait, par exemple, que l'on ait une fonction de qualité de la représentation obtenue. Le but serait alors de trouver, parmi toutes les solutions possibles, celle qui optimise cette fonction.

Malheureusement l'ensemble de "toutes les solutions possibles", même si l'on se limite à la recherche d'une simple partition et non d'une hiérarchie, devient vite très grand avec le nombre d'objets à classer et l'on ne peut les essayer toutes, même avec de puissants ordinateurs. Cependant, des chercheurs de la Maison des Sciences de l'Homme (Paris) se sont engagés dans cette voie mais en limitant le nombre des solutions essayées : on part d'une partition initiale choisie arbitrairement et l'on cherche dans le voisinage de cette partition une autre partition qui améliore la fonction de qualité qu'on s'est fixé . On considère généralement comme voisinage d'une partition, l'ensemble des partitions qui ne diffèrent de la partition donnée que par l'affectation d'un ou deux éléments ;

c'est pourquoi les méthodes de ce type sont appelées "algorithmes d'échange" car elles consistent à essayer de changer de classe un élément à la fois, ou bien d'échanger l'appartenance de deux éléments de classes différentes.

On réitère le processus jusqu'à ce que l'on obtienne une partition qui ne puisse plus être améliorée. Il est évident qu'on n'obtient ainsi qu'un optimum local de la fonction-critère ; les résultats sont d'autant meilleurs que cette fonction est bien choisie et que les voisinages envisagés sont plus grands.

Tout autres sont les algorithmes d'agglomération autour de centres mobiles que monsieur Diday exposera ensuite avec sa "méthode des nuées dynamiques", et dont le but est aussi de faire une partition de l'ensemble étudié.

Parlons enfin des constructions ascendantes de hiérarchies. Ce sont les procédures les plus anciennes donc les plus connues et les plus couramment utilisées. Elles ont l'avantage d'être rapides.

Ces méthodes supposent donné l'ensemble des distances - mesures de dissemblance - entre les objets à classer. On procède alors par fusions successives d'objets, deux à deux : la première fusion concerne les deux objets les plus proches - les plus ressemblants - et l'on considère après cette première étape qu'ils ne forment plus qu'un. Il faut alors mettre à jour le tableau des distances inter-individuelles puisque les deux lignes, et les deux colonnes, de ce tableau correspondant aux objets fusionnés doivent être remplacées par une ligne, et une colonne, représentant les distances entre le groupe résultant de la fusion et les autres objets. On réitère ensuite ce processus jusqu'à ce que tous les objets aient été agglomérés ensemble.

La difficulté d'emploi de ces algorithmes réside dans le choix de la formule des distances inter-individuelles initiales ainsi que de la formule de mise à jour, à chaque étape, de ces mêmes distances. Soient i et j les objets (ou les groupes) fusionnés pour former le groupe h, le problème consiste à calculer la nouvelle distance d'(h,k) entre le nouveau groupe h ou un autre objet (ou groupe) k, compte tenu des anciennes distances d(i,k), d(j,k). Les formules sont aussi nombreuses que les chercheurs mais l'une d'entre elles, qui consiste à prendre les moyennes des anciennes distances, donne généralement de bons résultats :

$$d'(h,k) = [p(i)\ d(i,k) + p(j)\ d(j,k)] / [p(i) + p(j)]$$

p(i) et p(j) désignent le nombre d'éléments des groupes i et j respectivement, nombre qui est égal à 1 si i et j sont des objets élémentaires comme c'est le cas au début de l'algorithme.

Citons pour mémoire la méthode de Williams & Lambert qui procède par dichotomies successives. Elle s'applique surtout à des données qualitatives, chaque objet étant décrit par un vecteur booléen ne comprenant que des zéros ou des uns.

A chaque étape de l'algorithme on cherche pour chaque groupe existant la variable la plus liée à toutes les autres, suivant un critère basé sur le χ^2 ; le groupe est alors subdivisé en deux suivant la valeur (0 ou 1) des objets du groupe pour la variable retenue. Cette méthode est valable lorsque la présence ou l'absence d'une qualité chez un individu ont autant d'importance, le χ^2 faisant jouer des rôles symétriques aux deux valeurs, 0 et 1, possibles. Elle est efficace dans le cas ou l'on a beaucoup d'objets décrits par un petit nombre de variables, car le temps de calcul augmente rapidement avec ce dernier.

En conclusion on voit qu'aucune des méthodes existantes ne peut garantir que le résultat possède un avantage décisif sur toute autre solution. C'est pourquoi nous préférons les considérer comme des auxiliaires des analyses multidimensionnelles dont monsieur Lebart parlera après l'exposé de monsieur Diday.

Bibliographie sommaire

Benzécri J.-P. & Collaborateurs : L'analyse des données. Tome 1 : La taxinomie Dunod, Paris, 1973.

Cole A. J. (Ed.) : Numerical Taxonomy. Academic Press. London, New-York, 1969.

De la Genière J. et De la Véga W. F. : A propos de la classification. Rapport interne du centre de calcul de la Maison des Sciences de l'Homme, Paris 1969.

De la Véga W. F. : Techniques de classification automatique utilisant un indice de ressemblance. Rev. Franç. de de Sociologie, Paris, 1967.

Sokal & Sneath : Principles of numerical taxonomy, Free & C°, San Francisco, London, 1963.

Williams W. T. & Lambert J. M. : Multivariate methods in plant ecology. 1 - Association analysis in plant communities. J. Ecol., 1959, 47, pp. 83-101.

L'APPROCHE STATISTIQUE
DE
LA RECONNAISSANCE DES FORMES

Résumé

Exposé des principes généraux permettant de définir et calculer, éventuellement par apprentissage, les discriminateurs en vue de la reconnaissance des formes.

Guy PERENNOU
(Université P. SABATIER, TOULOUSE)

L'APPROCHE STATISTIQUE

DE

LA RECONNAISSANCE DES FORMES

-:-:-:-:-:-:-:-:-:-:-:-:-:-:-:-:-:-:

Résumé

Exposé des principes généraux permettant de définir et calculer, éventuellement par apprentissage, les discriminateurs en vue de la reconnaissance des formes.

I - INTRODUCTION

Lorsque l'on aborde un problème de reconnaissance des formes de type statistique, on est classiquement amené à distinguer deux types de traitements :

- le prétraitement,
- la classification.

Le prétraitement a pour but :

- la réduction du nombre de paramètres décrivant les formes ;
- l'adaptation de la description des formes à la méthode de classification choisie.

La classification a pour but :

- d'associer à chaque forme ainsi traitée une classe, de manière à reproduire le plus fidèlement une classification naturelle donnée.

La figure 1 schématise ceci.

Les formes avant prétraitement constituent l'ensemble F, après prétraitement l'ensemble G. Un élément $x \in F$ est transformé en un élément $y \in G$ puis, après classification par l'organe appelé discriminateur, transformé en l'une des classes $\gamma(y)$ d'un ensemble $C = \{1, 2, \ldots, K\}$

La classe naturelle j de x est comparée à $\gamma(y)$ lorsque l'on veut "régler" le discriminateur ou en vérifier les performances.

L'objet de l'exposé est la classification, c'est-à-dire

.../...

l'étude des discriminateurs.

D'autres exposés sont consacrés plus particulièrement au problème du prétraitement et de la taxinomie.

Notons simplement ici que le problème de la reconnaissance des formes est inverse de celui de la taxinomie (ou classification automatique) car dans le premier les classes sont fixées et il faut rechercher une méthode permettant d'associer à chaque forme sa classe tandis que dans le second un critère d'appréciation d'ensembles classés est donné et il faut rechercher le meilleur classement.

II - METHODES DE CLASSIFICATION

Dans ce qui suit l'espace F ne joue aucun rôle. Les probabilités envisagées sur G sont les images des probabilités correspondantes sur F. Nous identifirons les éléments de G avec les formes.

Notations et hypothèses

q_i : probabilité a priori d'occurrence d'un élément de classe $i \in C$

$Q(.)$: probabilité a priori sur G
Elle peut être interprétée comme une probabilité d'occurrence des formes de G ;

$Q(.|i)$: probabilité associée à la classe i ($i \in C$) ;

$Q(i|y)$: probabilité d'association de la classe i avec y.

Nous faisons deux hypothèses, classiques dans ce problème :

H1) G est un espace vectoriel de dimension M que l'on identifie à $\mathbb{R}^M$;

H2) $Q(.)$ et $Q(.|i)$ possèdent des densités de probabilité $q(.)$ et $q(.|i)$ respectivement.

.../...

Un exemple : la représentation des voyelles par la position des formants F_1 et F_2

La figure 2 donne la répartition des voyelles dans les plans des fréquences F_1 et F_2. Nous constatons que le prétraitement a ici réduit les formes sonores à deux paramètres numériques F_1 et F_2.

G est donc l'ensemble des couples (F_1,F_2) correspondant aux diverses émissions de voyelles.

La figure 2 montre que les voyelles particulières se répartissent dans des régions plus restreintes. Mais des régions d'incertitude existent. Par exemple, dans la région commune aux deux voyelles a et o un même point y correspondra parfois à un a et parfois à un o.

On peut alors interpréter :

$Q(.|i)$ comme la distribution aléatoire de la voyelle numéro i dans le plan (F_1,F_2)

$Q(i|y)$ la probabilité que le son caractérisé par le point y soit interprété comme la voyelle numéro i,

q_i comme la probabilité a priori d'apparition de la voyelle numéro i.

II.1. - Le discriminateur théorique

Supposons que pour chaque couple de classes (i,j) nous soyons en mesure d'associer un nombre $\lambda(i,j)$ représentant le coût ou la perte lorsque une forme $y \in G$ est classée en j alors que sa classe est i pour l'occurrence considérée.

Il est alors possible pour la forme y d'évaluer le coût moyen du classement de y en j par :

$$\ell(y,j) = \sum_i \lambda(i,j) Q(i|y)$$

la formule de Bayes

$$Q(i|y) = q(y|i)(q_i\ q(y))$$

entraîne

$$\ell(y,j) = \frac{1}{q(y)} \sum_i \lambda(i,j) q(y|i) q_i \qquad \text{(II.1)}$$

on pose alors

$$L(y,j) = \sum_i \lambda(i,j) q(y|i) q_i \qquad \text{(II.2)}$$

.../...

Définition : on appelle discriminateur théorique tout dispositif qui associe à la forme y ($\forall y \in G$) la classe $\gamma(y)$ réalisant le minimum de $L(y,j)$ ($L(y,\gamma(y)) \leq L(y,j)$ pour tout $j \in C$)

Autrement dit le discriminateur théorique (voir fig. 3) réalise une classification γ donnant le coût moyen minimum (au sens de λ). En ce sens il est optimal.

Un cas particulier fondamental :

Lorsque le coût $\lambda(i,j)$ représente les performances en pourcentage de mauvaises classifications, la fonction λ se définit par :

$$\begin{cases} \lambda(i,j) = 1 & \text{si } i \neq j, \\ \lambda(i,i) = 0. \end{cases}$$

Il est facile de voir que :

$$L(y,j) = 1 - q(y|j)q_i.$$

Le discriminateur théorique se particularise alors par une classification γ vérifiant :

$$q(y|\gamma(y))q_{\gamma(y)} \geq q(y|j)q_i, \text{ pour tout } j \in C$$

Si de plus toutes les classes sont équiprobables, une simplification supplémentaire est possible puisque γ est alors définie par :

$$q(y|\gamma(y)) \geq q(y|j), \text{ pour tout } j \in C.$$

Dans ce cas $\gamma(y)$ est la classe la plus vraisemblable.

En pratique le discriminateur théorique ne peut s'utiliser car les distributions de probabilité $q(y|j)$ et q_j ne sont pas connues.

.../...

Il est nécessaire de procéder à des observations pour les estimer explicitement ou implicitement.

Ce sont les modalités d'estimation et les variantes imposées à ce modèle théorique pour tenir compte de coûts auxiliaires tels que "implémentation" qui déterminent les discriminateurs usuellement utilisés.

II. 2. - Méthodes basées sur l'estimation des densités q(y|

Dans ce paragraphe les distributions de probabilités doivent être estimées à partir d'un échantillon.

II.2.-1- Le discriminateur empirique

Soit $G_e = \{(y_k, i_k) | y_k \in G, i_k \in C, 1 \leqslant k \leqslant n\}$ un échantillon d'observations de formes associées avec leurs élasses.

A toute forme y de G il est possible d'associer un sous-échantillon $G_e(y)$ de G_e par l'une des deux procédures suivantes :

a) Une distance étant donnée, soit ν un nombre entier positif fixé ; les éléments de G_e sont alors rangés par distances croissantes à y ; $G_e(y)$ est constitué par les ν premiers ; (en résumé $G_e(y)$ est déterminé par les ν éléments de G_e les plus voisins de y) ; on pose $r_\nu(y)$ la distance de y au νème élément.

b) Une distance étant donnée, soit r un nombre positif fixé ; $G_e(y)$ est constitué par les formes de G_e dont la distance à y est moindre que r.

Quel que soit la procédure choisie il est alors possible de choisir pour estimateur de la fonction de densité q(y|i) :

.../...

$$\hat{q}(y|i) = \frac{n_i(y)}{n_i\ V(y)},$$

où

$n_i(y)$ (resp. n_i) est le nombre des éléments de $G_e(y)$ resp. G_e) associés à la classe i,

V(y) la mesure ordinaire (ou volume) de la boule de rayon $r_\nu(y)$ ou r selon que la procédure choisie est la première ou la seconde.

Ceci peut se justifier en se basant sur les travaux théoriques de Loftsgarden et Quesenberry (1965) et de Rosenblatt (1956).

Le discriminateur empirique se définit alors par la substitution, dans la formulation du discriminateur théorique de $\hat{q}(y|i)$ à $q(y|i)$ ($y \in G$, $i \in C$). Le principe de la classification réalisé par le discriminateur empirique est donc le suivant (voir fiture 4) :

à tout y de G est associée une classe (y) :
vérifiant, pour tout j de C,

$$\Sigma_i n_i(y).\lambda(i,j) \geqslant \Sigma_i n_i(y).\lambda(i,\gamma(y)).$$

En pratique la difficulté est de choisir r ou ν. De plus la qualité des résultats dépend de la distance choisie.

II.2-2- Fonctions orthogonales et noyaux

L'estimation des fonctions de densité $q(y|j)$ peut être envisagée par d'autres méthodes non paramétriques.

Donnons le principe de deux d'entre elles. On suppose ici, pour simplifier les notations, que nous recherchons une seule fonction de densité.

.../...

Développement en fonctions orthogonales

Soit $\{\psi_k | k > 0\}$ une famille de fonctions orthogonales.

Le développement de la fonction de densité f s'écrit :

$$f = \sum_{k \geqslant 0} C_k \cdot \psi_k \qquad \text{où} \qquad C_k = \int_{-\infty}^{+\infty} \psi_k(y) f(y) dy$$

C_k est donc l'espérance mathématique de ψ_k lorsque la probabilité est définie par la densité f.
Il est alors possible d'estimer C_k par la moyenne empirique :

$$\hat{C}_k = \frac{1}{n} \sum_{\ell=1}^{n} \psi_k (y_\ell)$$

Dans cette formule $\{y_\ell | 1 \leqslant \ell \leqslant n\}$ est un échantillon correspondant à la densité f.
Nous pouvons donc estimer le développement limité à m termes f_m de f par :

$$\hat{f}_m(y) = \frac{1}{n} \sum_{k=0}^{m} \left(\sum_{\ell=1}^{n} \psi_k (y_\ell) \right) \psi_k(y)$$

Pour cette méthode une difficulté est de rendre cohérente l'erreur $|f - f_m|$ qui dépend de m et l'erreur $|f_m - \hat{f}_m|$ qui dépend de l'échantillon.

L'estimateur de Parzen (1962)

Cette méthode utilise des noyaux K, ou fonctions numériques vérifiant

a) $\sup_x |K(x)| < \infty$,

b) $\int |K(x)| dx < \infty$;

c) $\text{Lim} |xK(x)| = 0$ quand $|x| \to 0$,

d) $\int K(x)\, dx = 1$.

.../...

Si l'on se donne pour tout n un nombre h(n) positif

$$\hat{f}_n(y) = \frac{1}{n.h(n)} \sum_{k=1}^{n} K\left(\frac{y-y_k}{h(n)}\right)$$

converge vers f(y) en moyenne quadratique quand $n \to \infty$ et $nh(n) \to \infty$.

De plus, $E(\hat{f}_n(y)) \to f(y)$ quand $n \to \infty$ et $h(n) \to 0$

Il est donc possible de définir par la formule précédente un estimateur pour f(y).

Cependant, dans la pratique il est délicat de déterminer des valeurs correctes pour h(n) quand on dispose d'un échantillon fixé de n éléments.

Remarquons que Murthy a généralisé les travaux précédents aux fonctions de densité à plusieurs variables.

II.2.-3- Méthodes paramétriques d'estimation des fonctions de densité

Il est parfois possible de faire des hypothèses simplificatrices sur la famille des densités q(.|i) envisageables pour le problème posé. Ceci peut se traduire par :

$$q(.|i) \in \{f_\theta(.|i) \mid \theta \in \Theta_i\}$$

ce qui conduit à remplacer la recherche de q(.|i) par la recherche de θ. Une famille quelconque pouvant toujours se paramétrer l'intérêt ne tient pas tellement à ce que l'on a ramené le problème à une recherche de paramètre θ mais plutôt au fait que la famille $\{f_\theta(.|i) \mid \theta \in \Theta_i\}$ est choisie de manière à simplifier la détermination de q(.|i).

Lorsque la méthode est applicable q(.|i) devient une fonction connue pour tout y une fois pour toute. C'est ce qui en fait l'intérêt.

.../...

Un exemple d'un usage fréquent est celui-ci. Pour toute classe i de C les distributions sont normales $N(\mu_i,\sigma_i)$ dans $\mathbb{R}^M$. Bien qu'en pratique M est de l'ordre de la dizaine supposons pour simplifier que M = 1.

Le problème est alors de déterminer le couple

$$\theta = (\mu_i,\sigma_i) \in \mathbb{R} \times \mathbb{R}_+ = \Theta_i$$

de telle manière que :

$$(1/\sqrt{2\pi}\sigma_i) \exp\ (-(y-\mu_i)^2 / 2\sigma_i^2)$$

soit une bonne estimation de $q(y|i)$.

On choisit alors pour valeur de μ_i (resp. σ_i^2) la moyenne (resp. la variance) empirique de l'échantillon.

II.-3- Recherche d'une classification dans une famille paramétrée

Dans ce paragraphe le problème de la détermination du discriminateur est ramené à celui de la détermination d'un paramètre indexant les fonctions de classifications. Bien entendu, l'intérêt de la méthode, comme en II.2-3, tient surtout au compromis entre les simplifications que le paramétrage engendre et le fait que la classification trouvée pourrait ne pas être optimale (toutes les fonctions de classifications ne figurant pas dans la famille paramétrée).

II.-3-1- Principe

Soit donné un espace auxiliaire de réponses, qui sera ici $\mathbb{R}^d$ (où d est petit devant M), muni d'une partition $\{C_i | i \in \{0\} \cup C\}$ (on suppose que $0 \notin C$).

.../...

Pour tout élément w d'un ensemble W nous supposons connue l'application $g_w : G \to \mathbb{R}^d$ qui pourra éventuellement être considérée comme une fonction de deux arguments ($w \in W$, $y \in G$) .

Associons à g_w la fonction de classification γ_w définie par :

$$\gamma_w(y) = j \text{ si et seulement si } g_w(y) \in C_j .$$

La classe 0 est introduite afin de permettre une réponse particulière en cas d'indécision. Dans ce cas on parle souvent de rejet.

Pour tout couple (y,i) de forme y associée à la classe i introduisons alors la fonction coût $\lambda(C_i, g_w(y), w)$ qui généralise celle des paragraphes précédents.

Le coût moyen pour y fixé est alors :

$$\ell(w,y) = \sum_i \lambda(C_i, g_w(y), w) . q(i|y)$$

et le coût moyen total

$$\ell(w) = \int_G \ell(w,y) . q(y) dy$$
$$= \sum_i q_i \int_G \lambda(C_i, g_w(y), w) . q(y|i) . dy$$

Finalement nous sommes amenés à formuler le problème de la manière suivante :

trouver $\bar{w} \in W$ tel que

pour tout $w \in W$: $\ell(\bar{w}) \leq \ell(w)$

Autrement dit il faut calculer $\bar{w}$ qui minimise ℓ sur W

Naturellement il n'est pas possible d'envisager des méthodes tant que des hypothèses plus précises ne sont pas faites sur W, g_w, λ.

.../...

II.3.-2- Un exemple fondamental

L'exemple suivant permet de comprendre les généralisations introduites au paragraphe précédent :

$\mathbb{R}^d = \mathbb{R}$ ($g_w(y)$ est un nombre réel)

$C = \{1,2\}$ (il y a 2 classes)

$C_1 = \{x \in R \mid x > 0\} = \mathbb{R}_+^*$

$C_2 = \{x \in R \mid x < 0\} = \mathbb{R}_-^*$

$C_o = \{o\}$

$G = W = \mathbb{R}^M$

$g_w(y) = \sum_{\ell=1}^{M} w_\ell \eta_\ell = (w,y)$

où $y = (\eta_1, \eta_2, \ldots, \eta_M)$, $w = (w_1, w_2, \ldots, w_M)$

$$\lambda(C_1, g_w(y), w) = \begin{cases} 0 \text{ si } g_w(y) > 0 \\ |g_w(y)| / \|w\| \quad \text{si } g_w(y) \leq 0, \end{cases}$$

$$\lambda(C_2, g_w(y), w) = \begin{cases} g_w(y) / \|w\| \quad \text{si } g_w(y) \geq 0, \\ 0 \text{ si } g_w(y) < 0, \end{cases}$$

(où $\|w\| = \sqrt{\sum_\ell w_\ell^2}$) G est alors diviser en deux demi-espaces séparés par le plan $g_w(y) = 0$.

La région correspondant à $g_w(y) > 0$ est associée à la classe 1 et la région correspondant à $g_w(y) < 0$ à la classe 2.

Dans cet exemple le plan optimal rendra minimum la distance moyenne au plan des formes mal classées. La figure 5 en donne un exemple quand M = 2.

.../...

Notons que si l'on remplace la fonction coût précédente λ par λ^{α} (où $\alpha > 0$) il est possible, ainsi que le montre la figure 6, par un choix convenable de α de faire porter le coût sur la distance des éléments mal classés au plan ou simplement sur la mauvaise classification.

II.3-3- La méthode du gradient appliquée à la recherche du minimum de $\ell(w)$ à partir d'un échantillon

Nous supposons maintenant que W est un espace vectoriel de dimension M' finie que nous identifions à $\mathbb{R}^{M'}$ ou un sous-ensemble de $\mathbb{R}^{M'}$.

Soit alors f une fonction de $\mathbb{R}^{M'}$ dans $\mathbb{R}$ possédant des dérivées partielles par rapport à toutes les variables. Nous posons :

$$\text{grad } f(w) = \left(\frac{\partial f(w)}{\partial w_1}, \frac{\partial f(w)}{\partial w_2}, \ldots, \frac{\partial f(w)}{\partial w_{M'}}\right).$$

fonction à valeurs dans $\mathbb{R}^{M'}$ appelée gradient de f.

Soit alors $G_e = \{(y_k, i_k) \mid k = 1,2,\ldots,n\}$ un échantillon donné de formes classées. Nous pouvons estimer $\ell(w)$ par :

$$n.\hat{\ell}(w) = \sum_{i=1}^{n} \gamma(C_{i_k}, g_w(y_k), w)$$

Basé sur le fait que le vecteur gradient est tangent à la ligne de plus grande pente, dans le sens des valeurs croissantes, l'algorithme suivant fournit une méthode pour rechercher la valeur optimale de w :

$w_{(0)}$ est arbitraire dans W

$$w_{(m+1)} = w_{(m)} - a_{(m)} \text{ grad } \hat{\ell}(w_{(m)})$$

où $a_{(m)}$ est un facteur positif de normalisation dépendant de ℓ et $w_{(m)}$ (dans le cas général). La méthode est particulièrement adaptée quand $\hat{\ell}$ est une fonction convexe et $W = \mathbb{R}^{M'}$. Si ces hypothèses ne sont pas vérifiées il peut apparaître des difficultés. En particulier si $\hat{\ell}$ n'est pas convexe $w_{(m)}$ peut converger

.../...

vers un minimum secondaire. On y rémédie en recommençant plusieurs fois à partir de valeurs initiales $w_{(o)}$ différentes.

Si W est un sous-ensemble convexe de $\mathbb{R}^{M'}$ il est possible de généraliser la méthode précédente par l'introduction de l'opérateur projection sur W. noté π_W (voir figure 7 donne une visualisation géométrique de cet opérateur). L'algorithme devient alors

$$w_{(o)} \text{ est arbitraire dans W}$$

$$w_{(m+1)} = \Pi_W\left[w_{(m)} - a_{(m)} \text{ grad } \hat{\ell}\,(w_m)\right]$$

La figure 8 montre géométriquement l'évolution de quelques itérations.

II.3-4- Méthodes d'approximation stochastique

La fonction $\ell(w)$ est l'espérance mathématique de λ par rapport à la distribution des couples (y,j) , où j est la classe associée à l'occurrence y considérée.

Il est alors possible d'envisager plusieurs méthodes d'approximations stochastiques pour minimiser $\ell(w)$. Donnons en une, appelée "descente stochastique".

Supposons à cet effet que l'application

$$w \to \lambda(E_j,\ g_w(y), w)$$

soit différentiable et formons une suite $(y_{(m)},\ j_{(m)})$ d'observations indépendantes de formes $y_{(m)}$ associées à leurs classes $j_{(m)}$. L'algorithme suivant peut alors être envisagé quand $W = \mathbb{R}^{M'}$

$$w_{(o)} \text{ arbitraire}$$

$$w_{(m+1)} = w_{(m)} - a_{(m)} \text{ grad}\lambda(C_{j_{(m)}},\ g_{w_{(m)}}(y_{(m)}),\ w_{(m)})$$

où $a_{(m)}$ vérifie : $\sum_{m>o} a_{(m)} = \infty$, $\sum_{m>o} a^2_{(m)} < \infty$

.../...

Nous ne donnons pas ici les conditions supplémentaires qui permettent de garantir la convergence car elles sont de formulation complexe. Mais la forme de l'algorithme appelle quelques commentaires.

a) La classification est définie par $w_{(m)}$ avant la $m^{ème}$ observation ;

b) Cette classification est rajustée à la $m^{ème}$ observation en fonction de l'ancienne valeur $w_{(m)}$, de la nouvelle observation $(y_{(m)}, j_{(m)})$ de forme classée, du gradient de la fonction coût ;

c) Ce rajustement est pondéré par $a_{(m)}$ qui tend assez vite vers 0 pour que $\sum_m a^2_{(m)} < \infty$, et assez lentement pour que $\sum_m a_{(m)} = \infty$;

Autrement dit, le discriminateur se modifie en tenant compte de l'expérience $(w_{(m)})$ et de l'information nouvelle $((y_{(m)}, j_{(m)}))$ avec une pondération en faveur de la nouvelle information qui décroit vers 0 à mesure que l'expérience se prolonge.

Si cette décroissance vers 0 est trop rapide $w_{(m)}$ risque de se stabiliser trop vite sur une valeur ne tenant pas suffisamment compte des observations faites pour m grand.

Si cette décroissance est trop lente $w_{(m)}$ fluctue trop en fonction des aléas d'observation.

Il va de soi que considérations intuitives n'ont d'intérêt que dans la mesure où elles correspondent à des démonstrations rigoureuses de convergence.

Quoi qu'il en soit l'appellation "apprentissage", souvent utilisée pour ce type d'algorithme est semble-t-il justifié.

.../...

III - C O N C L U S I O N

Dans cet exposé nous nous sommes surtout attachés à dégager les idées essentielles sur les discriminateurs. Il va de soi que nous ne prétendons pas avoir été exhaustifs dans un domaine où tant de travaux ont été faits au cours de ces dernières années.

Les discriminateurs décrits ont été largement utilisés en reconnaissance de la parole. Il convient pourtant de dire qu'i apparaissent de plus en plus comme des parties d'un modèle plus général où il faut tenir compte des variations de durées, des relations reliant les éléments d'un même mot, puis des mots d'une même phrase.

Mais ceci, loin de leur enlever de l'importance, en limitant leur utilisation à leur Domaine de validité, contribue à leur conserver un caractère fondamental.

-:-:-:-:-:-:-:-:-

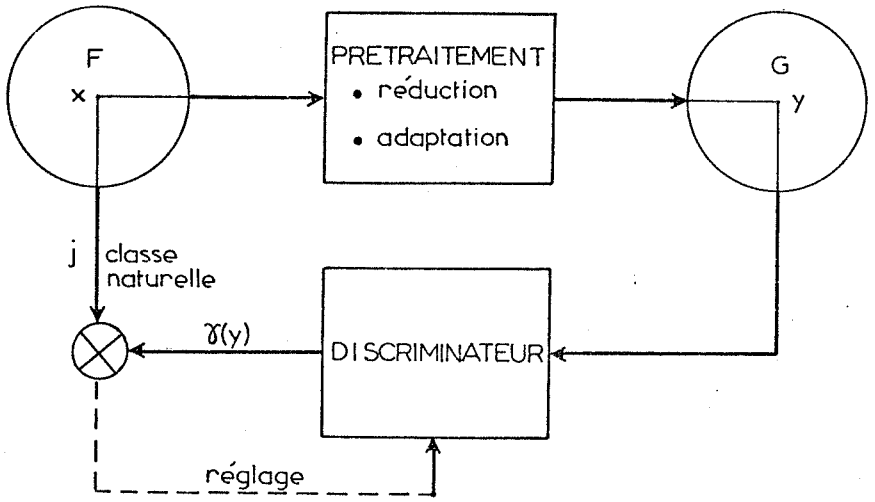


Figure 1. Modèle de reconnaissance des formes dans l'approche statistique.

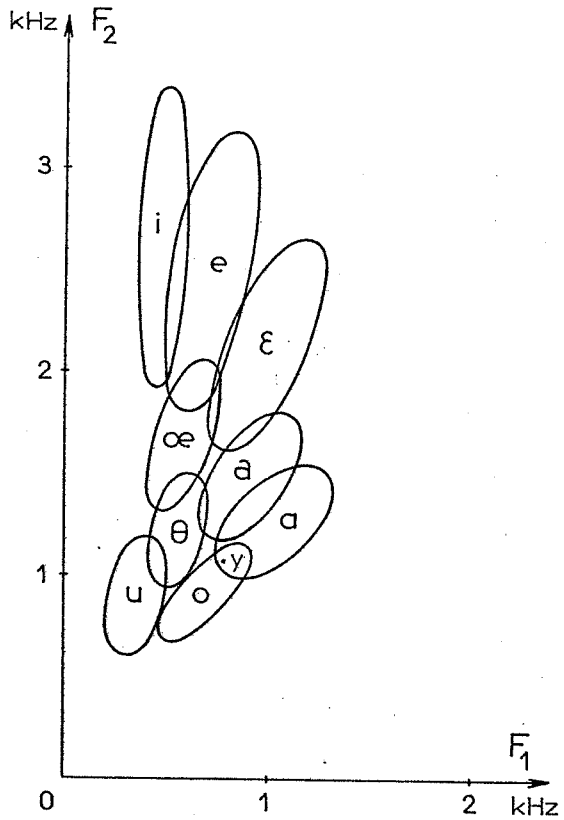


Figure 2. Distribution des voyelles en fonction des formants F_1 et F_2 (d'après BARNEY et PETERSON).

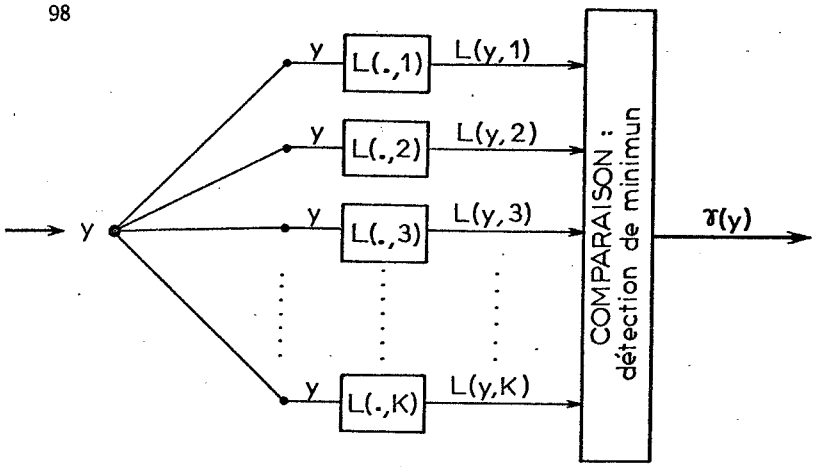


Figure 3. Schéma de principe du discriminateur théorique.

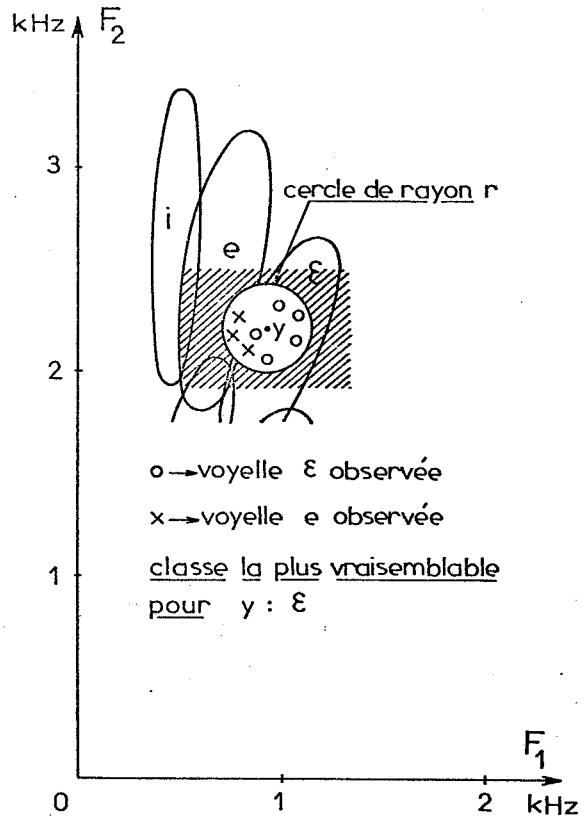


Figure 4. Le discriminateur empirique.

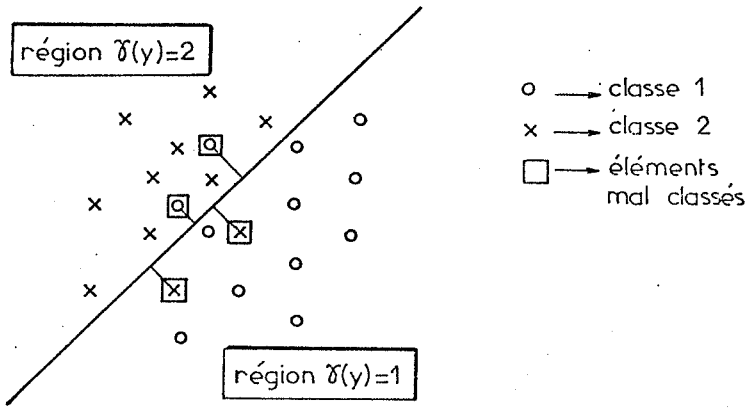


Figure 5. Séparation rendant minimale la distance moyenne des éléments mal classés au plan(=droite dans $\mathbb{R}^2$) de séparation.

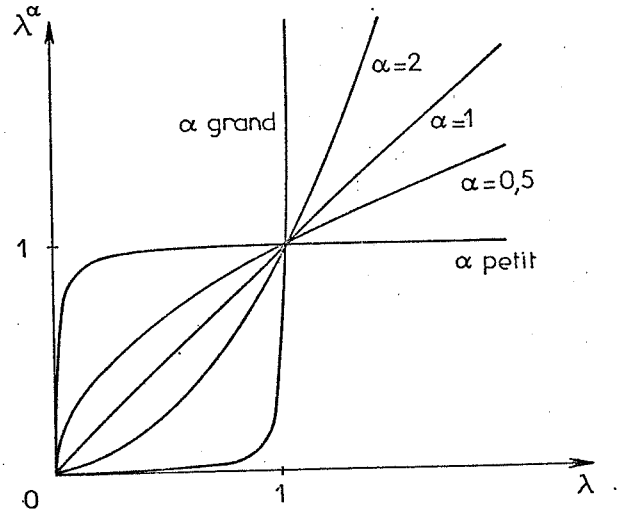


Figure 6. Lorsque α est grand les erreurs au-dessous de 1 donnent un coût λ^α négligeable. Inversement, quand α petit, et dès qu'il y a erreur, celle-ci est voisine de 1. On peut en remplaçant λ^α par $(\lambda/C)^\alpha$ faire jouer à C(C>0) le rôle joué par 1 ici.

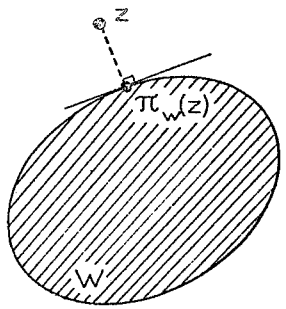


Figure 7. Projection de z sur W.

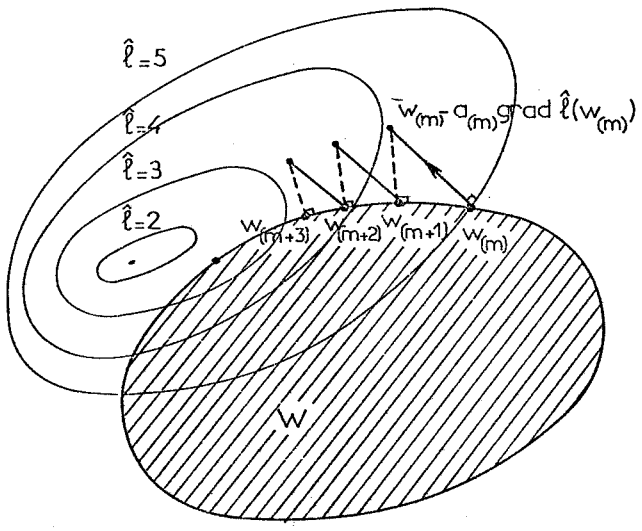


Figure 8. Principe de la méthode du gradient avec projection sur W.

COMMENTAIRES BIBLIOGRAPHIQUES

-:-:-:-:-:-:-:-:-:-:-:-:-:-:-

Au cours de la rédaction de cet exposé nous avons été amenés à consulter plus particulièrement :

1) - Cover, T.M., "A hiérarchy of probability density fonction estimates",

dans "Frontiers of Pattern Recognition" édité par Satosi Watanabe - Académie Press (1972) pp 83-98

2) - Harry, C.A., "Introduction to Mathematical Techniques in Pattern Recognition",

- Wiley Interscience - (1972)

On peut y trouver une bibliographie très complète.

-:-:-:-:-:-:-:-:-:-:-:-:-:-:-

Discussion.

C. ROCHE : Vous avez parlé d'apprentissage du sujet d'une méthode itérative de détermination des coefficients d'une fonction linéaire. L'opérateur utilisé est un opérateur linéaire dont la présence n'a pas été décidée automatiquement, mais par le programmeur.

Par ailleurs, est-il suffisant de ne considérer que des opérateurs linéaires, ou même arithmétiques pour le calcul des distances inter-éléments utilisées en classification. En particulier qu'advient-il lorsque les variables explicatives sont alphanumériques, ou plus généralement ont des états sur lesquels n'existe aucune relation d'ordre significative ?

G. PERENNOU : J'ai parlé d'apprentissage dans le sens où le discriminateur s'adapte en tenant compte du passé qui est résumé en $w_{(m)}$ et des observations nouvelles. De plus, cette adaptation correspond à un comportement qui s'optimise. Le point important reste toutefois que nous avons une méthode pour déterminer le discriminateur optimal.

En ce qui concerne les problèmes de reconnaissance des formes décrites d'une manière non numérique, qui ne s'intègrent pas dans le modèle statistique, bien que très importants, ils n'ont pu être intégrés à cet exposé.

Il est possible de dire ceci : toute étude de problèmes de reconnaissance des formes est précédée d'une analyse dont le but est de déterminer le modèle adapté à sa résolution.

Il est clair que le modèle statistique n'est pas toujours le modèle le plus adapté et quelquefois il n'est qu'un des éléments du modèle retenu.

Cependant, votre question sous-entend, sans doute, que l'on sait dire beaucoup de choses sur les modèles statistiques et qu'il reste beaucoup de travail à faire sur les autres : je suis d'accord avec ce point de vue.

PERCEPTION AUDITIVE ET ANALYSE SPECTRALE.

Résumé

Après avoir rappelé le caractère essentiellement fréquentiel de l'audition, l'auteur propose un modèle de l'oreille composé de deux parties, l'"oreille physique" et les organes collecteurs nerveux. L'oreille physique est analogue à un banc de filtres dont les réponses impulsionnelles inversées dans le temps sont représentables par une "fonction de mémoire", fonction de l'instant de l'écoute et de la fréquence. L'oreille physique qui reçoit un signal d'une forme temporelle donnée le "lit"sous la forme d'un signal temporel différent comprenant un paramètre fréquentiel - le spectre variable en fonction de l'instant de l'écoute - essai de synthèse réduisant à néant l'opposition traditionnelle entre la forme temporelle d'un signal et son analyse fréquentielle.

Ensuite, la discrétisation des maximums spectraux par les terminaisons nerveuses collectrices expliquerait l'effet de masque fréquentiel, qui ne nuit en rien à la résolution temporelle de l'audition.

Enfin, l'effet de masque, en augmentant fortement la résolution fréquentielle de l'oreille, lui permet d'atteindre le débit d'information réel de la fonction auditive.

Une formulation mathématique des courbes de masque, déduite de la forme de la fonction de mémoire, est proposée.

Summary

The author points out the essential frequency aspect of the hearing process and suggests a model of the ear with two components : the "physical" ear and the nervous collecting organs. The physical ear is similar to a filterbank whose impulse responses, inverted in the time scale, can be represented by a "memory function" which is function of the frequency and the instant of listening. The physical ear receiving a signal of a given temporal form "reads" it like a different time signal, with a frequency parameter, which is the running spectrum.

This constitutes a attempt to abolish the traditional dichotomy between the time domain and the frequency domain representation of a signal.

The discretisation of the spectral maxima by the nervous collecting organs would then explain the masking effect in the frequency domain, which does not affect the temporal resolution in the hearing process.

Finally, since it increases greatly the frequency resolution of the ear, the masking effect allows the auditive process to reach its information rate.

A mathematical formalisation of the masking curves based on the memory function form is as well proposed.

Jean BOSQUET
(Université Libre de Bruxelles)

PERCEPTION AUDITIVE ET ANALYSE SPECTRALE

Jean BOSQUET

1. INTRODUCTION

L'hypothèse très féconde émise dès 1968 par T. KORN, l'identification de la courbe de masque de l'oreille et du spectre d'un son pur dépourvu de durée est à l'origine de la présente communication.

Le rôle de l'analyse spectrale en acoustique se justifie par le fait qu'une certaine description fréquentielle des signaux correspond mieux à leur effet audible que la pure description temporelle oscillographique. Cette opinion a d'ailleurs trouvé sa première formulation dans la théorie auditive de HELMOLTZ qui a proposé de considérer l'oreille comme un analyseur spectral composé de résonateurs acoustiques, seuls dispositifs sélectifs imaginables à cette époque.

Les progrès de l'électronique ont permis la construction d'analyseurs spectraux fort évolués, aussi bien analogiques que digitaux, mais cette perfection technologique a mis en évidence la faiblesse du principe même de l'analyse spectrale, et, notamment, son faible débit d'information.

On a constaté que si les analyseurs physiques fournissent des données satisfaisantes concernant les signaux quasi-stationnaires ou statistiquement stationnaires (tels que certains bruits mécaniques), ils se sont révélés incapables de traiter en "temps réel" les signaux de communication à grand débit d'information : parole et musique. Cette faiblesse de l'analyse spectrale a été soulignée par certains spécialistes de l'analyse et de la synthèse de la parole, qui ont constaté qu'il est impossible de reconstituer fidèlement la parole à partir des données fournies par les analyseurs spectraux actuels (ATAL et HANAUER 1971),

Le fait que le débit d'information de signaux observés au travers d'analyseurs est de loin plus faible que celui des mêmes signaux captés auditivement, a amené certains chercheurs à supposer que, contrairement à la théorie de HELMOLTZ, le fonctionnement de l'oreille devait s'appuyer sur un procédé d'observation qui puise l'information, par exemple, directement à partir de la forme temporelle du signal, en refusant à

l'oreille tout rôle d'analyseur spectral. Dans cette conception, le problème est déplacé en sa totalité au niveau du système nerveux central, où il échappe jusqu'à présent à toute formulation scientifique.

Je crois donc qu'en premier lieu, il n'est pas inutile de rappeler un certain nombre de faits incontestables, basés sur des tests auditifs, capables de mettre en évidence le principe de l'observation des signaux par l'oreille.

2.1. La sensation de timbre et la phase des harmoniques.

En 1863, HELMOLTZ a formulé sa célèbre loi expérimentale : "le timbre de la portion musicale d'un son dépend seulement du nombre et de l'intensité des sons partiels, mais non de leurs différences de phase."

Il existe donc une infinité de formes temporelles différentes qui produisent exactement la même sensation globale, représentée par le même spectre. Cette loi est confirmée depuis toujours par la pratique musicale. Citons notamment les jeux de mutation des orgues, utilisés depuis le XVème siècle au moins. Remarquons que si le son a une durée suffisante, une oreille exercée parvient à déceler la présence de tel ou tel harmonique dans l'ensemble qui lui est présenté : il s'agit là d'une opération qui s'effectue au niveau du cerveau, grâce à un effort d'attention, sans que la sensation globale soit modifiée. On peut aussi composer un timbre en ajoutant successivement à un fondamental un certain nombre d'harmoniques (démonstration) que l'on distingue aisément grâce à cette procédure, mais si, dès que le timbre est constitué, on joue d'autres notes, la globalisation de la sensation est évidente. (démonstration)

Au § suivant, nous examinons le cas de signaux brefs.

2.2. La mesure de la fréquence dans le domaine temporel et spectral.

On sait que la notion de fréquence est différente selon qu'un signal est observé au travers d'un analyseur spectral, ou à l'aide de sa forme temporelle supposée connue ou accessible. Dans ce dernier cas, la mesure de la fréquence consiste éventuellement en un comptage d'alternances ou de passages par zéro, tandis que sa mesure dans le domaine fréquentiel est basée sur la localisation du sommet spectral sur l'axe des fréquences. Dans le cas de signaux qui ne sont pas très brefs, les résultats obtenus par ces deux méthodes

sont identiques, mais il n'en est pas ainsi dans le cas de signaux de relativement courte durée.

Considérons par exemple un signal

$$s(t) = \Pi(t) \cos(2\pi \nu_0 t + \varphi) \qquad (1)$$

constitué d'une enveloppe rectangulaire $\Pi(t)$ de durée τ, lu par un analyseur effectuant une simple transformée de FOURIER. Le maximum spectral de ce signal est fonction de φ. On calcule aisément que si la fenêtre rectangulaire comprend un nombre entier N d'alternances, soit si l'on a l'égalité

$$\nu_0 \tau = N \qquad (2)$$

le maximum spectral se situe, lorsque $\varphi = 0$, à une fréquence ν_1 inférieure à ν_0, et, lorsque $\varphi = \pi/2$, à une fréquence ν_2 supérieure à ν_0. On a approximativement

$$\Delta\nu = \nu_2 - \nu_1 = 3/\pi^2 N\tau \qquad (3)$$

Pour que ν_1 et ν_2 soient audibles, il faut que N ait une valeur suffisante, mais pour que $\Delta\nu$ ait une valeur appréciable, il faut que τ et N soient assez petits. Un compromis consiste à poser $N = 6$, $\tau = 20\,ms$, soit $\nu_0 = 300\,Hz$. On a alors $\Delta\nu = 2{,}5\,Hz$.

La démonstration se heurte à deux difficultés : la première, purement technique, est la difficulté de produire un signal tel que φ vale exactement 0 ou $\pi/2$. La seconde tient au fait que (ceci sera précisé plus loin) l'oreille n'effectue pas une simple analyse de FOURIER, quoique pour un signal bref comme celui-ci on ne soit pas loin de la réalité. Nous avons enregistré un signal comportant une phase φ aléatoire : avec un peu de chance, on perçoit des changements de coloration tonale du signal, ce qui prouve que l'oreille y est sensible, alors que la période T_0 du signal mesurable en fonction des passages par zéro de la sinusoïde vaut toujours $T_0 = 1/\nu_0$. (démonstration) (Celle-ci devrait être effectuée dans une chambre sourde, et non dans un local réverbérant.)

Cette expérience prouve également que la phase d'un des harmoniques d'un son timbré très bref a théoriquement une influence sur la hauteur perçue. HELMOLTZ s'en est rendu compte lorsqu'il écrit que sa loi rappelée en 2.1. s'applique à la "portion musicale" d'un

d'un son, ce qui implique une durée suffisante du signal.

2.3. Les sons résultants

Une objection que l'on formule parfois à la théorie spectrale de l'audition a pour origine le fait que, lors d'une communication téléphonique par exemple, on entend des fréquences graves qui sont inférieures à la limite inférieure de la bande passante du canal : c'est perdre de vue le fait que l'oreille moyenne et l'oreille interne possèdent une certaine non-linéarité. L'oreille reconstitue ainsi le fondamental de sons timbrés dont on ne lui transmet que des harmoniques : il s'agit du phénomène bien connu des sons différentiels, découverts dès 1700 par des musiciens (SORGE, organiste allemand, TARTINI, violoniste italien célèbre), phénomène expliqué pour la première fois également par HELMOLTZ et appliqué notamment, de nos jours dans la réalisation des "basses acoustiques" des orgues d'encombrement et de prix réduits. Démonstration : on émet un son de 80 Hz et son octave de 160 Hz, puis l'on ajoute la quinte du fondamental (120 Hz) : le son différentiel de 160-120 = 120-80 = 40 Hz apparaît immédiatement, à l'octave inférieure du son le plus grave de 80 Hz donné. Un sonomètre à bandes passantes d'1/3 d'octave permet de s'assurer du fait que ce son différentiel n'a pas d'existence objective.

La perception de ces sons différentiels dûs à la non-linéarité de l'oreille confirme donc la nature fréquentielle des sensations auditives, mais nous allons voir que loin de s'opposer, la forme temporelle d'un signal et son analyse spectrale ne sont que deux maillons d'une chaîne unique.

3.1. Un modèle simple de l'oreille physique. Sa fonction de mémoire.

Dans ce qui suit, nous présentons une hypothèse de travail volontairement simple, en comparaison par exemple du modèle mathématique de la membrane basilaire de FLANAGAN . Elle s'inspire d'ailleurs des conceptions déjà anciennes de cet auteur, ainsi que de celles de FANO (1951) et de NOLL (1964),
et se résume comme suit :

a) l'oreille se compose de deux parties. La première, que faute d'une meilleure expression, nous appelerons l'oreille physique, est l'ensemble des organes (oreille externe, oreille moyenne et une partie de l'oreille interne) dont au peut considérer qu'il fonctionne comme un banc de filtres dont le nombre se

situerait entre 5.000 et 25.000 (PIMONOV 1962)
Cet ensemble est suivi de la seconde partie, constituée d'organes collecteurs dont le rôle sera précisé plus loin.

b) la réponse percussionnelle du filtre de numéro d'ordre ℓ est de la forme (+)

$$h(t) = m(t) \begin{matrix} \sin \\ \cos \end{matrix} \omega_\ell t \qquad (4)$$

où m (t) est une fonction qui est nulle si $t \leq 0$, qui croît ensuite rapidement, passe par un maximum puis tend vers zéro lorsque t croît. Pour fixer les idées, nous supposerons dans ce qui suit que cette fonction est de la forme

$$m(t) = M t^n e^{-\alpha_\ell t} \qquad (5)$$

où M est une constante, le décrément α_ℓ une fonction de la fréquence $\omega_\ell / 2\pi$ et où n est un entier. α_ℓ et n sont à déterminer expérimentalement, nous verrons plus loin comment. La forme d'enveloppe (5) correspond approximativement au cas où chaque filtre du banc est composé de n + 1 circuits indépendants en cascade, accordés sur la même fréquence mais d'amortissements individuels éventuellement différents. Ce point sera également précisé plus loin.

c) on suppose que chaque filtre possède deux "sorties" correspondant aux réponses percussionnelles m (t) $\sin \omega_\ell t$ et m (t) $\cos \omega_\ell t$, mais ceci est une hypothèse d'importance secondaire.

d) En conséquence, si un signal descriptible par une fonction quelconque du temps s (t) est capté par le banc de filtres, les "lectures" effectuées aux sorties de chacun d'eux à un certain instant θ seront données par les expressions

$$\tilde{s}_c(\omega_\ell, \alpha_\ell, \theta) = \int_{-\infty}^{\theta} s(t)\, m(\theta - t) \cos \omega_\ell (\theta - t)\, dt \qquad (6)$$

$$\tilde{s}_s(\omega_\ell, \alpha_\ell, \theta) = \int_{-\infty}^{\theta} s(t)\, m(\theta - t) \sin \omega_\ell (\theta - t)\, dt \qquad (7)$$

d'où l'on déduit

$$\tilde{s}(\omega_\ell, \alpha_\ell, \theta) = \tilde{s}_c + j\tilde{s}_s = e^{j\omega_\ell \theta} \int_{-\infty}^{\theta} s(t)\, m(\theta - t)\, e^{-j\omega_\ell t}\, dt \qquad (8)$$

ainsi que la densité de puissance spectrale, somme des carrés des lectures (6) et (7) :

(+) Dans ce qui suit, nous supposons que les amplitudes des signaux sont suffisamment faibles pour que le modèle puisse être considéré comme linéaire.

$$\left| \tilde{s}(\omega_\ell, \alpha_\ell, \theta) \right|^2 = \left| \mathcal{F}[s(t) m(\theta - t)] \right|^2 \qquad (9)$$

où le symbole $\mathcal{F}$ désigne l'opérateur transformé de FOURIER. L'expression (8) est le produit de convolution de s (t) par m (t) $\exp(j\omega_\ell t)$, tandis que $m(\theta - t)$, enveloppe de la réponse percussionnelle du filtre considéré inversée dans le temps, est une fenêtre temporelle que nous désignons par le nom de fonction de mémoire du filtre.

Le mécanisme de la partie physique de l'écoute peut donc être représenté par le graphique de la fig. 1. A chaque instant postérieur au début de l'écoute de s (t), le spectre, variable en fonction de θ , est celui du produit de convolution de s (t) par m (t). Il est important de souligner ici que les opérations (6) à (9) sont de simples convolutions qui font correspondre à une fonction du temps quelconque s (t) une série d'autres fonctions temporelles $\tilde{s}(\omega_\ell, \alpha_\ell, \theta)$ dont l'ensemble constitue la réponse du système considéré. Chacune de ces fonctions temporelles est toutefois fonction d'un paramètre d'amortissement α_ℓ ainsi que d'un paramètre fréquentiel ω_ℓ qui confère automatiquement à la réponse la forme d'une transformée de FOURIER, impliquant donc la résolution fréquentielle du système. Soulignons également le fait que ce mécanisme est strictement causal : la fonction de mémoire n'est différente de 0 que pour les époques antérieures à l'instant d'écoute θ .

On pourrait, bien entendu, imaginer un modèle tel que la réponse impulsionnelle des éléments ne comporte pas de facteur $\sin\omega_\ell t$ ou $\cos\omega_\ell t$ Un tel système, composé d'éléments apériodiques, ne présenterait aucun intérêt, car il serait incapable de rendre compte de la résolution fréquentielle de l'oreille qui est une incontestable réalité. Comme, d'autre part, il me paraît difficile d'admettre que l' "oreille physique" n'existe pas, le modèle présenté m'a paru en définitive difficile à contester, au moins dans son principe, à savoir l'existence d'une fonction de mémoire du type général (4).

Cela étant établi, une conséquence capitale s'en déduit : la soi-disant opposition entre la forme temporelle d'un signal et son analyse fréquentielle n'existe plus : un signal d'une forme temporelle donnée est nécessairement lu par l'oreille sous la forme d'un signal temporel différent comprenant un paramètre fréquentiel, - le spectre variable en fonction du temps d'écoute -, base de la fonction auditive, dont je souligne le caractère bi-dimensionnel : θ et ω_ℓ .

Remarquons également que ce spectre (9) n'a rien à voir avec la transformée de FOURIER qui est à base du théorème de convolution classique

$$\mathcal{F}(s_1).\mathcal{F}(s_2) = \int_{-\infty}^{+\infty} e^{-j\omega\theta} d\theta \int_{-\infty}^{+\infty} s_1(t) s_2(\theta - t)\, dt = \mathcal{F}(s_1 * s_2) \qquad (10)$$

qui s'étend à la totalité de la durée des signaux $s_1(t)$ et $s_2(t)$ considérés. Dans (10), la variable θ , après l'intégration, a disparu. On notera cependant que la transformée inverse

$$s_1 * s_2 = \mathcal{F}^{-1}\left[\mathcal{F}(s_1)\,\mathcal{F}(s_2)\right] \qquad (11)$$

constitue, dans certains cas favorables, un procédé de calcul, classique d'ailleurs, des intégrales de (6) et (7), à condition de connaître les transformées de FOURIER (ou de LAPLACE) de s_1 et s_2 , donc ici de s (t) et de m (t) $\begin{matrix}\sin \\ \cos\end{matrix}\omega_1 t$

On notera également que si l'on peut considérer que le signal s (t) a commencé d'exister à un instant identique pour tous les filtres du système, qu'el'on prend pour origine du temps, la limite inférieure des intégrales (6), (7) et (8) sera 0 et non -∞. Dans le cas plus général où tous les filtres ne seraient pas atteints par le signal au même instant, cette limite inférieure serait un nombre $t(\omega_\ell)$ fonction de ω_ℓ , ce qui permettrait de tenir compte de la vitesse finie de propagation le long de la membrane basilaire.

3.2. Forme et durée de la fonction de mémoire.

Si la fonction de mémoire était infiniment brève (mesure de DIRAC) la lecture serait identique à la fonction s (t) et correspondrait à l'observation temporelle oscillographique du signal, sans aucune résolution fréquentielle.

A l'autre extrême, une mémoire infinie (réponse impulsionnelle en fonction escalier) donnerait une résolution fréquentielle égale à celle de la trasformée de FOURIER limitée à l'instant de lecture θ , mais ne permettrait aucune résolution temporelle : cette mémoire est purement accumulative, ce qui ne présente qu'un intérêt médiocre, à moins d'interrompre volontairement l'opération à des instants à fixer en fonction de critères extérieurs. Une telle procédure ne se prête donc pas à la réception de communications auditives composées de messages successifs et est irréalisable par un mécanisme autonome.

Une mémoire finie (cas intermédiaire) ou, ce qui revient au même pratiquement, facteur d'une exponentielledécroissante comme en (5) comporte une résolution temporelle automatique. Elle fournit

cependant une résolution fréquentielle nécessairement moins bonne que la mémoire infinie. Cela apparaît clairement dans le cas particulier simple de l'analyse d'un signal sinusoïdal stationnaire $A\exp[j\omega_0 t]$ (ce qui revient à ne pas tenir compte des termes en $\omega_0 + \omega$ dans le spectre, mais ces termes sont généralement négligeables). L'application de la formule (9) donne (en omettant, à partir d'ici, l'indice 1-

$$|\tilde{s}(\omega)| = AM\frac{n!}{\alpha^{n+1}}\left[1 - \frac{n+1}{2}\left(\frac{\omega-\omega_0}{\alpha}\right)^2\right] \qquad (12)$$

soit, pour un n donné, un spectre d'autant plus large que α est grand. On remarque que ce spectre ne présente qu'un seul maximum (absence de rebondissements) ce qui n'est pas le cas de ceux obtenus à l'aide de fonctions de prélèvement ou de "segmentation" de durée limitée (HANNING, HAMMING, etc.) ou présentant des discontinuités analytiques (LANDERCY)

3.3. Analyse de signaux de durées diverses.

Dans le but de mettre en évidence certaines propriétés fondamentales de l'oreille, il est nécessaire d'examiner comment une fonction de mémoire du type (5) "écoute" des signaux de durées diverses.

Pour fixer les idées et ne pas être conduits à des calculs trop lourds, considérons un signal du type

$$s(t) = A\,t^n e^{-\beta t + j\omega_0 t} \qquad (13)$$

Sa durée pratique sera fonction de la valeur de β.
Lorsque $\beta > \alpha$ il s'agit d'un signal relativement bref, et inversement si $\beta < \alpha$.

Appliquons la formule générale (8) en posant $u \equiv t/\theta$:

$$\tilde{s}(\omega,\theta) = AM\theta^{2n+1}e^{[j(\omega-\alpha)]\theta}\int_0^1 u^n(1-u)^n e^{-(\beta-\alpha)\theta u - j(\omega-\omega_0)\theta u}\,du \qquad (14)$$

et examinons d'abord ce qui se passe tout au début de la rencontre du signal et de la fonction de mémoire. En développant en série l'exponentielle figurant dans l'intégrale, on voit que pour les très petites valeurs de θ, le spectre est très large : c'est un bruit blanc croissant proportionnellement à θ^{2n+1} mais, progressivement, le maximum spectral $\omega = \omega_0$ s'accuse et, aux environs de ce maximum, le spectre vaut, si $\beta > \alpha$, aux termes du 3e ordre près,

$$|\tilde{s}(\omega,\theta)| \simeq AM\frac{(n!)^2}{(2n+1)!}\,\theta^{2n+1}e^{-\alpha\theta}\times$$

$$\left[1-(\beta-\alpha)\theta+\frac{4n+7}{4(2n+3)}(\beta-\alpha)^2\theta^2-\frac{(\omega-\omega_o)^2\theta^2}{4(2n+3)}\right]^{\frac{1}{2}} \qquad (15)$$

Lorsque $\beta<\alpha$, cette expression êut être écrite en intervertissant α et β, en raison de la symétrie de (9) (y remplacer u par $1-\nu$)

Lorsque θ croît, l'existence du facteur θ^2 de $(\omega-\omega_o)^2$ montre que le spectre se rétrécit progressivement ou, en d'autres termes, que le maximum spectral s'accuse.

Lorsque θ croît de plus en plus, cette approximation (15) n'est plus valable. L'intégration de (14) n'est pas pratique dans ce cas, aussi nous retournons à la formule générale (8) et prenons en considération le fait que, dans les deux hypothèses envisagées - signal bref ou signal long - l'une des deux fonctions s (t) ou m $(\theta-t)$ varie vite tandis que l'autre varie lentement, à partir d'une valeur suffisante de θ, soit $t_{s_1}=n/\beta$ l'époque à laquelle l'amplitude du signal est maximum (fig. 2). Développons $m(\theta-t)$ en série de TAYLOR à partie de $t = t_{s_1}$. Nous obtenons, en première approximation $[m(\theta-t)\simeq m(\theta-t_{s_1})]$,

$$\tilde{s}(\omega,\theta)\simeq m(\theta-t_{s_1})e^{j\omega\theta}\int_0^{\theta}s(t)e^{-j\omega t}\,dt \quad \forall\, t>t_{s_1}+\theta_i \qquad (16)$$

où θ_i est une valeur suffisante de θ, par exemple aux environs du premier point d'inflexion de $m(\theta-t)$. On voit par (16) que dans le cas envisagé du signal bref, la forme enveloppe de la f. de m. n'influence que très peu la forme du spectre du signal mais est pratiquement proportionnelle à son amplitude. On obtient la forme du spectre aux environs de son maximum $\omega-\omega_o$ en remplaçant s (t) dans (16) par l'expression (13) et en développant $exp[-j(\omega-\omega_c)t]$ jusqu'au terme du second ordre. A partir de l'instant (environ $3t_{s_1}$) où $exp(-\beta\theta)$ est pratiquement négligeable, on obtient, avec $\delta\equiv\omega-\omega_o$,

$$|\tilde{s}(\omega,\theta)|\simeq AM\frac{n!}{\beta^{n+1}}(\theta-t_{s_1})^n e^{-\alpha(\theta-t_{s_1})}\left[1-\frac{n+1}{2}\left(\frac{\delta}{\beta}\right)^2\right] \qquad (17)$$

soit un spectre d'autant plus large que β est grand, ce qui est d'ailleurs trivial.

On peut donc résumer ce qui précède en disant qu'un signal plus bref que la f. de m. est en quelque sorte "absorbé" par celle-ci en fournissant rapidement un spectre qui reste large, d'autant plus faiblement coloré que le signal est bref.

L'amplitude maximum du spectre est atteinte à peu près à l'instant

$$\theta_m = t_1 + \frac{n}{\alpha} = n\left(\frac{1}{\alpha} + \frac{1}{\beta}\right) \qquad (18)$$

Auditivement, l'écoute effectuée avec $n = 2$, α de l'ordre de 120 à 140 s^{-1} , à la fréquence de 1000 Hz, montre que la sensation est unique, c'est-à-dire ne comporte aucune impression de durée. (démonstration) On peut donc dire que la durée pratique de la fonction de mémoire de l'oreille est l'"épaisseur du présent", expression imagée due à Abraham A. MOLES (1960).

Examinons maintenant ce qui se passe lorsque le signal est relativement long ($\beta < \alpha$). Dans ce cas, la fonction m (t) varie plus vite que s (t) et l'on obtient, par le même procédé de calcul que ci-dessus :

$$|\tilde{s}(\omega,\theta)| = AM\frac{n!}{\alpha^{n+1}}(\theta - t_m)^n e^{-\beta(\theta - t_m)}\left[1 - \frac{n+1}{2}\left(\frac{\delta}{\alpha}\right)^2\right] \qquad (19)$$

où $t_m = n/\alpha$ est l'abscisse du maximum de la fonction $t^n \exp(-\alpha t)$ (fig. 3). Il en résulte que l'on devrait entendre, après le transitoire du début, c'est-à-dire dès que l'enveloppe du signal a pratiquement absorbé la fonction de mémoire, un son dont le spectre à la forme (19) qui, à l'enveloppe près, est identique à (12). Quoique ce spectre ait une certaine largeur, la sensation auditive est celle d'un son pur qui dure en croissant puis en s'éteignant progressivement. Avant de tirer la conclusion de cette constatation, voyons ce qui se passe lorsque le signal et la f. de m. ont des durées identiques ($\alpha = \beta$)

4.1. Un analyseur spectral exceptionnel : l'oreille

Appliquons la formule générale (14) dans l'hypothèse d'un signal de même durée que la f. de m. On obtient, comme en (15) mais avec $\alpha = \beta$, aux environs du maximum spectral,

$$|\tilde{s}(\omega,\theta)| \simeq AM\frac{(n!)^2}{(2n+1)^2}\,\theta^{2n+1} e^{-\alpha\theta}\left[1 - \frac{\delta^2\theta^2}{8(2n+3)}\right] \qquad (20)$$

formule valable quel que soit θ tel que $\delta^2\theta^2 \ll 8(2n+3)$

Le maximum spectral ($\delta = 0$) atteint sa plus grande amplitude à l'instant qui rend maximum le facteur d'enveloppe $\theta^{2n+1} e^{-\alpha\theta}$, soit

$$\theta_M = \frac{2n+1}{\alpha} \quad . \tag{21}$$

A cet instant, le spectre possède, aux environs de son maximum, la forme

$$|\tilde{s}(\omega, \theta_M)| \div 1 - \frac{(2n+1)^2}{8(2n+3)} \left(\frac{\omega - \omega_0}{\alpha} \right)^2 , \tag{22}$$

tandis que la sensation auditive est celle d'un son pur instantané, donc sans durée perceptible. Les essais effectués avec les valeurs numériques signalées ci-dessus confirment cette constatation. (Démonstration).

Le fait que, dans les deux cas précédents, on entend un son pur implique que ces spectres lus par l'oreille sont de loin plus étroits que leur largeur physique, le cas extrême étant celui de la formule (22). Ce résultat, assez surprenant, est dû au procédé de discrétisation par l'effet de masque. Cet effet ne peut être que la conséquence de l'action de la deuxième partie de l'oreille, au delà de ce que nous avons appelé l'oreille physique, il semble être dû à un phénomène d'inhibition, par les terminaisons nerveuses les plus sollicitées, des terminaisons voisines. (Signalons ici que LESUISSE (1970), sous la direction de KORN, a réalisé un discrétiseur élémentaire analogique au moyen d'amplificateurs à gain variable, fonctionnant aux fréquences 1150 et 1200 Hz.)

Ce processus n'empêche pas l'oreille de conserver la durée finie de sa fonction de mémoire, dont le rôle a été esquissé ci dessus et dont la mesure fera l'objet du paragraphe suivant.

Il résulte de ce qui précède que la forme de la courbe de masque, pour une fréquence donnée, aux environs immédiats de son maximum, serait donnée par (22).

4.2. Mesure de la longueur de la fonction de mémoire de l'oreille.

La longueur de la fonction de mémoire peut être étudiée à partir de la résolution temporelle à la sortie de l'analyseur, en utilisant des signaux de forme et de durée connue. Considérons par exemple une paire de stimuli élémentaires décalés dans le temps :

$$s_a(t) \div t^n e^{-\alpha t} \cos \omega_0 t \tag{23}$$

$$s_b(t) \div (t-\tau)^n e^{-\alpha(t-\tau)} \cos \omega_0 t \tag{24}$$

représentés à la fig. 4, ayant une valeur telle que ces stimuli écoulés individuellement ne donnent pas l'inpression de durée. L'expérience consiste à comparer les résolutions temporelles visuelle (oscillographique) et auditive, soit, dans chaque cas, l'apparition de deux maximums successifs.

Le signal composé de la somme des stimuli (23) et (24) présente visuellement deux maximums successifs si τ est supérieur à l'instant n/α où le premier stimulus atteint son maximum d'amplitude. Auditivement, le maximum d'amplitude du premier stimulus est atteint à l'instant défini par (21). Il en résulte que lorsque $n/\alpha < \tau < (2n+1)/\alpha$, la sensation auditive est encore toujours unique, ce qui constitue l'effet de masque temporel de l'audition. Ce n'est que lorsque τ est supérieur à θ_n que l'impression auditive est constituée de deux coups ou d'un effet de répétition. (Démonstration : son de 1000 Hz, n = 2, $\alpha = 133\,s^{-1}$, $t_1 = 2/\alpha = 15$ ms, $\theta_n = 5/\alpha = 37{,}5$ ms. On a choisi τ_1 (séparation visuelle) = 20 ms et τ_2 (séparation auditive) = 50 ms

La réverbération de la salle atténue malheureusement fortement l'effet escompté, très net en chambre sourde).

Des expériences systématiques, portant sur des signaux élémentaires individuels ou sur des stimuli répétés devraient permettre d'évaluer α à chaque fréquence utilisée, éventuellement pour d'autres valeurs de n.

A cet égard, on sait que FLANAGAN (1962) a estimé que la f. de m. de la <u>membrane basilaire</u> pourrait être de la forme $t^2 \exp(-\omega/2.t) \sin \omega t$, comportant un amortissement très important ($\omega/2 = 3140\,s^{-1} \simeq 1$ kHz), donc très supérieur en ordre de grandeur, à la valeur de α estimée ci-dessus. Cette constatation impliquerait qu'au delà de la membrane basilaire, notre modèle de l'oreille physique comprendrait une rangée supplémentaire de circuits résonnants très sensiblement moins amortis puisque, d'une façon générale, c'est le circuit le moins amorti d'une série qui impose sa loi d'amortissement à l'ensemble. Les formules (17) et (19) sont d'ailleurs une illustration de ce fait. Une deuxième conséquence en découle ipso facto : l'exposant de la f. de m. globale qui fait l'objet de nos préoccupations devrait valoir au moins 3 et non 2.

Pour en décider, outre des expériences directes sur des stimuli isolés ou répétés (comme indiqué ci-dessus), il conviendrait de calculer exactement l'expression (14), dans l'hypothèse d'un signal de type (13), dont le décrément β serait égal à la valeur α_o du décrément de la f. de m. pour la fréquence masquante $\nu_o = \omega_o/2\pi$,

ce qui donnerait pour le spectre une fonction de n, ω, α et ϑ. La valeur de ϑ qui rend maximum cette fonction, pour le spectral $\omega = \omega_o$, $\beta = \alpha_o$, n'est autre que ϑ_M. En remplaçant ϑ par ϑ_M dans l'expression générale, on obtient le spectre du son auditivement pur mais sans durée, qui doit, si notre modèle est valable, pouvoir s'identifier aux courbes de masque connues.

Pratiquement, cette identification donnera une série d'équations transcendantes qui permettront, en liaison avec les expériences directes, de déterminer $\alpha(\omega)$ et de fixer la valeur la plus adéquate de l'exposant n.

Une complication supplémentaire est à craindre : comme signalé en fin du § 3.1., la limite inférieure de l'intégrale ne peut sans doute pas être 0, mais bien un nombre $t(\omega)$, à déterminer également (ou à supputer à priori d'après les connaissances que l'on possède concernant la mécanique de la membrane basilaire). (*)

Dans l'hypothèse où (14) est applicable, nous donnons ci-après les grandes étapes du calcul lorsque n = 2, $\vartheta_M = 5/\alpha_o$, $t(\omega) = 0$.

Tout d'abord, le sommet spectral vaut, d'après (20),

$$|\tilde{s}(\omega_o, \vartheta_M)| = \frac{AM_o}{30}\left(\frac{5}{\alpha_o e}\right)^5 \qquad (25)$$

Ensuite, simplifions le calcul et les notations en posant

$$v = u - 1/2\,,\quad x = 5\left(1 - \alpha/\alpha_o\right),\ y = 5\,\frac{\omega-\omega_o}{\alpha_o},\ z = x + jy\,. \qquad (26)$$

Il vient alors

$$\tilde{s}(\omega, \vartheta_M) = AM\left(\frac{5}{\alpha_o}\right)^5 \exp\left[-2{,}5\left(1+\frac{\alpha}{\alpha_o}\right)\right] \int_{-0{,}5}^{+0{,}5} \left(1/4 - v\right)^2 e^{-3v}\, dv\,. \qquad (27)$$

L'intégrale vaut :

$$4z^{-5}\left[\left(z^2+12\right) sh\, z/2 - 6z\, ch\, z/2\right] \qquad (28)$$

et, enfin, on obtient pour le carré du module de $\tilde{s}$:

$$\begin{aligned}|\tilde{s}(\omega, x, y)|^2 = {} & 16A^2M^2\left[(\alpha-\alpha_o)^2 + (\omega-\omega_o)^2\right]^{-5} \exp\left[-5\left(1+\frac{\alpha}{\alpha_o}\right)\right] \times \\ & \left[\left(12 + x^2 - y^2\right)^2 + 4x^2y^2\right)\left(sh^2 x/2 + \sin^2 y/2\right) - 6\left(12 + x^2 - y^2\right) \times \\ & (x\, sh\, x + y \sin y) + 12xy\,(x \sin y - y\, sh\, x) \\ & + 36\left(x^2+y^2\right)\left(sh^2 x/2 + \cos^2 y/2\right)\Big]\,. \end{aligned} \qquad (29)$$

(*) De plus, il conviendrait de tenir compte du fait que, conformément aux courbes de Fletcher et Munson, l'amplitude M de la f. de m. dépend de la fréquence et de l'amplitude A du signal.

Si nous désignons par X la dernière expression entre crochets de (29), nous obtenons, pour le rapport du carré de l'amplitude du sommet spectral (25) au carré de l'amplitude du spectre pour une fréquence quelconque (29), l'équation transcendante

$$\sigma(\omega_0,\omega) = \left|\frac{\mathcal{F}(\omega_0)}{\mathcal{F}(\omega)}\right|^2 = \frac{(x^2+y^2)^5}{14400\, e^{3x}\, X}\left(\frac{M_0}{M}\right)^2 \qquad (30)$$

qui doit permettre de déterminer x, donc α, en fonction de valeurs connues de ω, donc de σ (courbe de masque expérimentale) et de y. Cette équation ne permet pas de déterminer α_0, car elle se réduit à l'identité 1 = 1 si l'on y fait $\alpha = \alpha_0$ ($x = 0$) et $\omega = \omega_0$ ($y = 0$). α_0 doit par conséquent être déterminé par la méthode décrite au début du présent paragraphe. Notons ici que les courbes de masque doivent être interpolées pour les fréquences voisines de celle du son masquant, les battements ayant altéré les mesures de WEGEL et LANE (1924) dans cette région.

5. Le phénomène des battements.

A titre d'application de la présente théorie, citons le cas des battements entre deux sons de fréquences voisines ν_1 et ν_2. Assez curieusement, HELMOLTZ donne une analyse purement temporelle du phénomène. Il est aisé de montrer - mais ceci dépasse les limites de la présente communication - que le battement est un bruit coloré (spectre assez large) dont l'amplitude varie périodiquement à la fréquence $\nu_2 - \nu_1$.

La sensation auditive confirme nettement cette prévision théorique. L'expérience est aisée à faire sur un orgue ou un harmonium en touchant deux notes à l'intervalle d'un demi-ton dans les premières octaves du huit pieds par exemple.

6. Le débit d'information de l'audition.

Il convient, au point où nous sommes arrivés, de revenir quelques instants sur la question du débit d'information de l'audition. Ce débit dépend de la résolution fréquentielle et de la résolution temporelle des résultats de l'analyse. Si l'organe auditif ne comprenait que l'oreille physique, ce débit serait très inférieur à ce qu'il est en réalité car, comme on l'a vu, le spectre physique (12) d'un signal sinusoïdal stationnaire est loin d'être la raie étroite d'un son pur. C'est grâce à l'effet de masque,

comme il a été souligné au § 4.1., que la résolution fréquentielle de l'oreille, fortement augmentée, lui permet d'atteindre le débit d'information réel de la fonction auditive.

Pour conclure, je formule le voeu que d'autres chercheurs s'intéressent au problème tel que j'ai cru intéressant de le poser et de le formuler.

REFERENCES

ATAL B.A., HANAUER Suzanne L. ., Speech Analysis and Synthesis by linear Prediction of the Speech Wave. J.A.S.A. 48.637-655 (1971)

FANO, R.M., Short-time Autocorrelation Functions and power spectra J.A.S.A. 22 . 546-550 (1950).

FLANAGAN, F.L., Models for approximating Basilar Membrane Displacement - part III. Bell System Techn. J. 41.959-1009 (1962a).

HELMOLTZ H. von . Théorie physiologique de la musique. éd. originale en langue allemande : 1863. Trad. française de GUEROULT. Paris 1874.

KORN, I., Theory of Audio Information. Acustica 22 . 6 (1970).

LANDERCY, A., Temporal segmentation - Influence of the envelope function on perception. J. of Speech and Hearing Research 14. 47-57 (1971)

LESUISSE, R., Réception des signaux à codage spectral. Travail de fin d'études. Université Libre de Bruxelles. 102 p. (1970).

MOLES ABRAHAM A., Les Musiques expérimentales. Ed. du Cercle d'Art contemporain. Paris - Zurich - Bruxelles - 166 p. (1960)

NOLL. A.M. Short-time Spectrum and "Cepstrum" Techniques for vocal Pitch Detection. J.A.S.A. 36. 296-302 (1964)

PIMONOV L., Vibrations en régime transitoire. Dunod Paris 357 p. (1962).

WEGEL R.L. et LANE C.E., The auditory Masking of one sound by another and its probable relation to the dynamics of the inner ear. Phys. Rev. 23. 266. (1924).

Fig. 1.

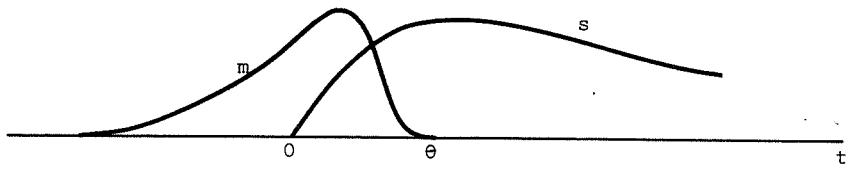


Fig. 2.

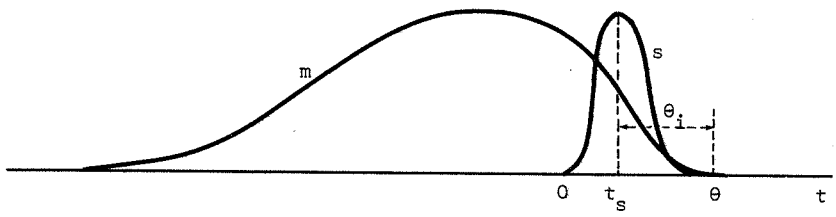
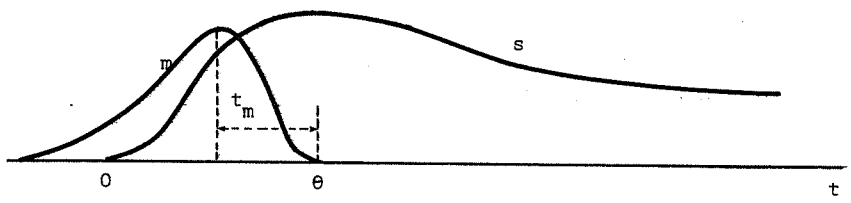


Fig. 3.



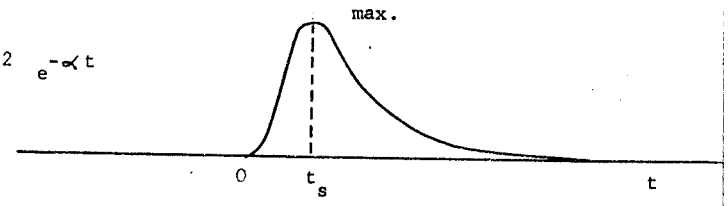
s : enveloppe d'un signal.

m : fonction de mémoire.

Fig.4 ENVELOPPES

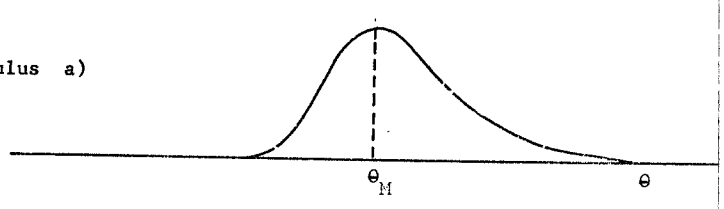
a) stimulus en $t^2 \; e^{-\alpha t}$

$t_s = \frac{2}{\alpha}$



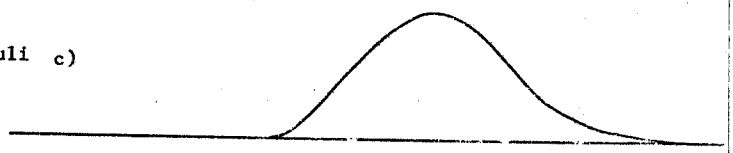
b) écoute du stimulus a)

$\theta_M = \frac{5}{\alpha}$

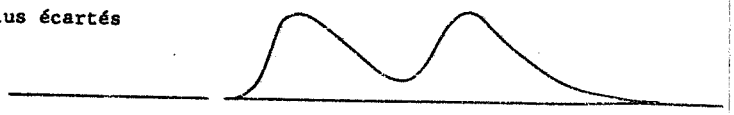


c) deux stimuli discernables visuellement

d) écoute des stimuli c)



e) deux stimuli plus écartés



f) écoute des stimuli e)

Discussion.

R. CARRE : a) Des travaux sur le rôle de la phase relative des composantes ont montré que la phase jouait un rôle important dans le cas où l'enveloppe temporelle E (t) subit des variations importantes au cours d'une période c'est-à-dire si $E(t)_{max}/E(t)_{min} > k$. Est-il possible d'expliquer ce phénomène ?
b) Certaines expériences ne peuvent être expliquées par le seul fonctionnement spectral de l'oreille. Une analyse de la forme temporelle semble être effectuée. Les deux types de fonctionnement n'existent-ils pas simultanément dans l'oreille ?

J. BOSQUET : a) Ce qui est dit au §3.3 éq (13) répond qualitativement à cette question. Si l'on considère un signal $s(t) = At^2 e^{-\beta t} \cos(\omega_0 t + \varphi) = \frac{1}{2} At^2 e^{-\beta t}\left[e^{j(\omega_0 t+\varphi)} + e^{-j(\omega_0 t+\varphi)}\right]$ dont l'enveloppe varie très vite (β très grand) le spectre (13) du premier terme étant très large, on ne peut plus négliger les termes en $\omega + \omega_0$ provenant du 2e terme; de plus, φ ne disparaît plus lorsque l'on prend le module. Il en résulte que le spectre audible du signal possède un maximum qui ne coïncide plus en général avec ω_0 et est, de plus, fonction de la phase φ.
b) Celle-ci figure en principe au §3.1.

D. ROSTOLLAND : Vous venez de nous faire écouter un son sinusoïdal modulé par une impulsion. Mais en réalité, ce que nous avons entendu ne correspond pas exactement au signal électrique tel qu'on peut le voir sur oscilloscope. Le signal acoustique comporte en lui-même une distorsion due à la "fonction masque" de toute la chaîne de reproduction sonore, y compris l'effet du temps de réverbération de cette salle. Si l'on fait varier la phase de la sinusoïde par rapport à l'impulsion, le spectre total est modifié, mais pensez-vous que cela puisse être perçu, avec le signal acoustique ainsi distordu, par une différence de timbre significative ?

Dans ces conditions, comment peut-on juger auditivement l'effet de la "fonction masque" de l'oreille elle-même ?

J. BOSQUET : Il conviendrait évidemment de faire les expériences en chambre sourde, avec des signaux dont on connaît avec certitude la forme aérienne.

La phase de la sinusoïde n'a pratiquement pas d'influence sur la hauteur perçue, lorsque la transformée de FOURIER de l'enveloppe du signal ne présente pas de rebondissements - ce qui est le cas ici - cf. l'équation (12) du § 3.2. et le commentaire qui l'accompagne.

MODELE DE PERCEPTION DE VOCOIDES SYNTHETIQUES ISOLES

Résumé

Les résultats de la plupart des expériences de reconnaissance de voyelles synthétiques semblent indiquer qu'une métrique euclidienne dans l'espace bidimensionnel Log F1/Log F2 permet d'estimer, au niveau d'un groupe de sujets, la probabilité qu'un stimulus x soit classé dans une catégorie phonologique y donnée : $P(x/y) = ae^{-bd2}$, d étant égal à

$$\left(\left(\text{Log } F1x - \text{Log } F1_{VNA}\right)^2 + \left(\text{Log } F2x - \text{Log } F2_{VNA}\right)^2\right)^{\frac{1}{2}}$$

VNA dans cette équation se rapporte au point de l'espace correspondant à une voyelle synthétique classée unanimement (à 100 %) par les sujets dans une catégorie, c'est à dire une voyelle non-ambiguë. Après description de l'algorithme de calcul des VNA, des résultats quantitatifs sont présentés, indiquant que l'on peut resynthétiser la matrice de classification à partir d'un nombre limité de paramètres tirés de ce modèle analogique. La notion de VNA et le rôle de F3 sont discutés à propos d'un essai de prévision quantitative des erreurs de classification pour des stimuli naturels. Une expérience de scaling permet d'émettre l'hypothèse que la distance euclidienne possède un corrélat psychologique. Il est alors possible d'envisager un modèle individuel, qui est brièvement présenté.

Summary

Most of the experiments dealing with vowel perception seem to support the view that a euclidean metric on a Log F1/Log F2 space allows the computation, at group level, of the probability that a stimulus x will be classified as an example of the phonological class y. Specifically $P(x/y) = ae^{-bd2}$, with d equal to

$$\left(\left(\text{Log } F1x - \text{Log } F1_{VNA}\right)^2 + \left(\text{Log } F2x - \text{Log } F2_{VNA}\right)^2\right)^{\frac{1}{2}}$$

In this equation, VNA refers to the spatial position of the synthetic vowel that would be classified unanimously as belonging to a single class, i.e. the vowel that is non-ambiguous. After a short description of the algorithm used to compute the position of the VNA, quantitative results are presented that indicate that it is possible to re-synthesize the classification matrix from a limited number of parameters

derived from this model. The concept of VNA and the possible role of F3 are discussed, following an attempt to predict a confusion matrix derived from a listening experiment using natural vowels. A scaling experiment reveals that the distance used in the model may have a psychological correlate. Finally, an individual model is presented, which is compatible with the previous group model.

François LONCHAMP
(Université de Nancy II).

MODÈLE DE PERCEPTION DE VOCOÏDES SYNTHÉTIQUES ISOLÉS

François LONCHAMP
Institut de Phonétique
Université de Nancy II

Ces quelques résultats concernent divers aspects d'un modèle de perception de la parole. Les spécialistes de l'enseignement du Français langue étrangère et les linguistes chargés de l'élaboration d'une stratégie d'acquisition de l'aptitude à la Compréhension Orale d'une langue étrangère ont souvent émis le voeu de disposer d'un outil permettant la prévision quantitative de la probabilité d'erreurs de classification de stimuli phonétiques. C'est dans cet esprit que l'Institut de Phonétique de Nancy II a décidé d'orienter ses recherches.

Un certain nombre de phonéticiens se sont déjà penchés sur ce problème et ont publié quelques études préliminaires. On peut trouver dans LONCHAMP (71) une analyse des contributions de POLITZER (1962) et de J. MEYER (1970). Il semble, pour résumer très grossièrement, que les pourcentages d'erreurs relevés à l'occasion d'un test de classification ne permettent qu'une prévision extrêmement grossière des résultats d'autres tests. Ceci est dû, sans aucun doute, à une grande variabilité dans la composition acoustique des stimuli utilisés.

Il paraît donc indispensable de formuler un modèle explicite de perception comportant une description de la façon dont les divers indicateurs acoustiques (cues) sont utilisés en vue de la classification.

Une première étape indispensable est l'étude du problème de la classification dans la langue maternelle. Pour des raisons de simplicité, nous n'étudions dans un premier temps que la classification de stimuli vocaliques. Il ne nous a pas paru indispensable de recourir dès l'abord à des expériences personnelles, en raison du grand nombre d'expériences relatées dans les diverses revues spécialisées.

En collationnant ces diverses expériences, on remarque le fait suivant : Les valeurs des Formants F1 et F2 des stimuli classés unanimement dans une catégorie phonologique par une population homogène ne correspondent pas aux valeurs moyennes de F1 et F2 pour un échantillon représentatif de

.../...

réalisations naturelles.

Nombreux sont les phonéticiens à avoir mentionné ce fait (cf., par exemple, p. DELATTRE (1951, 1952) et O. FUJIMURA (1967)). Ceci permet de prédire qu'une expérience de classification de stimli naturels comportera une part non négligeable d'erreurs de classification. FAIRBANKS & GRUBB (1961) obtiennent 26 % d'erreurs. KLEIN, PLOMP & POLS (1970) obtiennent un même pourcentage.

On constate d'autre part que les valeurs moyennes de F1 et F2 sont presque toujours centralisées par rapport aux valeurs de F1 et F2 des stimli synthétiques reconnus de façon non ambiguë. La Fig. 1 dessine à partir des données de MAJEWSKI & HOLLIEN (1967) en est un exemple. Les cercles se trouvent à l'intersection des valeurs moyennes de F1 et F2 calculées sur des échantillons de 28 voyelles. Les croix représentent les stimli synthétiques reconnus par le plus grand nombre de sujets.

Au vu des matrices de classification, il semble exister une relation précise entre la position d'un stimlus dans un espace bidimensionnel Log F1 / Log F2 et la probabilité que ce stimlus soit classé dans une catégorie phonologique.

Nous définissons dans un tel espace un critère de distance tel que la probabilité qu'un stimlus soit classé dans une catégorie soit fonction de la distance euclidienne séparant ce stimlus du point de l'espace correspondant à un stimlus reconnu unanimement (à 100 %). Ce point, définissant la voyelle non ambiguë, est noté VNA sur la figure 2 correspondant à l'éq. 1

$$(1)\quad d = \left(\left(\log F1_{x_1} - \log F1_{VNA} \right)^2 + \left(\log F2_{x_1} - \log F2_{VNA} \right)^2 \right)^{\frac{1}{2}}$$

La probabilité que xl soit classé dans une certaine catégorie dépend de la distance qui sépare xl de la position de la voyelle non ambiguë VNA.

Le tracé de la courbe donnant la probabilité de classification en fonction de la distance suggère l'équation 2, p (x | VNA) étant la probabilité que x soit classé dans la catégorie considérée.

$$(2)\quad p(x \mid VNA) = a\, e^{-b\, d^2}$$

.../...

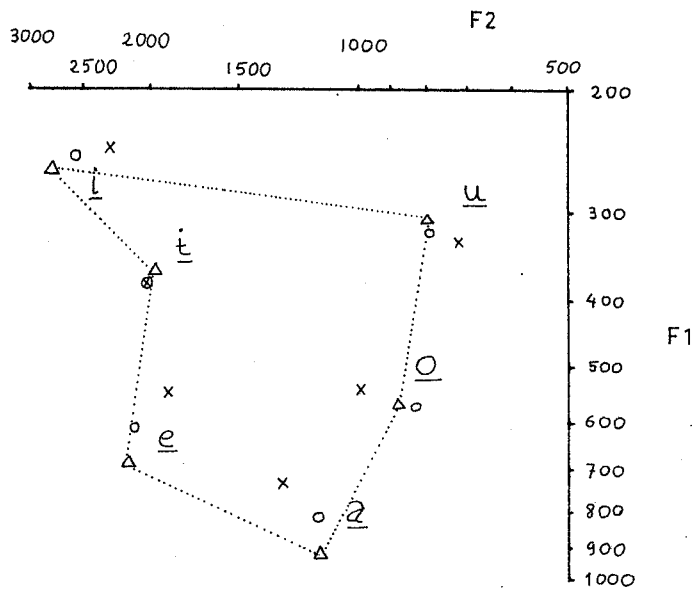


FIG 1 : MAJEWSKI-HOLLIEN (67)
× spectro
o synthétique
△ vna

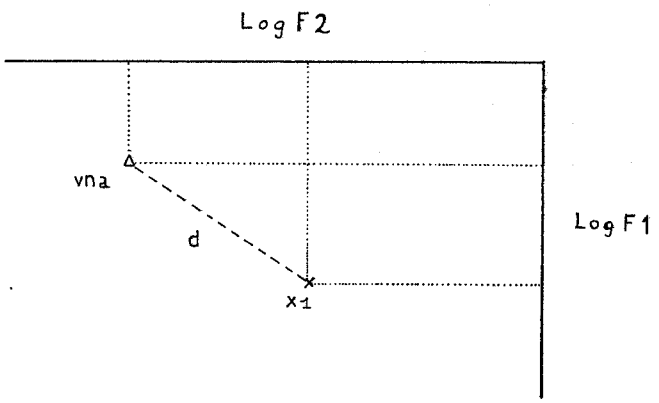


FIG 2 : VNA

Ce modèle est certainement faux au niveau de l'individu. L'activité de classification pour un sujet est une décision " discrète ". Les probabilités dont nous parlons ne sont que l'expression au niveau du groupe de choix individuels contradictoires.

Ce modèle de groupe a été utilisé pour analyser des données diverse sur le CII 100 70 du Centre de Calcul de l'Université de Nancy à l'aide de la procédure suivante :

(1) - exploration systématique de la région de l'espace où les réponses sont les plus unanimes. Pour un grand nombre de points, calcul de b par la relation (tirée de (2)) pour chacun des points représentant les vocoïdes.

$$\frac{\text{Log } P - \text{Log } a}{d^2} = - b$$

(2) Calcul de la variance de b pour chacune de ces estimations de la VNA. La position de la VNA est le point de l'espace qui minimise la variance de b. On se heurte à un problème de minima : on voit par exemple que si l'on prend une position présumée de VNA éloignée des divers points x, les estimations de - b tendent toutes vers 0 et la variance est faible.

(3) Pour éviter ce problème, calcul simultanément de p (x | VNA) en utilisant la moyenne des estimations de b. Le critère de minimisation de la valeur absolue de / pb expérimentale - pb calculée / conduit toujours à un minimum local de la variance de b.

La table 1 est le résultat de l'analyse d'une partie des données de SCHOLES (1967, 1968) concernant la voyelle [æ] de l'anglais américain. Nous avons rassemblé en un même tableau les résultats de deux expériences distinctes utilisant les mêmes stimuli synthétiques car la stabilité des réponses était remarquable.

Il s'agit du sous-ensemble des réponses " le stimulus correspond au phonème [æ] de l'anglais américain ", pour les valeurs de F1 et F2 indiquées. Le chiffre à gauche de chaque case est le résultat expérimental ; le chiffre à droite est la réponse recalculée à l'aide de (2). On trouve sous ce tableau les distances calculées à l'aide de (1).

La fig. 3 reproduit le tableau 1.

.../...

Hz

2600	2400	2200	2000	1800	F2/F1
8/9.7	14/13.9	18/17.5	18/18.6	8/15.8	650
16/16.1	21/23.1	29/29.2	33/31.0	28/26.3	750
23/17.3	29/24.9	31/31.4	33/33.4	30/28.3	850

Hz

TABLE 1 : SCHOLES (1967,1968) /æ/ VNA F1 = 820 Hz
F2 = 2037 Hz

a = 34 b = -11.07 % d'erreurs : 5.7

d(2600,650) = 0.337	d(2600,750) = 0.260	d(2600,850) = 0.247
d(2400,650) = 0.284	d(2400,750) = 0.187	d(2400,850) = 0.168
d(2200,650) = 0.245	d(2200,750) = 0.118	d(2200,850) = 0.085
d(2000,650) = 0.233	d(2000,750) = 0.091	d(2000,850) = 0.040
d(1800,650) = 0.263	d(1800,750) = 0.152	d(1800,850) = 0.129

Hz

2600	2400	2200	2000	1800	1600	1400	1200	F2/F1
27/19.5	42/23.1	25/25.2	15/24.8	12/21.2	13/15.0	18/ 8.1	0/	450
59/46.0	60/54.4	67/59.4	74/58.6	63/50.1	56/35.4	50/19.2	19/ 7.1	550
67/62.9	72/74.5	76/81.4	81/80.1	75/68.5	64/48.4	26/26.2	4/ 9.7	650
62/61.5	70/72.8	81/79.6	80/78.4	71/67.0	52/47.4	8/27.5	0/	750
51/48.4	58/57.3	63/62.7	69/61.7	34/52.8	10/37.3	0/	0/	850

TABLE 2 MAJEWSKI & HOLLIEN (1967) /e/ VNA F1 = 690 Hz
F2 = 2125 Hz

a = 84 b = -6.54 % d'erreurs : 10.1

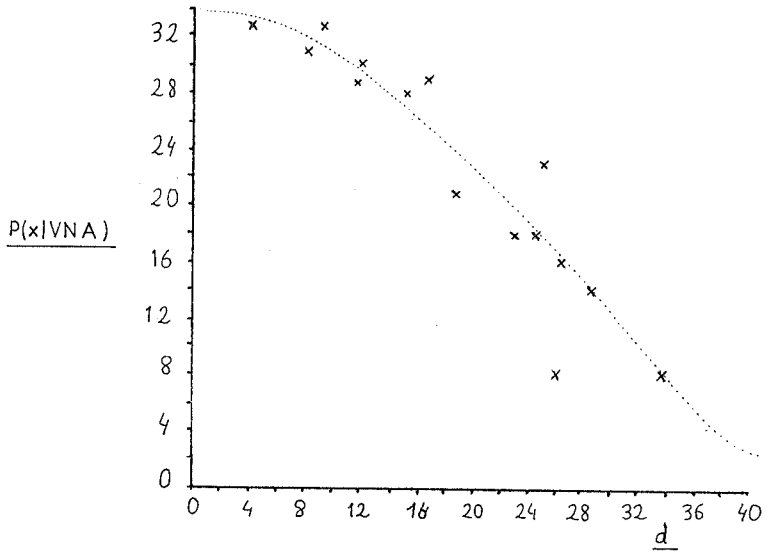


FIG 3 : SCHOLES (67,68) [æ]

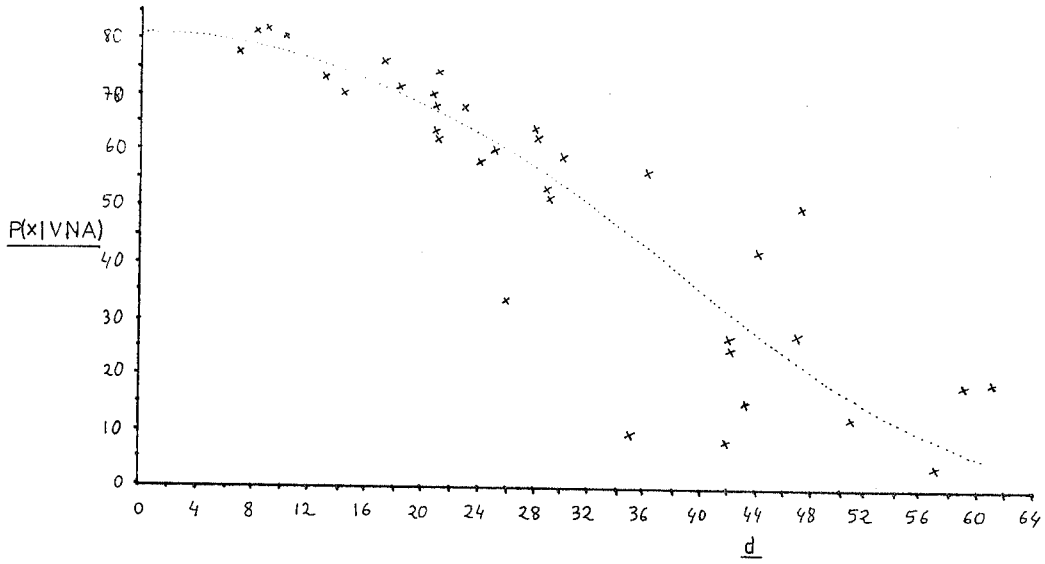


FIG 4 : MAJEWSKI - HOLLIEN (67) [e]

Le tableau 2 et la figure 4 sont le résultat de l'analyse des données de MAJEWSKI & HOLLIEN (1967) concernant le [e] polonais. Le chiffre à gauche dans chaque case est dans ce cas le nombre de réponses pondérées par un " coefficient de certitude " fourni par les sujets. Cet exemple donne un pourcentage d'erreurs assez élevé. Pour les 8 voyelles du polonais (dont les VNA sont portés sur la fig. 1) le pourcentage de discordance entre valeurs expérimentales et valeurs calculées s'établit ainsi (moyenne 8,8 %).

(i) : 5,7 %
(I) : 11,6 %
(e) : 10,1 %
(a) : 10,3 %
(o) : 9,9 %
(u) : 5,0 %

On remarque que le fort pourcentage de discordance pour [e] est dû (fig. 4) à la mauvaise estimation de p pour les stimuli éloignés de la VNA.

Ce fait indique les limites du modèle de groupe utilisé (fig. 5).

Un stimulus quelconque se trouve en fait plus ou moins proche de plusieurs VNA. Si l'on considère deux stimuli , x1 et x2 tels que d1 et d'1 soient égaux, notre modèle prévoit que p sera identique pour ces deux stimuli. Or x1 est plus éloigné de VNA [ɛ] que x2 de VNA [ɔ]. Ceci implique que p (x1 / VNA [ɛ]) est inférieur à p (x2 / VNA [ɔ]). Il en résulte qu'il est vraisemblable que p (x1 / VNA [a]) sera supérieur à p (x2 / VNA [a]). Ce fait est confirmé par l'étude du tableau 1. Pour le stimulus 850/2600 il y a 23 réponses [æ] au lieu des 17.3 prévues. Ce stimulus est loin de toutes VNA. Pour le stimulus 650/1800 au contraire, p expérimentale vaut 8 alors que p calculée vaut 15.8. Ce dernier stimulus est proche de la VNA [ɒ].

Ce fait nous a conduit à utiliser le nouveau critère

$$\left| \frac{p\ exp. - p\ calculée}{a - p\ exp.} \right|$$

dans le calcul de la VNA. Le problème est alors le critère arbitraire de cette pondération. Les valeurs des VNA obtenues avec cette procédure ne diffèrent guère de celles obtenues sans pondération.

En résumé, nous pouvons maintenant recalculer les matrices de classification de stimuli synthétiques à partir d'un nombre limité de paramètres : position de la VNA, valeurs de F1 et F2 pour chaque stimulus, valeur du coefficient b. L'erreur moyenne est inférieure à 10 %. Il est à noter que b ne semble pas être invariant, bien qu'il soit possible qu'une réduction du nombre de stimuli éloigné de la VNA diminue la dispersion.

b (i) = - 13.2
b (I) = - 21.1
b (e) = - 6.5
b (a) = - 9.3

moyenne : - 11.9

b (o) = - 12.3

b (u) = - 9.7

Sans faire une discussion complète du " concept " de VNA, on peut noter que :

(1) Les VNA sont presque toujours plus éloignées de la voyelle neutre [ ], que les points obtenus à partir des moyennes de F1 et F2 pour des échantillons de voyelles naturelles. Il est impossible de prononcer une voyelle qui ait pour F1 et F2 les valeurs de la VNA.

En utilisant une statistique assez ancienne de LISKER (1948) concernant la voyelle [ɤ] de l'anglais américain (F1 = 742 Hz, σ = 35 ; F2 = 1695 Hz, σ = 46), on trouve un écart réduit de 7.5 pour F2 de la VNA.

Cette constatation rejoint une notation de MOL (1970). " Si l'on demande à des sujets de produire le système vocalique de leur langue maternelle en ajustant les paramètres F1 et F2 d'un synthétiseur à formants, on observe qu'ils utilisent la totalité de la gamme de variation de F2. Ils créent un système qu'il leur serait impossible de réaliser vocalement ".

(2) Les VNA semblent relativement stables. On peut s'en convaincre en étudiant en détail les deux expériences de SCHOLES.

(3) Les VNA sont définies sans référence à F3. On peut noter que les expériences de HANSON (1967), SINGH et WOODS (1971), POLS, Van der KAMP et PLOMP (69), révèlent des facteurs que l'on peut attribuer à F1 et F2, mais beaucoup plus difficilement à F3. Nous reprendrons ce point à propos de l'expérience de KLEIN, PLOMP et POLS (1970).Quant à l'expérience de FUJIMURA (1967) sur le rôle de F2 et F3 dans la perception des voyelles d'avant, nous ne partageons pas les conclusions de l'auteur ; "... il est très improbable que l'on puisse représenter le timbre des voyelles dans un espace bidimensionnel,tel celui défini par F1 et F2 ". Cette conclusion s'appuie sur une expérience de classification utilisant des stimli synthétiques dont les formants F2 et F3 étaient beaucoup trop proches, leur enlevant toute ressemblance avec des vocoïdes naturels. (pour une critique complète cf. LONCHAMP (1971) pages 23-25).

.../...

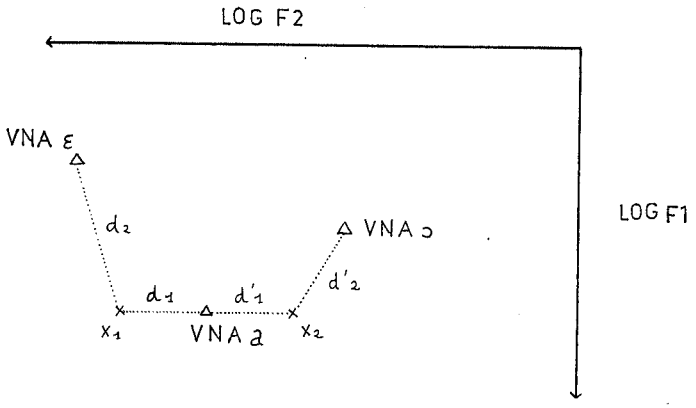
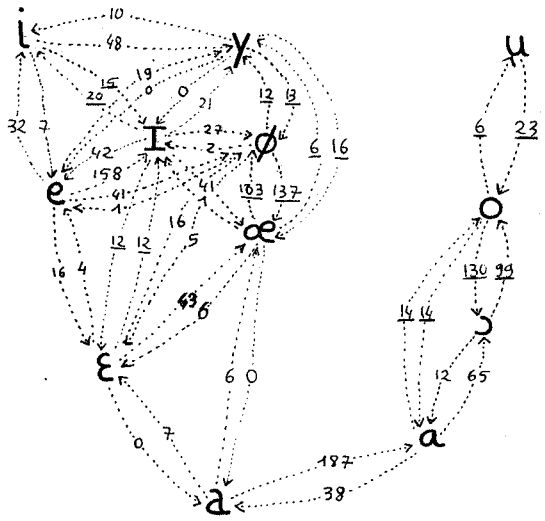


FIG 5 : $d_1 = d'_1$ $d_2 > d'_2$



i – I	: 15, 20	17.5
I – ɛ	: 12, 12	12
ɔ – o	: 130, 99	114.5
ɑ – o	: 14, 14	14
ø – y	: 12, 13	12.5
œ – ø	: 137, 103	125
o – u	: 6, 23	14.5
y – œ	: 16, 6	11

transferts égaux

FIG 6 : KLEIN, PLOMP, POLS (70)

(4) Les expériences de LINDNER (1966) portant sur la catégorisation de stimuli synthétiques à différent Fo semblent indiquer que la position des VNA varie avec Fo. Pour Fo = 110 Hz, la VNA [ɑ] de l'allemand s'établit à 740 / 1175 Hz ; pour Fo = 220 Hz, la VNA se trouve à 831 /1175 Hz. Ceci rejoint une expérience de FOURCIN (1968) sur la perception des transitions formantiques à des fréquences fondamentales différentes. La position des VNA semblent se centraliser quand Fo diminue. Une étude détaillée de ces problèmes, ainsi que l'ensemble des résultats obtenus en ce qui concerne les VNA seront publiés dans le 1er rapport d'activité de l'Institut de Phonétique de Nancy, à paraître en Décembre 1973.

A partir de ce qui précède, il est possible de prévoir, au prix de certaines approximations, certaines relations quantitatives fournies par les matrices de confusions obtenues lors de la classification de stimuli vocaliques naturels. Parmi le grand nombre de matrices publiées, nous avons choisi celles de KLEIN, PLOMP, et POLS (1970), reproduite ici sous forme graphique (fig. 6). La figure 7 présente de façon schématique l'étude réalisée. X1 et X2 représentent les moyennes des paramètres F1 et F2 pour deux populations de voyelles différentes dont les VNA sont inconnues. En première approximation, on peut dire que la probabilité que les éléments de la population 1 soit classée comme appartenant à la classe 2 est fonction de d1, distance qui n'est guère différente de d2. Il faut pour cela que X1 soit une bonne approximation des différentes positions des voyelles de la catégorie 1 et que d4 soit petit devant d1 et d2. La figure 6 montre qu'il y a 23 couples de transferts significatifs (par transfert, on entend la classification d'un stimulus dans une catégorie différente de celle voulue par le locuteur). Sur ces 23 couples, seuls 8 sont approximativement égaux. Or, d'après le modèle présenté, tous les transferts devraient être égaux. La figure 8 est le graphe de l'importance des transferts en fonction de la distance pour les 8 transferts égaux. Il semble exister une relation précise entre ces deux paramètres.

Ce fait est le premier résultat tangible concernant directement l'un des buts de cette étude, à savoir la précision quantitative des erreurs de classification dans la langue maternelle.

Trois points sont à souligner :

(1) Il n'y a transfert entre voyelles d'avant et voyelles d'arrière qu'au niveau des phonèmes [ɑ] et [a].

.../...

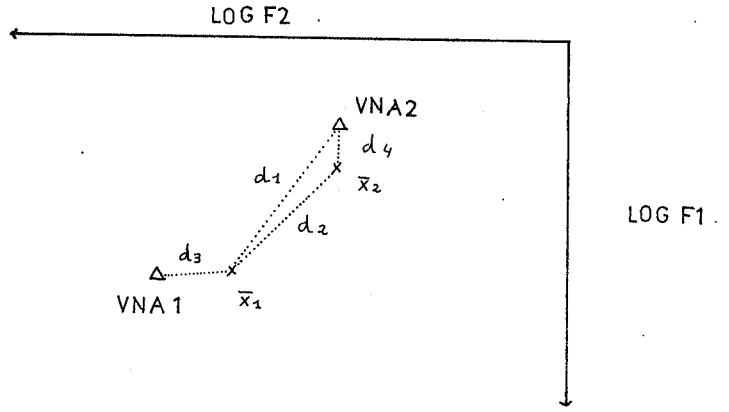
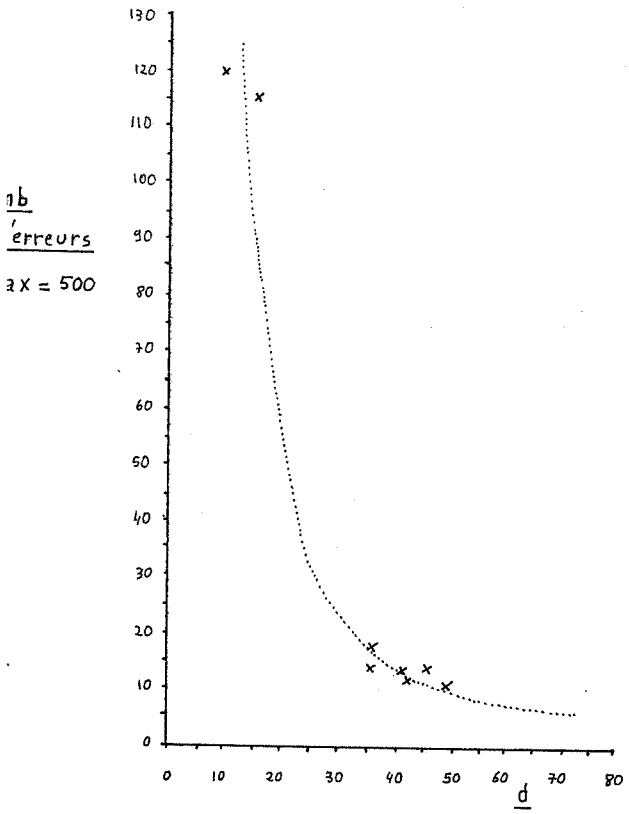


FIG 7 : $d_1 \simeq d_2$; $d_3, d_4 \ll d_1, d_2$



$d(i,I) : 0.360$
$d(I,\varepsilon) : 0.429$
$d(a,o) : 0.363$
$d(\supset\text{-}o) : 0.145$
$d(o\text{-}u) : 0.457$
$d(y\text{-}\phi) : 0.425$
$d(y\text{-}\alpha) : 0.496$
$d(\phi\text{-}\alpha) : 0.083$

FIG 8 : KLEIN, PLOMP, POLS (70) transferts égaux

(2) Il n'y a aucun transfert entre voyelles arrondies et voyelles d'arrière.

(3) Les transferts entre voyelles arrondies et voyelles d'avant sont faibles (moyenne : 2) alors les transferts entre voyelles d'avant et voyelles arrondies sont plus importants (moyenne : 2 5).

La caractéristique des voyelles arrondies est qu'elles possèdent une valeur de F3 relativement différente de celle des voyelles non arrondies. Nous faisons l'hypothèse que F3 joue le rôle de " marque " au moment de la classification. La valeur de F3 ne jouerait un rôle qu'au niveau de la classification dans les catégories arrondies / non arrondies. A l'intérieur de la catégorie arrondie, les distances définies à partir de F1 et F2 suffisent pour rendre compte des erreurs de classification.

Trois raisons peuvent expliquer les transferts inégaux :

(1) Le modèle est inadéquat.

(2) Les hypothèses simplificatrices (d1 $\simeq$ d2, d3 $\ll$ d1 et d2, dispersion faible devant d1 et d2) ne tiennent pas pour certains couples.

(3) D'après les auteurs eux-mêmes, 3 voyelles du Néerlandais sont longues et légèrement diphtonguées : [o], [ø], [e]. Leur normalisation en durée (150 ms) a pu supprimer deux indicateurs acoustiques pertinents : durée et variations formantiques.

Notons enfin que le chiffre le plus important de certains transferts correspond à la valeur que lui prédit la courbe de la figure 8; c'est le cas de [a - œ], [e - ɛ], [e - a] parmi d'autres. Nous arrivons alors à une prédiction quantitative correcte dans 50 % des cas.

Ce modèle de groupe permet également une analyse intéressante des résultats obtenus lorsqu'on demande à des sujets d'estimer, par une méthode de scaling la <u>similarité</u> entre des couples de voyelles. Cette expérience a été réalisée sur 40 couples de voyelles françaises naturelles de durée et Fo identiques (150 ms, Fo = 140 Hz). Les résultats soulignent le caractère prédictif de la notion de distance associée à la notion de VNA, et permettent d'émettre l'hypothèse que la distance telle que nous l'avons définie possède un certain corrélat psychologique. L'examen de la fig. 9 donnant le résultat brut du scaling (ms) en fonction de la distance entre les 40 couples, permet de constater que pour un sujet déjà, les ré-

.../...

sultats étant plus nets encore au niveau du groupe, il apparaît une relation précise entre ces deux mesures. Nous avons séparé les 20 premières réponses des 20 dernières car le critère utilisé par ce sujet ne semble s'être stabilisé qu'au milieu de l'expérience, impliquant une phase d'apprentissage assez longue. Il semble également que pour les couples de stimuli identifiés individuellement sur la fig. 9 la relation ne tienne pas. On notera que ces couples contiennent soit la voyelle [a], soit une voyelle arrondie [y] ou [ø]. Ce phénomène pourrait être dû pour [a] à une mauvaise estimation des valeurs de F1 et F2. En ce qui concerne les voyelles arrondies, il faudrait vérifier, par des expériences ultérieures, si l'introduction de F3 n'entraîne pas une amélioration de la prédiction. Cette expérience préliminaire semble pourtant indiquer que la notion de similarité entre les éléments d'un couple de voyelles est réductible à une mesure objective, pour les voyelles non arrondies du moins.

Il est évident que le modèle de groupe esquissé ne peut rendre compte de l'activité de classification d'un sujet. On peut néanmoins envisager un modèle individuel compatible avec ce modèle de groupe. On peut faire l'hypothèse que l'acquisition des données s'accompagne d'une estimation des distances séparant ce stimulus des VNA proches. S'il y a 2 VNA concurrentes, un critère de décision possible est : si d1 (distance de x à VNA 1) est inférieur à d2 (distance de x à VNA 2) le stimulus est classé dans la catégorie correspondant à la VNA la plus proche, c'est-à-dire VNA 1. Ce critère, très simple, ne permet pas d'expliquer les divergences de classification entre sujets. Une seconde hypothèse est que les mesures d1 et d2 sont entachées d'erreurs. Si le sujet procède à un grand nombre d'estimations $\delta 1$ et $\delta 2$ de d1 et d2, ces estimations se distribuent de façon normale autour des moyennes d1 et d2 avec des écarts types σ_1 et σ_2 inconnus (cf. fig. 10). Il est clair que pour le cas de figure où d1 est inférieur à d2, $\delta 2$ peut être inférieur à $\delta 1$ pour une estimation particulière, d'où classification différente. Afin de tester ce modèle, nous travaillons à des estimations empiriques de $\sigma 1$ et $\sigma 2$ en fonction de d à partir des expériences de classification publiées, avec l'hypothèse secondaire que σ est une fonction monotonement croissante de d.

Pour conclure, nous dirons que ce modèle nous semble être un premier pas vers l'élaboration de procédures de prévisions quantitatives utiles aux pédagogues, mais qui pourrait également présenter un intérêt dans le cadre d'un système de décodage automatique de la parole.

Développer à partir d'expériences portant sur les voyelles, un test " puissant " de la validité de ce modèle sera son efficacité en ce qui concerne la précision des stratégies de classement des consonnes. Un indice

.../...

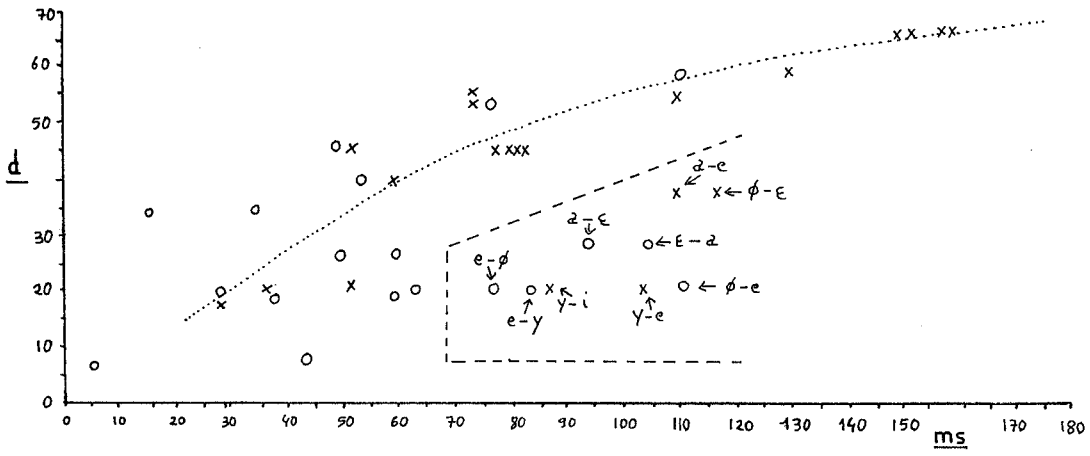


FIG 9 : SCALING : 40 paires

o : 20 premières paires
x : 20 dernières paires

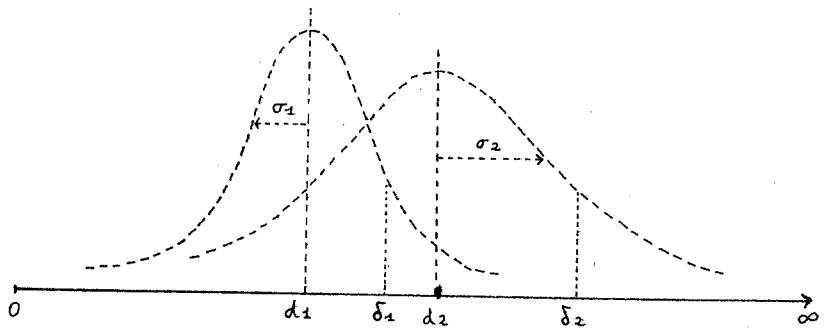


FIG 10 $p(x|VNA_2) = p(\delta_2 < \delta_1)$

d_1 : d de x à VNA 1
d_2 : d de x à VNA 2
$x|VNA_1$ si $\delta_1 < \delta_2$

minime, mais encourageant, est fourni par les expériences de WINITZ, SCHEIB et REEDS (1972) montrant que la position fréquentielle de la tache d'explosion des occlusives a une influence considérable sur le pourcentage d'identification correcte.

REFERENCES BIBLIOGRAPHIQUES

(1) DELATTRE, P., LIBERMAN, A.M., et COOPER, F.S..- Voyelles synthétiques à deux formants et voyelles cardinales. Le Maître Phonétique 96, 30 (1951).

(2) DELATTRE, P., LIBERMAN, A.M., COOPER, F.S. et GERSTMAN, L..- An experimental study of the acoustic determinant of vowel colour. Word No 8, 195 (1952).

(3) FAIRBANKS, G., et GRUBB, P..- A psychophysical investigation of vowel formants. Journal of Speech Hearing Research No 4, 203-219 (1961).

(4) FOURCIN, A.J..- Formant transition perception with different Fo frequencies. Zeitschrift für Phonetik No 21, 89-93 (1968).

(5) FUJIMURA, O..- On the second spectral peak of Front Vowels (a perceptual study of the role of F2 and F3). Language and speech No 10, 181-193 (1967).

(6) HANSON, G..- Dimensions in speech Sound Perception : an experimental study of vowel perception. Ericsson Technik No 23, 3-175 (1967).

(7) KLEIN, W., PLOMP, R., et POLS, L.C.W..- Vowel spectra, vowel spaces and vowel identification. J.A.S.A. No 48 (4), 999-1009 (1970).

(8) LINDNER, G..- Beurteilung synthetish erzeugter vokalartiger Klänge durch deutschsprachige Hörer. Zeitschrift für Phonetik No 19, 45-90 (1966).

(9) LISKER, L..- The distinction between [æ] and [ɛ] : a problem in acoustic analysis. Language No 24, 397-407 (1948).

(10) LONCHAMP, F..- Un modèle probabiliste de la parole. Mémoire de Maîtrise NANCY (1971).

(11) MAJEWSKI, W. et HOLLIEN, H..- Formant frequency regions of Polish vowels. J.A.S.A. No 42 (5), 1031-1037 (1967).

(12) MOL, H..- Fundamentals of Phonetics : Part II. Mouton (1970), page 87.

(13) POLS, L.C.W., Van der KAMP, L.J.Th., et PLOMP, R..- Perceptual and Physical Space of Vowel Sounds. J.A.S.A. No 46 (2), 458-467 (1969).

.../...

(14) SCHOLES, R. J. - Phoneme categorization of synthetic vocalic stimuli by speakers of various languages. Language and speech N° 10 (1), 46 - 68 (1967).

(15) SCHOLES, R.J..- Phonemic interferences as a perceptual phenomenon. Language and Speech n° 11 (2), 86-103 (1968).

(16) SINGH, S. et WOODS, D.R..- Perceptual structure of 12 American English vowels. J.A.S.A. No 49 (6), 1861-1866 (1971).

(17) WINITZ, H., SCHEIB, M.E. et REEDS, J.A..- Identification of stops and vowels from the burst portion of / p, t, k / isolated from conversational speech. J.A.S.A. No 51 (4), 1309-1317 (1972).

-=-=-=-=-=-=-

Discussion.

M. CAkTIER : Pourriez-vous préciser les points suivants :

1. Quel était le choix offert aux auditeurs ?
2. Les références des auditeurs étaient-elles naturelles, ou apprises en début d'expériences ?

F. LONCHAMP : SCHOLES (1967) utilise la méthode du choix forcé entre [æ], [ɛ], [I] [ɑ], [ʌ], [ʊ] et [-] (non-américain) pour les matrices utilisées (Tables 5 et 9). La méthode utilisée par MAJEWSKI & HOLLIEN (1967) est identique : choix forcé entre la représentation orthographique des 6 voyelles du polonais et la réponse/ non-polonais/ . L'orthographe des voyelles du polonais est totalement phonologique. Les références des auditeurs étaient donc naturelles, sans apprentissage préalable.

J-Y. GRESSER :

1. Quel est l'objectif effectivement poursuivi : élaboration d'un nouveau modèle ou application de modèles connus au français ?
2. Envisagez-vous l'extension des voyelles isolées aux voyelles en contexte et à d'autres types de son (consonnes par ex. liquides, plosives, etc..)?
3. Pensez-vous qu'une telle étude soit intéressante pour la reconnaissance vocale, qui utilise des modèles analogues sur le même sujet depuis 20 ans ?

F. LONCHAMP : L'objectif est bien l'élaboration d'un modèle de perception PAR L'HOMME des structures acoustiques de la parole. En simplifiant à l'extrême, nous pensons que ce modèle se révèlera adéquat s'il permet de prévoir la direction et l'importance quantitative des erreurs de classification par l'homme. L'extension aux voyelles en contexte et aux consonnes est souhaitable. Ceci n'a pas encore été fait car nous ne disposons pas d'un synthétiseur.

Ce n'est pas à moi de dire si ce modèle est , en dernière analyse, intéressant pour la reconnaissance automatique de la parole. Même si cela n'est pas le cas, ce modèle, s'il est juste, est d'un intérêt capital pour le psycho-phonéticien.

Il me semble néanmoins totalement abusif de dire qu'un modèle de ce type est utilisé depuis 20 ans. Aucun modèle de reconnaissance vocale, à ma connaissance, n'utilise de données extérieures au corpus utilisé (dans le domaine phonétique). Le modèle proposé définit un INVARIANT (les VNA) à partir duquel la reconnaissance s'effectue. Cet invariant semble ne dépendre que de la langue et de Fo. Il est donc extérieur au corpus. Voici, nous le pensons, l'originalité foncière de ce modèle.

D. ROSTOLLAND : Vous avez parlé de "voyelles non ambiguës" reconnues à 100 % pour une population homogène. Dans quelles conditions expérimentales peut-on définir un tel critère ? Que les essais de reconnaissance se fassent à partir de sons vocaliques naturels ou synthétiques, n'a-t-on pas intérêt à normaliser chaque son en intensité et en durée ? Je ne pense pas, à propos de l'intensité, qu'il soit indifférent d'entendre des voyelles parlées fortes ou des voyelles parlées faibles amplifiées électroniquement au même niveau. Nous étudions actuellement le cas de voyelles criées et, dans ce cas limite, nous avons trouvé des différences importantes. Cela est en relation avec le déplacement du triangle vocalique dans le plan F1/F2 lorsque l'on passe progressivement de la voix "basse" à la voix "criée".

F. LONCHAMP : Les VNA sont, par définition, les stimuli synthétiques reconnus à 100 % par une population homogène. Les conditions expérimentales sont les conditions habituelles en psychophonétique : contrôle strict des paramètres "auxiliaires" (durée, Fo, intensité....), listes balancées pour éviter le facteur d'ordre, contrôle du rapport S/B, écoute à un niveau confortable, sujets à audition normale, etc... Les stimuli synthétiques doivent correspondre fidèlement, au niveau des paramètres acoustiques tels que l'intensité relative des formants et leurs largeurs de bande, à ceux que l'on trouve pour des stimuli naturels. En ce qui concerne les voyelles "criées", il est certain que des différences importantes existent pour ces deux derniers paramètres par rapport aux voyelles "normales". Ces différences font que le modèle présenté ne peut s'appliquer immédiatement.

J-P. HATON: Il serait sans doute très intéressant de reprendre vos expériences en utilisant des méthodes de reconnaissance automatique plus évoluées, c'est-à-dire finalement en se servant de systèmes de reconnaissance existant actuellement.

F. LONCHAMP : Je pense, en effet, que ce serait très intéressant, mais peut-être un peu prématuré. Nous ne disposons, à ce jour, que de très peu d'informations en ce qui concerne le français. D'autre part, l'extraction des formants est une condition sine qua non, et je ne suis pas sûr que cela puisse être fait aujourd'hui à un coût raisonnable.

L'EFFET D'ASYMETRIE LATERALE DANS L'ECOUTE MULTIPLE

Résumé

Différentes interprétations essayent de rendre compte de l'association entre la supériorité de l'oreille droite dans l'écoute dichotique de matériel verbal et la dominance de l'hémisphère gauche pour le langage. On peut se demander si cet effet d'asymétrie latérale résulte de meilleures connexions neurales avec l'oreille droite, dues à la supériorité des voies auditives contralatérales, ou d'un meilleur traitement des messages provenant de la droite du sujet. Avec P. Bertelson, nous avons essayé de séparer les effets de la position spatiale de ceux de l'oreille d'entrée, en utilisant des présentations de deux messages simultanés à travers des haut-parleurs. Avec des haut-parleurs situés l'un à droite et l'autre à gauche du sujet, le message provenant du côté droit était le mieux rappelé. Avec des haut-parleurs situés l'un sur l'autre des côtés et l'autre en face, c'était le message provenant du milieu le mieux rappelé. Ainsi, les différences en rapport avec la position des sources dans l'espace ont été enregistrées, mais aucune conclusion définitive concernant les mécanismes de l'asymétrie latérale ne peut être avancée sur la base de ces résultats.

Summary

Interpretations which have been proposed for the association between right ear superiority in dichotic listening to verbal material and left hemisphere dominance for language are briefly reviewed. The question is asked whether this lateral asymmetry reflects better neural connections of the right ear because of the contralateral auditory pathways superiority, or better processing of messages coming from the right. P. Bertelson and I have tried to separate the effects of spatial position from those of ear of entry by use of presentation of two simultaneous messages over loudspeakers. With the loudspeakers situated to the left and to the right of the subject, the right side message was better recalled. With one loudspeaker in front of the subject and the other on one side, the message from the middle was better recalled. Differences associated with the relative locations of the sources can thus be observed, but no definite conclusion regarding the mechanism of lateral asymmetry can be reached on the basis of these results.

José MORAIS
Université Libre de Brxelles.

L'EFFET D'ASYMETRIE LATERALE DANS L'ECOUTE MULTIPLE

José Morais
Laboratoire de Psychologie Expérimentale
Université Libre de Bruxelles

Tout le travail sur les effets de latéralité dans la perception de messages auditifs a été lancé par une découverte rapportée en 1961 par Doreen Kimura. Kimura a trouvé, en situation d'écoute dichotique et avec du matériel verbal, que les messages présentés à l'oreille contralatérale à l'hémisphère dominant pour le langage sont mieux reproduits que ceux présentés à l'oreille ipsilatérale.

La situation expérimentale était la même que celle utilisée dans les expériences d'empan dichotique de Broadbent (1954) : les sujets, qui portaient des écouteurs, recevaient trois paires de chiffres en succession rapide, les deux chiffres de chaque paire étant présentés simultanément, l'un à l'oreille droite et l'autre à l'oreille gauche; après présentation des six chiffres, ils devaient les reproduire dans n'importe quel ordre.

Les sujets utilisés par Kimura étaient partagés en deux groupes : un groupe de sujets à dominance hémisphérique gauche établie ou supposée (plus de 90 % des droitiers manuels seraient des gauchers cérébraux) et un groupe de sujets à dominance hémisphérique droite, établie à l'aide du test de l'amytal de sodium.

Pour le groupe de sujets à dominance hémisphérique gauche le message présenté à l'oreille droite était mieux reproduit que celui présenté à l'oreille gauche. Par contre, le groupe de sujets à dominance hémisphérique droite montrait une supériorité dans le rappel du message présenté à l'oreille gauche.

L'existence d'une relation croisée entre cet effet de latéralité auditive et la latéralisation cérébrale des fonctions du langage est par conséquent assez évidente. Si cela est bien établi depuis Kimura, la nature du mécanisme qui fait que la relation soit justement croisée soulève encore des discussions.

L'interprétation de Kimura (1961, 1967) est la suivante :

Il y a des données physiologiques (Rosenzweig, 1951) montrant que les voies qui lient chaque oreille à l'hémisphère contralatéral sont plus nombreuses ou plus efficaces que les voies ipsilatérales. Quand la voie ipsilatérale et la voie contralatérale à l'hémisphère dominant transportent des informations différentes qui se recouvrent dans le temps, la supériorité de la voie contralatérale deviendrait encore plus importante par la suppression partielle, aux points de rencontre des deux voies, des influx transportés le long de la voie ipsilatérale. Ceci conduirait à un plus grand nombre d'erreurs dans la reproduction du message présenté à l'oreille ipsilatérale à l'hémisphère qui traite les stimuli.

Une nouvelle version de cette interprétation structurale a été établie à la suite des études sur les sujets commissurectomisés, c'est-à-dire, dont le corps calleux et les autres commissures neocorticales ont été sectionnés par intervention chirurgicale. Sparks et Geschwind (1968), Milner, Taylor et Sperry (1968), ont montré que ces sujets, confrontés à des situations d'écoute dichotique de matériel verbal, réalisaient une très bonne performance pour l'oreille droite à côté d'une performance nulle ou quasi nulle pour le message présenté à l'oreille gauche. D'autre part, en présentation monaurale toutes les deux oreilles se sont révélées très efficaces, ce qui montre que la voie ipsilatérale gauche et la voie contralatérale droite étaient intactes et fonctionnelles.

Les patients commissurectomisés en situation d'écoute dichotique réalisent ainsi les conditions de l'interprétation de Kimura: compétition entre la voie ipsilatérale et la voie contralatérale à l'hémisphère dominant avec exclusion de tout autre influx pouvant atteindre indirectement ce dernier. Les résultats indiquent une suppression totale ou presque totale des voies ipsilatérales par les contralatérales.

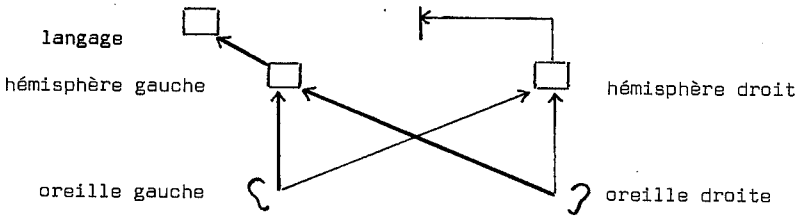


Fig. 1 - Compétition entre la voie contralatérale et la voie ipsilatérale chez les sujets commissurectomisés. (cf. Krashen, 1972)

Ainsi, le transfert de l'information de l'hémisphère droit à l'hémisphère gauche à travers le corps calleux doit intervenir dans une très large mesure dans les performances habituelles de l'oreille gauche; la différence entre les oreilles chez les sujets normaux serait due alors au fait que le message qui prédomine au niveau de l'aire auditive primaire gauche va atteindre les centres du langage en meilleur état et plus rapidement que le message qui prédomine au niveau de l'aire auditive primaire droite.

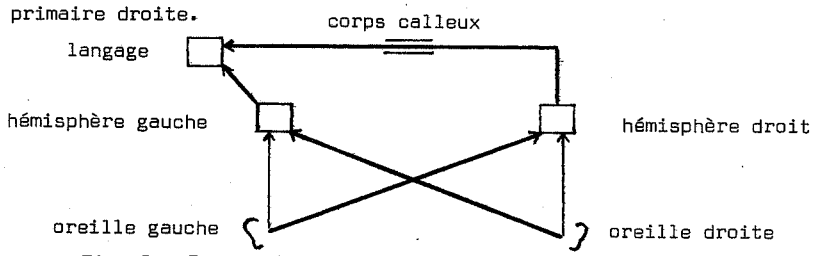


Fig. 2 - Interprétation structurale de l'effet de latéralité en termes de rapport de voies (cf. Krashen, 1972).

Cette interprétation, basée sur des propriétés structurales du sytème auditif, suppose donc que ce qui est critique dans la différence de performance entre les deux messages c'est le fait d'atteindre ou de ne pas atteindre l'oreille plus directement connectée avec les centres de traitement. Cependant, la supériorité dans la reproduction du message présenté à l'oreille droite pourrait être due non pas à de meilleures connexions neurales entre l'oreille droite et les centres de traitement du langage mais à un meilleur traitement du message localisé par le sujet comme provenant de sa droite. A cet égard les situations d'écoute dichotique (et d'écoute monaurale) sont ambiguës, puisqu'un message qui est présenté dans une seule oreille apparaît au sujet comme provenant du côté correspondant de l'espace. La question se pose donc de savoir si l'effet de latéralité dans la perception de messages auditifs est un effet d'oreille d'entrée ou de la position spatiale des sources.

L'hypothèse que le facteur relevant soit la position spatiale des sources est impliquée dans une autre interprétation des différences latérales dans la perception, proposée par Kinsbourne (1970) et selon laquelle l'activation unilatérale d'un hémisphère détermine une orientation involontaire de l'attention vers la moitié contralatérale de l'espace. Dans l'état d'expectative qui précède la présentation dichotique de matériel verbal, l'attention du sujet serait biaisée vers la moitié droite de l'espace.

Nous avons réalisé une expérience qui pourrait permettre d'écarter les interprétations en termes de position spatiale des sources, en utilisant une situation d'écoute diotique de deux messages verbaux simultanés, envoyés au sujet à travers des haut-parleurs (Morais et Bertelson, en préparation). Si aucun effet de latéralité n'était observé lorsque les messages provenaient de la droite et de la gauche du sujet, une interprétation en termes de position spatiale deviendrait intenable. L'expérience comportait deux autres conditions, dans lesquelles l'un des messages provenait d'un haut-parleur situé face au sujet et l'autre message provenait, soit de la droite, soit de la gauche. Certains patterns de résultats, tels que, par exemple, une performance pour la position du milieu intermédiaire des performances pour les positions des extrêmes, seraient incompatibles avec une interprétation en termes d'oreille d'entrée.

A chaque essai, trois paires de syllabes CV simultanées étaient présentées au taux de 2 paires par seconde. Trois haut-parleurs étaient situés en face, à droite et à gauche du sujet, à 1 m. de l'oreille la plus proche. Les sujets, des droitiers manuels, gardaient la tête immobile pendant les essais en mordant dans une empreinte dentaire. Avant chaque essai, ils étaient informés des haut-parleurs qui seraient utilisés. La tâche des sujets était de rappeler immédiatement après l'essai autant de syllabes que possible, dans n'importe quel ordre.

Nous avons observé les résultats suivants : le message provenant de droite était mieux rappelé que le message provenant de gauche; et le message provant du milieu était mieux rappelé que celui de droite et que celui de gauche. Toutes ces différences étaient statistiquement significatives avec un risque d'erreur inférieur à .01.

Tableau I

Pourcentage de syllabes correctement rappelées

Condition	Position		
	G	M	D
D - G	33.9	-	39.3
M - D	-	38.4	32.2
G - M	30.0	40.0	-

Des effets d'asymétrie latérale peuvent donc être observés en situation d'écoute diotique, associés avec les positions des sources dans l'espace.

L'hypothèse, proposée par Kinsbourne, d'une orientation de l'attention vers la droite, déterminée par l'activation unilatérale de l'hémisphère gauche, peut rendre compte des résultats de la condition gauche-droite mais un autre principe serait nécessaire pour rendre compte de la supériorité du milieu. Cette interprétation n'est pas la seule compatible avec un effet de position spatiale; en particulier, à partir de données physiologiques (Rosenzweig, 1954) suggérant une relation entre l'amplitude de la réponse de chacun des hémisphères à un signal et la localisation apparente de ce signal, il serait possible d'élaborer une autre forme assez simple d'interprétation structurale qui rendrait compte de nos résultats. D'autre part, il faut admettre que toutes les versions d'une interprétation en termes d'oreille d'entrée ne sont pas incompatibles avec les résultats obtenus puisque les différences de temps ou d'intensité entre les deux messages au niveau de chaque oreille pourraient jouer un rôle. Quant au délai entre l'atteinte d'une oreille par le message provenant du côté de cette oreille et le message provenant du côté opposé, il ne dépasserait pas 0,7 msec. en cas de synchronisation parfaite des deux messages; mais dans notre matériel les erreurs de synchronisation sont le plus souvent de 5 à 10 msec et par conséquent les différences de temps entre les deux messages au niveau de la même oreille ne peuvent produire aucun effet systématique. Quant à la possibilité que les différences d'intensité suscitent une compétition entre les messages favorisant le plus intense, celle-là ne peut pas être écartée. La supériorité du côté droit dans la condition droite-gauche pourrait être liée au fait que le message de droite atteint l'oreille de ce côté à un plus haut niveau d'intensité que le message de gauche. Pour rendre compte des résultats de la condition milieu-droite il faudrait en plus admettre que cette compétition des messages en fonction des différences d'intensité est aussi relevante au niveau de la voie la plus longue.

Actuellement, nous réalisons un ensemble d'expériences dont le but est de savoir si les différences d'intensité entre les messages au niveau de la même oreille sont ou non une condition nécessaire de l'apparition d'effets d'asymétrie latérale en écoute diotique. En cas de réponse négative, une interprétation structurale de la latéralité en termes de rapport de voies deviendra inadéquate.

Nous voudrions ajouter deux remarques concernant la supériorité de la position du milieu. Premièrement : ce résultat confirme des données obtenues par Treisman (1964) selon lesquelles il est plus facile de répéter mot à mot un message irrelevant envoyé dans l'écouteur droit que de répéter un message envoyé à droite en présence d'un message binaural irrelevant. Treisman n'interprète pas son résultat; nous pouvons simplement ajouter, en fonction de notre expérience, que le facteur critique de l'asymétrie observée par Treisman ne se situe pas au niveau des conditions différentes de présentation des deux messages (stimulation binaurale contre stimulation monaurale). Deuxièmement : quel que soit le mécanisme sousjacent, la supériorité de la position médiane a une valeur adaptative évidente puisqu' on donne priorité aux sons provenant de la personne ou de l'objet que l'on regarde en vision centrale.

BIBLIOGRAPHIE

Broadbent, D. E., 1954, "The role of auditory localization in attention and memory span", J. Exp. Psychol., 47, 191-196.

Kimura, D., 1961, "Cerebral dominance and the perception of verbal stimuli", Canad. J. Psychol., 15, 166-171.

Kimura, D., 1967, "Functional asymmetry of the brain in dichotic listening", Cortex, 3, 163-178.

Kinsbourne, M., 1970, "The cerebral basis of lateral asymmetries in attention", in Attention and Performance III, Ed. A. F. Sanders (North-Holland Publishing Company, Amsterdam), 193-201.

Krashen, S., 1972, "Language and the left hemisphere", Working Papers in Phonetics, n° 24, University of California, Los Angeles.

Milner, B., Taylor, L. B.,et Sperry, R. W., 1968, "Lateralized suppression of dichotically-presented digits after commissural section in man", Science, 161, 184-185.

Morais, J. et Bertelson,P., "Laterality effects in diotic listening", en préparation.

Rosenzweig, M. R., 1951, "Representation of the two ears at the auditory cortex", Amer. J. Physiol., 167, 147-158.

Rosenzweig, M. R., 1954, "Cortical correlates of auditory localization and of related perceptual phenomena", J. Comp. Physiol. Psychol., 47, 269-276.

Sparks, R. W., et Geschwind, N., 1968, "Dichotic listening in Man after section of neocortical commissures",Cortex, 4, 3-16.

Treisman, A. M., 1964, "The effect of irrelevant material on the efficiency of selective listening",Am.J. Psychol., 77, 533-546.

Discussion.

J. BOSQUET : Soit plusieurs sources de son groupées (un ensemble de violons, par exemple) donnant la même note. Les champs créés par ces sources interfèrent en un point d'écoute. L'écoute monaurale de ce champ sera donc "rugueuse" ("harsh"), désagréable, car s'y ajoute des fluctuations inévitables d'intensité. Par contre, l'écoute binaurale restitue à l'ensemble une qualité indéniable.

De même l'écoute monaurale d'un orchestre entier est infiniment moins satisfaisante que son écoute binaurale. On constate une très importante réduction de l'intensité globale et une altération de la qualité du message sonore.

Existe-t-il une explication de ces constatations ?

J. MORAIS : L'écoute binaurale est plus satisfaisante que l'écoute monaurale probablement parce qu'elle procure à celui qui écoute les indices (différences interaurales de phase et d'intensité) qui permettent la localisation précise des sources et, dans le cas de plusieurs sources émettant simultanément, leur discrémination dans l'espace. Quant au phénomène de sommation binaurale d'intensité, il résulte en partie, probablement, d'un processus de recrutement neural, c'est-à-dire d'une extension progressive du nombre de neurones excités.

G. NOIZET : Dans votre expérience, la dominance hémisphérique des sujets a-t-elle été contrôlée et comment les groupes étaient-ils constitués de ce point de vue ? Les groupes étaient-ils appareillés en ce qui concerne les traitements expérimentaux ?

Le tableau des résultats que vous avez présenté fait apparaître que les performances relatives aux sources "droite" et "gauche" chutent toutes les deux lorsque ces dernières entrent en compétition avec la source "milieu" , mais elles le font de manière inégale. La diminution est de 7.1 pour la source droite et de 3.9 pour la source "gauche". Cette différence est-elle significative et, si elle l'est, avez-vous une explication à ce sujet ?

J. MORAIS : La dominance hémisphérique des sujets a été contrôlée sur la base de la haute corrélation existant entre la dominance manuelle droite et la dominance hémisphérique gauche. Après l'expérimentation,

chaque sujet devait répondre au questionnaire de dominance manuelle de Oldfield (1969), adapté au français, et tous les sujets qui avaient un "quotient de latéralité" inférieur à + 70 (-100 correspondant à un gaucher parfait, et + 100 à un droitier parfait) n'étaient pas considérés dans l'analyse des résultats. Notre échantillon serait donc composé, en grande majorité, de gauchers cérébraux.

Tous les sujets passaient toutes les conditions expérimentales; ces conditions étaient réparties de façon aléatoire parmi les essais.

La source "droite", supérieure par rapport à la source "gauche", doit normalement chuter fort quand elle entre en compétition avec la source la plus puissante, tandis que la source "gauche" confrontée dans chaque cas à une source prioritaire par rapport à elle doit normalement chuter moins fort. Il faut surtout noter que, si au point de vue de la performance globale les conditions "milieu-gauche" et "milieu-droite" ne se distinguent pas (35 % et 35.3 %), au point de vue de la différence entre les positions la différence "milieu"-"gauche" est plus importante que la différence "milieu"-"droite" ($p < 0.01$), ce qui veut dire que la droite est un rival plus sérieux de la source du milieu que la gauche.

Oldfield, R.C., 1969, "Handedness in musicians", Brit. J. Psychol., 60, 91-99.

ASYMETRIE LATERALE ET PERCEPTION DE L'ORDRE AU SEIN D'UNE PHRASE PARLEE

Résumé

Dans la situation expérimentale où l'on demande à un sujet d'estimer la position où un clic a été produit pendant l'audition d'une phrase parlée, Fodor & Bever ont montré que le clic est jugé survenir plus tôt quand il est dirigé vers l'oreille gauche, et la phrase vers l'oreille droite, qu'avec la disposition inverse. On passe en revue une série d'expériences menées pour analyser ce nouveau cas d'asymétrie perceptive plus à fond. Des expériences combinant des présentations binaurales et monaurales ont montré que le facteur critique est la position respective dans l'espace auditif des sources apparentes du clic et de la phrase, et non le fait d'atteindre ou non une oreille particulière. Cette conclusion a été confirmée par les résultats d'une expérience plus récente où la position apparente a été manipulée en utilisant des haut-parleurs. L'expérience principale de la série a consisté à tester des sujets israéliens bilingues avec des phrases en français et en hébreu. Les différences latérales obtenues avec le matériel hébreu sont des images en miroir de celles obtenues avec le matériel français, lesquelles sont identiques à celles que fournissent des francophones de naissance. L'existence d'une telle liaison entre le sens de l'asymétrie perceptive et le sens de lecture et/ou d'écriture de la langue employée dans le test suggère l'intervention d'une forme d'imagerie graphique pendant l'écoute de la parole. Le phénomène ne serait pas limité à la situation de la tâche de localisation de clics, car on observe une asymétrie latérale même dans une situation où le sujet n'a pas été prévenu de l'arrivée du clic.

Summary.

In the experimental situation in which a listener is asked to estimate the position where a click has been superimposed on a spoken sentence, Fodor & Bever have found that the click is judged as occuring earlier when it is delivered monaurally

to the left ear, and the speech to the right ear, than with the opposite arrangement. A series of experiments is reviewed, which was carried out to analyze this new case of perceptual asymmetry further. Experiments combining monaural and binaural presentations showed that the crucial factor in determining the effect is the relative position in auditory space of the apparent sources of the speech and of the click, and not the fact of reaching or not reaching a particular ear. This conclusion is supported by the results of a more recent experiment where spatial origin was manipulated by having the components of the task come on loudspeakers with different angular separations relative to the head of the listener. The main experiment in the series is one where bilingual Israeli subjects were tested with both Hebrew and French sentences. The lateral differences obtained with the Hebrew material were exact mirror images of those obtained with the French material, which were identical to those showed by native speakers of French. The existence of such an association between the direction of the lateral asymmetry and the direction of reading and/or writing of the language used in the test suggests the operation of a form of graphical visualization during speech perception. This phenomenon would not be specific of the click location task, since the lateral asymmetry effect is observed also in a situation where the subject has not been warned beforehand of the occurrence of the click.

P. BERTELSON
(Université Libre de Bruxelles)

ASYMETRIE LATERALE ET PERCEPTION DE L'ORDRE AU SEIN D'UNE PHRASE PARLEE

Notre propos est d'examiner si la manière selon laquelle des adultes alphabétisés écoutent une séquence parlée peut être influencée par les propriétés spatiales de la représentation graphique correspondante. Des études récentes ont fourni de nombreux exemples d'une influence des propriétés acoustiques de la parole sur la perception et la mémorisation de matériel verbal présenté par écrit : par exemple, les erreurs commises dans la rétention à court terme de lettres dépendent des qualités sonores des noms utilisés pour désigner ces lettres (Conrad, 1964) ; des syllabes sans signification présentées au tachistoscope sont mieux reconnues quand elles sont facilement prononçables (Gibson, Pick, Osser & Hammond, 1962) .Il y a, à ma connaissance,beaucoup moins d'exemples de la relation inverse.

Nos arguments expérimentaux proviennent de la situation, introduite en 1960 par Ladefoged & Broadbent, où un signal sonore étranger, un clic par exemple,est produit pendant qu'un auditeur écoute une séquence parlée et est prié de localiser ce clic au sein de la séquence. La tâche est typiquement très difficile. Des erreurs d'estimation sont faites qui atteignent plusieurs mots. La taille de ces erreurs suggère plusieurs hypothèses intéressantes sur la façon dont les mécanismes de perception de l'ordre temporel sont mis en jeu dans la perception de la parole (Bertelson & Tisseyre, 1970; Reber & Anderson, 1970; Warren & Obusek, 1971). La situation est surtout connue, toutefois par l'usage qu'en ont fait Fodor et Bever (1965) et le groupe des psycholinguistes du MIT pour argumenter au sujet du rôle de la structure grammaticale dans la perception de la parole. Les problèmes posés par les interprétations de ce groupe sortent du cadre du présent exposé et ne seront pas discutés. Dans leur expérience bien connue de 1965, Fodor & Bever utilisèrent des présentations dichotiques par écouteurs d'oreilles : ils présentaient une phrase à une oreille et un clic à l'oreille opposée. Cette procédure leur permit d'observer incidemment un phénomène curieux : le clic était jugé comme arrivant plus tôt quand il atteignait l'oreille gauche et la phrase l'oreille droite, qu'avec l'arrangement inverse : les erreurs consistant en préposition du clic représentaient 61 % de l'ensemble des erreurs dans le premier cas, 43 % dans le second. L'observation a été confirmée dans des études ultérieures du même groupe (Bever, Lackner & Kirk, 1969; Bever, Lackner & Stolz, 1969). Elle constituera notre point de départ. Apparemment , un

nouveau cas doit s'être ajouté à la liste des asymétries latérales perceptives, dont l'exemple le plus connu est l'effet découvert en 1962 par Kimura : du matériel verbal présenté à l'oreille opposée à l'hémisphère cérébral dominant est mieux reconnu que du matériel présenté simultanément à l'autre oreille (Morais, 1973).

Nous avons entrepris une série d'études visant à analyser le phénomène. Ces études s'organisent selon deux volets.

Le premier concerne les conditions d'apparition de l'asymétrie latérale. A première vue, on aurait affaire à un phénomène de précédence du matériel atteignant une oreille sur celui atteignant l'autre oreille. Mais avant de conclure que le facteur critique est le fait d'atteindre ou non une oreille particulière, il faut considérer que du matériel présenté par écouteur à une oreille semble provenir de la moitié correspondante de l'espace : on pourrait donc aussi avoir affaire à une précédence du matériel provenant de la gauche. Dans une série d'expériences menées avec Françoise Tisseyre, nou avons pu séparer les deux possibilités en utilisant des combinaisons de présentations monaurales et de présentations binaurales (Bertelson & Tisseyre, 1972). Considérons une condition où le clic atteint l'oreille gauche et la phrase les deux oreilles (condition $Ph_{Bi}Cl_{G}$). Si le facteur critique est le fait d'atteindre l'oreille gauche, les deux composantes de la situation le font et doivent donc bénéficier de la précédence. La direction des erreurs observées ne doit pas différer de la condition opposée où le clic atteint les deux oreilles et la phrase l'oreille gauche seulement (condition $Ph_{G}Cl_{Bi}$). En fait, on observe des erreurs significativement différentes. Si on mesure les erreurs individuelles en nombre de syllabes séparant la position estimée de la position objective, et qu'on convient d'appeler négatives les erreurs consistant à situer le clic plus tôt que sa position objective, on obtient pour la condition $Ph_{Bi}Cl_{G}$ une erreur moyenne de - 0.86 syllabes et pour la condition $Ph_{G}Cl_{Bi}$ une moyenne de - 0.53 (2). Il faut maintenant considérer qu'un message présenté de façon binaurale donne l'impression de venir d'une source située au centre de la tête. La source apparente du clic est donc située à gauche de celle de la phrase dans la condition $Ph_{Bi}Cl_{G}$, à droite dans la condition $Ph_{G}Cl_{Bi}$.
Il semblerait donc que le facteur responsable de l'effet observé avec présentation dichotique soit non pas l'oreille d'entrée, mais la position de la source apparente dans l'espace auditif.

D'autres expériences décrites dans l'article cité (Bertelson & Tisseyre, 1972) ont montré par ailleurs que la relation entre erreur moyenne d'estimation et séparation des sources dans l'espace auditif n'est pas monotone : quand le clic vient à gauche de la phrase, il est jugé comme suivant d'autant plus tôt que l'angle de séparation est plus grand; mais quand il vient à droite de la phrase, le degré de séparation n'a plus d'importance et l'erreur moyenne semble en première approximation se maintenir au même niveau que dans les conditions sans séparation spatiale.

Nous n'avons pour le moment aucune explication à offrir pour la forme de cette relation.

Dans l'expérience que nous venons d'évoquer, les impressions de localisation spatiale étaient toujours créées en dirigeant les composantes de la situation (la phrase et le signal) soit vers une oreille particulière , soit simultanément vers les deux oreilles. Dans une expérience plus récente, nous venons de vérifier que les mêmes effets sont obtenus quand on a recours à des présentations par haut-parleurs situés respectivement à droite, à gauche ou en face du sujet, dans le plan médian de la tête.

La question principale qu'on est amené à se poser concerne évidemment l'origine de l'effet. Nous abordons ici le deuxième volet de nos travaux. Les autres cas analysés d'asymétrie latérale se sont révélés liés à deux groupes de déterminants : d'une part, les asymétries structurelles telles que les spécialisations des hémisphères cérébraux pour certaines opérations, d'autre part les asymétries fonctionnelles découlant de la polarité de l'écriture et de la lecture. Les deux groupes ne doivent évidemment pas être considérés comme mutuellement exclusifs. Dans un cas au moins, celui de la supériorité d'un hémi-champ visuel sur l'autre dans la reconnaissance d'un matériel présenté brièvement, il est bien établi que les différences observées résultent à la fois de la spécialisation hémisphérique et des habitudes de lecture (Orbach, 1966).

Nous avons mis à l'épreuve l'hypothèse d'un lien avec la direction de lecture, en testant des sujets dans une langue se lisant de droite à gauche. Nous avons pour cela eu recours à une groupe d'étudiants de nationalité israélienne inscrits à l'Université de Bruxelles. Comme ils étaient tous bilingues hébreu-français, nous avons pu les tester dans les deux langues. Nous avons employé des phrases en hébreu et des phrases en français, enregistrées par le même locuteur, un interprète hébreu-français à la cadence rapide de 7 syllabes/seconde (la même cadence que dans les phrases employées dans les expériences citées précédemment). Trois conditions de présentation sur écouteurs ont été utilisées : Ph_DCl_G, Ph_GCl_D et $Ph_{Bi}Cl_{Bi}$. En plus des 27 sujets israéliens qui ont été testés en français et en hébreu, un groupe-contrôle de 15 sujets ayant le français comme langue maternelle a été testé sur les phrases françaises uniquement. Les résultats qui apparaissent au tableau 1 sont très clairs. Les francophones se comportent comme les sujets des expériences antérieures : ils font des erreurs plus négatives dans la condition Ph_DCl_G que dans les conditions $Ph_{Bi}Cl_{Bi}$ ou Ph_GCl_D. Les Israéliens se comportent comme les francophones en français , mais en hébreu ils fournissent la configuration inverse : ils font des erreurs plus négatives dans la condition Ph_GCl_D que dans les deux autres.

Tableau 1.

Erreurs moyennes, en syllabes, sur les matériels hébreu et français.
Une erreur est appelée négative quand la position estimée est antérieure à la position réelle.

Sujets	Matériel	Conditions		
		$Ph_D Cl_G$	$\overline{Ph_{Bi} Cl_{Bi}}$	$Ph_G Cl_D$
Francophones	Français	-.82	-.49	-.52
Israéliens	Français	-.62	-.27	-.37
	Hébreu	-.21	-.16	-.64

Il apparaît donc une association entre les erreurs de localisation du clic et la direction de lecture de la langue employée dans le test. Avant d'interpréter ce résultat une difficulté devait d'abord être rencontrée. Dans toutes les expériences citées jusqu'ici , les sujets avaient répondu en transcrivant le passage qui leur semblait correspondre à l'endroit où le clic s'était produit et en indiquant sa position par un trait vertical. On pouvait se demander si l'effet d'asymétrie ne résultait pas de cette façon particulière de donner la réponse, le sujet déplaçant son trait dans la direction d'où le clic est venu. Cette éventualité nous a amené à mener une expérience-contrôle où des sujets francophones ont été testés en donnant leurs réponses de deux façons différentes : par écrit comme dans les expériences précédentes, ou oralement. L'effet d'asymétrie se manifeste aussi bien avec réponse orale qu'avec réponse écrite (Bertelson & Tisseyre, 1973). L'hypothèse d'un artéfact lié au mode de réponse semble donc pouvoir être écartée.

On aurait donc affaire à un effet situé au niveau perceptif . Il semblerait qu'à un certain stade de traitement, celui en particulier où l'information relative à la position du clic est obtenue, la phrase soit codée sous une forme qui présente les propriétés directionnelles de la représentation graphique correspondante. On aurait par exemple intervention d'une espèce d'imagerie graphique. Il semble toutefois que peu d'auditeurs aient conscience d'une intervention de ce genre.

Un dernier problème que nous avons abordé concerne la généralité d'une telle intervention du système graphique dans la perception de la parole. L'intervention se produit-elle généralement lorsque nous écoutons la parole ou est-elle propre à la situation expérimentale utilisée? Il faut considérer en effet, que quand on demande au sujet d'exprimer la position temporelle du clic, il n'a guère à sa disposition d'autre façon de coder son impression qu'une traduction en symboles graphiques, c'est-à-dire en termes de lettres. L'effet pourrait donc être créé artificiellement par la tâche de localisation. La seule façon de répondre à la question était d'obtenir des jugements de localisation de sujets non prévenus. Nous avons réalisé l'expérience dans des établissements d'enseignement secondaire de Bruxelles (Athénées), où nous avons testé des élèves des années supérieures, d'âge allant de 16 à 20 ans.

L'expérience se déroulait en collectif, dans des locaux où on avait disposé deux haut-parleurs de part et d'autre du groupe, en arrangeant les tables de telle façon que la moitié des élèves recevaient les phrases à droite et les clics à gauche et l'autre moitié l'inverse. Au départ , on expliquait aux sujets que leur tâche était d'écouter attentivement chacune des phrases présentées pour pouvoir ensuite l'identifier . Après l'audition de chaque phrase , ils devaient tourner la page d'un livret et choisir parmi deux phrases qui leur étaient proposées , celle qui correspondait à la phrase entendue.

On présentait d'abord deux phrases pour lesquelles les sujets accomplissaient effe tivement cette tâche. La troisième phrase était accompagnée d'un clic intense et dès la fir l'expérimentateur faisait remarquer qu'un clic s'était produit et demandait aux sujets d'indiquer la position du clic en traçant une barre verticale à l'endroit correspondant sur la phrase proposée. On annonçait ensuite qu'il faudrait à nouveau juge de la position d'un clic sur une nouvelle phrase et on recueillait ainsi un jugement dans une condition comparable à celle des expériences antérieures. Les phrases qui avaient été présentées dans la condition sans avertissement pour un groupe étaient présentées avec avertissement pour un autre groupe et vice-versa, de façon à contrôler les effets des particularités des phrases individuelles.

Comme il apparaît au tableau 2, un effet d'asymétrie latérale se manifeste clairement dans la condition sans avertissement .Il est même plus fort que dans la condition avec avertissement. On constate aussi que l'erreur tend à devenir positive sans avertissement . C'est un résultat intéressant mais sans rapport direct avec les problèmes discutés ici. Pour ce qui concerne l'asymétrie latérale, il semble donc bien que l'intervention des propriétés graphiques lors de la perception de la parole ne soit pas limitée à la situation où le sujet s'attend à devoir faire un jugement de position.

Tableau 2.

Erreurs moyennes pour un clic inattendu et un clic attendu (écarts-types entre parenthèses).

	Condition de présentation	
	$Ph_D Cl_G$	$Ph_G Cl_D$
Clic inattendu	+ 0.21	+ 1.26
	(2.80)	(3.02)
Clic attendu	- 1.04	- . 55
	(1.48)	(1.56)

Notes.

1. Les recherches décrites ici ont été menées en collaboration avec l'équipe de Max Wajskop, du Laboratoire de Phonétique expérimentale, dans le cadre des contrats 612 et 10.152 avec le Fonds National de la Recherche fondamentale collective. La technique de positionnement des signaux sur les phrases a été mise au point par A. Landercy, et est basée sur l'usage du segmentateur électronique du laboratoire de Phonétique (Landercy, Sylin & Wajskop, 1969).
Les enregistrements de la comparaison hébreu-français ont été réalisés par E. Reichert, du Centre National des Hautes Etudes juives de cette université, et les autres enregistrements par Monique Wajskop.

2. Dans la majeure partie des expériences publiées utilisant la tâche de localisation des clics, une tendance générale aux erreurs négatives, donc à la préposition du clic, se manifeste. Nous avons discuté la portée de ce phénomène ailleurs (Bertelson & Tisseyre, 1970). Pour une raison inconnue, le phénomène n'est pas apparu dans les expériences de Fodor & Bever (1965), Bever, Lackner & Stolz (1969) et de Bever, Lackner et Kirk (1969) mais bien dans les expériences plus récentes de Bever.

References.

Bertelson, P. Listening from left to right versus/to left. Perception, 1972,1, 161-165.

Bertelson, P. & Tisseyre, F. Perceiving the sequence of speech and non-speech stimuli. The Quarterly Journal of Experimental Psychology, 1970, 22, 653-662.

Bertelson, P. & Tisseyre, F. Lateral asymmetry in the perceived sequence of speech nonspeech stimuli. Perception and Psychophysics, 1972, 11, 356-362.

Bertelson, P. & Tisseyre, F. Lateral asymmetry in judgments of click location : not an artifact of reporting mode. Perceptual and Motor Skills, 1973, 36, 849-850.

Bever, T.G., Kirk, R. & Lackner, J. An automatic reflection of syntactic structure. Neuropsychologia, 1969, 7, 23-28.

Bever, T.G., Lackner, J. & Kirk, R. The underlying structure of sentences are the primary units of immediate speech processing. Perception & Psychophysics, 1969, 5, 225-234.

Bever, T.G., Lackner, J. & Stolz, W. Transitional probability is not a general mechanism for the segmentation of speech. Journal of Experimental Psychology, 1969, 70, 387-394.

Conrad, R. Acoustic confusions in immediate memory. British Journal of Psychology, 1964, 55, 75-83.

Fodor, J. & Bever, T.G. The psychological reality of linguistic segments. Journal of Verbal Learning & Verbal behavior, 1965, 4, 414-420.

Gibson, E.J., Pick, A.D., Osser,H. & Hammond, M. The role of grapheme-phenomene correspondence in the perception of words. American Journal of Psychology, 1962, 75, 554-570.

Kimura, D. Cerebral dominance and the perception of verbal stimuli. Canadian Journal of Psychology, 1961, 15, 166-171.

Ladefoged, P. & Broadbent, D.E. Perception of sequence in auditory events. The Quarterly Journal of Experimental Psychology, 1960, 12, 162-170.

Landercy, A., Sylin, G. & Wajskop, M. Etude et réalisation d'un segmentateur électronique et de son organe de commande. Revue d'Acoustique, 1969, 2, 31-36.

Morais, J. L'effet d'asymétrie latérale dans l'écoute multiple (ce volume).

Orbach, J. Differential recognition of Hebrew and English words in right and left visual fields as a function of cerebral dominance and reading habits. Neuropsychologia, 1967, 5, 127-134.

Reber, A.S. & Anderson, J.R. The perception of clicks in linguistic and nonlinguistic messages. Perception & Psychophysics, 1970, 8, 81-89.

Warren, R.M. & Obusek, C.J. Speech perception and phonemic restorations. Perception & Psychophysics, 9, 358-362.

Discussion

J. VAISSIERE : Avez-vous pris en considération dans vos expériences, le degré d'accoutumance du sujet à l'usage de la seconde langue ? Il semble en effet que l'usage quotidien d'une seconde langue affecte sensiblement les mécanismes de compréhension, lecture et écriture. Avez-vous, également pensé à réaliser cette expérience avec des non-bilingues ? Les résultats auraient peut-être été différents.

P. BERTELSON : Comme le résultat obtenu avec des bilingues est très clair, nous n'avions pas tellement de raison de souhaiter étendre l'étude à des unilingues. Le fait, d'ailleurs, que les sujets israëliens fournissent en français à peu près la même configuration de résultats que les francophones de naissance suggère que l'asymétrie latérale n'est pas tellement influencée par le degré de maîtrise de la langue.

M.ROSSI : Que pensez-vous de l'hypothèse selon laquelle cet effet de latéralité serait dû à une spécialisation de chacun des hémisphères ou de chacune des deux oreilles : l'oreille droite favorisant les stimuli de parole, l'oreille gauche favorisant les stimuli acoustiques insignifiants ?

L'expérience citée avec un (s) segmenté à la place du clic n'est pas concluante sur ce point, car ce stimulus ne peut plus être considéré comme un stimulus de parole, mais comme un simple bruit.

P. BERTELSON : Je pense que les résultats des expériences combinant des présentations binaurales et monaurales sont incompatibles avec toute interprétation en termes de précédence, éventuellement spécifique d'une catégorie de stimuli, liée à une oreille particulière. Il reste la possibilité d'un effet lié à la spécialisation des hémisphères dans le traitement des stimuli venant de la moitié contralatérale de l'espace. Les stimuli verbaux atteindraient plus facilement l'hémisphère gauche, spécialisé dans leur traitement, quand ils proviennent de la droite, et vice-versa pour les stimuli "non-verbaux". En raisonnant de cette façon, on s'attendrait à ce que dans ce cas on ait des erreurs de localisation du clic plus positives. Or, c'est le contraire qui est observé. Je pense que le genre de mécanisme que vous évoquez explique les différences latérales dans la reconnaissance, c'est-à-dire le genre d'effet dont José Morais vient de parler. Il me semble peu probable qu'on trouve un

mécanisme qui explique à la fois la meilleure reconnaissance de matériel verbal quand il est présenté à droite, et la précédence relative dans le jugement d'ordre du matériel présenté à gauche. Pour le moment, je pense qu'on a affaire à deux phénomènes distincts, résultant de mécanismes différents.

En ce qui concerne l'expérience de Ladefoged & Broadbent, j'admets que le caractère linguistique d'un (s) isolé est très discutable.

F. SCHEELINGS : Dans quelle mesure tenez-vous compte de la force des clics présentés dans vos expériences ?

P. BERTELSON : Dans une expérience préliminaire, nous avons vérifié que dans la gamme où nous travaillons, ni le sens ni la taille des erreurs ne sont influencés par l'intensité relative du clic et de la phrase. A la suite de ce résultat, il est apparu inutile, dans les conditions où une des composantes est présentée de façon binaurale, d'atténuer son intensité pour contrôler l'effet de la sommation binaurale.

M. CARTIER : Les effets observés avec des clics sont-ils les mêmes qu'avec des segments de parole ?

P. BERTELSON : La seule étude, à ma connaissance, où l'on ait superposé des sons linguistiques à une phrase est la seconde expérience de Ladefoged & Broadbent. Les résultats ne suggèrent pas que ce genre de signal, (il s'agit d'un (s) isolé) donne lieu à d'autres erreurs que des clics, mais évidemment la possibilité d'asymétrie latérale n'a pas été considérée.

S. CASTAN : Avez-vous étudié l'influence des droitiers et gauchers dans le phénomène d'asymétrie latérale ?

P. BERTELSON : Dans la plupart des expériences dont j'ai parlé, les sujets ont été interrogés au sujet de leurs habitudes manuelles, après avoir passé l'épreuve. Les résultats dont j'ai parlé sont toujours ceux des sujets nettement droitiers. Nous avons analysé séparément les résultats des sujets non nettement droitiers. Ce dépouillement ne suggère aucune liaison entre latéralité manuelle et asymétrie perceptive. Il s'agit bien sûr d'une analyse a posteriori de données recueillies de façon systématique. Ce résultat ne permet pas d'éliminer l'hypothèse d'une liaison. Disons qu'il n'encourage pas à explorer cette possibilité.

J. GUIBERT : 1. Envisagez-vous des expériences du type de celles faites avec des sujets israéliens, mais avec des sujets écrivant verticalement ?

2. Peut-on envisager, dans le même domaine, et plus généralement, en ce qui concerne les phénomènes de latéralisation, des tests sur des sujets analphabètes, de façon à voir jusqu'à quel point l'influence culturelle joue un rôle dans ces phénomènes ?

P. BERTELSON : Pour la seconde question, nous devons en effet examiner des populations moins habituées à la lecture que les étudiants d'université ou d'enseignement secondaire que nous avons uniquement étudiés jusqu'ici. Nous envisageons notamment de procéder à une étude de la genèse de l'effet, avec des enfants de différents âges.

En ce qui concerne les langues s'écrivant verticalement, il y a certainement là une possibilité intéressante.

NORMALISATION FREQUENTIELLE DE LA PAROLE

Résumé

Les différences observées dans le plan temps-fréquence entre les voix de divers locuteurs peuvent être rangées dans les catégories suivantes : netteté d'élocution, niveau sonore, timbre, et distorsions du squelette informatif. Cette dernière classe comprend notamment une transformation linéaire de l'échelle fréquentielle, définie par un coefficient caractérisant chaque locuteur. On montre expérimentalement que les particularités individuelles de ce type peuvent être compensées par l'analyse en agissant sur les fréquences centrales des filtres.

Summary

The acoustical properties of individual voices can be classified into the following classes : speech clearness, sound level, timbre and distortions of the informative frame. The latter includes particularly a linear transformation of the frequency scale, defined by a coefficient characterizing each speaker. It is shown experimentally that individual particularities of that kind can be balanced right from the analysis by adjusting the central frequencies of the filters.

J.S LIENARD, J. SAPALY, M. MLOUKA.

Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur.

NORMALISATION FREQUENTIELLE DE LA PAROLE

J.S. LIENARD (*), J. SAPALY (**), M. MLOUKA (*)

La reconnaissance automatique de la parole connaît certains succès, quelle que soit la méthode employée, lorsqu'elle s'applique à un seul locuteur ou à un petit groupe de locuteurs sélectionnés. Pour qu'elle puisse s'appliquer à un locuteur quelconque, sous réserve qu'il s'agisse d'une même langue et d'un même accent, il faut pratiquer une normalisation des données issues de l'analyse. Cette normalisation est explicite si elle consiste en une transformation visant à compenser les différences entre locuteurs, ou implicite si la reconnaissance opère sur des paramètres choisis pour leur invariance d'un locuteur à l'autre. La difficulté est de savoir sur quels aspects du message parlé doit porter la normalisation.

Nous allons étudier rapidement les principales différences acoustiques entre locuteurs, observables dans le plan temps-fréquence, avant d'approfondir quelque peu l'une d'entre elles que nous nommons anamorphose fréquentielle de la parole, et d'examiner ses conséquences pour la reconnaissance automatique.

I- LES PRINCIPALES DIFFERENCES ACOUSTIQUES ENTRE LOCUTEURS.

I-1- La netteté d'élocution

Selon le soin apporté par le locuteur à son articulation, l'image temps-fréquence du message est plus ou moins nette, au sens graphique du terme: Les formants sont plus ou moins bien marqués, les explosions plus ou moins accentuées, etc... Ce facteur, d'une importance vitale en reconnaissance, est extrêmement difficile à contrôler, et même à estimer. Il dépend du tonus musculaire, de la conformation et des habitudes articulatoires du sujet. Peu étudié jusqu'à présent, il est cependant bien connu empiriquement par les professionnels de la diction, mais souvent confondu

(*)- Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI) - C.N.R.S, ORSAY (Pr. MALAVARD).
(**)-Section d'Electronique de l'Institut de Mécanique Théorique et Appliquée (IMTA) de l'Université PARIS VI- St. Cyr L'Ecole (Pr. SIESTRUNCK).

avec les facteurs purement prosodiques que sont le rythme, l'intonation et l'accentuation.

I-2- Le niveau sonore de la parole

L'extraordinaire faculté d'adaptation de l'oreille fait que l'on ne remarque pas, dans la conversation courante, les variations de niveau sonore tenant au locuteur.

Ces variations, souvent considérables, doivent être compensées autant que possible lors des enregistrements destinés à la reconnaissance automatique. Encore faut-il distinguer entre le niveau moyen d'un enregistrement, assez facile à normaliser en machine, et le niveau instantané, dont la régulation est d'autant plus délicate que l'on considère des intervalles de temps plus petits. D'un locuteur à l'autre les disparités sont manifestes dans l'évolution du niveau instantané: tel locuteur fournit sur telle explosion un niveau de vingt décibels plus élevé que ne le fait tel autre, sans que cette différence soit perçue de manière consciente par l'auditeur.

Le niveau sonore reflète surtout l'intensité des premiers harmoniques de la parole, et n'est pas un paramètre porteur d'information sémantique, ce qui n'est pas le cas pour le timbre de la voix.

I-3- Le timbre de la voix

Le timbre de la voix est en relation assez floue avec la parole, puisque celle-ci est elle-même constituée par la variation de l'énergie spectrale dans le temps. Cependant, si l'on ne s'intéresse qu'aux aspects particuliers du timbre que sont la nature du spectre et la répartition moyenne de l'énergie spectrale, on peut caractériser le type de voix adopté et, dans une certaine mesure, le locuteur.

Le spectre peut être quasi-harmonique (voix normale, ou voisée) ou continu (voix chuchotée). Mais en général les deux types de spectres existent simultanément, dans des proportions variables. Un locuteur peut adopter toutes les nuances intermédiaires entre la voix chuchotée faible, à peine perceptible à une vingtaine de centimètres, et la voix criée qui porte à plusieurs centaines de mètres. En dehors du niveau sonore croissant, les types de voix intermédiaires se caractérisent essentiellement par:

-L'atténuation, puis la disparition des bruits d'explosion et de friction.

-L'apparition progressive du spectre de raies: en voix semi-voisée, par exemple, la partie inférieure du spectre se compose du fondamental et de quelques harmoniques, alors que la partie supérieure a la même structure

qu'en voix chuchotée.

- L'augmentation, en voix forte ou criée, de la fréquence moyenne du fondamental, ainsi que l'enrichissement relatif du spectre.

D'un locuteur à l'autre ces mêmes caractéristiques existent, et s'ajoutent aux caractéristiques individuelles. Pour comparer le spectre moyen de deux locuteurs, il faut être sûr que ceux-ci ont adopté le même type de voix, ce qui impose quelques précautions lors de l'enregistrement.

Sans entrer dans le détail, signalons que ces problèmes sont souvent ignorés en reconnaissance automatique, dont ils conditionnent pourtant le taux de réussite.

I-4- Les distorsions du squelette informatif

Nous appelons squelette informatif (ou sémantique) de la parole la structure constituée dans le plan temps-fréquence par les traces des formants, les bruits d'explosion et de friction, dans la mesure où ces éléments convoient une information sémantique, ce dont on peut s'assurer par synthèse.

Conformément aux théories gestaltistes et structuralistes, le squelette informatif est susceptible de subir certaines déformations, ou anamorphoses, sans perdre son caractère de totalité (1) (2). L'anamorphose peut affecter la dimension de temps, la dimension de fréquence, ou les deux à la fois (Fig 1); elle peut être simple ou complexe, mais elle est exprimée dans tous les cas par une transformation mathématique, portant sur les échelles de temps et de fréquence (3).

a) Les anamorphoses temporelles

Ce sont les plus évidentes, celles dont la reconnaissance de la parole s'occupe depuis une dizaine d'années. La même phrase prononcée plus ou moins vite par le même locuteur donne lieu en première approximation à une anamorphose régulière, c'est à dire à une simple dilatation de l'échelle de temps, facile à compenser lors d'un traitement numérique: il suffit de normaliser par rapport à la durée totale. Mais ceci n'est valable qu'en première approximation, pour un seul locuteur. En pratique les distorsions localisées de l'échelle de temps sont les plus gênantes, dans la mesure où elles sont liées au rythme de la parole, facteur prosodique transportant peu d'information sémantique.

Pour avoir quelque chance de succès, la reconnaissance automatique doit en premier lieu s'affranchir de ces distorsions. Ceci peut être obtenu de différentes manières ; nous avons, de notre côté, effectué diverses

expérimentations (3) (4), les unes en utilisant la fonction de stationnarité de la parole, les autres en développant une méthode apparentée à la programmation dynamique (méthode des correspondances).

b) Les anamorphoses fréquentielles

Un même locuteur est capable de modifier d'un instant à l'autre son échelle fréquentielle (Fig 2). Ceci permet de donner à la voix chuchotée une certaine intonation, que nous pouvons appeler intonation "formantique" pour la différencier de l'intonation "mélodique" liée à l'évolution du fondamental en voix voisée (1).

D'un locuteur à l'autre on peut constater qu'il existe de telles distorsions de l'échelle fréquentielle, qui se ramènent pour l'essentiel à de simples affinités. La proposition classique selon laquelle une voyelle est définie par la fréquence de ses formants est inexacte. Il est plus raisonnable de caractériser une voyelle par les rapports des fréquences de formants, ceci dans certaines limites et en première approximation seulement.

Mais la notion d'anamorphose ne s'applique pas aux seules voyelles: c'est l'ensemble du squelette informatif qui peut subir une dilatation de l'échelle fréquentielle. Ce fait, que nous avons observé en analyse et démontré par synthèse (3), va nous servir à normaliser en fréquence, dès l'entrée d'un système de reconnaissance, la parole d'un locuteur donné.

II - NORMALISATION FREQUENTIELLE DE LA PAROLE

II-1- L'analyseur - anamorphoseur

Cet appareil a été construit en vue de pratiquer dès l'analyse l'anamorphose nécessaire pour que l'échelle de fréquence du locuteur X soit sensiblement identique à celle du locuteur R considéré comme référence.

Il s'agit d'un banc de 32 filtres, couplé à un ordinateur, dont l'étude a été menée dans le cadre de contrats DRME et CRI (Fig 3);les filtres peuvent être réglés en fréquence centrale, en largeur de bande et en gain. Une commande unique permet en outre de multiplier toutes les fréquences centrales par un même coefficient (coefficient d'anamorphose, ou d'affinité fréquentielle).

II-2- Détermination du coefficient d'affinité

Lors d'une étude antérieure nous avons mis au point un programme (ADAPT) qui permet de calculer le coefficient d'affinité existant entre deux séquences numérisées X et R, de même contenu phonétique, normalisées en temps (3)(4). Le principe consiste à rechercher le coefficient d'affinité

fournissant le maximum de ressemblance (au sens d'un certains critère de distance) entre les séquences X et R.

Nous avons utilisé ce même programme pour vérifier que l'anamorphose inverse effectuée sur l'analyseur lui-même permettait de ramener n'importe quelle voix à une même échelle fréquentielle.

II-3- Processus expérimental

Un même ensemble de mots et de phrases a été enregistré par une dizaine de locuteurs hommes, femmes et enfants, sans précautions particulières quant à la prononciation. Les différentes réalisations d'une même phrase ("Lucie et Chantal sont à l'école pour deux heures"), choisie pour sa représentativité sur le plan phonétique, ont été analysées avec les mêmes réglages de l'analyseur (filtres répartis linéairement entre 0 et 5 KHz, largeur de bande 200 Hz).

Les analyses obtenues (tableaux de nombres analogues à des sonagrammes) ont été normalisées en temps manuellement, de façon à être comparables dans cette dimension. En d'autres termes on a prelevé dans les tableaux de nombres les colonnes représentant les spectres jugés significatifs du point de vue phonétique. Cette opération pourrait être effectuée de manière automatique (3)(4), avec cependant un léger risque d'erreur qui nous a fait préférer ici une procédure manuelle. Le coefficient d'affinité attaché à chaque locuteur, par rapport à l'un des locuteurs considéré comme référence, a ensuite été déterminé par le programme ADAPT. Puis une autre phrase ("Le clairon réveille le soldat") a donné lieu à analyse, mais cette fois l'analyseur était réglé pour compenser exactement l'anamorphose calculée précédemment sur chaque locuteur. Un passage du programme ADAPT sur ces nouvelles données, avec le même locuteur - référence, devait en principe fournir la valeur 1 pour chaque locuteur. La figure 4 montre que ce résultat est atteint avec une bonne approximation.

III - CONCLUSION

Nous avons tenté de classer selon leur nature acoustique les aspects du message parlé caractérisant chaque locuteur, qui sont autant de sources d'erreur pour la reconnaissance automatique de la parole. Nous avons donc distingué la netteté d'élocution, le niveau sonore, le timbre, et les distorsions spectro-temporelles du squelette informatif. Les

anamorphoses fréquentielles appartiennent à cette dernière catégorie. Elles peuvent en général être considérées comme de simples affinités, et chaque locuteur est caractérisé sous cet aspect par un coefficient calculé à partir de l'analyse d'un mot-clé, qui le situe par rapport à un locuteur de référence. Ce point a été vérifié expérimentalement au moyen d'un appareillage, l'analyseur anamorphoseur, prévu pour assurer dès l'analyse ce type d'adaptation au locuteur.

IV- BIBLIOGRAPHIE

(1) - E. LEIPP, M. CASTELLENGO, J. SAPALY, J.S. LIENARD.
Structure physique et contenu sémantique de la parole.
Colloque sur la parole organisé par le GALF à Grenoble, avril 1967.
La Revue d'Acoustique n° 3-4, 1968.

(2) - E. LEIPP, M. CASTELLENGO, J. SAPALY, J.S. LIENARD.
Les anamorphoses de la parole et leur simulation sur ordinateur.
Comptes-rendus du 7e Congrès International d'Acoustique, Budapest, août 1971.

(3) - J.S. LIENARD - Analyse, synthèse et reconnaissance automatique de la parole.
Thèse d'état, Université de PARIS VI, avril 1972.

(4) J.S. LIENARD, M. CASTELLENGO, E. LEIPP, M. MLOUKA, G. RENARD, J. SAPALY, D. TEIL. Quelques idées directrices en reconnaissance automatique de la parole.
Automatisme, T XVIII, n° 3, mars 1973.

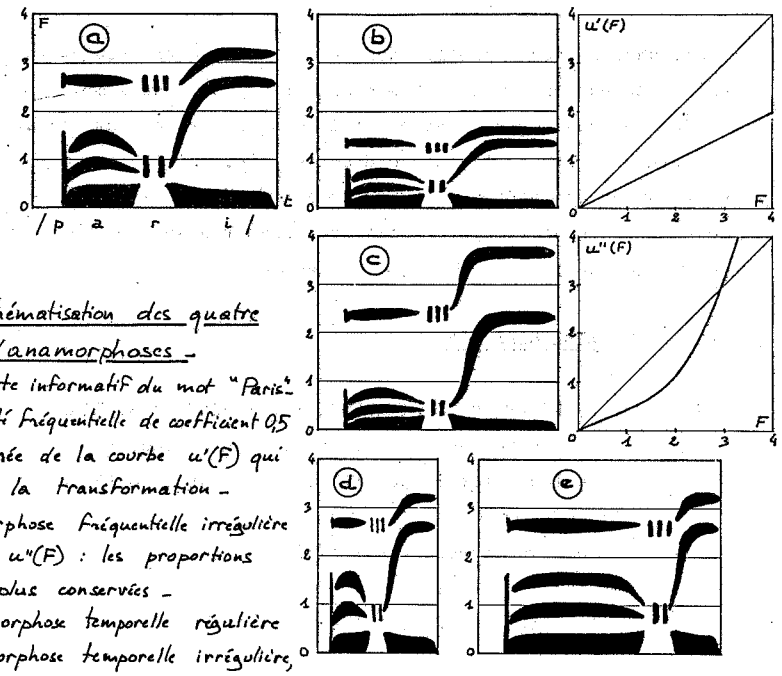


Fig 1

Schématisation des quatre types d'anamorphoses -

a) Squelette informatif du mot "Paris" -
b) Affinité fréquentielle de coefficient 0,5 accompagnée de la courbe u'(F) qui représente la transformation -
c) Anamorphose fréquentielle irrégulière et courbe u"(F) : les proportions ne sont plus conservées -
d) Anamorphose temporelle régulière
e) Anamorphose temporelle irrégulière, ou anamorphose rythmique -

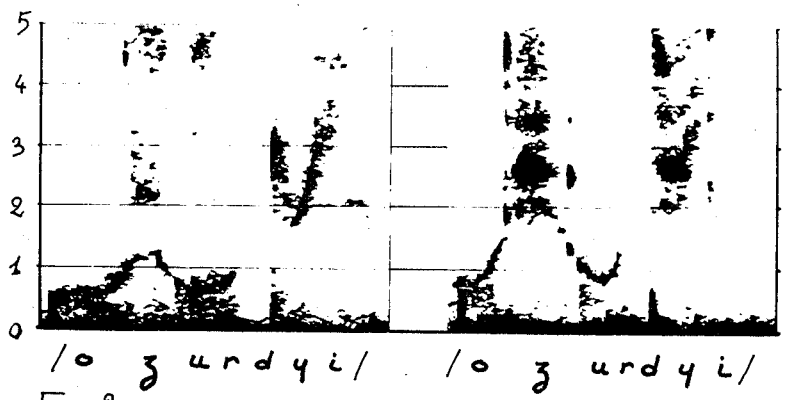


Fig 2

Anamorphose fréquentielle d'une voix réelle

Le mot " aujourd'hui " est ici prononcé par le même locuteur, en voix chuchotée, de deux manières différentes - L'anamorphose n'est pas tout à fait régulière -

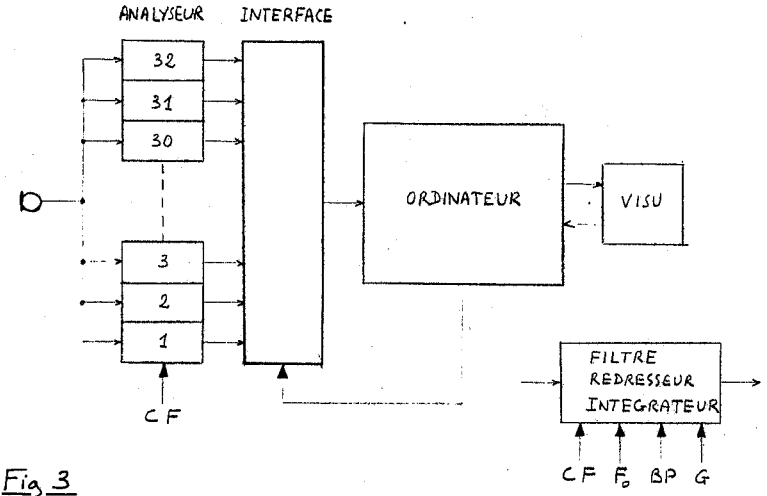


Fig 3

Couplage de l'analyseur-anamorphoseur avec l'ordinateur et l'écran de visualisation -

Les paramètres suivants des filtres d'analyse sont réglables :

f_0 : Fréquence centrale

CF : commande d'affinité, commune à tous les filtres

BP : bande passante

G : gain

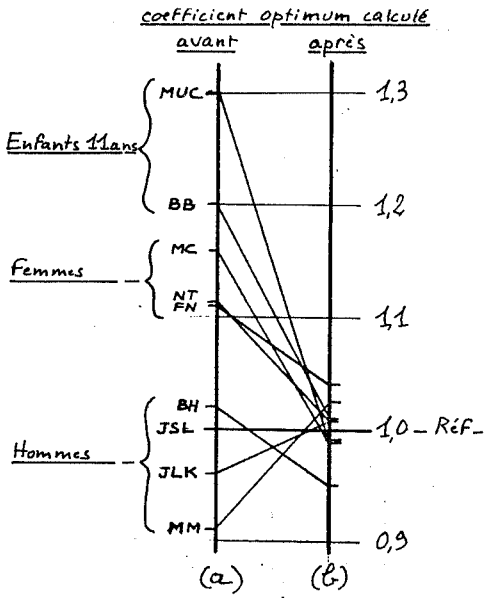


Fig 4

Réduction de la dispersion des échelles fréquentielles propres à plusieurs locuteurs, par rapport à une même référence -

a) Sans anamorphose

b) Avec anamorphose, calculée d'après l'analyse d'un mot-clé.

Discussion.

R. DE MORI : Pensez-vous que la reconnaissance de la parole passera nécessairement par une normalisation du type que vous proposez ?

J-S. LIENARD : Oui, dans la mesure où l'on cherche à appliquer la reconnaissance à un locuteur quelconque. Les anamorphoses que j'ai mentionnées sont des faits d'observations, les normalisations correspondantes pourront prendre d'autres formes que celles que nous proposons, explicites ou implicites, mais il est impossible d'ignorer ce problème.

J-P. HATON : Ne pensez-vous pas qu'il faudrait définir un coefficient d'anamorphose variable avec la fréquence, et non pas un coefficient fixe pour tout le spectre, ce qui reviendrait à définir une transformation plus complexe du spectre ?

J-S. LIENARD : En effet, c'est ce que nous avons appelé "anamorphose fréquentielle irrégulière". Un exemple en est donné sous forme schématique dans le figure 1 c, et la transformation est définie par une fonction reliant l'échelle fréquentielle du locuteur X à celle du locuteur R. Pour être complet il faudrait même prendre en compte l'évolution de cette fonction avec le temps. Mais nos relevés expérimentaux montrent que cette fonction peut être considérée comme une droite en première approximation; cette assimilation au premier ordre est d'ailleurs naturelle dans la mesure où la détermination des fréquences de formants est très imprécise.

G. PERENNOU : Est-ce que l'anamorphose, qui préserve l'intelligibilité d'un mot entier, conserve également l'intelligibilité des voyelles constituant le mot ?

J-S. LIENARD : Je pense que vous faites allusion aux tests d'intelligibilité effectués avec la voix synthétique (Icophone). Nous n'avons pas fait de mesures sur les voyelles isolées, celles-ci nous paraissant dénuées d'intérêt pratique. Les voyelles isolées, parfaitement stables, n'apparaissent pas dans la parole réelle, le mécanisme de leur perception n'est pas nécessairement identique à celui qui nous intéresse.

C. ROCHE : Comment est calculée l'anamorphose utilisée par M. Liénard pour sa normalisation en fréquence ?

J-S. LIENARD : Le coefficient caractérisant chaque locuteur par rapport à un locuteur considéré comme référence est calculé à partir de l'analyse d'un mot-clé. On fait évoluer l'échelle fréquentielle de l'un par rapport à l'autre jusqu'à trouver un maximum de coïncidence entre les sonagrammes numériques, préalablement normalisés en temps. Le coefficient correspondant est alors considéré comme optimum.

Question de M. GRESSER : voir réponse de M. Maissis.

NORMALISATION DES PARAMETRES PHONEMIQUES EN RECONNAISSANCE AUTOMATIQUE DE LA PAROLE.

Résumé.

Cet article présente une méthode visant à réduire la dispersion du signal vocal par rapport aux occurrences et aux locuteurs. Des paramètres de reconnaissance phonémiques déduits d'un modèle approximatif de la membrane basil aire, sont normalisés par un ensemble de paramètres spectraux. Une procédure élémentaire de reconnaissance des formes illustre alors la qualité des résultats par ses taux de reconnaissance élévés.

Summary.

This paper introduces a normalisation method for reducing the dispersion of speech signals versus occurences and speakers. The recognition parameters are deduced from an approximate model of basilar membrane. They are then normalized by a set of spectral parameters. A simple procedure provides high recognition rates for phoneme recognition.

A.H. MAISSIS
Ecole Nationale Supérieure des Télécommunications.

Introduction :

La diversité du signal vocal pour un seul locuteur et encore plus pour plusieurs locuteurs est très grande . L'idée de normaliser des paramètres extraits du signal en vue d'une reconnaissance automatique de la parole s'introduit ainsi naturellement . Le problème qui se pose est le suivant " quels paramètres normaliser , et par rapport à quoi ". Le présent papier essaie de donner une réponse possible . L'idée centrale est que les paramètres extraits d'un modèle de l'oreille humaine peuvent servir à la reconnaissance s'ils sont normalisés par rapport à une information de nature spectrale. Cette idée est d'une certaine façon la conséquence d'hypothèses physiologiques suivant lesquelles l'excitation du système neural par les déplacements verticaux d'un point sur la membrane basilaire est modulée par le voisinage du point considéré et surtout par le point de résonance le plus proche.

1. PARAMETRISATION

La méthode de prétraitement utilisée se caractérise par l'élaboration de deux séries de paramètres :

- des paramètres de reconnaissance issus d'un modèle approximatif de la membrane basilaire.
- des paramètres de normalisation déduits d'une analyse spectrale.

1.1. Rappels sur les paramètres de reconnaissance

Le système de prétraitement du signal vocal a déjà été exposé par ailleurs (/1/,/2/,/3/) et est représenté par la figure 1. Après un découpage initial du signal temporel en éléments de 10 ms (segments minimaux j), on effectue la séparation des sons sonores et sourds par un algorithme que nous avons développé pour la détection du pitch /4/ . Ces segments sont ensuite analysés par un modèle de la membrane basilaire . Ce modèle est basé sur les travaux de Bekesy et est simulé sur ordinateur d'après la méthode de FLANAGAN (/5/ , /6/).

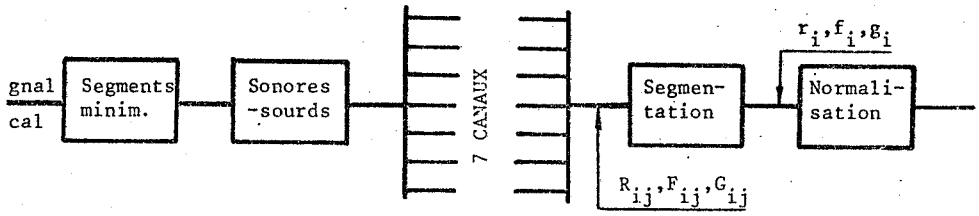


Figure 1 : Prétraitement du signal.

On peut admettre que la réponse d'un point de cette membrane (déplacement verticaux $x(t,l)$ à la distance l de la fenêtre ovale) à l'excitation par un signal sonore peut être simulée par un filtre à large bande caractérisé par :

- une fréquence de résonnance dépendant du point considéré .
- une bande passante à 3 db proportionnelle à cette fréquence .
- un gain maximal .

La méthode de paramètrisation basée sur ces données utilise la réponse temporelle $x(t,l)$ en sept points de la membrane simulée et les dérivées spatiales correspondantes définies comme suit :

i - les sept canaux admettent pour fréquences de résonance : 300,470,800, 1200,2100,3600,4600, Hz de façon à couvrir une part importante du spectre de la parole . (Figure 2).

ii - l'approximation de $\partial x(t,l)/\partial l$ est réalisée par la création d'une deuxième série de 7 filtres du type précédent de fréquences : 260, 410, 720, 1200,1900,3350,4200 Hz correspondant à un Δl de 0,5 mm . (Figure 4).

Pour le segment minimal j , on détermine le maximum du signal à la sortie du filtre (i), soit A(i,j) ; et le maximum de la différence des sorties des filtres (i) et (ii), soit A'(i,j) (figure 3). On dispose ainsi d'un paramètre de reconnaissance constitué par le rapport:

$$R_{ij} = \frac{A(i,j)}{A'(i,j)} \qquad i = 1 \ldots 7$$

On notera que cette méthode réalise la conjonction de données spectrales et temporelles grâce à l'analyse du signal par un système régi par des équations aux dérivées partielles (propagation des vibrations le long de la membrane).

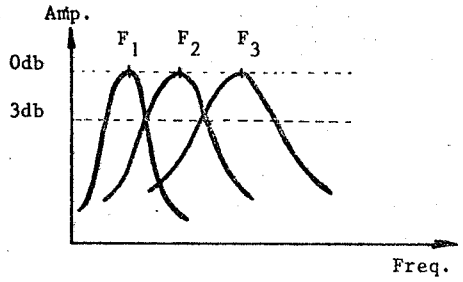


Fig 2: Choix des canaux

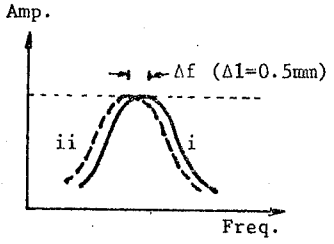
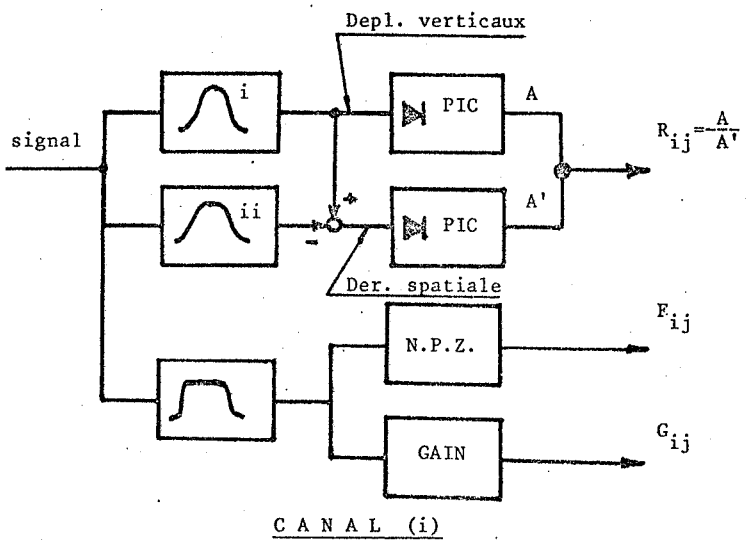


Fig 4: Filtres (i) et (ii)



CANAUX (i)	(ii)
300	260 HZ
470	410 "
800	720 "
1300	1200 "
2100	1900 "
3600	3350 "
4600	4200 "

Figure 3 : Extraction des paramètres

1.2. Paramètres de Normalisation

Parallèlement au calcul de ces paramètres principaux , on détermine pour chaque canal une fréquence dominante F_{ij} et une atténuation G_{ij} Ces paramètres de normalisation sont extraits du signal par passage dans un filtre rectangulaire de même bande passante que le filtre principal (i).

La détermination du nombre de passagess à zéro de la sortie de ce filtre fournit une approximation de la fréquence dominante dans cette bande soit F_{ij}.

L'atténuation G_{ij} permettant d'estimer l'importance de cette fréquence est définie comme le rapport sur un segment minimal des maxima des signaux de sortie et d'entrée.

A l'issue de la phase de prétraitement on dispose donc de 21 paramètres pour chaque segment minimal . L'étude des régions de stabilité relative de ces paramètres permet d'agréger les segments minimaux en

éléments de reconnaissance qui apparaissent très proches des phonèmes . Un lissage de ces régions fournit une représentation phonèmique par 21 paramètres :

$$R_{ij} , F_{ij} , G_{ij} \qquad i = 1...7$$

2. NORMALISATION.

2.1. Les données

Nous avons , dans toutes nos expériences, travaillé sur les sons sonores (voyelles et consonnes). Les occurrences de ces phonèmes sont tirées d'une liste de 235 mots du français courant prononcés par 2 locuteurs à voix très différentes . Les caractéristiques de ces données sont :

- locuteur N°1 fréquence fondamentale moyenne 110 Hz
- " N°2 " " " 180 Hz
- 2 enregistrements (de la même liste) par le locuteur 1
  1 enregistrement par le locuteur 2
- nombre de phonèmes : 22
- L'échantillon d'apprentissage est constitué par le premier enregistrement des locuteurs 1 et l'enregistrement du locuteur 2.
- nombre d'occurrences par phonème et par enregistrement très variable allant de 11 à 60.

2.2. Normalisation par rapport aux occurrences

Les paramètres (R,F,G) , par leur nature même , ne sont pas indépendants entre eux , des corrélations plus ou moins importantes pouvant être discernées. Nos premières expériences sur ce sujet se réfèrent à la première définition du système (/1/,/2/) quand le nombre de paramètres extraits était égal à 10 ; soient les 7 paramètres R_i et les trois premiers formants F_i . Des résultats expérimentaux , basés sur un seul locuteur, ont montré l'existance d'une relation approximativement linéaire entre les R_i et F_i pour diverses occurrences d'un même phonème . La figure 5 donne un exemple relatif au phonème 'A' (/2/).

Une étude théorique menée en assimilant la membrane basilaire à une membrane élastique mince enroulée en spirale , a permis de montrer /7/ qu'une relation approchée réalisant la normalisation dans un large domaine est du type :

$$R = \frac{a}{\sqrt{F^2 - b}} \qquad (1)$$

L'introduction d'un deuxième locuteur à voix très differente de celle du premier, permet de montrer que la relation précédente n'est plus valable . C'est pour affiner cette relation que les paramètres supplémentaires ont été introduits (7 de type F au lieu de 3 et 7 de type G).

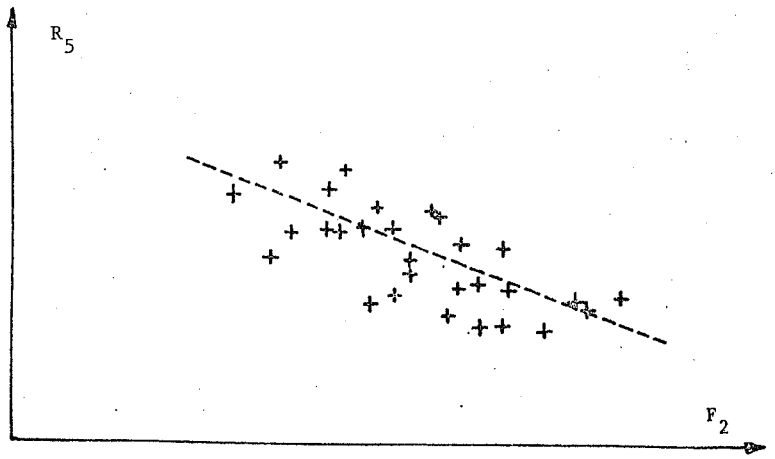


Figure 5 : Relation R_5 et F_2 pour le phonème 'A'

2.3. Normalisation par rapport aux locuteurs

Les résultats que nous venons de présenter nous conduisent à envisager la normalisation par rapport aux occurrences et aux locuteurs. L'idée de base reste toujours la même : l'excitation neurale résultant des déplacements verticaux d'un point sur la membrane basilaire dépend d'une part de ces déplacements (exprimés par nos paramètres de type R) et d'autre part du point résonnant dans le voisinage du point considéré (exprimé par nos paramètres spectraux F,G). C'est ainsi que vient de façon naturelle l'idée de chercher une relation entre les paramètres R et les paramètres F,G.

Nous allons , dans les deux chapitres suivants , proposer deux approximations possibles de cette relation , Nous allons étudier plus profondément la deuxième approximation et donner à la fin un exemple d'utilisation de la normalisation en reconnaissance phonèmique.

2.3.1. Normalisation des R_i par rapport aux F_i et G_i.

Nous avons cherché à mettre en évidence une relation entre les Ri Fi Gi du type :

$$f (R_i , F_i , G_i) = 0 \qquad (2)$$

traduisant pour chaque canal , l'existence,dans l'espace à 3 dimensions, d'une surface contenant approximativement les points représentatifs des occurrences d'un phonème donné . Remarquons que si l'existence de telles surfaces peut être confirmée , le problème de reconnaissance se réduit à la représentation de chaque classe par 7 surfaces et la prise de décision par un calcul de la distance du point , représentant le phonème à reconnaître par rapport à ces surfaces .

On cherchera donc à approximer la relation (2) par un developpement polynomial de type :

$$f(R,G,F) = R + \sum_{k=1}^{N} a_k . F^k + \sum_{k=1}^{M} b_k . G^k = 0 \qquad (3)$$

{ a_k , b_k : coefficients à déterminer }

et cela pour chaque canal.

Il est interessant de remplacer, dans la formule (3), les R,F,G par leurs variations par rapport aux moyennes respectives.

$$\begin{aligned} R : \quad & \Delta R = R - \bar{R} \\ F : \quad & \Delta F = F - \bar{F} \\ G : \quad & \Delta G = G - \bar{G} \end{aligned} \qquad (4)$$

On obtient ainsi la surface f (ΔR, ΔF, ΔG) = 0 qui passe par le point ($\bar{R},\bar{F},\bar{G}$). Considérons la figure 7,où nous avons tracé cette surface. Soit 1 le point correspondant à une occurrence quelquonque (R,F,G). Considérons la surface f_1 passant par 1 et obtenue par translation verticale de δR de la surface f. Le point 2 (sur la surface f_1 et d'abscisse $\bar{F},\bar{G}$) sera considéré comme le point 1 normalisé. L' ordonnée R_n est définie comme étant le paramètre R normalisé.

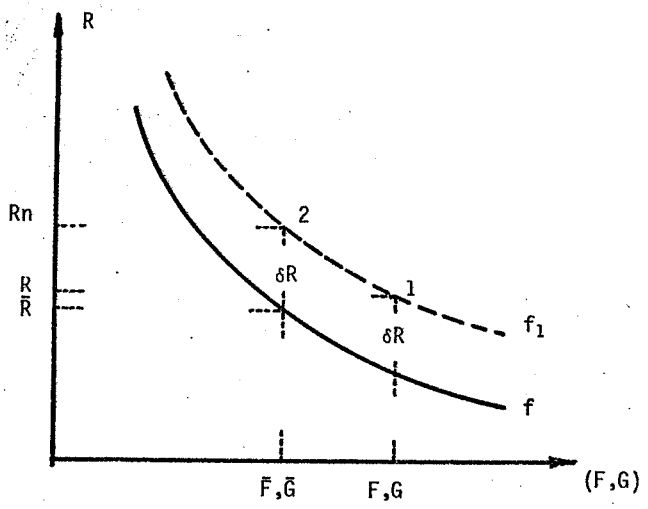


Figure 7

Nous pouvons alors écrire la formule (3) :

$$R_n = R + \sum a_k . \Delta F^k + \sum b_k . \Delta G^k \qquad (5)$$

Le calcul des coefficients a_k et b_k est réalisé par la méthod des moindres carrés appliquée à la variance des paramètres normalisés ; c'est à dire en minimisant sur l'ensemble de K occurrences la quantité:

$$Q = \frac{1}{K} \sum^{K} (R_n - \bar{R}_n)^2 = E(R_n - E(R_n))$$

{ $\bar{R}_n$: moyenne de R_n }

On aura alors à résoudre le système linéaire défini par :

$$\partial Q / \partial a_k = 0 \quad (k = 1...N) , \quad \partial Q / \partial b_k = 0 \quad (k = 1..M) \qquad (6)$$

En travaillant de cette façon sur chaque canal et sur chaque classe , lors de l'apprentissage , cette dernière peut donc être

caractérisée par :

- 7 ensembles de coefficients a_k, b_k
- 7 moyennes $\bar{F}_i$ et $\bar{G}_i$
- le centre de gravité $\underline{\bar{R}}_n$ (7 coordonnées) du nuage normalisé.
- la matrice de covariance de la distribution .

$$\Sigma = E\{ (\underline{R}_n - \underline{\bar{R}}_n).(\underline{R}_n - \underline{\bar{R}}_n)^T \}$$

$$= \{ \sigma_{ij} \}$$

$$\sigma_{ij} = E\{ (R_n(i) - \bar{R}_n(i)).(R_n(j) - \bar{R}_n(j)) \}$$

Résultats - Remarques.

Cette normalisation aboutit à une réduction importante de la dispersion du nuage comme le montre le tableau 1 . Le gain obtenu est en règle générale de 1:2 à 1:5 .

- la formule (3) étant d'ordre supérieur à 1 (dans ce cas concret on a pris N = M = 3) , cette normalisation ne saurait être atteinte par des approches linéaires telles que l'analyse factorielle etc...
- la représentation de chaque classe devient très simple et , comme le montre l'exemple de reconnaissance donné plus loin , très efficace.

Cependant on a dû constater certains inconvénients:

- pendant la phase de constitution des classes nous avons dû séparer certaines classes en sous-classes parce que la dispersion obtenu était très grande (voir par exemple N4 sur le tableau 1). Cette situation traduit tantôt un phénomène réel (phonème au début du mot et au milieu

TABLEAU 1

REDUCTION DE LA DISPERSION DU NUAGE

Phonème	Canal	Critère C		
		Avant	(1) Après	(2)
A	2	0.225	0.178	0.145
	3	0.393	0.178	0.129
	4	0.672	0.216	0.171
N	2	0.430	0.290	0.236
	3	0.625	0.309	0.318
	4	0.930	0.432	0.301
U	2	0.306	0.215	0.178
	3	0.462	0.292	0.171
	4	0.929	0.090	0.069
	5	0.551	0.260	0.262
	6	0.381	0.283	0.266
EU	2	0.326	0.138	0.062
	3	0.653	0.252	0.181
	4	0.729	0.248	0.186
	5	0.566	0.263	0.179

$$C = \frac{R_{max} - R_{min}}{R_{moy}}$$

(1) Normalisation R_i par F_i , G_i

(2) " R_i " $F_{i-1}, F_i, F_{i+1}, G_{i-1}, G_i, G_{i+1}$

par exemple) , tantôt les faiblesses de la méthode adoptée .

- Certains problèmes de convergence de la formule (3) (stabilisation des coefficients a_k et b_k quand le nombre d'occurrence K augmente).

2.3.2. Normalisation des R_i par rapport aux F_{i-1}, G_{i-1}, F_i , G_i , F_{i+1}, G_{i+1} .

En essayant toujours d'exprimer la relation entre R et F , G nous avons procédé à une extension de la formule (3) et à une étude systématique des propriétés de la normalisation obtenue.

Ce deuxième pas est la suite naturelle de ce qui précéde ; au lieu de chercher la relation entre les paramètres issus d'un seul canal on cherchera la relation entre le paramètre principal R du canal i et les paramètres de normalisation calculés sur les trois canaux successifs i-1, i, i+1. La raison en est que les paramètres spectraux issus des trois canaux successifs peuvent mieux exprimer ce qu' on pourrait appeler "point resonnant dans le voisinage du point i" .

Ainsi l' équation (2) devient:

$$f_i (R_i, F_{i-1}, F_i, F_{i+1}, G_{i-1}, G_i, G_{i+1}) = 0 \qquad (7)$$

L' approximation polynomiale (5) devient:

$$R_{in} = R_i + A(i-1) + A(i) + A(i+1) \qquad (8)$$

avec : $$A(i) = \sum a_k(i) . F_i^k + \sum b_k(i) . G_i^k$$

Cette formule étant appliquée pour i= 1..6, on aboutit à une réduction du nombre de paramètres de 21 à 5.

En posant:

$$N = M$$

$$\underline{C}_i^T = \{ a_1(i-1),\ a_1(i),\ a_1(i+1),\ b_1(i-1),\ b_1(i),\ \ldots\ldots \ldots a_N(i-1),\ \ldots b_N(i+1) \}$$

$$\underline{X}_i^T = \{ \Delta F(i-1),\ \Delta F(i),\ \Delta F(i+1),\ \Delta G(i-1),\ \Delta G(i),\ \Delta G(i+1),\ \ldots \ldots \Delta F^N(i-1),\ \ldots \Delta G^N(i+1) \}$$

L' équation (8) s' écrit :

$$R_{in} = R_i + \underline{C}_i^T . \underline{X}_i^T \qquad (9)$$

Le calcul des coefficients $\underline{C}$ se fait avec la même méthode , c'est à dire en minimisant la quantité :

$$Q = E\ (\ R_{in} - E(R_{in})\ )^2 \qquad (10)$$

alors:

$$E\ (R_{in}) = E(R_i) - \underline{C}_i^T . E(\underline{X}_i)$$

et $$Q = E\ (\ R_i - E(R_i) + \underline{C}_i^T . (\ \underline{X}_i - E(\underline{X}_i)\ )$$

le vecteur $\underline{C}_i$ est la solution du système lineaire :

$$\partial Q / \partial \underline{C}_i = 0$$

c'est à dire :

$$E\ (\ \underline{X}_i . \underline{X}_i^T - E(\underline{X}_i) . E(\underline{X}_i^T)\ ) .\ \underline{C}_i = = E(R_i) . E(\underline{X}_i) - E(R_i . \underline{X}_i) \qquad (11)$$

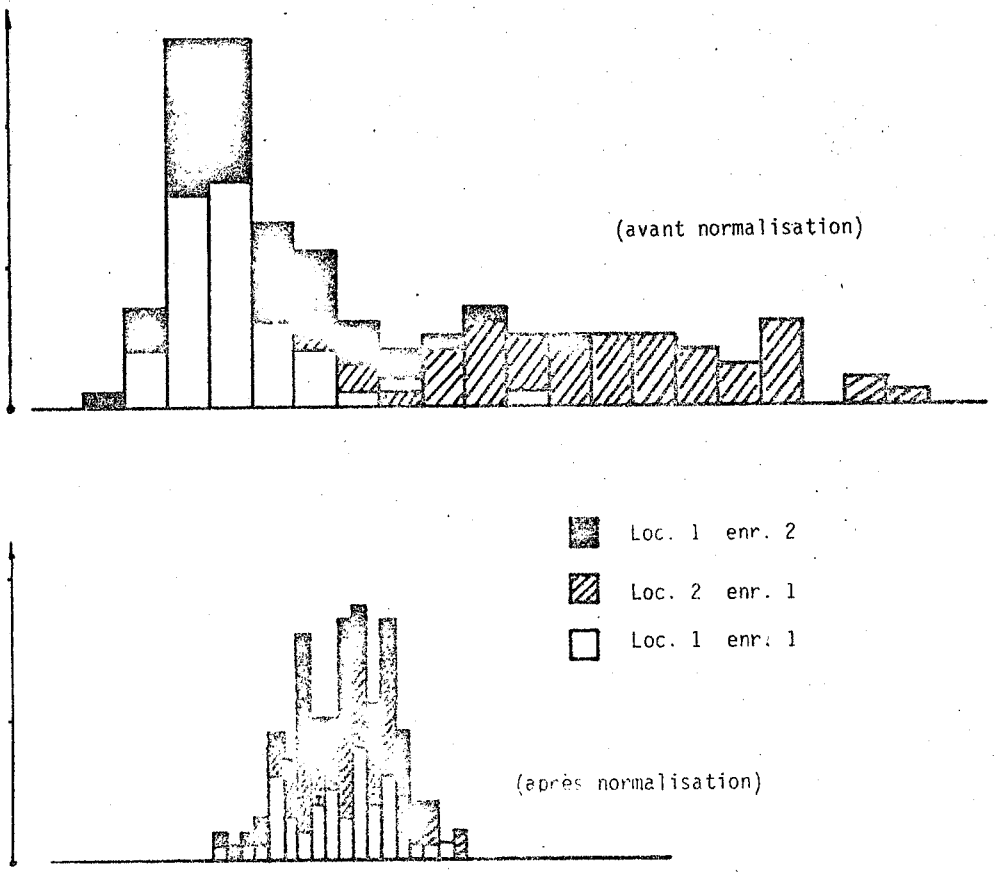
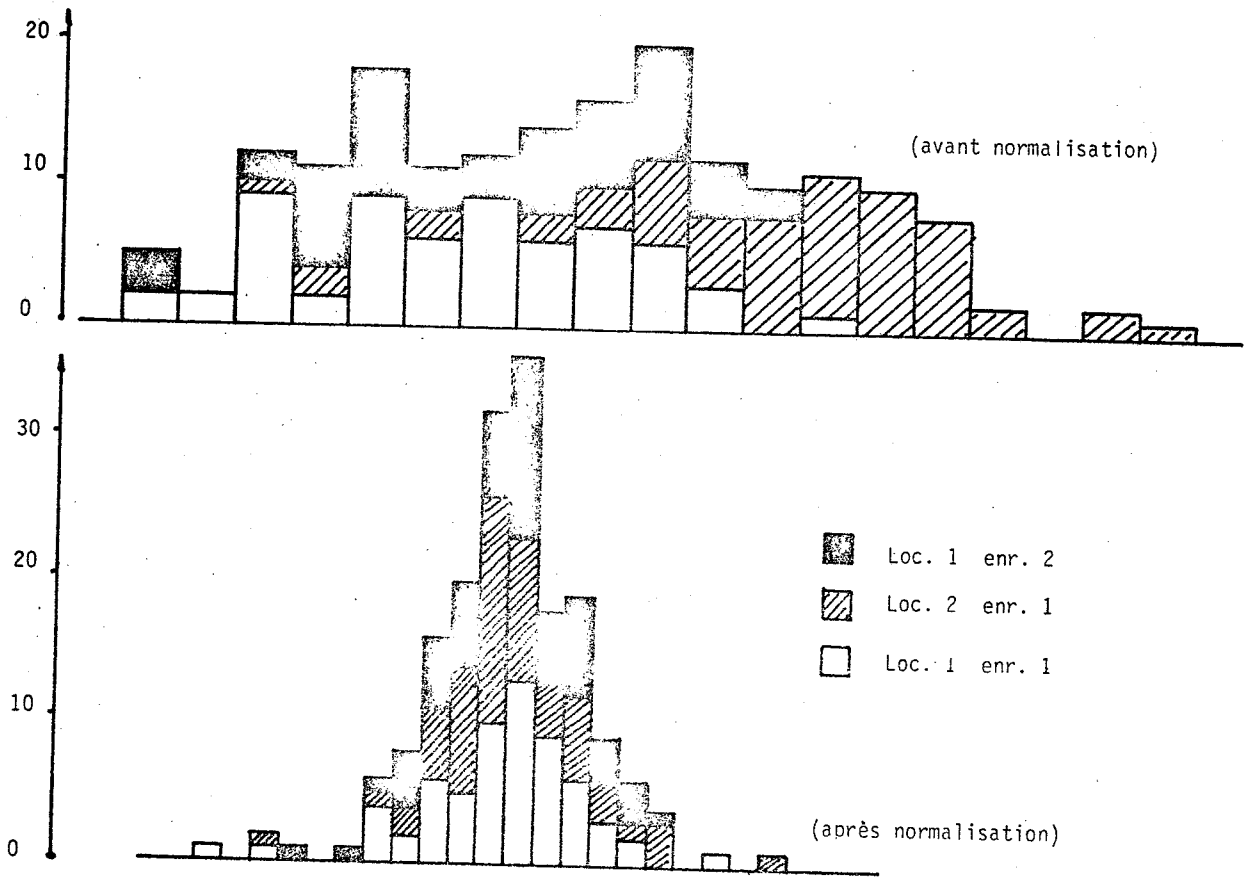


Figure 8 : Histogramme du paramètre R_4 (phonème 'N') avant et après la normalisation.



Fifure 9 : Histogramme du paramètre R_3 (phonème 'M') avant et après la normalisation

Propriétés de la normalisation .

Sur les figures 8 et 9 nous avons rapporté , sur une même échelle l'histogramme des variations d'un paramètre R (R_4 et R_3) avant et après la normalisation . Il est à noter :

a . La réduction importante de la dispersion (rapport environ 1:3 dans les deux cas). Le tableau 1 donne quelques exemples supplémentaires en établissant en même temps une comparaison entre cette normalisation et la précédente . La supériorité de la deuxième normalisation est manifeste d'autant plus que l'ordre de l'approximation ici est égal à 2 tandis qu'à la première il est égal à 3 .

b. Les occurrences des deux locuteurs sont effectivement mélangés. Dans les deux exemples cités, la séparation des deux locuteurs est très prononcée avant la normalisation tandis qu'après elle a complétement disparue .

c. La distribution des occurrences , après normalisation, rapproche de la normale, ce qui permet l'utilisation de toute la théorie des distributions Gaussièennes lors de la reconnaissance.

Convergence

Le problème posé est de savoir si la formule (8) a des chances de converger , autrement dit si le vecteur $\underline{C}$ ne varie pas beaucoup quand le nombre d'occurrences augmente .

Le nombre d'occurrences de chaque phonème qu'on dispose étant variable (de 10 à 60 par phonème et enregistrement) et le nombre de locuteurs étant limité à deux les résultats présentés ici possèdent un caractère plutôt probable que sûr .

Si $\underline{C}(K)$ est le vecteur $\underline{C}$ calculé par le système d'équation (11) pour K occurrences, nous définissons pour le canal i le critère de

convergence suivant :

$$c_i(K) = (\underline{C}(K) - \underline{C}(K'))^T . (\underline{C}(K) - \underline{C}(K')) \qquad (12)$$

et pour tous les canaux :

$$c(K) = c_i(K)$$

Les figure 10 et 11 donnent l'évolution du critère global pour divers phonèmes . Il est à remarquer la bonne convergence de la normalisation pout certains phonèmes (A,N,V,..e.t.c), tandis que au contraire pour d' autres (I,AN,EU,..e.t.c) la convergence est mauvaise.

En règle générale on peut constater que la formule converge pour une centaine d'occurrences plus ou moins représentatives.

3. EXEMPLE DE RECONNAISSANCE .

L'intérêt des normalisations précédentes peut être démontré par l'application directe d'une procédure de reconnaissance d'une extrême simplicité . L'économie du temps de calcul , jointe à des taux de reconnaissance élevés , est en effet l'un des critères essentiels de qualité pour un système de reconnaissance de la parole .

En utilisant la représentation des classes introduite au paragraphe 2.3.1. , on peut , lors de la reconnaissance , travailler d'une façon très simple . Un phonème inconnu est représenté par ses paramètres R_i, F_i , G_i . Pour chaque classe existante , l'application de la formule de normalisation avec ses coefficients specifiques fournit les R_{in}. On détermine alors la distance de l'occurrence par rapport au centre de gravité de chaque classe par la formule :

$$d = \{ \sum (\frac{R_{in} - \bar{R}_{in}}{s_i})^2 \}^{\frac{1}{2}} \qquad (13)$$

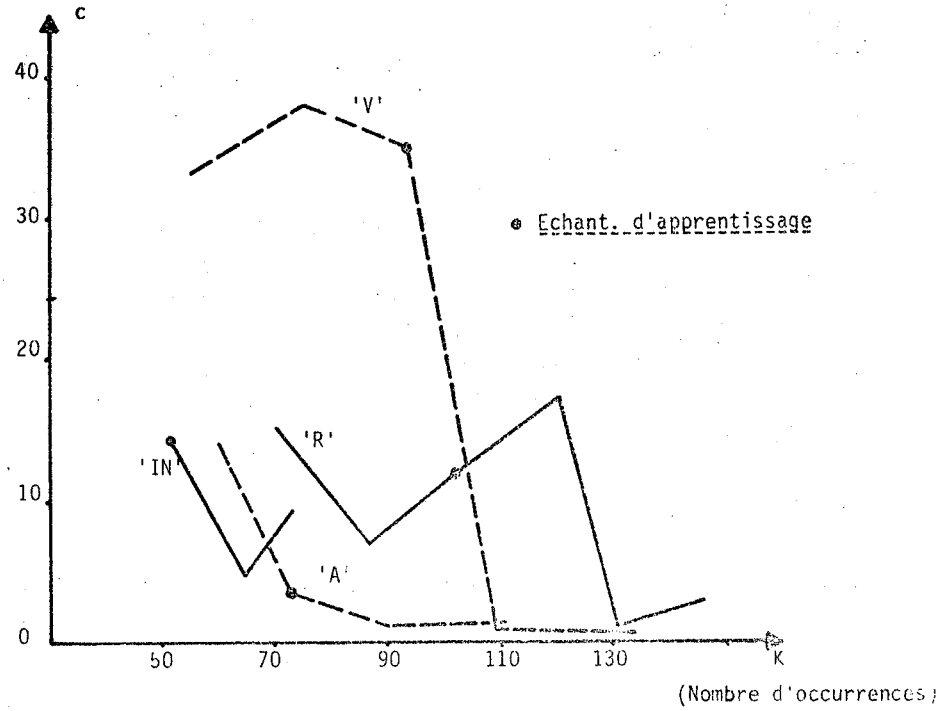


Figure 10 : Convergence de la normalisation.

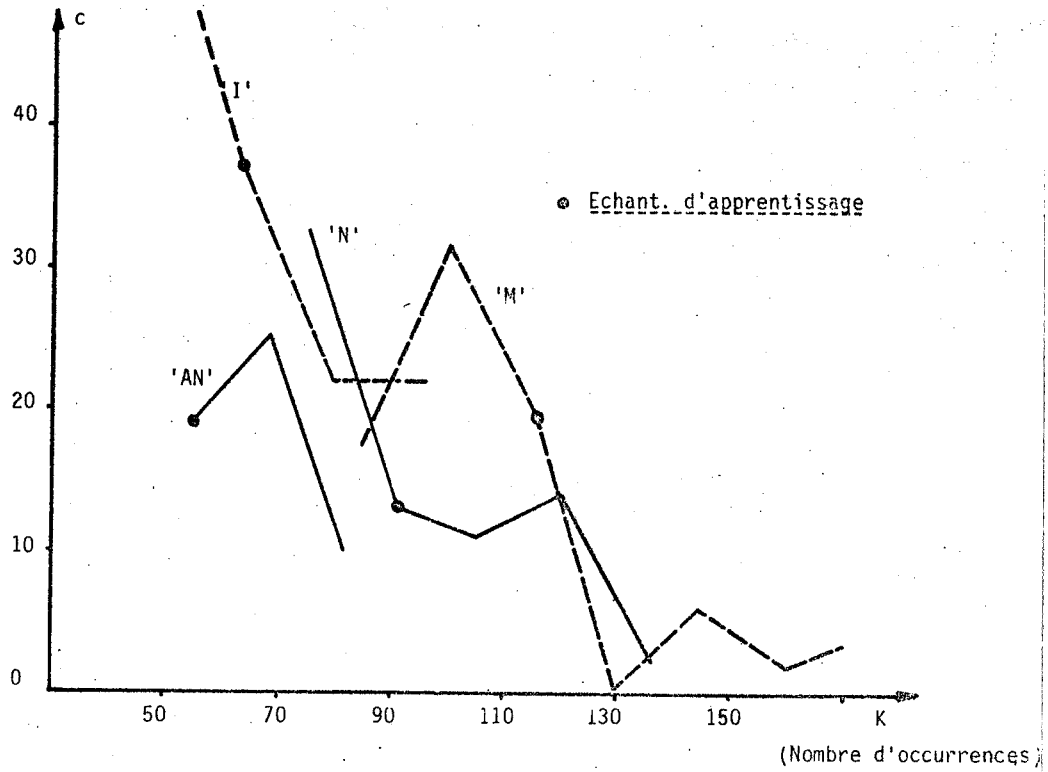


Figure 11 : Convergence de la normalisation.

avec $s_i = \sqrt{\sigma_{ii}}$: écart type de la distribution des points par rapport au centre de gravité du nuage (sur l'axe i) . La décision est finalement prise sur le critère de la distance minimale .

Cette procédure simpliste a été appliquée sur des paramètres normalisés par la première formule (3) et a donnée des résultats excellents (/3/) ; dans le cas d'un seul locuteur le taux de reconnaissance moyen sur 22 classes était 83% et pour deux locuteurs le taux était encore de 72% . Dans les deux cas , l'échantillon de test est celui d'apprentissage .

La deuxième normalisation (formule (8)) appliquée sur l'échantillon d'apprentissage a donné des résultats un peu meilleurs que la première . Sur l'échantillon de test , comme on s'y attendait , le taux de reconnaissance variait beaucoup de phonème à phonème selon que la convergence de la normalisation était bonne ou mauvaise .

Toute l'expérience acquise sur la normalisation / reconnaissance nous permet d'envisager le problème de la reconnaissance phonèmique de manière hiérarchique sur 3 étapes :

1. Première sélection sur la base des paramètres G :

Les seuls paramètres G permettent une première sélection parce qu'ils donnent une estimation très grossière du spectre du phonème . Pour une comparaison rapide entre les G du phonème à reconnaître et les $\bar{G}$ des classes on peut rejeter les classes qui sont trop éloignées du phonème. Cette sélection permet de réduire considérablement le temps de calcul puisqu'on aura à travailler sur quelques classes (5 à 10) et non sur les 22 .

2. Normalisation / deuxième sélection.

En appliquant la procédure simple déjà exposée , on peut , sur la base de la distance de l'occurrence à reconnaître par rapport aux centres de gravité des classes retenues pendant la première phase , établir un ordre de ces classes (de la plus proche à la plus éloignée du phonème).

Il est à remarquer qu'étant donné la distribution très proche de la normale des paramètres R_n il vaut mieux, au lieu d'utiliser la distance définie par (13) utiliser la distance:

$$d = (\underline{R}_n - \underline{\bar{R}}_n)^T . \Sigma^{-1} . (\underline{R}_n - \underline{\bar{R}}_n)$$

avec : Σ : matrice de covariance des paramètres normalisés
$\underline{\bar{R}}_n$: centre de gracité de la classe
$\underline{R}_n$: paramètre normalisés du phonème à reconnaitre

Les résultats de cette phrase permettrons de prendre soit une décision définitive soit de passer à la troisième phase. En effet , l'expérience montre que certains phonèmes sont très bien reconnus pendant cette deuxième phase (par exemple phonèmes A, IN , Y). Si une de ces classes arrive en tête de la liste établie avec, de plus, une distance importante entre les deux premières classes, on peut prendre une décision définitive . Par contre si ces deux conditions ne sont pas remplies on passe à la troisième phase en retenant les classes les plus probables (2-3).

3. Discrimination finale

Pëndant cette dernière phase on a recours à des méthodes de discrimination très puissantes , telle que la densité de probabilité locale . Cette procédure , qui demande en général un temps calcul élevé et une place en mémoire importante pésera pas beaucoup sur le temps global puisqu'elle sera appliquée pour choisir une classe parmi 2 ou 3 . A ce stade là on peut recourir soit aux densités de probabilité locales établies sur le nuage normalisé (on aura donc à faire avec seulement 5 paramètres); soit aux densités de probabilité locales des paramètres bruts (21) . Cela est quelquefois indispensable quand il s'agit de choisi entre des phonèmes par nature très proches l'un de l'autre (par exemple

/n/ et /m/ ou /oe/ et /ə/ .

Cette procedure de reconnaissance n'a pas encore été entièrement appliquée parce que l'échantillon d'apprentissage n'est pas complet.

CONCLUSION

Ces méthodes de normalisation présentent le double avantage de faire disparaître la particularité de chaque locuteur ainsi que de permettre dans une large mesure , la représentation de chaque classe par son centre de gravité . Les problèmes de convergence ne sont pas en général graves puisqu'il semble bien que lorsque le nombre d'occurrences dépasse la centaine les coefficients $\underline{C}$ se stabilisent . Enfin, nous pensons qu'une combinaison de la méthode de normalisation avec plusieurs procédures de reconnaissance utilisés de façon hiérarchisée , peut donner une très bonne solution du problème de la reconnaissance automatique de la parole au niveau phonèmique .

REFERENCES

/1/ MAISSIS A. , WALRAVE P. :

Traitement des signaux aléatoires.
Tome IV : Paramètrisation et reconnaissance phonèmique .
Contrat DGRST 70.02.186 Juin 1970

/2/ MAISSIS A. , PAU L.F , GUEGUEN C.J. :

Paramètrisation et procédures de reconnaissance de la parole
Automatique , Mars 1972 (p 65-68)

/3/ GUEGUEN , MAISSIS A,PAU L.F :

Communication homme machine sur support vocal
L'echo des Recherches ; Octobre 1972

/4/ MAISSIS A. : Une méthode d'extraction du fondamental
l'Onde Electrique , Mars 1973 (p;110 .112)

/5/ FLANAGAN J.L :

Speech Analysis Synthesis and perception
Springer-Verlag , Berlin , Heidelberg , 2è édition 1972

/6/ FLANAGAN J.L

Models for approximating basilar membrane displacement .
Bell System Tech .J. 41 , Part II , (p 959-1009)
1962

/7/ PAU L.F

Méthodes statistiques de réduction et de reconnaissance des formes . Normalisation des paramètres phonèmiques .
Application à la reconnaissance de la voix.
Thèse de Docteur-Ingénieur . Fac Sciences Orsay Mai 7

Discussion.

J-Y. GRESSER : J'ai l'impression que MM. Liénard et Maissis se contredisent. Maissis pense-t-il applicable aux transformations décrites par Liénard, sa méthode de normalisation.
I.2. Les expériences de Maissis ont-elles eu lieu en temps réel ?
1.3. Les résultats annoncés ont-ils été obtenus sur 20 ou plus phonèmes ?

II. Pensez-vous qu'il y ait un niveau privilégié de normalisation ?

A. MAISSIS : I.1. Je ne le pense pas. A mon avis, il y a deux différences essentielles entre les deux méthodes.
- Chez Liénard, il y a normalisation seulement par rapport au locuteur dont les caractéristiques sont extraites en temps différé par l'analyse d'un mot clef spécialement prononcé. Chez nous, par contre, les paramètres de normalisation étant intrinsèques et continuellement extraits, permettent une normalisation par rapport au locuteur et aux conditions spécifiques dans lesquelles le phonème examiné a été prononcé.
- Chez Liénard la normalisation se fait par changement des fréquences centrales des filtres analyseurs, donc au niveau des capteurs. Chez nous, par contre, la normalisation s'opère au niveau de la reconnaissance des formes.
1.2. La paramétrisation du signal se fait actuellement par des moyens numériques et par conséquent les expériences de normalisation/reconnaissance ont eu lieu en temps différé. Toutefois, l'application de la première normalisation ($Ri = fi\ (Fi, Gi)$) nécessite environ 150 ms (CPU) par phonème reconnu. La deuxième normalisation est plus longue mais l'application de la reconnaissance hiérarchisée ainsi que l'optimalisation des programmes permettront d'atteindre dans la plupart des cas, le temps réel.
1.3. Nous avons travaillé sur 22 phonèmes dont 13 voyelles, 2 semi-voyelles et 7 consonnes.

II. Nos travaux concernent essentiellement la reconnaissance phonémique et par là la reconnaissance des mots. Entre ces deux niveaux je pense que la normalisation au niveau phonémique est préférable puisqu'elle permet une représentation automatique normalisée des mots. On peut, en effet, définir un mot comme une suite de phonèmes, chaque phonème étant représenté par ses paramètres normalisés (R_n).

J-P. HATON : Vous effectuez une normalisation sur vos paramètres R exclusivement. Pourquoi ne normalisez-vous pas les paramètres G qui sont directement liés au spectre, ce qui rejoindrait la normalisation fréquentielle de M. Liénard ?

A. MAISSIS : Nous avons choisi la normalisation des paramètres R puisque d'après notre conception, l'information principale se trouve justement contenue dans ces paramètres. Cette information doit pourtant être rapportée à des conditions "standard" suivant la valeur des F et G. De plus, les paramètres G donnent une estimation spectrale aussi grossière qu'incomplète puisqu'ils se complètent avec les paramètres F.

Evidemment sur le plan mathématique l'opération suggérée est parfaitement envisageable. Mais je pense que, même dans ce cas là, on ne peut vraiment parler d'un rapprochement des deux méthodes : chez Liénard, le spectre est représenté par 32 points, donc avec une grande précision, tandis que chez nous les 7 paramètres G, extraits d'un filtrage par des filtres à large bande passante, ne donnent une image du spectre que très approximative.

ANALYSE DE LA PAROLE PAR FILTRAGE OPTIMAL

Résumé

Cet article présente une méthode d'estimation d'un nombre réduit de paramètres caractérisant le signal vocal, par identification de la fonction de transfert du système de phonation. Cette identification est réalisée au sens du minimum de variance des paramètres par application du filtrage optimal de KALMAN. L'algorithme récursif correspondant fournit un modèle stable et permet de réaliser en ligne à la fois la détection synchrone du fondamental, l'identification du système et le calcul a posteriori de la source excitatrice correspondante . Certaines applications en reconnaissance de la parole sont présentées.

Summary

This paper introduces a new method for an acuate speech analysis based on the identification of the vocal tract transfer function by Kalman's optimal filtering. This method provides recurrently a stable parametric model, adjusted on a minimum variance criterion, and fitted to on line parameter estimation, synchroneous pitch detection, and a posteriori source wave computation. Some applications on the speech recognition problem are emphasized.

C.J. GUEGUEN et G. CARAYANNIS
Ecole Nationale Supérieure des Télécommunications.

1. LES METHODES D'ANALYSE DE LA PAROLE.

La complexité indéniable du signal vocal a suscité l'introduction de nombreuses méthodes d'analyse. Le but était toujours d'extraire un nombre suffisamment restreint de paramètres caractérisant le signal de manière pertinente . Ce soucis répond à la nécessité d'une réduction de la redondance de la parole destinée à garder seulement l'information utile. La qualité de ce traitement est en effet essentielle dans tout système de synthèse vocale, de reconnaissance automatique de la parole ou d' identification des locuteurs, où la mise en oeuvre des algorithmes se heurte au volume prohibitif des données brutes.

En reconnaissance de la parole en particulier, la pertinence des paramètres apparaît essentielle dès que le traitement d'un vocabulaire de grande dimension est envisagé. Les méthodes actuelles de la reconnaissance structurelle destinées à enchainer entre eux des éléments reconnus (phonèmes par exemple) pour former des unités linguistiques d'ordre supérieur (mots par exemple) se révèlent faiblement efficaces quand les taux de reconnaissance élémentaires ne sont pas très élevés.

Par ailleurs, la recherche fondamentale sur la parole est encore conditionnée par la disponibilité de puissants outils d'analyse destinés a être implantés sur calculateur , même au prix d'une certaine complexité de mise en oeuvre et, donc, d'un temps de calcul important.

C'est dans ce sens que s'est developpé l'utilisation de l'analyse spectrale appliquée au signal vocal. L'introduction de la Transformée de Fourier Rapide (F.F.T.) a donné une impulsion considérable à ces techniques, en permettant une analyse à la fois précise et rapide du fait de la création de " hardwares"spécialisés. Peu après, l'élaboration par SCHAFFER et RABINER /1/ d'un algorithme destiné à accentuer les maxima du spectre (chirp z-transform) a permis de lever certaines ambiguités dans l'extraction et la poursuite des formants . La quasi-periodicité du signal vocal (modulation du spectre par le fondamental) et le caractère convolutionnel de sa production (sources excitant le canal vocal) ont

donné lieu à des techniques particulières - cepstre par exemple - relatives au lissage du spectre et à la séparation des contributions individuelles de la source et du canal vocal /2/. L'appel à plusieurs transformées de Fourier augmente parfois le temps de calcul de manière prohibitive. De plus, le signal vocal étant par nature non stationnaire c'est la définition d'un spectre à court terme qui est requise . Celui-ci établi globalement sur plusieurs périodes fondamentales, demeure d'une interprétation délicate (fenêtre d'estimation spectrale).

Une voie de recherche particulièrement fructueuse est constituée par l'ensemble des méthodes qui réalisent l'ajustement des paramètres a_i d'un modèle et peuvent être regroupées sous le concept d'identification. Le processus d'ajustement utilise directement le signal s et est illustré par la figure 1. Différents types de critères ont été considérés:

i - Moindres carrés :

$$C = \text{Min} \sum_{i=1}^{N} \varepsilon_i^2$$

SCHROEDER,ATAL,HANAUER /3/,/4/
LOZVOSKY /5/

ii - Maximum de vraisemblance :

$$C = \text{Max } p(s / \underline{a})$$

SAÏTO , ITAKURA /6/ /7/.

Mais du fait des hypothèses introduites , (i) et (ii) se ramènent à la résolution du même système d'équations définissant les a_i . C'est encore aux mêmes équations qu'aboutit la méthode du filtrage inverse proposée par MARKEL /8/ . Toutes ces techniques se caractérisent par une approche globale -ie : sur une période fondamentale entière - et ne differant que par le rpocessus de résolution adopté . Le modèle obtenu se révèle parfois instable et nécessite un recalage des pôles.

La méthode proposée ici , réalise de manière récurrente l'identification de la fonction de transfert du système de phonation sur un critère de :

iii - Minimum de variance de l'estimation $\hat{a}$.

$$C = \text{Min Trace } \left[E (a - \hat{a}) (a - \hat{a})^T \right]$$

Le modèle optimal est alors obtenu par application du filtre de Kalman et s'affirme stable à la convergence . Une nouvelle estimation étant disponible à chaque échantillon , la détection synchrone du début de chaque période fondamentale peut être réalisée sur un critère intrinsèque - ie : déduit du modèle -qui se révèle très performant . De plus, tenant compte d'hypothèses usuelles, une séparation source excitatrice -canal vocal peut être envisagée sur une base temporelle.

2. MODELISATION DU SYSTEME DE PHONATION.

Le choix d'un problème adéquat est toujours une étape importante dans la résolution d'un problème d'identification . Dans le cas du système de phonation , un grand nombre de modèles paramètriques plus ou moins réalistes au point de vue physiologique ont été proposés /9//10/ . On adoptera intentionnellement ici un modèle de forme très générale (fonction de transfert en z) car c'est plus la caractérisation abstraite du signal qui est recherchée ici , que la modélisation des organes de phonation . Un tel modèle couramment utilisé considère le signal vocal comme la réponse d'un système acoustique (conduit vocal , dérivation nasale, radiation buccale ....) à une excitation de type fixé . Dans le cas des sons sonores , l'entrée est ainsi constituée par une série d'impulsion quasi-périodiques émanant des cordes vocales , et dans le cas des sons sourds , par un bruit blanc dû aux turbulences de l'air dans les constrictions du canal vocal. Les deux sources d'excitation peuvent de plus coexister . L'identification consiste alors à ajuster les paramètres du modèle pour le rendre aussi voisin que possible du système réel au sens d'un critère donné.

Supposant le système de phonation linéaire ,il pourra être représenté pour les nécessités de traitement numérique par une équation recurrente d'ordre k, liant les valeurs échantillonnées de l'entrée e_n (excitation) et de la sortie s_n (signal vocal) :

$$a_o s_n + a_1 s_{n-1} + \ldots + a_k s_{n-k} = b_o e_n + b_1 e_{n-1} + \ldots + b_k e_{n-k} \qquad (1)$$

soit , appliquant la transformée en z (conditions initiales nulles) :

$$S(z) = H(z).E(z) \quad \text{avec} \quad H(z) = \frac{b_o + b_1 z^{-1} + \ldots + b_k z^{-k}}{a_o + a_1 z^{-1} + \ldots + a_k z^{-K}} \qquad (2)$$

Ce modèle peut être particularisé en assimilant le canal vocal à une série de résonnateurs en cascade . On admettra que la fonction de transfert H(z) ne comporte alors que des pôles . La mise en oeuvre de la dérivation nasale introduit des zéros dans H(z) mais ceux-ci (sous réserve de stabilité) peuvent être approximés par un nombre convenable de pôles. Il en résulte qu'un modèle suffisant est caractérisé par $b_i = o$ pour $i \neq o$ dans (1) et (2).

La relation entrée-sortie peut être explicitée, en l'absence de conditions initiales , sous forme matricielle :

$$b_o \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_k \\ \vdots \\ e_n \end{bmatrix} = a_o \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_k \\ \vdots \\ s_n \end{bmatrix} + \begin{bmatrix} s_o & o & \cdots & o \\ s_1 & s_o & \ddots & \vdots \\ \vdots & \ddots & \ddots & o \\ s_{k-1} & \cdots & & s_o \\ \vdots & & & \vdots \\ s_{n-1} & \cdots & & s_{n-k} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix} \qquad (3)$$

$$b_o \; \underline{e} = a_o \; \underline{s} + S \; \underline{a}$$

Pour tenir compte d'éventuelles conditions initiales , il suffit de n'écrire la relation (3) qu'à partir de l'indice k. La matrice S est alors complète et s_i , i = o ; k-1 jouent le rôle de conditions initiales. Le choix d'un tel modèle résulte d'un compromis entre l'exactitude et la complexité de la représentation . La complexité du processus de phonation introduit certaines difficultés dans son identification :

i - L'entrée du système n'est connue que très approximativement (le modèle bruit-impulsion est un peu simpliste) un certain nombre de sources intermédiaires coexistent à l'interieur même du canal vocal (constrictions en particulier).

ii - Le système est variable dans le temps et en conséquence la période d'identification doit être nécessairement courte . Le canal vocal est un conduit acoustique à paramètres répartis dont on cherche à rendre compte globalement par un modèle linéaire entrée-sortie . L'ordre de ce modèle dépendant de la propagation des ondes dans ce conduit , est donc essentiellement variable.

iii - Le signal acoustique capté est toujours entaché de bruit , ce qui limite la précision de l'identification et , pour les ordres élevés , augmente la variabilité du modèle.

3. FORMULATION DE LA METHODE.

Le modèle du système de phonation est exprimé par l'équation :

$$s_n + a_1 s_{n-1} + \ldots + a_k s_{n-k} = b_o e_n$$

avec l'hypothèse que e_n est un bruit pseudo-blanc tel que :

$$E(e_n) = o \,, \quad E(e_i e_j) = o \text{ si } i \neq j \,, \quad E(e_i e_j) = \frac{\sigma^2}{b_o} \qquad (4)$$

Il peut être reformulé par des équations d'état :

$$\begin{cases} \underline{a}_{n+1} = I\,\underline{a}_n \\ s_n = \underline{s}_n^T\,\underline{a}_n + W_n \end{cases} \qquad \begin{aligned} \underline{a}_n^T &= a^T = \left[a_1, \ldots, a_k\right] \\ \underline{s}_n^T &= \left[s_{n-1}, \ldots, s_{n-k}\right] \end{aligned} \tag{5}$$

avec I matrice unité $k \times k$, $W_n = b_o e_n$

Le problème est donc de déterminer une estimation optimale du vecteur d'état $\underline{a}_n$ de (5) . Soit $\hat{a}_n$ l'estimation à l'instant n , et P_n la matrice de covariance correspondante:

$$P_n = E\left[(\underline{a} - \hat{a}_n)\,(\underline{a} - \hat{a}_n)^T \right]$$

Il est naturel d'adopter comme critère :

$$C_n = \underset{\underline{a}_n}{\text{Min}} \left[\text{Trace } P_n \right] \tag{6}$$

Si l'on se restreint à un estimateur linéaire, on est conduit à définir la nouvelle estimation à l'instant n, résultant de la connaissance de s_n , par la relation de recurrence :

$$\hat{\underline{a}}_n = \hat{\underline{a}}_{n-1} + \underline{K}_n\,(s_n - \hat{s}_n)\ ;\ \hat{s}_n = \underline{s}_n^T \cdot \hat{\underline{a}}_{n-1} \tag{7}$$

où le coefficient vectoriel K_n pondère les influences respectives de l'ancienne estimation et de la différence entre la sortie réelle s_n et la sortie attendue : $\hat{s}_n$. Il est à prévoir que plus la confiance dans l'estimation précédente $\hat{\underline{a}}_{n-1}$ sera faible, plus le gain $\underline{K}_n$ permettant de tenir compte de la nouvelle mesure devra être augmenté . C'est la détermination gain optimal $\underline{K}_n$ qui est l'objet du filtre de Kalman /11/ .

Les équations récurrentes définissant la solution se réduisent ici à l'ensemble :

$$\underline{K}_n = - P_{n-1} \underline{s}_n \left(\underline{s}_n^T P_{n-1} \underline{s}_n + \sigma^2 \right)^{-1} \qquad (8)$$

$$P_n = \left(I + \underline{K}_n \underline{s}_n^T \right) P_{n-1} \left(I + \underline{K} \underline{s}_n^T \right)^T$$

ou encore :

$$P_n = P_{n-1} + \underline{K}_n \underline{s}_n^T P_{n-1} \qquad (9)$$

On montre que l'estimation est sans biais si l'estimation initiale en est aussi dépourvue, que l'estimateur linéaire (7) est l'optimal absolu dans le cas gaussien . L'apport d'information correspondant à l'arrivée de s_n , se traduit , d'après (9) , par

$$\text{Trace} \left[P_n \right] \leqslant \text{Trace} \left[P_{n-1} \right] \qquad (10)$$

4. MISE EN OEUVRE DE L'ALGORITHME

L'algorithme se prête à un grand nombre de variantes par le jeu des paramètres relatifs à l'initialisation /12/.

Les valeurs initiales $\underline{\hat{a}}_o$ et P_o dans (7) et (8) permettent d'intégrer dans la méthode des informations a priori . En l'absence de telles informations,on aura recours à une initialisation du type:

$$a_o = o \qquad ; \qquad P_o = \alpha^2 I \qquad \alpha \text{ grand}$$

En cours de traitement , on pourra adopter des valeurs déduites de segments précédents . On a montré que la convergence est notablement accélérée par le choix d'une matrice P_o sophistiquée /13/:

$$P_o = \left\{ p_{ij} : \exp - (|i - j| / \theta) \right\}$$

Le système de phonation variant dans le temps, il convient d'initialiser périodiquement l'algorithme pour actualiser le modèle. Cette opération peut, en particulier , être déclanchée automatiquement en synchronisme avec les périodes fondamentales . Un décalage systèmatique peut être introduit pour n'appliquer l'algorithme que sur des segments où l'excitation peut être considérée de moyenne nulle (hypothèse (4)). Pour les sons non voisés, une durée arbitraire de 15 ms est retenue . Une mise en oeuvre asynchrone a été réalisée, par ailleurs, en pondérant le signal par une fenêtre de hamming glissante.

L'écart type σ du bruit W_n constitue aussi une information a priori à ajuster . On a constaté une relative insensibilité de l'algorithme aux variations de ce paramètre .

Un exemple de mise en oeuvre est donné par la figure 2 pour une fréquence d'échantillonnage de 10 KHz et un ordre k = 12 . La convergence de l'algorithme est, en moyenne , réalisée après 30 échantillons (soit 30 ms , durée inférieure à la période fondamentale). Le modèle obtenu est stable sauf éventuellement sur des transitions pour des voix de fondamental très élevé . Le temps de calcul demeure en tout cas important ($\simeq$ 100 temps réel en fortran IV sur 10070 CII).

La détection des périodes fondamentales est réalisée par l'évolution après convergence des grandeurs :

$$\varepsilon_n = s_n - \hat{s}_n \ , \quad \delta_n = \underline{a}_n \underline{a}_n^T - \underline{a}_{n-1} \underline{a}_{n-1}^T \ , \quad \gamma_n = (a_n - a_{n-1}) (a_n - a_{n-1})^T$$

Les critères intrinsèques δ_n et γ_n , seuls accessibles par une méthode récurrente, sont très sensibles. Si cette extraction est seule recherchée, un ordre $K \leqslant 4$ est suffisant et l'algorithme atteint le temps-réel.

Le spectre du signal peut être aisemment déduit du modèle identifié. Il s'agit d'un spectre lisse et à très court terme (inférieur à la période fondamentale). Ce spectre est doté de maxima accentués bien

adaptés à l'extraction automatique des formants . Ce phenomène, conséquence de la proximité des pôles et du cercle unitaire , révèle, cependant , un amortissement non-réaliste des formants.

Un aspect intéressant de la méthode , dans la mesure où la convergence est atteinte après une initialisation sur la partie finale d'une période fondamentale, est le calcul a posteriori de l'entrée, le modèle étant alors figé . La séparation source-canal est donc ici envisagée sous une forme temporelle . L'interprétation de l'excitation calculée est, cependant , précaire du fait de l'abscence supposée de zéros dans la fonction de transfert.

5. CONCLUSION : CONTRIBUTION DE LA METHODE A LA RECONNAISSANCE DE LA PAROLE.

La méthode présentée constitue un outil d'analyse de la parole à la fois performant et souple . Certains aspects particuliers la rendent bien adaptée à une paramètrisation du signal en vue de la reconnaissance.

La qualité des paramètres a_i obtenus peut, en effet, être aisemment vérifiée directement par une synthèse immédiate . L'unité de réponse vocale est alors essentiellement constituée par un filtre numérique ajustable dont la réalisation ne soulève aucun problème technologique , ou qui peut être simplement simulé sur le calculateur. La seule transmission des a_i , de la puissance et du fondamental (à l'exclusion de l'erreur de prédiction) produit une parole de qualité.

Les systèmes de reconnaissance de la parole sont par nature destinés à fonctionner à partir de données acoustiques bruitées. Le système d'analyse étant constitué par un filtre stochastique adapté , l'influence des bruits sur les mesures est minimisée. L'ensemble analyse-synthèse est susceptible d'augmenter le rapport signal à bruit, et d'extraire la parole du bruit ambiant.

L'utilisation directe des paramètres a_i pour la reconnaissance n'est pas à recommander car ceux-ci contiennent des informations relatives au locuteur. Une méthode de normalisation /14/ /15/ serait appliquée avec profit. D'autre part , on a signalé qu'à défaut d'un "hardware " spécialisé , la méthode d'analyse ne peut être , pour l'instant, appliquée en temps réel . La difficulté peut cependant être tournée avantageusement par l'utilisation , en parallèle , d'une batterie de modèles . Une séquence de signal est ainsi confrontée avec toutes les relations de récurrence caractéristiques de chacun des phonèmes .Le système constitue ainsi un " vocoder à canaux " dont chaque canal est adapté à la discrimination d'un phonème particulier . Des procédures de reconnaissance de type déjà connu /16/ peuvent être alors appliquées avec des résultats prometteurs .

C'est dans cette optique que sont dirigés à l'heure actuelle les études sur la méthode proposée . Mais nous pensons, de plus , que l'application de telles techniques de modélisation adaptative et à auto-apprentissage , dont le filtre de Kalman n'est qu'un exemple , doivent trouver dans la parole un domaine d'application privilégié s'étendant aux problèmes fondamentaux de segmentation /17/ et de reconnaissance.

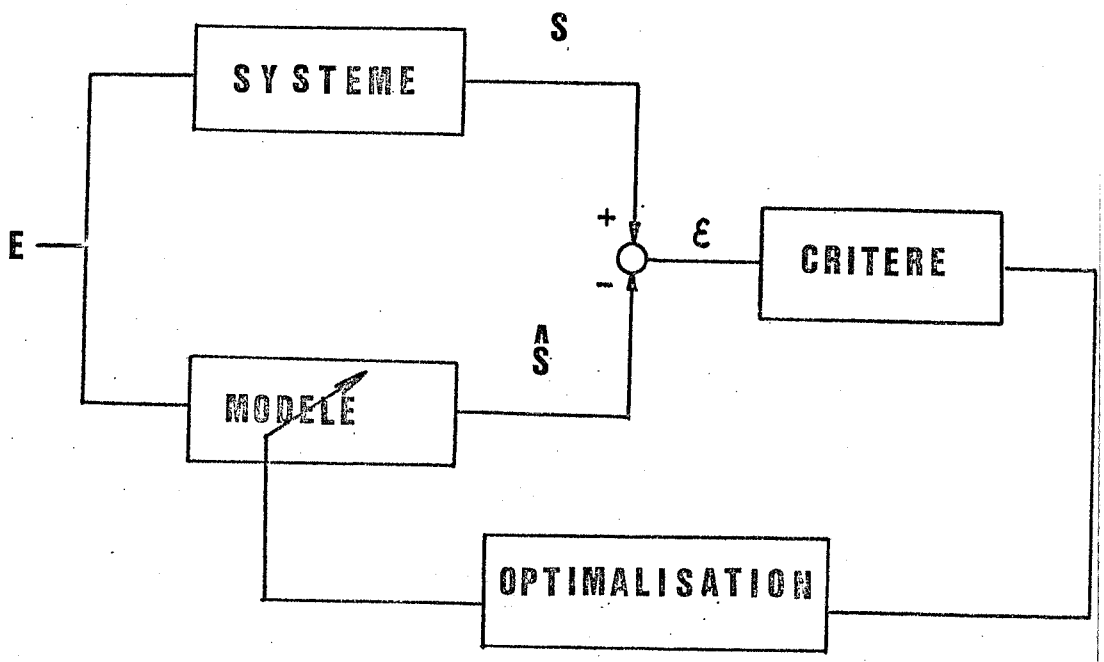


FIGURE 1 : Processus d'identification.

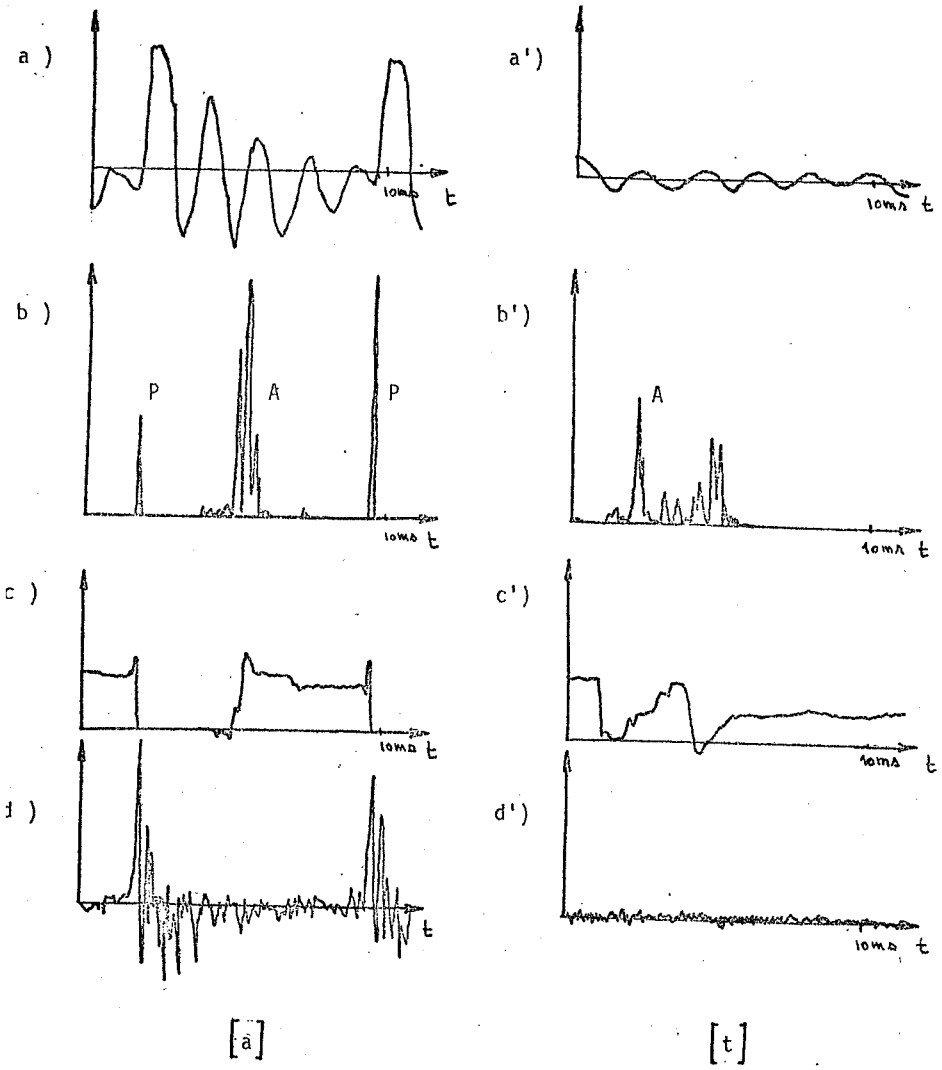


FIGURE 2 : Mise en oeuvre de l'algorithme.

a) Son sonore a
b) Variation du critère γ_n.
c) Evolution du paramètre a_3.
d) Entrée calculée .

a') Son sourd t
b') Variation du critère γ_n.
c') Evolution du paramètres a_3
d') Entrée calculée

Pics P : détection du pitch

Pics A : initialisation de l'algorithme.

REFERENCES.

/1/. L. RABINER , R. SCHAFER , C. RADER : The Chirp z transform Algorithm and its Application.
Bell System Technical Journal, June 1969.

/2/. M.NOLL : Cepstrum Pitch Determination
Journal of the Acoutical Society of America , Vol41,pp 293-309,1967.

/3/. B.ATAL , S. HANAUER : Speech Analysis and Synthesis by linear prediction of the speech Wave .
Journal of the Acoustical Society of America , Vol 50,pp 637-655,197

/4/. B.S.ATAL , M.RM SCHROEDER : Adaptive predictive coding of speech signals.
B.S.T.J. Vol49 N°8 pp 1973-1986 - 1970.

/5/. V.S. LOZOVSKY : Analysis and Synthesis of Speech based on its z - description.
Vych.Syst.issue 44 . Novosibirsk . 1971

/6/. F. ITAKURA , S.SAITO : A statistical method for estimation of speech spectral density and formant frequencies.
Electronics and communications in Japan Vol 53A N°1 -1970

/7/. F.ITAKURA , S.SAITO : An analysis -synthesis telephony based on maximum likelihood method.
Proc. Int . Congr. Acoust. C-5-5 Tokyo - 1968.

/8/. J.D. MARKEL : Digital inverse filtering - A new tool for formant trajectory estimation .
IEEE Trans on audio Vol 20 N°2 pp 129-137 -1972.

/9/. G.FANT : Acoustic theory of speech production
Mouton - the Hague . Paris 1970.

/10/. J. FLANAGAN : Speech Analysis,Synthesis and Perception.
second Edition . Springer - Verlag New York , 1972.

/11/. R. KALMAN : A new approach to linear Filtering and Prediction Problems.
ASME . Trans. , Journal of Basic Eng . 35 , 1960.

/12/. C.GUEGUEN , G.CARAYANNIS : Analyse de la parole par filtrage optimal de Kalman.
Automatisme . Tome 18 N°3 .1973

/13/. A. BARRAUD : Contribution à l'identification des séquences de pondération des systèmes dynamiques.
Thèse de Docteur Ingénieur , Univ. de Nantes , 1971.

/14/. L.F PAU : Optimisation d'une métrique en reconnaissance des formes .
4ème journées du Groupe " Communication parlée " du GALF.
Bruxelles 1973.

/15/. A. MAISSIS : Normalisation non-linéaire des paramètres phonémiques ; Application à la segmentation automatique.
4ème Journées du Groupe " Communication parlée " du GALF
Bruxelles 1973.

/16/ C.GUEGUEN, A. MAISSIS, LF PAU , G.CARAYANNIS : Communication homme-machine sur support vocal .
Rapport final et annexes techniques contrat CRI 71-14 .1972

/17/ G.CARAYANNIS : Modelisation des transitions phonemiques ; Application à la segmentation automatique .
4ème Journée du groupe " Communication parlée " du GALF.
Bruxelles 1973.

---===0===---

Discussion

Fr. LHOTE : Je suis assez sceptique sur la validité du raisonnement qui consiste à substituer dans la fonction de transfert des pôles additionnels aux zéros représentatifs des antirésonances. Pourquoi ne pas introduire ces zéros, qui ne se traduisent que par l'existence d'un second membre dans l'équation de récurrence ?

Par ailleurs, quel était l'ordre de la récurrence utilisée dans votre algorithme, ainsi que la fréquence d'échantillonnage ?

Cl. GUEGUEN : Le fait de n'avoir pas accès à l'entrée du système à identifier, rend impossible la détermination des coefficients du second membre de la relation de récurrence définissant le système. On est donc, dans ce cas, réduit à utiliser des hypothèses additionnelles sur la nature de l'entrée. L'hypothèse ici utilisée est que l'entrée est à moyenne nulle sur un segment suffisant à la convergence. Cette hypothèse est communément admise sur la fin des périodes fondamentales. Une méthode en cours d'étude à l'ENST, qui donnerait une identification du numérateur de H(z), est l'identification par coïncidence des spectres du signal et du modèle (utilisation de la F.F.T. et de la programmation non linéaire).
Les résultats présentés sont relatifs à :
- ordre du système : k = 12
- fréquence d'échantillonnage : F = 10 kHz.

A. NEMETH : Vous avez dit que vous étiez tenté d'exciter le filtre ? Est-ce la réduction du taux d'information qui vous a tenté ? Si oui, quel taux d'information était nécessaire pour coder la fonction d'excitation ?

Cl. GUEGUEN : Aucune expérience systématique n'a été entreprise sur la réduction du taux d'information. On pourra se référer aux résultats de ATAL, orientés essentiellement vers la synthèse, pour apprécier ce taux, la différence essentielle étant dans la méthode d'établissement du modèle. On notera, cependant, que dans les expériences de synthèse entreprises, on n'a pas eu recours à la transmission de l'erreur de prédiction comme pour ATAL. Il en résulte que les paramètres de commande se réduisent au fondamental, la puissance, la proportion de bruit.

R. DE MORI : a) Le modèle de ATAL porte souvent sur des modlèes instables. Quel est le degré de sécurité de votre modèle à ce propos ?
b) Comment excitez-vous le système en synthèse ?
c) Je crois que dans le modèle de la dernière figure il faudrait considérer aussi les effets de coarticulation ?

Cl. GUEGUEN : a) Pour un ordre suffisant, le modèle s'est toujours révélé stable, sauf sur les transitions rapides pour des voix féminines où l'algorithme n'a pas eu le temps de converger. L'aspect du spectre déduit du modèle montre, cependant, des pics très prononcés à l'emplacement des formants (amortissement non-réaliste). Ce phénomène traduit une proximité des pôles et du cercle unitaire en z qui constitue la limite de stabilité. Dans ces conditions, le calcul du spectre se comporte comme une "shirp-z transform" qui n'est pas dénuée d'intérêt si le modèle reste stable comme c'est le cas.
b) En synthèse, le modèle est excité par un mélange impulsion-bruit dont les paramètres sont : le fondamental, la puissance de sortie, la proportion de bruit. Ces quantités se sont révélées à l'expérience assez peu critiques. Il semble que, du fait de l'ordre élevé, une partie de l'entrée soit prise en compte dans le modèle.
c) Le système de reconnaissance présenté par la dernière ligne est sous cette forme, quelque peu simpliste. Il est cependant nouveau, car seule la disposition d'un modèle de qualité permettait de réaliser cette idée simple de réson ateurs accordés sur les divers phonèmes. Il faut considérer que le schéma présente un nouveau "vocoder à canaux adaptés" qui doit être suivi, pour tenir compte en particulier de la coarticulation, d'un système complet de reconnaissance. Les intérêts d'un tel système sont : l'analyse en temps réel (la méthode d'analyse fine, ici présentée, n'est appliquée qu'à l'apprentissage), un prétraitement élaboré (adaptation à l'objet à reconnaître).

OPTIMISATION D'UNE METRIQUE EN RECONNAISSANCE DES FORMES.

Résumé

L'article compare différentes métriques déterministes et probabilisées, sur le critère du taux de reconnaissance et sur celui du temps unitaire de reconnaissance. Les formes considérées dans l'application numérique sont des formes phonémiques à 10 dimensions, reconnues par la procédure généralisée du voisin d'ordre k. La distance L1 se révèle être la plus performante.

Dans une seconde partie, on tente d'optimiser une transformation non-linéaire des images des formes considérées. Cette transformation est définie par un développement limité, dont les coefficients sont ajustés par une méthode du gradient afin de maximiser le taux de reconnaissance. Il s'avère que les transformations optimales trouvées s'interprètent fort bien à partir des définitions des formes phonémiques étudiées.

Le choix de la métrique L1 permet de gagner 1-4 %, et la transformation non-linéaire ultérieurement 1-6 %, selon les ensembles de formes d'apprentissage et de formes à tester. Les taux de reconnaissance obtenus atteignent 89 % à 10 dimensions, et 92 % après compression, sur des formes distinctes des formes d'apprentissage.

La présente étude est basée sur les références 1 et 11, et a été réalisée en collaboration avec le Laboratoire d'Automatique et de l'ENS des Télécommunications, Paris.

Summary.

This paper compares several deterministic and probabilistic distance measures, on the basis of the recognition scores obtained, and of the recognition time by pattern. Those patterns to be considered numerically are 10-dimensional phonemic patterns, recognized by the generalized nearest neighbor rule. The L1 metric appears to perform best.

In the second part, an attempt is made in order to optimize a non-linear transformation of the images of our patterns. This tranformation is characterized

by a limited development, the coefficients of which are adjusted by means of a gradient algorithm in order to maximize the recognition score. The optimal transformations found may very well be interpretated on the basis of the definitions of the phonemic patterns studied.

Selecting the L1 metric yields extra 1-4 %, and the non-linear transformation moreover adds up 1-6 %, depending on the sets of learning or testing patterns used. Recognition scores of 89 % have been reached on 10-dimensional patterns, and 92 % after feature extraction and compression, both on patterns differing from the learning patterns.

The present study is based upon the references [1] , [11] , and has been completed in collaboration with the Laboratoire d'Automatique, Ecole Nationale Supérieure des Télécommunications, Paris.

L-F. PAU
(I.M.S.O.R. - Danemark).

1. COMPARAISON DE METRIQUES EN RECONNAISSANCE DES FORMES

11. INTRODUCTION

Soit Ω un ensemble d'observations , supposé métrisable par la norme $\|.\|$, et ayant également la structure d'espace de probabilité $(\Omega, \sigma(\Omega), P_\Omega)$, P_Ω étant la mesure de probabilité sur la σ -algèbre $\sigma(\Omega)$. Ω sera également appelé espace contrasté [1] .

Dans le présent rapport , nous ne considèrerons que les images $X \in \Omega$ des formes $\phi \in \Omega^F$ définies à partir d'un ensemble de fond F comme des homomorphismes de F dans Ω [1] . Nous désignerons par $Z \in \Omega$ l'image dans Ω d'une forme de référence fixée λ, et soit Λ un ensemble donné de ces formes de référence . Une forme de référence (aussi appelée forme forte) , est une forme qui caractérise seule ou partiellement, une classe donnée de formes ; cette forme de référence n'est pas nécessairement identifiable avec une forme d'apprentissage (cf [14]).

Nous avons étudié dans [1] de manière théorique les règles de décision $s^* : \Omega^F \longrightarrow \Lambda$ maximisant la discrimination entre classes d'équivalence de formes $\phi \in \Omega^F$. On rappelle que s^* dépend étroitement entre autres de la métrique d dans Ω^F et de P_Ω , la métrique d étant elle-même dans une large mesure fixée par la donnée d'une métrique d_Ω dans Ω . Nous avons discuté dans [1] un ensemble de métriques d, d_Ω , et souligné qu'il s'imposait de les comparer expérimentalement.

On donne :

- $L = \{l\}$ ensemble des formes d'apprentissage ;
- $\bar{L} = \{\bar{l}\}$ ensemble de formes non apprises , et destinées aux tests;
- $\Lambda = \{\lambda\}$ ensemble de formes de référence , structuré en classes;
- s règle de décision non-paramétrique $X \in \Omega \longrightarrow s(X) \in \Lambda$
- P_Ω mesure de probabilité dans Ω , estimée à partir de L.

Cette première partie se propose donc de comparer , pour différentes métriques d_Ω , les taux de reconnaissance pondérés $T(d_\Omega)$ (voir 136.) obtenus en appliquant la règle de décision s à $\bar{L}$, avec P_Ω et la métrique d_Ω entre les images des formes dans Ω . L'auteur estime en effet que des combinaisons convenables des grandeurs citées , et en particulier le choix de la métrique d_Ω , doivent permettre d'augmenter les performances d'un système de reconnaissance des formes .

Après avoir fait le catalogue des métriques d_Ω envisagées , nous spécifierons la nature des exemples test , et nous discuterons les résultats .

12. METRIQUES d_Ω DANS L'ESPACE CONTRASTE Ω

Nous dressons ici un catalogue de quelques distances d_Ω , dans le cas où Ω est isomorphe à $\mathbb{R}^m$. Nous noterons (x_i , i:1,n) les n coordonnées réelles de l'image $X \in \mathbb{R}^m$. De plus , nous poserons :

- $P_\Omega(X^l)$ est la probabilité marginale de l'image X^l d'une forme ϕ^l , et est estimée à partir des formes d'apprentissage de L ainsi que du tableau de contingence représentant P_Ω (déduit de $[x_k^l , k:1,n \quad l \in L]$).

$$P_\Omega(X^l) = \left(\sum_{k=1,m} x_k^l\right) \Big/ \left(\sum_{l,k} x_k^l\right) \qquad (1)$$

- p_k k:1,n est la probabilité marginale de la mesure k , et est estimée dans le tableau de contingence représentant P_Ω :

$$p_k = \left(\sum_{l \in L} x_k^l\right) \Big/ \left(\sum_{l,k} x_k^l\right) \qquad (2)$$

Nous étendrons la définition de $P_\Omega(X^l)$ $l \in L$, aux images X de formes appartenant à $\bar{L}$, en posant :

$$P_\Omega(X) = \left(\sum_{k=1,m} x_k\right) \Big/ \left(\sum_{l,k} x_k^l\right) \qquad (3)$$

121. Distances L^p de MINKOWSKI

On obtient une mesure de probabilité symmétrique d_Ω entre les images X^1, X^2 de deux formes ϕ^1 ϕ^2 par :

$$d_\Omega = \varphi^{-1}\left(\sum_{k=1,n} \varphi\left(\left|x_k^1 - x_k^2\right|\right)\right)$$

où la fonction réelle f vérifie :

a) monotonie : si $x_k^2 \geq x_k^3 \geq x_k^1$ ou si $x_k^2 \leq x_k^3 \leq x_k^1$,
alors :

$$\varphi\left(\left|x_k^1 - x_k^2\right|\right) \geq \varphi\left(\left|x_k^1 - x_k^3\right|\right)$$

Cette condition de monotonie est plus générale que l'inégalité triangulaire .

b) $\varphi\left(\left|x_k^1 - x_k^2\right|\right) = 0 \Leftrightarrow x_k^1 = x_k^2$

Les mesures de similarité de MINKOWSKI d'ordre p s'obtiennent en posant :

$$\varphi\left(\left|x_k^1 - x_k^2\right|\right) = \left|x_k^1 - x_k^2\right|^p$$

d'où :

$$d_\Omega = \left(\sum_{k=1,n} \left|x_k^1 - x_k^2\right|^p\right)^{1/p} \quad (4)$$

Si $p \geq 1$, ces mesures vérifient l'inégalité triangulaire , et deviennent une métrique véritable .

Si toutes les coordonnées sont booléennes , toutes les mesures de similarité de MINKOWSKI se réduisent à la distance de HAMMINGS . Les aléas statistiques de cette dernière distance comme classifieur entre fonctions de JACCARD , sont discutés dans [15] . Toujours dans le cas des coordonnées booléennes , on rencontre également :

$$x_k^1, x_k^2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad d_\Omega = \left[\sum_{k=1,n} \left(\left|x_k^1 - x_k^2\right| - \mathrm{Log}\, P_\Omega(X^2)\right)^p\right]^{1/p} \quad (5)$$

On pourra ici interpréter $P_\Omega(X^2)$ comme la probabilité a priori , estimée dans L , pour qu'une forme quelconque appartienne à la classe définie par la forme dont l'image est X^2 .

122. Distance euclidienne L^2

$$d_\Omega = \left(\sum_{k=1,n} \left|x_k^1 - x_k^2\right|^2\right)^{1/2} \quad (6)$$

123. Distance L^1, ou de MINKOWSKI d'ordre p : 1

$$d_\Omega = \left(\sum_{k=1,n} \left|x_k^1 - x_k^2\right|\right) \quad (7)$$

Cette métrique correspond à des surfaces séparatrices iso-L^1 non-linéaires, à savoir linéaires par morceaux . BATCHELOR [2] a étudié cette distance L^1, et semble indiquer que les taux de reconnaissance obtenus avec son aide sont meilleurs ou au moins comparables à ceux résultant de l'usage d'une autre métrique LP $p > 1$. On retiendra que si les observations x_k k:1,n sont non-corrélées , et si L^2 donne un bon taux de reconnaissance , la distance L^1 donnera un taux au moins aussi bon tout en étant plus rapide à calculer . Si l'on dispose d'une règle de décision à métrique euclidienne L^2, on aura souvent à pondérer sensiblement les images X^1,X^2 par leurs probabilités a priori comme dans la formule (5) . Il peut alors être plus simple et moins long de remplacer L^2 par la métrique L^1 qui exploitera mieux les observations tout en permettant d'inclure une information sur le contexte comme dans (5) (voir aussi [4]) .

TOUSSAINT [3] compare , pour des images binaires X , des fonctions discriminantes linéaires , la distance L^1 et la règle de décision bayésienne . Il confirme ainsi théoriquement les résultats expérimentaux de RAVI [4] , qui montre l'importance d'un équilibrage convenable entre l'information de mesure X et le contexte $P_\Omega(X)$.

124. Divergence de KULLBACH

$$d_{\Omega}=\sum_{k=1,m}\uparrow_k\left[\frac{x_k^1}{P_{\Omega}(X^1)}-\frac{x_k^2}{P_{\Omega}(X^2)}\right]\mathrm{Log}\left[\frac{x_k^1}{P_{\Omega}(X^1)}\Big/\frac{x_k^2}{P_{\Omega}(X^2)}\right]$$

Voir KULLBACH [5], KAILATH [6].

125. Distance variationnelle de KOLMOGOROV

$$d_{\Omega}=\sum_{k=1,m}\uparrow_k\left|\frac{x_k^1}{P_{\Omega}(X^1)}-\frac{x_k^2}{P_{\Omega}(X^2)}\right|$$

126. Distance de BHATTACHARYAA [7]

$$d_{\Omega}=\sum_{k=1,m}\uparrow_k\left[\frac{x_k^1\,x_k^2}{P_{\Omega}(X^1)\,P_{\Omega}(X^2)}\right]^{1/2}$$

Ceci n'est qu'une pseudo-distance, car $d_{\Omega}(X^1,X^1)\neq 0$. Nous proposons donc la distance symmétrique modifiée :

$$d_{\Omega}=\sum_{k=1,m}\uparrow_k\left[\frac{1}{2}\left(\frac{x_k^1}{P_{\Omega}(X^1)}+\frac{x_k^2}{P_{\Omega}(X^2)}\right)-\sqrt{\frac{x_k^1\,x_k^2}{P_{\Omega}(X^1)\,P_{\Omega}(X^2)}}\right]\geqslant 0$$

127. Information discriminante, ou entropie conditionnelle de KULLBACH [5]

$$d_{\Omega}=\left|\sum_{k=1,m}\left(\uparrow_k\frac{x_k^1}{P_{\Omega}(X^1)}\,\mathrm{Log}\left(\frac{x_k^1}{P_{\Omega}(X^1)}\Big/\frac{x_k^2}{P_{\Omega}(X^2)}\right)\right)\right|$$

d_{Ω} n'est pas symmétrique sous la forme ci-dessus ; d_{Ω} peut être considérée comme l'espérance du logarithme du rapport de vraisemblance (utilisé en théorie de la détection), pondéré par les probabilités conditionnelles par rapport à l'une des images. Cette mesure de similarité est plus particulièrement conçue pour le cas de deux classes de formes.

128. Distance distributionnelle du Chi-2 de BENZECRI [8]

$$d_{\Omega}^2=\sum_{k=1,m}\uparrow_k\left[\frac{x_k^1}{\uparrow_k P_{\Omega}(X^1)}-\frac{x_k^2}{\uparrow_k P_{\Omega}(X^2)}\right]^2$$

Cette distance est basée sur la définition du rapport de correspondance ($x_k / p_k\,P_{\Omega}(X)$), c.a.d. le rapport de la probabilité du couple (X,k) sur le produit des probabilités marginales des éléments X et k. Ce rapport de correspondance apparaît en théorie de l'information dans l'expression de la trans-information (PICINBONO [9]).

129. Remarque

TOUSSAINT [10] donne une comparaison des bornes inférieures de d_{Ω} pour X^1,X^2 donnés, entre les distances suivantes :

- divergence de KULLBACH (124.)
- distance variationnelle de KOLMOGOROV (125.)
- distance de BHATTACHARYAA (126.)

13. EXEMPLE DE COMPARAISON NUMERIQUE DE METRIQUES

La comparaison numérique envisagée est effectuée sur un ensemble de formes phonémiques à n: 10 coordonnées par forme. Ces phonèmes s'entendent après segmentation, et pour un locuteur unique ; 21 classes de phonèmes sont considérées, et pour l'interprétation des symboles utilisés pour représenter ces classes, on se reportera à [11].

131. L : ensemble des formes d'apprentissage

L est constitué par la première partie du fichier GOB1 de formes phonémiques constitué par MAISSIS [11] [12] . Les 7 premières mesures brutes sont les rapports R relatifs au phonème considéré , et les trois dernières mesures sont les 3 fréquences fondamentales détectées dans ce phonème .

Nous envisagerons deux types d'images X des formes phonémiques ci-dessus :

a) X^1 : x_k^1 = k-ième mesure brute sur le phonème d'apprentissage 1;

b) Y^1 : $y_k^1 = \mathrm{Log}\ (x_k^1)$, où x_k^1 est défini en a) ; on justifiera plus loin cette transformation .

Dans les planches , nous désignerons par N_i i:1,21 le nombre total de phonèmes d'apprentissage de la classe i considérés pour constituer L .

132. $\overline{L}$: ensemble de formes non-apprises à tester

$\overline{L}$ est une partie du complémentaire de L dans le fichier GOB1 . On appliquera bien entendu aux formes de $\overline{L}$ l'une ou l'autre des transformations a) ou b) , en sorte que les images de toutes les formes considérées soient homogènes dans un test de reconnaissance donné . Nous désignerons par N_i' i:1,21 le nombre total de phonèmes de la classe i utilisés pour constituer $\overline{L}$.

133. Λ : ensemble des formes de référence

Deux définitions ont donné lieu à des tests numériques :

a) $\Lambda = L$, c.a.d. que toute forme d'apprentissage est une forme de référence dont on connaît la classe d'appartenance .

b) Λ = ensemble des centres de gravité G_i i:1,21 des formes d'apprentissage $\ell \in L$ telles que ℓ appartienne à la classe i de phonèmes ; les mesures relatives à G_i sont donc calculées sur la seule base des mesures relatives aux formes d'apprentissage appropriées .

Pour obtenir l'image Z_i des formes de référence , on appliquera bien entendu selon le cas l'une ou l'autre des transformations a) ou b) du 131 ..

134. s : règle de décision non-paramétrique

La règle de décision s considérée est celle du voisin d'ordre v_i "généralisée" , dont les propriétés et les performances sont données dans PAU [13] [14] . Les probabilités a priori des différentes classes sont identiques :

$$X \in \text{classe } i \iff \frac{v_i}{(m_i+1)\,\varphi_i(X, v_i, d_\Omega)} = \max_{j=1,21} \frac{v_j}{(m_j+1)\,\varphi_j(X, v_j, d_\Omega)}$$

v_j : entiers donnés $\geqslant 1$; si $v_j > (n_j - 1)$, on pose $v_j = (n_j - 1)$.

$\varphi_j(X, v_j, d_\Omega)$: volume du plus petit voisinage de X , contenant exactement v_j images des formes de référence $\lambda \in \Lambda$ de la classe j ; ces voisinages de X sont limités par des surfaces $d_\Omega(X, Z(\lambda)) = \text{constante}$; le volume φ_j est calculé en utilisant la distance d_Ω , et ceci dans l'espace Ω à n dimensions ; par définition des voisinages de X , le volume en est proportionnel à $[d_\Omega(X, Z(\lambda))]^n$

n_j : nombre total de formes de référence de la classe j j:1,21 .

Pour le choix 133 b). de Λ , on aura $v_j = n_j = 1$, et la règle s ci-dessus sera équivalente à la règle dite du centre de gravité le plus proche .

135. P_Ω : mesure de probabilité dans Ω estimée à partir de L

P_Ω est estimée à partir des formes d'apprentissage de L , comme indiqué au début du paragraphe 12.. Ici encore , on tiendra compte de la transformation choisie en 131 ..

136. $T(d_\Omega)$: taux de reconnaissance pondéré

A partir d'une règle de décision s (voir 134 .) , et des formes $\overline{L}$ à tester , on obtient la matrice de confusion correspondante S : $[s_i^j (d_\Omega) , i :1,21 \; j:1,21]$ où ;

$s_i^j(d_\Omega)$: fréquence avec laquelle une forme de $\overline{L}$ appartenant à la classe i , est affectée à la classe j lorsqu'usage est fait de la règle $s(d_\Omega)$ dépendant de la métrique d_Ω.

Nous définissons alors le taux de reconnaissance pondéré comme :

$$T(d_\Omega)=\left(\sum_{i=1,21} N_i' \, s_i^i(d_\Omega)\right)\Big/\left(\sum_{i=1,21} N_i'\right)$$

Les taux de reconnaissance $s_i^i(d_\Omega)$ sont en effet pondérés par le nombre d'échantillons considérés de la classe i . Nous soulignerons que seul le score $T(d_\Omega)$ ci-dessus est correctement exprimé du point de vue statistique . L'auteur demande qu'on normalise de toute urgence la définition d'un taux de reconnaissance "moyen" , même sila définition normalisée donne des résultats à première vue moins bons que ceux publiés antérieurement .

14. RESULTATS

Les résultats numériques sont consignés dans les planchæ jointes , dans lesquelles les métriques envisagées sont repérées comme suit :

Metrique 1	Distance euclidienne L^2	122.
Métrique 2	Divergence de KULLBACH	124.
Métrique 3	Distance variationnelle de KOLMOGOROV	125.
Métrique 4	Distance distributionnelle du Chi-2	128.
Métrique 5	Distance L^1	123.

141. Choix des images de formes phonémiques

La transformation y_k = Log (mesure no k) ou 131 b). donne des résultats bien meilleurs que x_k = (mesure no k) , pour des raisons évoquées plus loin . On atteint un taux pondéré de 85 % pour 21 classes , et des formes à n:10 coordonnées . Les résultats qualitatifs suivants sont néanmoins quasi-indépendants du choix de la transformation .

142. Comparaison des métriques 1 à 5 sur la base des taux de reconnaissance pondérés $T(d_\Omega)$

a) Lorsque les formes de référence sont les formes d'apprentissage (133 a)), on obtient la classification suivante sensiblement indépendante des v_i :

$5 \succ 1 \succ (4 \succcurlyeq 2) \succ 3$ où $i \succ j$ signifie que la métrique i est meilleure que la métrique j .

La distance L^1 est de loin la meilleure . Plus les nuages partiels des classes sont étendus (c.a.d. plus v_i doit être grand , selon PAU [13]) , plus se creuse le fossé entre les métriques L^1 et L^2 . Dans les nuages partiels étendus , la distance du Chi-2 semble intéressante .

b) Lorsque les formesde référence sont les centres de gravité des classes individuelles (133 b)) , on trouve :

x_k : mesure k — $(4 \gtrsim 5) > 1 > 2 > 3$

y_k : Log(mesure k) — $5 > 1 > (4 \gtrsim 2) > 3$

Nous concluerons comme au a) , à ceci près que la métrique distributionnelle du Chi-2 peut être meilleure que la métrique L^1 dans certains cas à cause de son anisotropie ellipsoïdale .

143. Comparaison des métriques 1 à 5 sur la base des temps unitaires de reconnaissance

Les temps unitaires de reconnaissance τ par forme phonémique croissent bien entendu rapidement avec le nombre IMAX1 de formes de référence , approximativement comme IMAX1 log(IMAX1) . Une nouvelle programmation de la règle de décision s (134) a été réalisée , et a permis de réduire considérablement τ par rapport aux valeurs de [13] , toutes choses égales par ailleurs . En particulier , pour un compilateur donné et IMAX1 donné , τ est rendu approximativement indépendant de v_i . Sur ordinateur 370-65 et en FORTRAN IV (G ou H) , τ est compris entre 5 et 19 millisecondes CPU selon la métrique d_Ω , avec le classement suivant par valeurs croissantes de τ :

$$5 > 1 > 3 > 4 > 2$$

Une analyse plus détaillée des temps de calcul requis par les différentes phases de la reconnaissance révèle que ce classement est identique à celui résultant du classement des durées requises pour le seul calcul des distances de l'image à reconnaître aux images des formes de référence ; pour la métrique 5 , ce temps est de 5-10 % du temps total τ ,contre 50 % pour la métrique 2 . Avec la programmation réalisée , les autres calculs ne dépendent que du nombre de classes et de IMAX1 , et non pas de la métrique choisie .

144. Comparaison des métriques 1 à 5 sur la base de l'écart-type σ du taux de reconnaissance pondéré $T(d_\Omega)$

Nous avons mis en evidence une relation expérimentale intéressante entre le rapport $\sigma(d_\Omega)/T(d_\Omega)$ et le taux de reconnaissance pondéré $T(d_\Omega)$, l'écart-type étant évalué dans les mêmes conditions . Cette relation est linéaire dans l'intervalle $T(d_\Omega) \in [50, 90\ \%]$, étant entendu que les taux de reconnaissance $s_1^i(d_\Omega)$ sont majorés par 100 % . Cette propriété est indépendante de la métrique d_Ω choisie pour v_i donné :

$$\sigma(d_\Omega)/T(d_\Omega) = 11{,}6\ T(d_\Omega) - 493{,}4 \qquad T(d_\Omega) \text{ en } \%$$

Ceci illustre bien la complexité des problèmes étudiés sur la voie de l'amélioration des performances d'un système de reconnaissance des formes , car il est bien évident qu'il est déplorable d'augmenter l'écart-type lorsqu'on tente d'améliorer le taux de reconnaissance pondéré : cette amélioration "déstabilise" en fait le système . Nous devons en déduire qu'il n'est pas possible de stabiliser $\sigma(d_\Omega)$ en optimisant la règle de décision ; une telle stabilisation requiert probablement plutôt une transformation/normalisation des formes observées en amont de la reconnaissance , une réduction du nombre de classes et celle de la dimension n des vecteurs de forme .

On note que $\sigma(d_\Omega)/T(d_\Omega)$ décroît très légèrement en fonction de v_i , ce qui semble indiquer que les écarts-types sont légèrement moindres dans les nuages étendus , ceci étant probablement lié à la nature des données exploitées .

2. OPTIMISATION D'UNE TRANSFORMATION NON-LINEAIRE DANS L'ESPACE CONTRASTE

21. INTRODUCTION

La première partie a mis en évidence :

a) l'intérêt fondamental de transformations des formes observées , essentiellement sous la forme de transformations Ψ des n mesures k:1,n relatives à chaque forme ϕ :

$$Y=\Psi(X)=\Psi([x_1,\dots,x_k,\dots,x_n]) \qquad (8)$$

Avec la terminologie de [1] et du paragraphe 11. , on peut donc dire que l'on cherche un espace contrasté $\mathfrak{Q}$ dont on accuse précisément les contrastes discriminants à l'aide d'une des transformations citées Ψ .

b) expérimentalement une transformation non-linéaire particulière Ψ , à savoir :

$$X \to Y=\Psi(X)=[\log(x_1),\dots,\log(x_k),\dots,\log(x_n)] \qquad (9)$$

dont les resultats du point de vue taux de reconnaissance pondéré sont meilleurs que ceux de Y : X (cf. 141.) .

Pour des raisons numériques , et à titre restrictif , nous ne considèrons par la suite que des transformations non-linéaires caractérisées par une fonction réelle d'une variable réelle positive , à savoir :

$$Y=\Psi(X)=[h(x_1),\dots,h(x_n)]=[y_1,\dots,y_k,\dots,y_n] \quad h:\mathbb{R}^+\to\mathbb{R} \qquad (10)$$

Sur la base des constatations a) et b) , et de l'hypothèse restrictive ci-dessus , nous nous proposons de tenter l'optimisation du taux de reconnaissance pondéré $T(d_{\mathfrak{Q}})$ par rapport à la transformation h du (10) ; dans ce qui précède L, $\overline{L}$, s , Λ , $P_{\mathfrak{Q}}$ sont définis comme au 11. , et les applications numériques réalisées sont conformes aux énoncés du 13. .

Le problème ci-dessus est l'optimisation d'une fonction scalaire dans un espace fonctionnel approprié E . Afin de rendre cette optimisation réalisable du point de vue numérique , et profitant de l'expérience du paragraphe 141. , les espaces fonctionnels choisis sont les espaces E_m définis comme suit à partir de la donnée de l'entier m :

$$E_m=\left\{h:\mathbb{R}^+\to\mathbb{R} \;\middle|\; \forall x\in\mathbb{R}^+ : h(x)=\log x + a_1 x + a_2 x^2 + \dots + a_m x^m \; ; \; a_1,\dots,a_m\in\mathbb{R}\right\} \qquad (11)$$

En d'autres termes , les transformations h considérées pour chaque mesure de forme , sont des fonctions admettant un développement obtenu en additionnant un polynôme de degré m à la fonction logarithme . Cette dernière a exclusivement été choisie dans (9) et (11) pour ses propriétés expérimentales dans le cas concret considéré (13.) . Il est bien évident que la méthodologie ici présentée est beaucoup plus générale , et nous verrons qu'elle se prête relativement bien à des applications numériques .

22. OPTIMISATION DE LA TRANSFORMATION $h \in E_m$

221. Définition du problème d'optimisation

Dans une optimisation vis-à-vis de fonctions $h\in E_m$, le but est d'optimiser le taux de reconnaissance pondéré $T(d_{\mathfrak{Q}})$: $T(d_{\mathfrak{Q}})(A_m)$ par rapport aux vecteurs scalaires A_m: $[a_1 \dots a_m]$. $T(d_{\mathfrak{Q}})(A_m)$ est bien entendu calculé sur la base de la donnée de $d_{\mathfrak{Q}}$, L , $\overline{L}$, s , Λ , $P_{\mathfrak{Q}}$ et du nombre de classes de formes , les images $Y\in\mathfrak{Q}$ des formes ϕ de L , $\overline{L}$ étant calculées par la formule (10) avec h : h_m .

222. Caractéristique du problème $\operatorname{Max}_{A_m} T(d_{\mathfrak{Q}})(A_m)$

Dans les cas pratiques rencontrés , les ensembles de formes L,$\overline{L}$ seront finis discrets , et le nombre de classes sera un entier borné . La fonction de décision s ayant ainsi ses images dans un ensemble fini discret , les taux de reconnaissance

Δ_L^i dans chaque classe seront donc des fonctions en escalier vis-à-vis de toute modification de l'image par s d'une forme de $\overline{L}$. Il en résulte que $T(d_\Omega)(A_m)$, combinaison linéaire des $s_i(d_\Omega)$ pour $\overline{L}$ borné , sera une fonction en escalier de A_m puisque s est une fonction explicite de A_m .

La caractéristique fondamentale du problème $\operatorname{Max} T(d_\Omega)(A_m)$ est celui de conduire à la maximisation sans contraintes d'une fonction scalaire étagée de A_m. En d'autres termes , $T(d_\Omega)(A_m)$ sera une constante dans certains pavés de $\mathbb{R}^m$. Une telle maximisation rappelle à bien des égards des problèmes de programmation non-linéaire en nombres entiers , dont on connaît la complexité . Nous présenterons deux algorithmes jugés aptes à réaliser une telle maximisation :

a) un algorithme de gradient modifié (223.)
b) un algorithme "branch and bound" (224.)

223. Algorithme de gradient modifié

Cet algorithme opère séquentiellement sur la dimension m du vecteur A_m , à savoir que l'on commence par fixer m:1 et par prendre comme solution initiale $A_1^{(0)} : a_1 : 0$. Soit $\hat{a}_1$ la solution optimale obtenue pour $h \in E_1$. On fixe alors $m=2$ et la solution initiale $A_2^{(0)} = [\hat{a}_1 \ , \ 0]$ avant de recommencer l'optimisation, et ainsi de suite jusqu'à ce que des limites externes (temps de calcul) soient atteintes . Pour m fixé , l'algorithme est le suivant :

(1) Itération i:0 Fixer la solution initiale $A_m^{(0)} = [\hat{a}_1 \ldots \hat{a}_{m-1} \ 0]$, et calculer $T(d_\Omega)(A_m^{(0)})$. Se donner les scalaires $H > 0$, $M > 0$. Communiquer au gradient conjugué la précision absolue requise sur la valeur du critère à l'optimum , et le nombre max. d'itérations . Indiquer la valeur estimée du maximum du critère.

(2) Itération i (21) Calculer $T(d_\Omega)(A_m^{(i-1)})$

(22) Appliquer (23) pour k:1,m ; aller ensuite en (24).

(23) Soit $A_k^* = [a_1^{(i-1)} \ a_2^{(i-1)} \ldots a_k^* \ a_{k+1}^{(i-1)} \ldots a_m^{(i-1)}]$

$$a_k^* = \begin{cases} (1+H)\, a_k^{(i-1)} & \text{si } a_k^{(i-1)} \neq 0 \\ H & \text{si } a_k^{(i-1)} = 0 \end{cases} \qquad (12)$$

Calculer

$$DT_k = \begin{cases} \left(T(d_\Omega)(A_k^*) - T(d_\Omega)(A_m^{(i-1)})\right) \Big/ \left(H\, a_k^{(i-1)}\right) & \text{si } a_k^{(i-1)} \neq 0 \\ \left(T(d_\Omega)(A_k^*) - T(d_\Omega)(A_m^{(i-1)})\right) / H & \text{si } a_k^{(i-1)} = 0 \end{cases} \qquad (13)$$

- si $DT_k \neq 0$ retourner en (23) .
- si $DT_k = 0$, remplacer H dans (12)(13) par $(2\,H)$; si cette dernière valeur est inférieure à M , reprendre le calcul de DT_k et a_k^* avec cette nouvelle valeur de H . Sinon reprendre (12)(13) avec H remplacé par $(-H)$, H ayant sa valeur initiale fixée en (1) ; procéder comme indiqué plus haut pour les valeurs positives des H, jusqu'à ce que la valeur négative en cours devienne inférieure à $(-M)$: on fixera alors $DT_k = 0$, et retourner en (23).

(24) Considérer (DT_k , k:1,m) obtenus en (22) comme l'approximation du vecteur gradient du critère $T(d_\Omega)$ en $A_m^{(i-1)}$. Appeler ici un algorithme du gradient conjugué (par exemple [16]) pour calculer la solution améliorée $A_m^{(i)}$ à partir de $A_m^{(i-1)}$ et dudit vecteur gradient .

- si la précision requise sur l'optimum de $T(d_{\varrho})$ est atteinte , aller en (4) .
- si la précision atteinte n'est pas requise , aller en (3) .
- si le nombre max. d'itérations du gradient est atteint , ou en cas d'erreur , aller en (4).

(3) Si $T(d_{\varrho})(A_m^{(i)})$ est suffisamment près , ousupérieur à la cible assignée au (1) , aller en (4) . Sinon , faire $i = i+1$ et aller en (2) .

(4) FIN ; si le diagnostic n'est pas un diagnostic d'erreur , considérer $\hat{A}_m = [a^{(i)}, \ldots a^{(i)}]$ comme la solution optimale , et passer éventuellement de m à 1 $(m+1)^m$ dimensions .

On remarquera que le principe retenu au (23) pour l'estimation des dérivées partielles du critère par rapport à un scalaire a_k , est de procéder à une exploration de la fonction critère de part et d'autre de la valeur centrale $a_k^{(i-1)}$. Le pas de l'exploration est proportionnel à cette dernière . Si cette exploration est modérée , on la poursuit jusqu'à quitter éventuellement le plateau de la fonction critère auquel appartient $a_k^{(i-1)}$: ceci accélère la convergence . La dérivée partielle n'est nulle que si ce plateau est suffisamment étendu de part et d'autre de $a_k^{(i-1)}$, sa plus petite largeur $2\,M\,a_k^{(i-1)}$ devant être proportionnelle à $a_k^{(i-1)}$. On pourra bien entendu adapter les valeurs de H et M aux ordres de grandeur courants des constantes a_k dans (12)(13) .

224. Algorithme branch-and-bound

La prise en compte de ce principe de branch-and-bound ([17]) est basée sur la considération suivante : $T(d_{\varrho})(A_m)$ étant une fonction en escalier de A_m moyennant les hypothèses du 222. , il existera une infinité de A_m donnant la même valeur du taux de reconnaissance pondéré . Il existera donc une partition de l'espace $\mathbb{R}^m$ en pavés élémentaires π_j disjoints , associés chacun à une valeur de $T(d_{\varrho})(A_m)$. On peut donc considérer le problème 222. comme étant celui de la maximisation de la fonction discrète $T(d_{\varrho})(\pi_j)$ sur l'ensemble discret de pavés $(\cup \pi_j)$, isomorphe à $\mathbb{N}$. Nous voilà donc dans le cadre de la programmation en nombres entiers . On en déduit l'algorithme suivant , par application du principe de branch-and-bound au problème 222. :

(1) Se donner une estimation $\tilde{T}$ du taux de reconnaissance pondéré optimal pour $h \in E_m$. Donner $i = 0$. Aller en (2) .

(2) Branch : Dénombrer les pavés $\pi_j^{(i)} \subset \mathbb{R}^m$ tels que $T(d_{\varrho})(\pi_j^{(i)})$ soit voisin de $\tilde{T}$; on caractérisera à chaque itération i $\pi_j^{(i)}$ par un vecteur $A_{mj}^{(i)} \subset \pi_j^{(i)}$; dans chacun de ces pavés étendus , le taux de reconnaissance sera en général multiforme . Aller en (3).

(3) Bound : Trouver le pavé j^* tel que $T(d_{\varrho})(\pi_{j^*}^{(i)})$ soit maximum sur l'ensemble $\{j\}$ des pavés considérés à l'itération i ; si cette valeur est proche , ou supérieure à $\tilde{T}$, aller en (5) , sinon aller en (4).

(4) Partition : Partitionner le pavé $\pi_{j^*}^{(i)}$ en pavés moindres désignés par $\pi_j^{(i+1)}$; aller en (2) avec $i = i+1$.

(5) FIN

Cet algorithme n'a pas donné lieu pour l'instant à des expérimentations numériques , en dépit de sa souplesse remarquable . L'exploration requise par le point (2) pourra être effectuée comme le point (23) de l'algorithme de gradient , avec cependant une adaptation de M,H pour chaque itération i , afin de tenir compte des tailles décroissantes des pavés

23. EXEMPLE NUMERIQUE D'OPTIMISATION DE $h \in E_m$

On a appliqué l'algorithme du gradient modifié (223.) pour $h \in E_1$, E_2 et E_3 , l'optimisation étant réalisée grâce aux formes de L , et le test sur les formes de $\overline{L}$ .Les images des formes de L ou $\overline{L}$ à $n = 10$ dimensions se déduisent des données brutes par la transformation h (131.) . Les formes de référence Λ sont les formes d'apprentissage (133 a)), Λ partitionné en 21 classes . La règle de décision s est celle du 134 ., avec $v_i = 1$ pour toutes les 21 classes . La métrique choisie est celle du L^1 , conformément aux recommandations du 14. , mais aussi à cause de la rapidité de calcul (143.) qui est ici fondamentale pour réduire le temps de calcul requis pour une optimisation à m fixé (typiquement 5-20 mn CPU sur 370/65). On n'a pas jugé nécessaire d'aller au-delà de $m = 3$ eu égard aux faibles gains résultant du passage à un nombre supérieur de variables , et aux coûts de calcul correspondants .

24. RESULTATS

241. Amelioration du taux de reconnaissance pondéré $T(d_\Omega)(A_\cdot)$

Il convient essentiellement de comparer , pour L et $\overline{L}$ donnés , les gains réalisés en passant récursivement de l'espace E_o (c.a.d. de la transformation $h(x) = \mathrm{Log}(x)$), aux espaces E_m supérieurs . Selon les tailles respectives IMAX,IMAX1 des ensembles L et $\overline{L}$, on obtiendra des gains plus ou moins accusés à chaque itération m . Les gains les plus importants sont évidemment réalisés pour IMAX,IMAX1 petits (de l'ordre de 150 tous deux) .

D'une manière générale ,on constatera pour L,$\overline{L}$ donnés , que $\left(T(d_\Omega)(A_3) - T(d_\Omega)(A_o)\right)$ est de 1 % à 6 % pour des valeurs initiales $T(d_\Omega)(A_o)$ de 78-85 % pour des formes à $n = 10$ dimensions . Le taux maximum atteint $T(d_\Omega)(A_2)$ fut de 89,6 % pour IMAX,IMAX1 de l'ordre de 200 et $h \in E_2$.

Pour un m donné , la valeur maximale $T(d_\Omega)(A_m)$ est en général atteinte en 1-6 itérations (i) ; un choix convenable des paramètres de l'optimisation (223.(1)) permettra de laisser l'algorithme explorer un domaine important de $\mathbb{R}^m$ afin de limiter les frontières des pavés Π_j . L'optimisation du taux de reconnaissance pondéré et une exploration sommaire requièrent en général 5-10 mn CPU sur ordinateur 370/65 .

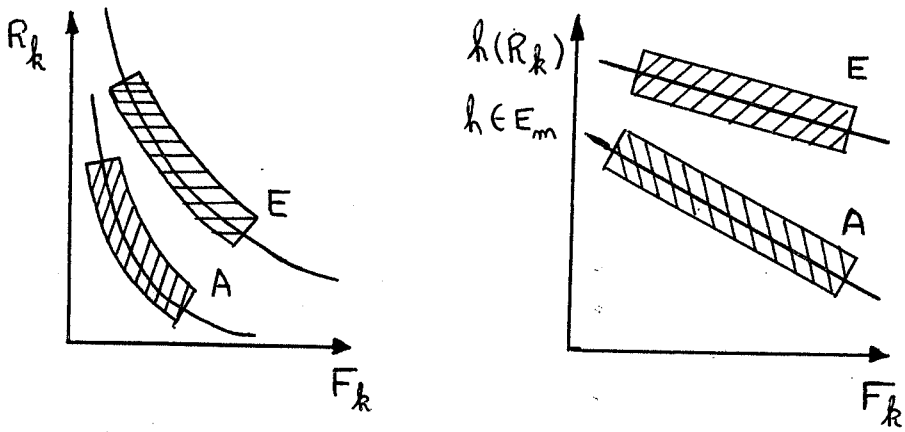
On trouvera dans les planches jointes quelques exemples d'application , et les valeurs optimales $\hat{A}_m$ obtenues dans chaque cas .

242. Interprétation des transformations optimales $h \in E_m$

Les propriétés expérimentales des transformations optimales $h \in E_m$ sont parfaitement conformes aux autres propriétés des formes phonémiques étudiées (cf. 131.) . On a d'abord mis en évidence expérimentalement (MAISSIS [11][12]) , puis justifié théoriquement par un modèle approché de la membrane basilaire (PAU [13][14]) , que les paramètres R_k étaient des fonctions décroissantes des formants correspondants F_k : voir la figure ci-dessous . L'allure de ces fonctions devrait être en $1/\sqrt{F_k - F_{ko}}$ selon [13] ; une telle relation peut être approximée par :

$$R_k = \mathrm{EXP}\left(-\mu\left(F_k - F_{ko}\right)\right) \qquad (14)$$

où u, F_{ko} sont des paramètres caractéristiques du canal k considéré et de la classe de phonèmes étudiée .



Par la transformation $h \in E_o$, on transforme les images des formes phonémiques en :

$$y_k = \text{Log } R_k = -u (F_k - F_{ko}) \qquad k = 1, m$$

qui est une expression linéaire par rapport à F_k , mais aussi par rapport à des paramètres caractérisant la classe de phonèmes considérée .

On peut en conclure que les transformations $h \in E_m$ ont pour rôle de transformer les formes phonémiques en des images linéairement séparables . Or on sait que la discrimination par séparatrices linéaires permet d'atteindre des taux de reconnaissance élevés . Les ajustements polynomiaux résultant de l'optimisation du vecteur A_m correspond à des ajustements de l'approximation précédente (14) . On remarquera que FLANAGAN recommandait déjà voici bien longtemps la normalisation logarithmique $h \in E_o$ de paramètres analogues aux paramètres R_k . Un projet américain en cours utilise cette même normalisation .

243. Conséquences pour un système opérationnel de reconnaissance phonémique

Soit G_r un opérateur de compression des données , transformant uniformément toutes les formes Φ de n dimensions à $r < n$ dimensions . Gr est défini sur la base des formes d'apprentissage de L . On peut alors reconsidérer le test de reconnaissance 221. , en remplaçant l'opérateur Ψ de (10) par l'opérateur composé $(G_r \circ \Psi)$.

A titre d'exemple , on a choisi pour opérateur G_r l'analyse factorielle des correspondances , comme dans PAU [13] [14] , la dimension de l'espace comprimé étant $r = 3$. Faisant choix de la transformation $h \in E_2$ optimale donnant $T(d_\Omega)(\hat{A}_2) = 89{,}6\ \%$ (cf. 241.) , on a obtenu dans l'espace des images tri-dimensionnelles un taux de reconnaissance pondéré de 92,8 % avec le même ensemble de formes d'apprentissage (IMAX = 200) ,mais un nombre double de formes à tester L (IMAX1 = 400) . Ce résultat peut paraître surprenant , mais peut s'expliquer à partir des propriétés énoncées au 242. , et des propriétés discriminantes de l'analyse factorielle des correspondances . Les temps unitaires de reconnaissance et de compression à 3 dimensions sont de l'ordre de 2-4 millisecondes CPU sur 370/65 .

3. CONCLUSION

Le trait commun aux deux parties de la présente étude est d'insister sur les aspects topologiques de la reconnaissance des formes discrètes stationnaires . On montre que des taux et des temps de reconnaissance améliorés peuvent être obtenus par une combinaison apprppriée d'un filtrage (métrique) et d'une déformation (transformation des images) .

On pourra objecter à la présente étude que ses résultats sont spécifiques des formes phonémiques analysées . Tel est certes le cas des taux de reconnaissance obtenus ; mais il n'en reste pas moins que l'on dispose maintenant de méthodes automatiques aptes à accélérer les études sur la normalisation des paramètres . Elles complètent heureusement les normalisations basées sur une meilleure connaissance physique des processus générant les paramètres servant à la reconnaissance .

REFERENCES

(1) L.F.PAU Common theoretical formulation of the pattern recognition , identification and detection problems ,IMSOR, Technical University of Denmark, march 1972 , 24 p.

(2) B.G. BATCHELOR Learning machines for pattern recognition, PhD thesis , Southampton , England , 1968

(3) G.D. TOUSSAINT On a simple Minkowski metric classifier, IEEE Trans. SSC, october 1970 , p.360-362

(4) J. RAVIV Decision-making in Markov chains applied to the problem of pattern recognition , IEEE Trans. Inform. Th., IT-13 , october 1967 , p. 536-551

(5) S. KULLBACH Information theory and statistics , WILEY , New-York , 1959

(6) T. KAILATH , The divergence and Bhattacharya distance measures in signal selection , IEEE Trans. COM-15 , february 1967

(7) B.P. BHATTACHARYA An approach to identification of piecewise linear control systems , IEEE Trans. Aut. Control., AC-11 , january 1966

(8) J.P. BENZECRI Distance distributionnelle du Chi-2 , Polycopié ISUP , Faculté des sciences Paris

(9) B. PICINBONO Eléments sur la théorie du signal, de la détection et de l'information , Revue du CETHEDEC , Vol 4, 1967-3 , no 11

(10) G.T. TOUSSAINT Some functional lower bounds on the expected divergence for multi-hypothesis pattern recognition , IEEE Trans. Syst. Sci. , SMC-1, no 4 , october 1971

(11) E.N.S. des TELECOMMUNICATIONS(Laboratoire d'automatique) Communication homme-machine à support vocal , Contrat CRI 71-14 , octobre 1972

(12) C.GUEGUEN , A. MAISSIS , L.F. PAU Communication homme-machine sur support vocal , Echo des recherches , no 70, octobre 1972

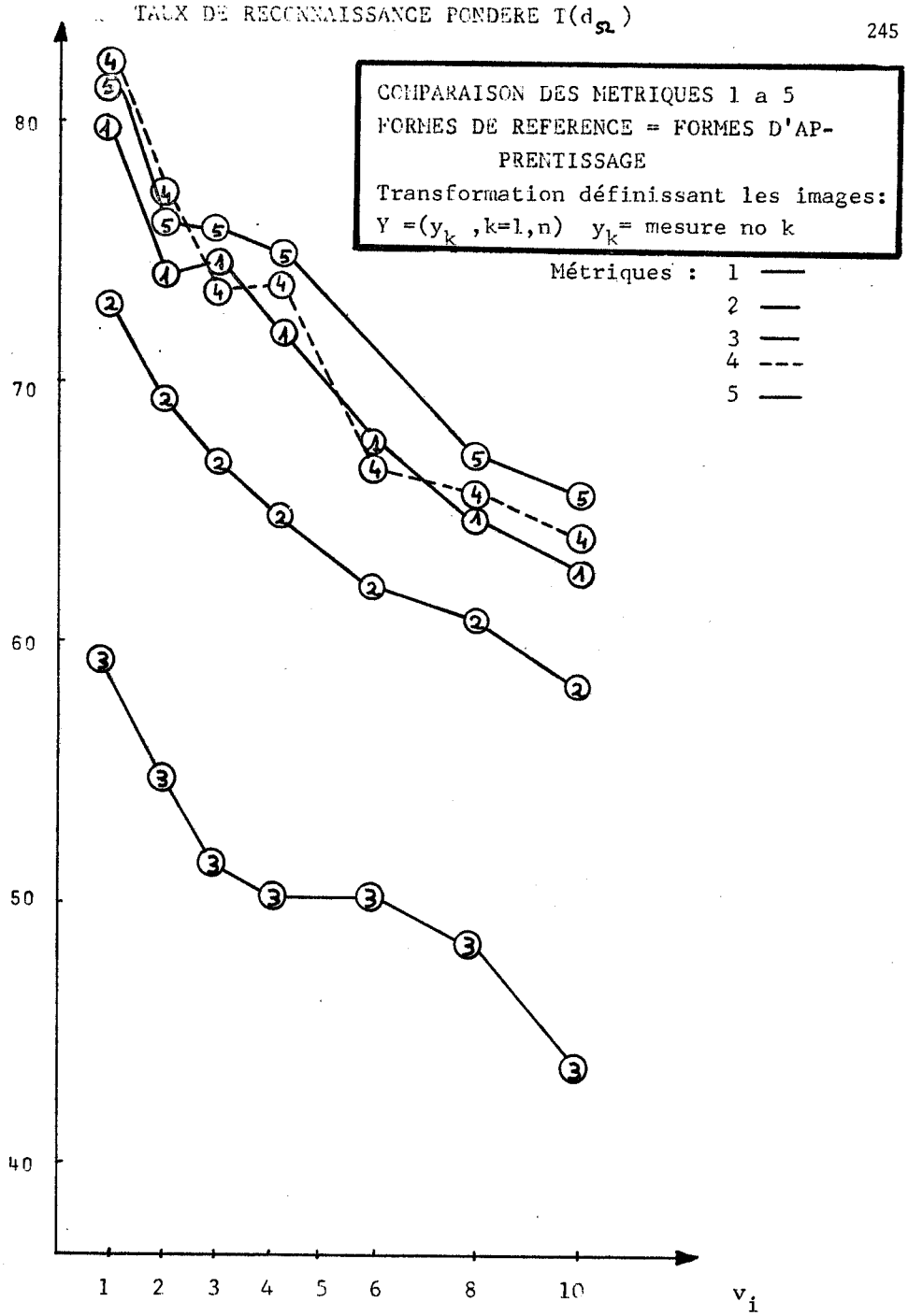
(13) L.F. PAU Méthodes statistiques de réduction et de reconnaissance des formes ; normalisations des paramètres phonémiques . Thèse , Université Paris-Sud(Orsay) , 29 mai 1972 ; voir aussi : Annexe technique 3 " Réduction et reconnaissance phonémiques" dans (11) .

(14) L.F. PAU Statistical reduction and recognition of speech patterns ; normalization of some phoneme parameters , " Machine perception of patterns and pictures" , Conf. series no 13 , Institute of Physics Publ. , London , 1972

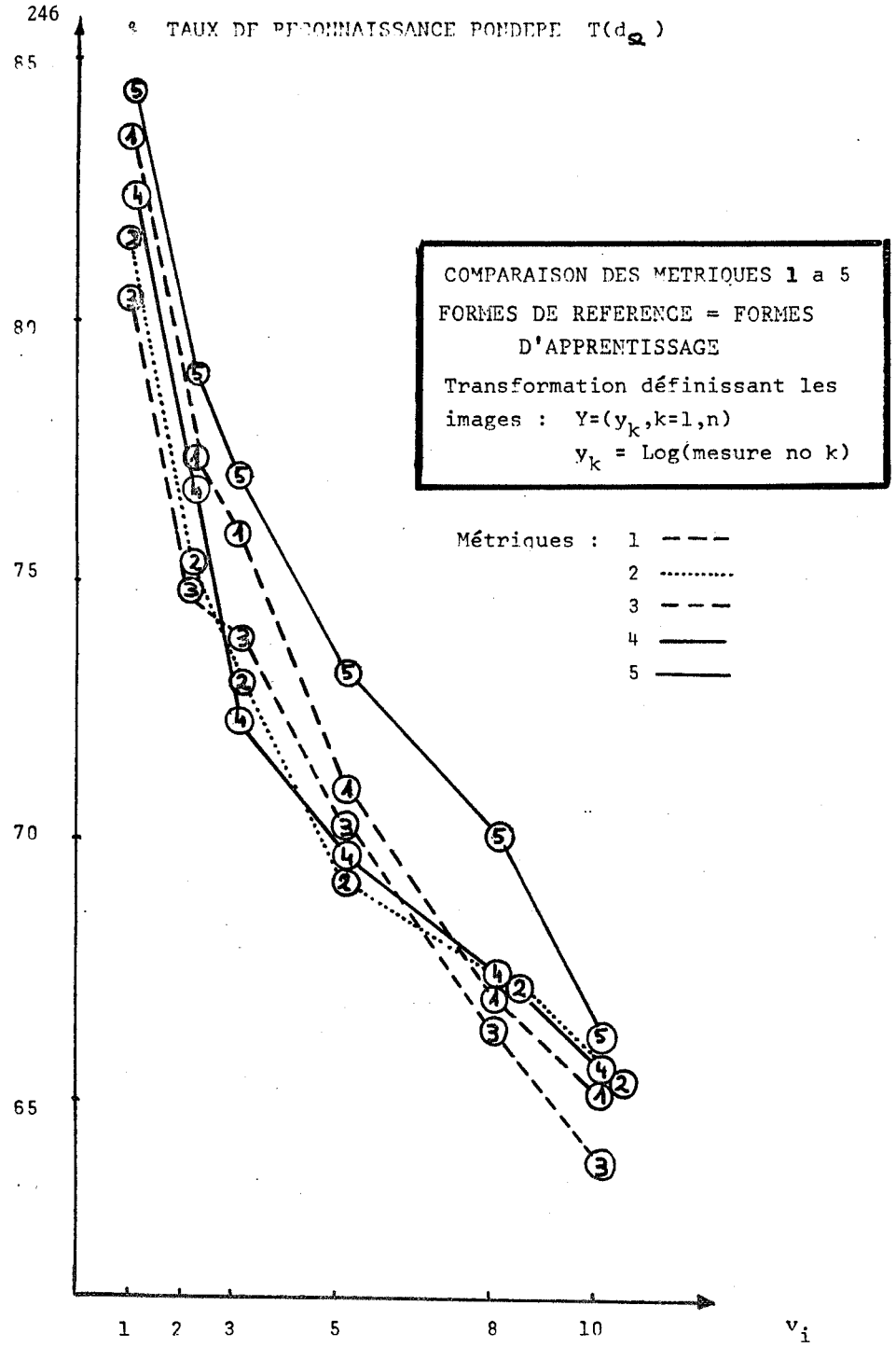
(15) D.N. JACKSON , I.J. WHITE , Effect of random errors on generalized distance computations , Pattern recognition, Vol. 4 , no 3 , october 1972

(16) R. FLETCHER , C.M, REEVES Function minimization by conjugate gradients , Computer J. , Vol. 7 , iss. 2 , 1964 , p. 149-154

(17) T.C. HU Integer programming and network flows , Addison-Wesley , 1969



Le paramètre v_i est ici le même pour toutes les classes i, dans un test donné . Les autres paramètres sont définis sur les planches relatives à la même comparaison entre métriques .



Le paramètre v_i est ici le même pour toutes les classes i , dans un test donné . Les autres paramètres sont définis sur les planches.

METRIQUE no. 1 - VOISINS LES PLUS PROCHES - TAUX DE RECONNAISSANCE EN MILLIEMES.

V_i	TAUX PONDERE	E	I	Y	O	A	AI	AN	ON	L	ZZ	OO	V	R	M	N	IN	Z	EE	J	W	EU
y = Log (x)																						
1	836	889	806	944	862	970	333	963	1000	622	889	895	821	594	842	903	1000	824	571	938	571	500
2	771	852	871	944	724	970	333	963	1000	405	778	842	821	531	711	806	1000	765	714	750	286	500
3	758	889	903	944	655	970	333	963	1000	405	778	789	714	531	744	744	1000	706	667	625	429	500
5	709	889	774	889	655	909	333	963	957	324	722	842	714	375	658	774	1000	765	571	625	143	0
8	671	926	839	889	586	848	333	963	957	216	611	842	643	281	632	742	1000	706	571	563	0	0
10	651	926	774	944	552	848	333	963	957	270	389	789	607	250	526	871	1000	647	571	500	0	0
y = x																						
1	796	815	839	1000	759	848	667	963	957	593	889	789	714	625	737	806	714	941	714	813	857	500
2	740	889	839	1000	621	818	333	963	957	486	611	737	679	500	737	806	714	647	714	813	714	500
3	744	926	839	1000	586	879	333	926	957	514	611	632	714	594	711	774	1000	588	714	813	571	500
4	718	852	839	1000	517	848	333	889	957	514	611	684	714	469	711	806	857	588	714	813	143	500
6	669	889	806	1000	552	818	333	889	957	432	333	632	464	344	711	806	857	588	429	813	143	500
8	656	926	774	1000	517	818	333	852	913	405	333	684	429	344	684	806	857	588	429	813	0	500
10	629	963	742	1000	379	818	333	852	1000	162	333	737	429	313	632	871	857	588	429	813	0	0
N'_i	450	27	31	18	29	33	3	27	23	37	18	19	28	32	38	31	7	17	7	16	7	2
N_i	583	33	33	28	29	33	21	28	28	33	32	23	33	33	33	33	25	27	24	29	11	14

METRIQUE no. 2 - VOISINS LES PLUS PROCHES - TAUX DE RECONNAISSANCE EN MILLIEMES.

V_i	TAUX PONDERE	E	I	Y	O	A	AI	AN	ON	L	ZZ	OO	V	R	M	N	IN	Z	EE	J	W	EU
y = Log (x)																						
1	816	889	871	1000	724	970	333	963	1000	622	889	842	821	594	763	806	1000	824	429	938	571	500
2	751	889	903	944	621	970	333	963	1000	405	833	737	750	469	711	710	1000	824	571	813	143	500
3	729	889	871	944	517	939	333	963	1000	459	778	789	679	500	692	742	1000	706	500	625	143	0
5	693	889	806	889	586	879	333	963	1000	324	722	789	679	375	632	742	1000	706	429	688	0	0
8	673	926	871	889	552	848	333	963	1000	297	722	842	643	188	500	806	1000	765	571	563	0	0
10	653	926	871	944	517	848	333	963	1000	270	389	842	607	281	421	871	1000	647	429	563	0	0
y = x																						
1	733	667	806	1000	655	909	667	889	913	622	833	737	607	688	605	774	857	647	425	625	714	0
2	691	704	774	944	655	848	333	889	913	568	778	684	536	406	579	806	857	706	429	750	286	0
3	667	778	774	1000	448	818	667	889	913	486	778	632	500	406	579	806	857	647	429	625	143	500
4	642	815	774	1000	483	848	333	889	913	459	556	526	500	406	579	774	857	588	286	500	0	500
6	618	815	839	1000	414	818	333	926	870	405	444	684	429	281	500	742	857	647	286	500	0	500
8	598	815	710	1000	345	848	0	852	913	297	278	737	393	250	474	871	857	647	286	688	0	500
10	578	815	645	1000	310	818	0	852	913	216	278	684	429	188	474	903	857	647	286	688	0	0
N'_i	450	27	31	18	29	33	3	27	23	37	18	19	28	32	38	31	7	17	7	16	7	2
N_i	583	33	33	28	29	33	21	28	28	33	32	23	33	33	33	33	25	27	24	29	11	14

METRIQUE no. 3 - VOISINS LES PLUS PROCHES - TAUX DE RECONNAISSANCE EN MILLIEMES.

V_i	TAUX PONDERE	E	I	Y	O	A	AI	AN	ON	L	ZZ	OO	V	R	M	N	IN	Z	EE	J	W	EU
y = Log (x)																						
1	804	852	871	1000	586	970	333	963	957	649	889	737	857	625	737	871	1000	824	714	875	286	500
2	747	852	871	944	586	939	333	963	957	459	833	684	821	469	711	710	1000	706	714	938	143	0
3	738	889	903	944	586	939	333	963	1000	486	778	789	714	438	718	710	1000	765	333	625	143	500
5	700	926	839	944	586	909	333	963	957	405	778	789	607	406	553	710	1000	706	429	688	0	500
8	667	926	806	944	552	848	333	963	957	351	500	737	643	281	553	839	857	647	286	688	0	0
10	636	926	774	944	448	848	333	963	957	270	389	789	643	156	474	871	857	647	429	625	0	0
y = x																						
1	593	370	677	944	448	879	333	852	565	432	611	632	429	406	684	516	714	765	429	563	571	0
2	544	370	774	944	345	848	0	852	609	324	500	579	357	281	632	645	429	588	143	563	143	0
3	513	370	677	833	310	788	0	778	478	378	500	579	250	344	526	613	571	588	286	688	0	0
4	509	333	677	883	276	788	0	778	522	378	389	632	286	219	553	742	571	588	143	563	0	0
6	500	444	387	833	345	758	0	815	696	324	389	526	250	188	474	806	714	647	143	688	0	0
8	484	481	355	833	310	788	0	778	652	270	278	474	321	125	474	806	714	647	143	688	0	0
10	436	444	226	889	276	788	0	741	609	189	56	474	214	31	368	839	857	647	143	88	0	0
N'_i	450	27	31	18	29	33	3	27	23	37	18	19	28	32	38	31	7	17	7	16	7	2
N_i	583	33	33	28	29	33	21	28	28	33	32	23	33	33	33	33	25	27	24	29	11	14

METRIQUE no. 4 - VOISINS LES PLUS PROCHES - TAUX DE RECONNAISSANCE EN MILLIEMES.

V_i	TAUX PONDERE	E	I	Y	O	A	AI	AN	ON	L	ZZ	OO	V	R	M	N	IN	Z	EE	J	W	EU
y = Log (x)																						
1	824	926	871	1000	793	970	333	963	1000	622	833	842	821	594	737	806	1000	824	714	1000	571	500
2	767	889	935	944	690	970	333	963	1000	405	889	789	786	469	684	742	1000	824	571	813	286	500
3	722	852	839	944	552	970	333	963	1000	405	722	789	714	469	692	710	1000	765	500	625	143	0
5	696	889	806	889	552	848	333	963	1000	378	722	842	714	375	632	710	1000	765	429	625	0	0
8	676	926	871	889	552	848	333	963	1000	270	722	842	643	219	553	774	1000	765	429	625	0	0
10	656	926	839	889	517	848	333	963	1000	270	444	789	607	281	447	871	1000	647	571	625	0	0
y = x																						
1	820	889	903	1000	793	939	667	963	1000	622	889	842	893	594	737	806	1000	765	571	875	429	500
2	776	889	871	1000	621	939	333	963	1000	514	833	842	893	500	711	774	857	706	714	813	286	500
3	733	852	806	1000	517	939	333	963	1000	405	778	789	821	469	632	774	857	706	571	813	429	0
4	736	926	839	944	586	909	333	963	1000	432	722	789	821	500	632	806	857	706	571	750	0	0
6	667	889	871	944	517	848	333	926	1000	324	500	789	679	281	474	774	857	706	429	813	0	0
8	656	926	806	944	552	848	333	963	957	270	333	789	714	156	474	871	857	765	429	688	0	500
10	629	889	839	944	552	848	333	926	957	297	222	789	714	94	421	871	857	706	286	500	0	0
N'_i	450	27	31	18	29	33	3	27	23	37	18	19	28	32	38	31	7	17	7	16	7	2
N_i	583	33	33	28	29	33	21	28	28	33	33	23	33	33	33	33	25	27	24	29	11	14

METRIQUE no. 5 - VOISINS LES PLUS PROCHES - TAUX DE RECONNAISSANCE EN MILLIEMES.

V_i	TAUX PONDERE	E	I	Y	O	A	AI	AN	ON	L	ZZ	OO	V	R	M	N	IN	Z	EE	J	W	EU
y = Log (x)																						
1	844	926	903	1000	793	970	333	963	1000	649	889	895	857	594	842	839	1000	824	714	938	571	500
2	789	852	871	1000	690	939	333	963	1000	405	833	789	821	563	763	806	1000	824	857	938	429	500
3	771	889	935	1000	621	939	333	963	1000	432	889	789	786	438	692	774	1000	824	667	875	429	500
5	733	889	839	944	621	939	333	963	1000	297	778	842	643	438	737	774	1000	765	429	813	286	500
8	702	926	871	944	586	879	333	963	1000	243	667	842	643	219	684	806	1000	765	571	813	143	0
10	664	926	839	944	517	848	0	963	1000	189	444	842	643	219	605	839	1000	706	714	625	0	0
y = x																						
1	818	815	839	1000	793	909	667	963	1000	676	833	789	750	625	789	806	857	941	714	813	857	500
2	760	852	839	1000	586	909	333	963	1000	486	667	737	750	469	816	806	857	706	714	813	714	500
3	758	963	839	1000	552	909	333	963	1000	514	667	789	714	500	711	839	857	706	714	813	429	500
4	749	963	806	1000	621	909	333	963	1000	486	611	684	750	500	684	806	1000	647	714	813	429	500
6	713	963	806	1000	586	879	333	926	1000	378	500	789	536	375	737	871	1000	647	714	813	0	500
8		963	806	1000	483	848	333	926	1000	270	333	684	643	281	658	903	1000	588	429	750	0	500
10	656	963	774	1000	448	848	333	889	1000	189	333	737	607	188	684	903	857	647	571	813	0	0
N'_i	450	27	31	18	29	33	3	27	23	37	18	19	28	32	38	31	7	17	7	16	7	2
N_i	583	33	33	28	29	33	21	28	28	33	32	23	33	33	33	33	25	27	24	29	11	14

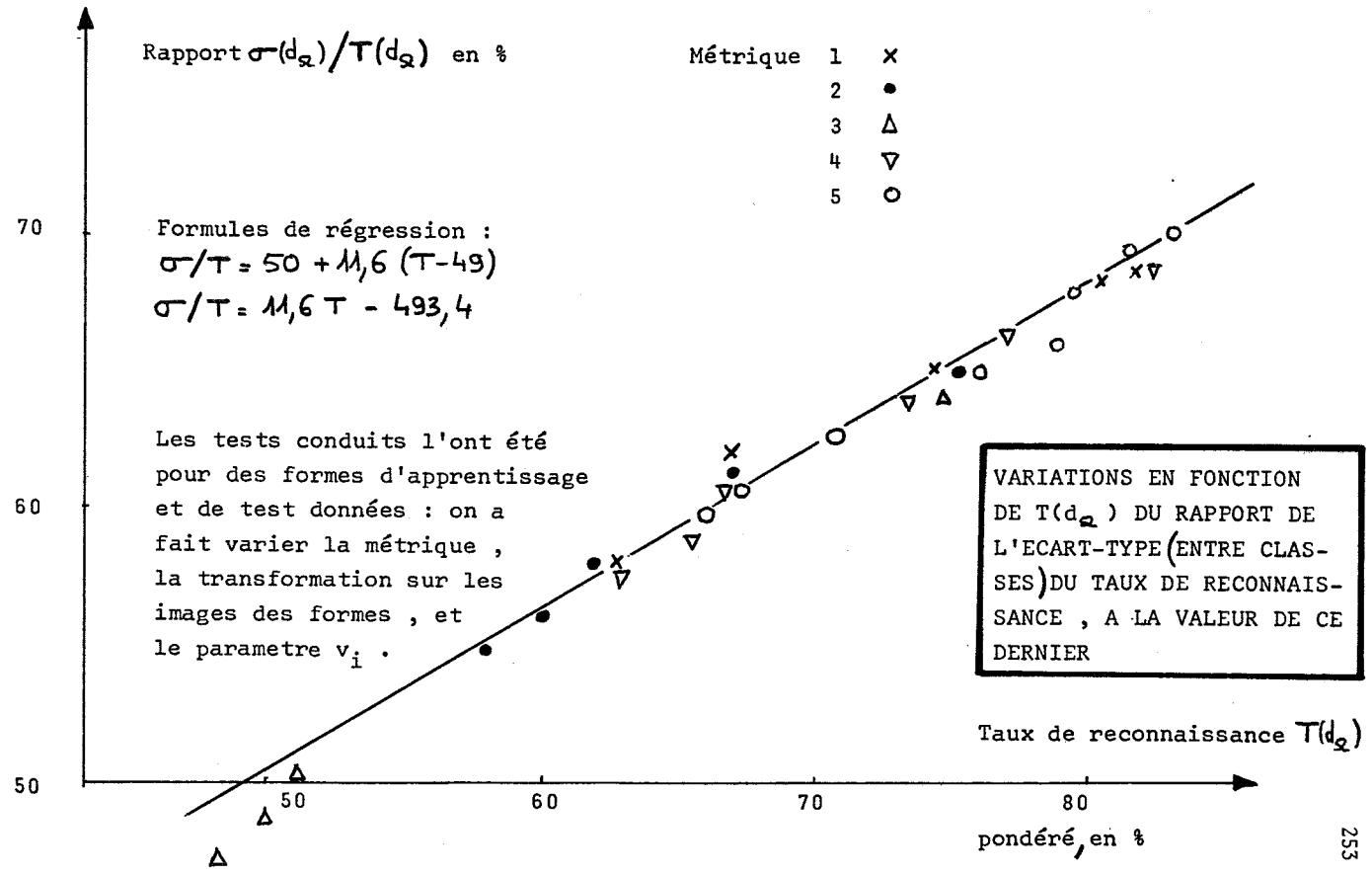
DISTANCES AUX CENTRES DE GRAVITE DES DIFFERENTES CLASSES - TAUX DE RECONNAISSANCE EN MILLIEMES

	$T_{\mu s}$	τ	σ	E	I	Y	O	A	AI	AN	ØN	L	ZZ	OO	V	R	M	N	IN	Z	EE	J	W	EU
y = x																								
METRIC#1	159	533	234	676	471	917	448	771	400	857	667	200	226	783	325	225	550	700	652	652	200	520	727	667
METRIC#2	635	481	186,5	618	294	750	207	771	400	893	704	225	290	478	225	300	500	750	783	696	250	360	182	333
METRIC#3	159	359	92	441	176	667	172	629	300	821	593	75	226	217	200	250	525	450	391	696	200	120	91	83
METRIC#4	317	588	287	853	412	875	276	829	500	929	926	275	323	696	550	275	525	750	696	739	350	560	545	750
METRIC#5	159	581	280	706	471	833	483	800	500	929	926	275	290	739	425	275	625	700	783	652	300	520	455	833
y = Log (x)																								
METRIC#1	159	623	320	912	529	875	379	829	500	929	926	275	516	696	550	325	675	600	783	696	400	560	636	833
METRIC#2	476	611	309	912	471	917	310	829	400	964	963	250	516	565	550	400	600	650	783	696	350	560	545	833
METRIC#3	159	591	290	853	412	875	345	829	550	964	963	275	452	522	550	325	625	650	826	696	250	440	273	833
METRIC#4	159	614	312	912	471	917	310	829	350	964	963	250	516	609	550	425	625	650	783	696	350	560	545	833
METRIC#5	159	638	335	853	471	875	448	829	550	964	926	250	484	696	550	450	675	625	913	696	500	560	636	833
N_i'		599		34	34	24	29	35	20	28	27	40	31	23	40	40	40	40	23	23	20	25	11	12

$T_{\mu s}$: temps unitaire CPU de reconnaissance par phonème , en microsecondes

τ : taux de reconnaissance pondéré , en millièmes

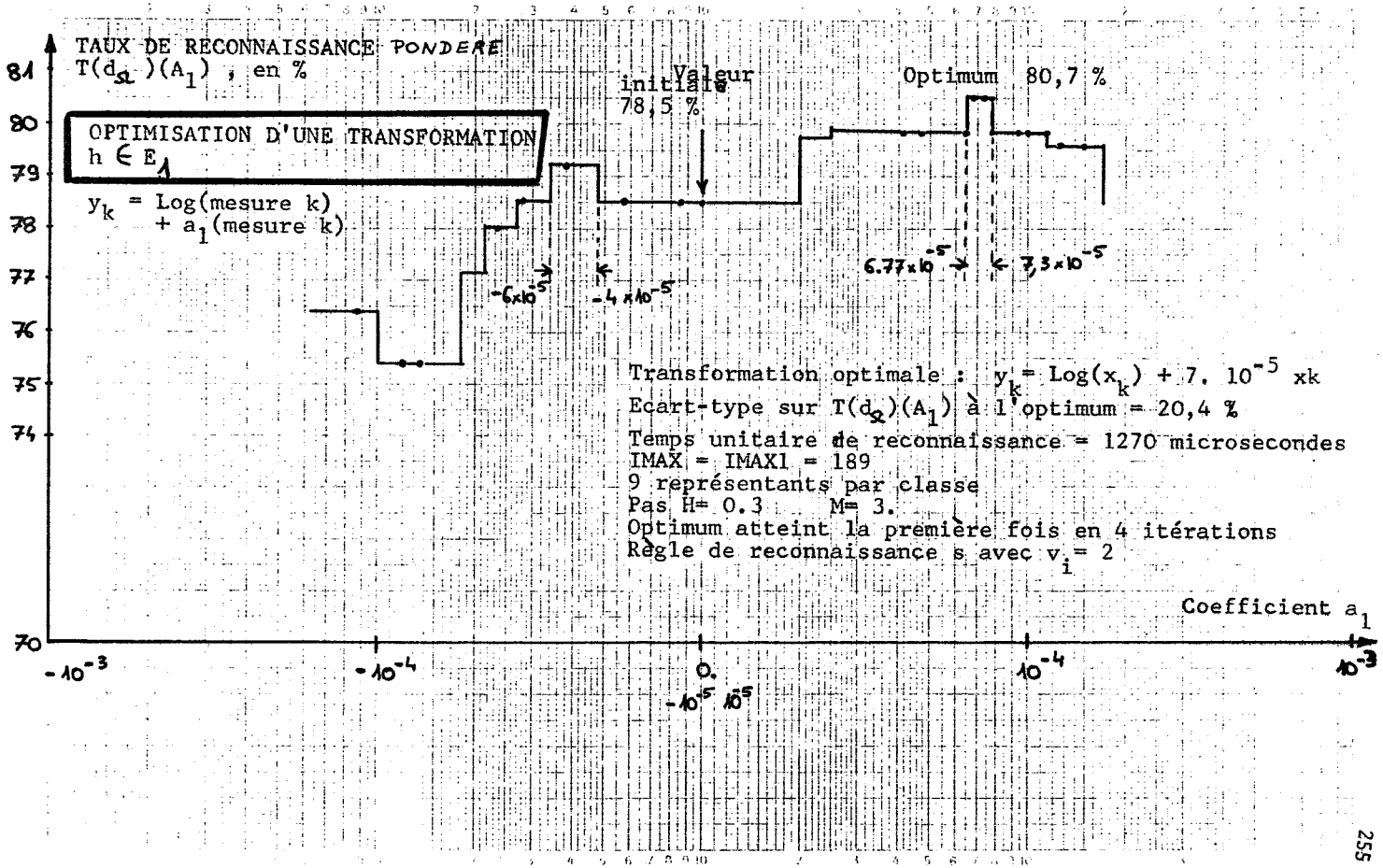
σ : écart-type sur le taux de reconnaissance pondéré , en millièmes

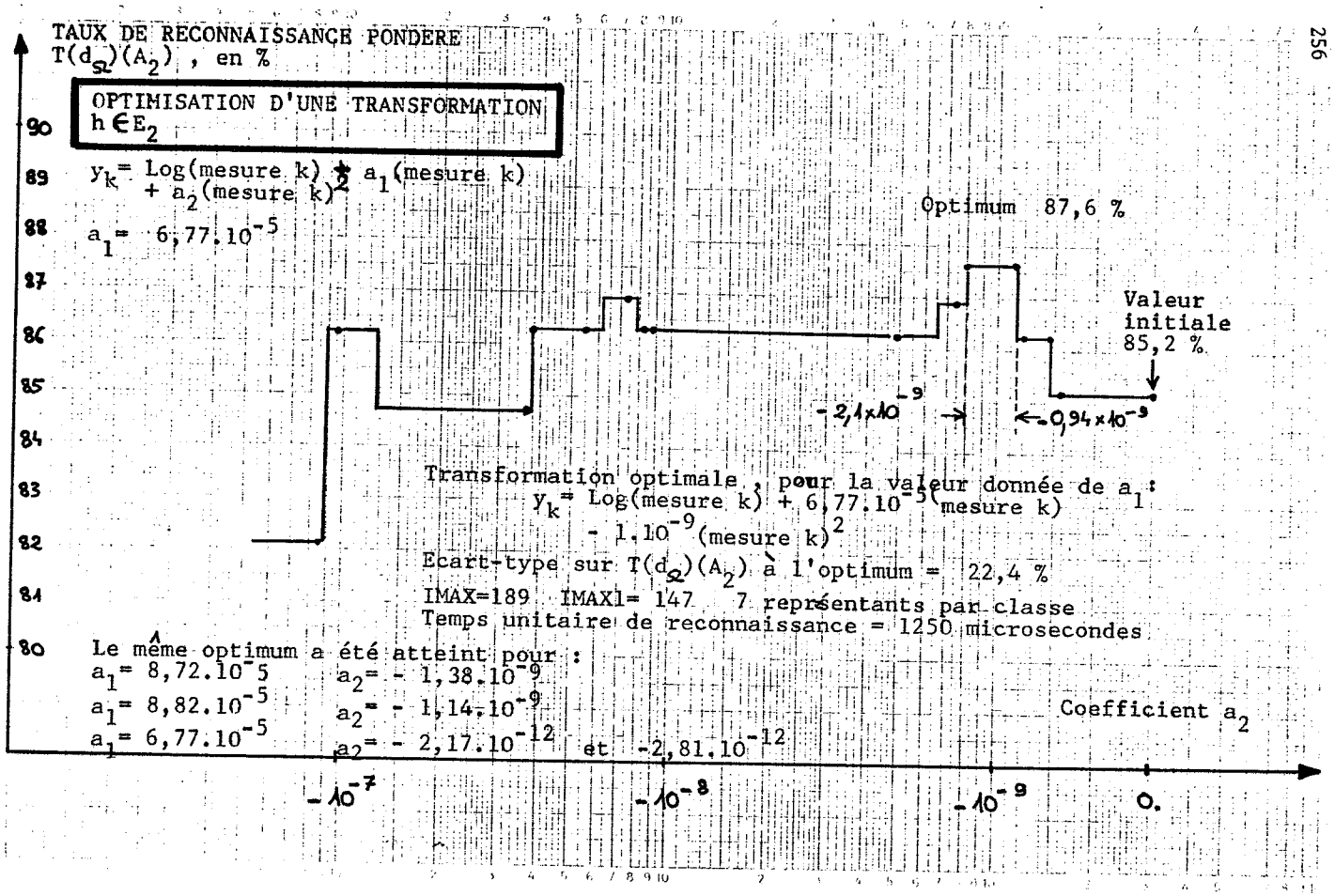

Rapport $\sigma(d_Q)/T(d_Q)$ en %
Métrique 1 2 3 4 5
70
60
50
Formules de régression :
$\sigma/T = 50 + 11,6\ (T-49)$
$\sigma/T = 11,6\ T - 493,4$
Les tests conduits l'ont été pour des formes d'apprentissage et de test données : on a fait varier la métrique , la transformation sur les images des formes , et le parametre v_i .
VARIATIONS EN FONCTION DE $T(d_Q)$ DU RAPPORT DE L'ECART-TYPE (ENTRE CLASSES) DU TAUX DE RECONNAISSANCE , A LA VALEUR DE CE DERNIER
Taux de reconnaissance $T(d_Q)$
50
60
70
80
pondéré, en %

TEMPS UNITAIRES DE RECONNAISSANCE POUR LES DIFFERENTES METRIQUES

en microsecondes CPU sur 370/65

IMAX=583 IMAX1=450 n=10 21 classes			METRIQUE 1	METRIQUE 2	METRIQUE 3	METRIQUE 4	METRIQUE 5
Y= LogX k=1,n							
v_i	Compilateur	Durée totale CPU du test					
2	G	6' 12''	7778	19 364	9841	9841	7142
3	H	8' 45''	7619	18 888	9365	9841	6984
5	G	8' 51''	7460	18 570	9999	10158	6984
8	G	8' 4''	6508	17 935	8730	9206	6190
10	G	8' 43''	7460	18 729	9047	10317	7301
Y = X k=1,n							
v_i	Compilateur	Durée totale CPU dutest					
1	G	8' 13''	6825	19 046	9682	9841	6666
2	G	7' 53''	6508	18 412	9047	9365	6666
3	H	6' 27''	4762	16 189	6666	7460	4920
4	G	8' 14''	6825	19 046	9682	9523	6825
6	H	6' 34''	4920	16 824	7142	7619	4920
8	H	6' 21''	4762	15 872	6666	7301	4444


TAUX DE RECONNAISSANCE PONDERE
$T(d_{\mathcal{L}})(A_1)$, en %
OPTIMISATION D'UNE TRANSFORMATION
$h \in E_1$
y_k = Log(mesure k) + a_1(mesure k)
Valeur initiale 78,5 %
Optimum 80,7 %
-6×10^{-5}
-4×10^{-5}
6.77×10^{-5}
$7,3 \times 10^{-5}$
Transformation optimale : y_k = Log(x_k) + 7. 10^{-5} xk
Ecart-type sur $T(d_{\mathcal{L}})(A_1)$ à l'optimum = 20,4 %
Temps unitaire de reconnaissance = 1270 microsecondes
IMAX = IMAX1 = 189
9 représentants par classe
Pas H= 0.3 M= 3.
Optimum atteint la première fois en 4 itérations
Règle de reconnaissance s avec v_i = 2
Coefficient a_1
81
80
79
78
77
76
75
74
70
-10^{-3}
-10^{-4}
0.
-10^{-5}
10^{-5}
10^{-4}
10^{-3}


TAUX DE RECONNAISSANCE PONDERE
$T(d_{S2})(A_2)$, en %
OPTIMISATION D'UNE TRANSFORMATION
$h \in E_2$
y_k = Log(mesure k) + a_1(mesure k)^2 + a_2(mesure k)
a_1 = 6,77.10^{-5}
90
89
88
87
86
85
84
83
82
81
80
Optimum 87,6 %
Valeur initiale 85,2 %
$-2,1\times10^{-9}$
$-0,94\times10^{-9}$
Transformation optimale , pour la valeur donnée de a_1 :
y_k = Log(mesure k) + 6,77.10^{-5}(mesure k) - 1.10^{-9}(mesure k)^2
Ecart-type sur $T(d_{S2})(A_2)$ à l'optimum = 22,4 %
IMAX=189 IMAX1= 147 7 représentants par classe
Temps unitaire de reconnaissance = 1250 microsecondes
Le même optimum a été atteint pour :
a_1= 8,72.10^{-5} a_2= - 1,38.10^{-9}
a_1= 8,82.10^{-5} a_2= - 1,14.10^{-9}
a_1= 6,77.10^{-5} a_2= - 2,17.10^{-12} et -2,81.10^{-12}
Coefficient a_2
-10^{-7}
-10^{-8}
-10^{-9}
0.

Discussion

J.P. HATON : Ne pensez-vous pas que la nature des paramètres utilisés, et la façon dont ils ont été choisis, influent sur l'évaluation de la distance optimale ?

L.F. PAU : Certes, tous les résultats qui viennent de vous être exposés doivent en toute rigueur être pris conditionnellement par rapport à la base de donnée utilisée. Un de mes rapports antérieurs a néanmoins montré, tout en introduisant l'analyse des correspondances en reconnaissance phonémique, que cette base de donnée était caractérisée par un recouvrement très important des classes entre elles. Sous la réserve précédente, je ne vois donc pas comment les formes des nuages individuels aient ici conduit à privilégier telle mesure de distance par rapport à telle autre. D'autre part, vous aurez le soin de remarquer que le classement de ces mesures est demeuré pratiquement invariant quand on appliquait aux formes brutes X, aussi bien la transformation convexe $Y = \text{Log } X$, que la transformation non monotone $Y = \text{Log } x + a_1 x + \ldots + a_n x^n$; or ces transformations transforment essentiellement les formes des nuages et leurs propriétés statistiques. L'invariance du classement des distances vis-à-vis de ces transformations contribue, me semble-t-il, à confirmer mon affirmation conditionnelle du début.

PRELIMINAIRES A L'ETUDE DES LANGAGES DE COMMUNICATION PARLEE :

CODAGE PHONETIQUE, ORTHOGRAPHE ET SYNTAXE.(++)

Résumé

A partir d'une transcription phonétique idéale est-il possible de restituer l'orthographe ? Nous nous sommes posés le problème pour une émission de voix (suite continue de caractères dont les pauses sont absentes). Un modèle markovien a donné des résultats médiocres. Un modèle déterministe, limité aux émissions de voix tirées de phrases affirmatives, a semblé satisfaisant. Le traitement des accords nécessite l'introduction d'une grammaire d'attributs manipulée comme une "sémantique" de la syntaxe initiale.

L'incorporation de l'analyseur syntaxique à un reconnaisseur acoustique est brièvement présentée.

Summary

Syntax described as a markovian process or as a context-free grammar, was used in an experiment of phoneme to grapheme translation of breath groups (in french). The deterministic model seems to be the only one sufficiently successfull. Final orthographic treatment is done as a semantic interpretation of the initial syntax.

We present a possible interface between an acoustic recognizer and our syntax analyzer.

L. MICLET (+), B. ROUGEOT (+), J-P. LE PAPE (+), J-Y. GRESSER.
Centre National d'Etudes des Télécommunications.

(++) Cette étude a été effectuée avec la participation du CRI dans le cadre de la convention de recherches 70.105/71.09

(+) Alors à l'ENSEEIHT (5e année).

En reconnaissance vocale la démarche analytique consiste à transcrire le flot de parole par une suite continue de pseudo-phonèmes. L'ingénieur suppose qu'il sera facile par la suite d'en extraire ce qui est utile à une bonne interprétation de la parole prononcée, à la limite d'une transcription intégrale correctement orthographiée.

Nous avons voulu éprouver la difficulté réelle du passage du "phonétique" à l'orthographe, de

(p.ex.) " mamă lɛptibato " à "maman, les petits bateaux",

et justifier ou infirmer l'optimisme habituel.

REFERENCE.

Commencée au CNET en 1971, poursuivie au CNET et à l'ENSEEIHT en 1972 et 1973, l'étude présentée reprend et complète, pour le français, un travail de D.R. REDDY et A.E. ROBINSON (1), dont Je rappelle les grandes lignes.

Le problème posé est la restitution de l'orthographe d'une phrase isolée présentée sous forme phonétique, c'est-à-dire écrite à l'aide d'un alphabet phonétique, sans indication de pause, ni d'élément prosodique.

Le vocabulaire est réduit à 1000 mots environ tirés parmi les plus fréquents d'un dictionnaire d'usage courant (DEWEY). Les phrases sont analysées à l'aide d'une grammaire de transitions comportant un petit nombre de règles. Ces transitions portent sur 10 catégories grammaticales regroupant les 1000 mots.

Exemple de règle : nom $\longrightarrow$ verbe,
adverbe
conjonction.

Quelques exemples sont donnés. Mais aucun résultat d'ensemble n'est chiffré. Les auteurs se déclarent cependant satisfaits. Le temps de calcul (sur ordinateur moyen) est inférieur à 1 s. Le programme permet une mise à jour facile du vocabulaire. Notre étude diffère de celle de REDDY et ROBINSON par quelques points mineurs (taille du vocabulaire plus importante, mots et grammaire français), et majeurs (organisation du lexique, essais sur plusieurs types de grammaires).

DEFINITIONS.

Nous considérons deux alphabets. L'un est l'ensemble des lettres de A à Z, l'autre est un ensemble de 36 caractères dits phonétiques.

Un mot est un triplet.

chaîne orthographique (construite par concaténation des lettres)	chaîne phonétique (construite par concaténation des caractères phonétiques)	attributs

Deux mots sont différents si l'une des deux chaînes diffère. Ainsi cheval - ʃfal
et cheval - ʃœval sont différents
(ʃ(ə)val n'existant pas).

Chaque chaîne est composée de deux sous-chaînes le radical et la désinence. Le nombre de désinences est très inférieur à celui des radicaux.

Une émission de voix est une séquence de mots contenue dans une phrase.

Ex. (kivɔ̃syrlo / qui vont sur l'eau)

(ki / qui) (vɔ̃ / vont) (syr / sur) (l / l') (o / eau)

Le problème abordé dans cette étude peut s'énoncer de la manière suivante :

une émission de voix étant partiellement donnée par la suite continue des chaînes phonétiques des mots qui la composent, restituer la forme orthographique correcte.

Ex. $\tilde{o}tizd\varepsilon\tilde{z}\tilde{a}b$ → ont-ils des jambes ?

C'est le genre de problème qu'il faudrait résoudre à l'étage ultime d'un reconnaisseur de parole disposant d'un analyseur phonétique idéal, capable de fournir une transcription intégrale, exempte d'erreur.

La simple concaténation entre chaînes phonétiques nous permet d'esquiver la question des liaisons et élisions qui seront traitées ultérieurement.

LE VOCABULAIRE.

Contenu: Nous avons tiré notre vocabulaire de base (1700 mots environ) d'un dictionnaire pour enfants. Certains mots caractéristiques du vocabulaire enfantin ont été éliminés, d'autres mots courants ont été rajoutés.

Le lexique est fractionné en un dictionnaire et une table.

- Le dictionnaire des radicaux est un répertoire de tous les mots du vocabulaire reconnu. Il contient les parties invariables de ces mots, sous forme phonétique et orthographique, et des pointeurs vers

- la table des désinences. Celle-ci contient toutes les parties variables pouvant être accolées à un radical pour former un mot français correct.

A la table et au dictionnaire, qui sont triés, sont associés des tableaux permettant un accès rapide aux informations qu'ils contiennent.

Une entrée du dictionnaire des radicaux a la forme :

Représentation phonétique	Représentation graphique	N° de désinence	Catégorie gram.
de 1 à 12 phonèmes, codée sur 3 mots	de 1 à 20 lettres, codée sur 5 mots	3 chiffres (1 mot)	3 ou 4 lettres

Une entrée du dictionnaire des désinences a la forme suivante :

N° de désinence	Ecr. phonétique	Ecr. graphique	Renseignements grammaticaux
3 chiffres	1 à 4 phonèmes	1 à 8 lettres	0 ou 3 lettres

Pour le vocabulaire considéré, il y a 1900 radicaux (compte tenu des conjugaisons) et 70 désinences.

DECOUPAGE PRELIMINAIRE.

On cherche, à partir d'une séquence phonétique représentant une émission de voix donnée, à découper celle-ci en chaînes ou mots phonétiques, de toutes les manières possibles, et à en donner toutes les orthographes possibles.

Représentation d'une séquence phonétique par un graphe.

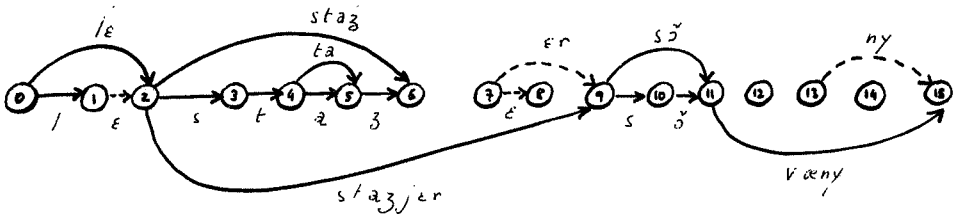
- chaque phonème est représenté par un noeud du graphe,

- il y a un arc entre le noeud i et le noeud j si et seulement s'il existe un mot phonétique commençant au phonème i (i non compris dans le mot), et se terminant au phonème j (j compris),

- le noeud 0 ne correspond à aucun phonème, et représente le noeud initial.

Exemple :

"les stagiaires sont venus"



Les arcs indiqués en pointillés correspondent à des mots effectivement présents dans la chaîne phonétique, mais qui ne doivent pas être pris en compte, puisqu'il n'existe pas de mot se terminant juste avant leur premier phonème.

Le découpage fonctionne en deux phases :

- une phase de construction du graphe,

- une phase d'exploration.

Tel quel, le nombre de solutions obtenues est très grand. Il faut le réduire par une analyse plus poussée de l'émission de voix.

Nous nous sommes limités à une analyse syntaxique, pour laquelle nous présentons dans ce texte deux types d'approche : statistique et déterministe.

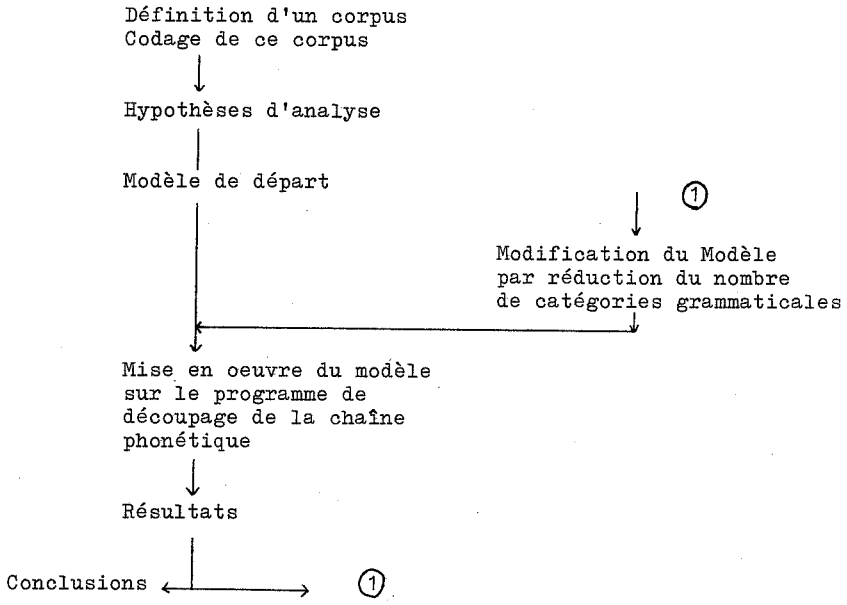
MODELE STATISTIQUE.

Les raisons du choix initial ont été les suivantes :

- un modèle statistique nous semblait moins restrictif qu'un modèle déterministe et nous pensions pouvlir éliminer un grand nombre de découpages erronés par l'application de critères simples,

- l'unité à découper étant une émission de voix et non une phrase complète, le modèle déterministe présentait quelques difficultés d'application.

La méthode d'élaboration du modèle a été la suivante :



Codage d'une émission de voix.

Nous avons, d'après la grammaire française traditionnelle, défini 33 catégories grammaticales.

Une émission de voix est une suite de mots de la langue française, dont chacun peut être codé par le nom de la catégorie grammaticale à laquelle il appartient. Le silence, qui délimite une émission de voix, a été représenté par la catégorie "DEB" qui signifie "absence de catégorie grammaticale".

A titre d'exemple, le codage de l'émission de voix "souvent, pour s'amuser, les hommes d'équipage" est, dans le cadre ainsi défini :

/DEB/ADV/PRE/PCP/VII/ARD/NCO/PRE/NCO/DEB

Corpus codé.

Nous avons ainsi codé 3145 émissions de voix, soit 22738 mots (y compris les silences).

Sur ces émissions de voix, environ 1100 sont tirées d'articles de politique étrangère du journal "Le Monde", le reste représentant le codage intégral du contenu du "Dictionnaire des 2000 phrases" de Henri Frei, à la Librairie DROZ, qui tente de donner un éventail aussi complet que possible de concepts, à l'aide de phrases variées du français courant.

Le corpus choisi n'est donc pas très homogène. Il va de soi qu'il ne prétend pas non plus être définitif. Il est composé d'une part d'"émissions de voix" théoriques extraites d'articles assez littéraires, d'autre part de véritables émissions assez courtes et de structure grammaticale variée qui sont effectivement du Français parlé. Il faut donc préciser que la variété et la relative minceur du corpus ne permettent pas d'émettre des conclusions formelles.

Hypothèses d'analyse.

L'émission de voix, c'est-à-dire la suite de catégories grammaticales, a été considérée comme une suite de réalisations d'une variable aléatoire discrète à autant d'états qu'il y a de catégories possibles.

Une émission de voix est donc un processus aléatoire discret et homogène, puisque l'instant d'émission des mots ne peut pas intervenir (la chaîne phonétique d'entrée est continue, et il s'agit de la découper).

Nous supposerons que cette suite de réalisations est une chaîne de Markov, c'est-à-dire que, dans la chaîne de mots, la probabilité pour un mot d'appartenir à une catégorie grammaticale donnée ne dépend que de la catégorie grammaticale du mot précédent.

Nous avons donc, à partir de cette hypothèse, analysé le corpus afin de le réduire dans une matrice appelée "Matrice des Suites Grammaticales", ITAB, et telle que : ITAB (I, J) est le nombre de fois où, dans le corpus, un mot appartenant à la catégorie grammaticale de rang I est suivi par un mot appartenant à la catégorie grammaticale de rang J.

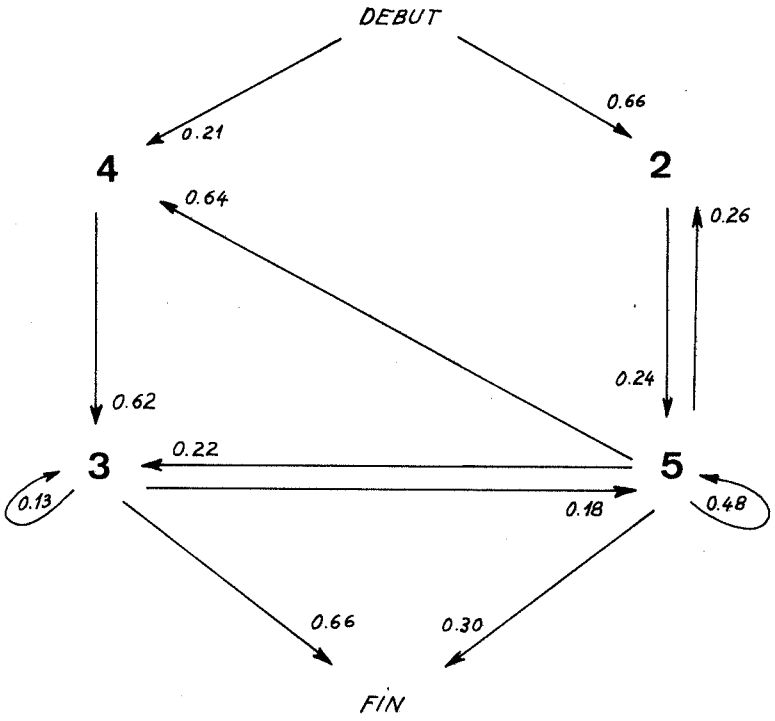
Par division de chaque élément de ITAB par la somme des éléments de la colonne où il se trouve, on obtient la Matrice de Transition de la chaîne de Markov considérée.

SIMPLIFICATION DU MODELE...

... Par réduction du nombre de catégories grammaticales.

Une distance est définie entre deux catégories. Deux catégories voisines voient leur comportement confondu. L'arrêt du processus de réduction se fait sur une variation brusque de la distance minimum considérée.

Dans notre étude nous nous sommes arrêtés à un modèle à 5 états.



Mise en oeuvre du modèle.

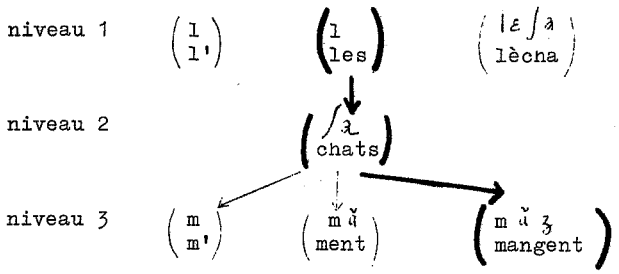
Après les initialisations, la mise en mémoire des dictionnaires et des tables qui les décrivent, le programme appelle la séquence de phonèmes à traiter.

La première étape est de rechercher à l'aide des dictionnaires tous les mots qui peuvent être le premier mot de la phrase issue de la chaîne de phonèmes.

Tous les mots trouvés forment ce que nous avons appelé le premier "niveau" du découpage de la chaîne de phonèmes. Ces mots sont répertoriés ; ils représentent les noeuds du premier niveau d'une structure d'arbre. De façon interne, chaque noeud est décrit par sa représentation orthographiée, le rang dans la chaîne de phonèmes de son phonème final, et sa catégorie grammaticale.

Ensuite, on repart du premier noeud afin de construire de la même façon les noeuds du "deuxième niveau".

Exemple :



Le modèle statistique intervient pour limiter le nombre de mots examinés à chaque niveau, éventuellement pour les classer par leur probabilité d'occurence.

Résultats (voir fig. 1, 2).

La caractérisation grammaticale ne permet pas de distinguer au cours du découpage les homonymes de même catégorie. Pour l'édition seul le premier homonyme est cité. Les accords sont ignorés.

Le progrès apporté par la modification du modèle de départ n'est pas important.

Il y a une perte de précision dans les résultats, puisqu'à un mot donné correspondent à priori plus d'homonymes de même catégorie grammaticale que dans le modèle à 33 catégories.

D'un autre côté, il y a moins d'émissions phonétiques non comprises : il n'y a pas, dans le modèle à 5 états, de suite impossible de deux catégories.

Le découpage proprement dit semble légèrement meilleur pour le modèle à 5 états car la structure plus rigide de l'autre modèle conduit parfois à un mauvais choix entre deux découpages.

Exemple AN K O R E' K R I R

pour le modèle à 5 états : ENCORE ECRIRE

pour le modèle original : EN CORPS ET QUE RIRE

Mais, d'une façon générale, le découpage n'est pas en gros progrès par le nouveau modèle.

Le temps de traitement d'une séquence de phonèmes est en général plus court qu'avec le modèle déterministe. Mais l'imprécision des résultats, la grande proportion de phrases erronées dans leur traduction sont des inconvénients très lourds de cette méthode dont l'efficacité est finalement assez médiocre.

MODELE DETERMINISTE.

Le sous-ensemble du Français que nous avons considéré est celui défini par la grammaire hors-contexte construite par M. ICHBIAH, H. VALABREGUE et F. LEVERY (2).

Cette grammaire a le mérite d'être à la fois peu étendue (une trentaine de terminaux et environ autant de non-terminaux), et suffisamment large pour pemettre la reconnaissance d'un assez grand nombre de types de phrases.

Les non-terminaux "inutiles" ont été supprimés, ce qui réduit d'autant le volume de la grammaire. Un non terminal peut être dit "inutile" lorsque la règle de production qui le définit ne contient dans sa partie droite qu'un seul élément du vocabulaire (terminal ou non terminal).

La méthode de reconnaissance du participe présent (consistant à accoler le terminal "ant" après le terminal représentant le verbe) nous a paru plus compliquée d'emploi, vu le problème que nous cherchons à résoudre, que celle qui consiste à créer plusieurs terminaux supplémentaires pour représenter les participes. Ces terminaux sont :

* patr (≡ vt ant) : participe présent transitif
* pain (≡ vi ant) : participe présent intransitif
* pade (≡ vde ant) : participe présent suivi de "de"
* paa (≡ va ant) : participe présent suivi de "a".

Représentation de la grammaire et procédure d'analyse.

La représentation choisie est due à CONWAY.

A chaque règle est associé un graphe dont chaque noeud représente un état syntaxique.

Le langage généré par la grammaire devant être analysé par un analyseur syntaxique descendant, les restrictions classiques pour ce genre d'analyseur doivent être respectées.

L'analyseur écrit est inspiré de la méthode de CONWAY et de la "parsing machine" de KNUTH.

Son but est, rappelons-le, de déterminer si une chaîne de catégories grammaticales fournie par le programme de découpage correspond à une phrase correcte selon la grammaire utilisée.

Mise en oeuvre.

Le modèle intervient dans le choix du mot suivant et dans la validation syntaxique de l'émission de voix analysée. Le mot choisi au niveau n appartient à une catégorie grammaticale qui peut suivre correctement celle du mot retenu au niveau n - 1 (le premier mot choisi a la longueur en phonèmes la plus grande).

En fait la syntaxe conduit l'ensemble du processus d'exploration, notamment les retours en arrière : après un échec l'analyse reprend au dernier niveau supérieur qui peut mener à une phrase ou une émission de voix syntaxiquement correcte.

Résultats et commentaires (fig. 3).

Comme on pouvait s'y attendre le fait de limiter la syntaxe admise par l'introduction d'une grammaire assure qu'une séquence phonétique correcte (c'est-à-dire n'utilisant que des mots du dictionnaire et obéissant à la syntaxe) sera bien découpée.

C'est ainsi qu'une phrase difficilement reconnaissable à l'oreille est reconnue par le programme.

Ex. : "ces six scies scient ces six cyprès".

Il reste cependant de nombreux cas d'ambiguités que l'on peut classer selon les faiblesses du modèle :

1. - l'absence de traitement des accords

. de genre, nombre, personne,
ex. : "nous avons fais un jolis voie y à je"
masc fém

. de mode,
ex. : "ont peut dire que le fond de l'air est frais"
↘emploi du mode indicatif au lieu de participe passé.
"ce col être haut étroit"
↘ verbe principal à l'infinitif.

2. - le manque de précision des catégories grammaticales

Le cas le plus frappant est l'absence des catégories "pronom personnel", ceux-ci appartenant à la catégorie nom propre qui paraît être la plus similaire.

Ex. : "dans les temps en sien" sien est substantif
en sien devient complément prépositionnel nominal.

"mon chat est à c'y depuis une heure"
c' nom propre
y adverbe

"la chat c'est ouverte" c' nom propre opposé à chat

3. - l'ambition de la grammaire

En effet des règles ont été ajoutées pour permettre la reconnaissance d'émissions de voix ne correspondant plus à des phrases mais à des groupes de mots c'est-à-dire à des non terminaux de la grammaire.

Ceci provoque des ambiguités.

Ex. : "ont peut dire que le fond de l'air est frais"
"es là les cheveux épais"

Ceci vient de la règle

5 → PRED qui se propose de reconnaître les prédicats seuls.

4. - l'absence de traitement sémantique

Ex. : "dans les temps anciens" et "dans l'étang ancien".

TRAITEMENT DES ACCORDS.

Celui-ci justifierait à lui seul un exposé. Nous en donnons ici les grandes lignes.

Des attributs sont donnés à chaque noeud au cours de l'analyse syntaxique. Ce sont des paramètres symboliques qui ne prennent leur valeur effective qu'une fois l'analyse syntaxique achevée.

A chaque production syntaxique sont associées des règles décrivant les manipulations à effectuer sur les attributs des éléments intervenant dans cette production.

La règle initiale est complétée par des règles du type :

$X_o \longrightarrow X_1 X_2 \ldots\ldots X_h \qquad A(X_i) \longrightarrow A(X_j)$

$A(X_l)\ R \quad A(X_i) \quad A(X_j)$

une règle faisant intervenir X_i, X_l, X_j étant activée au niveau de la rencontre de X_j (si $j > i$, l) dans le diagramme de X.

Cette méthode impose la construction et la conservation en mémoire d'une certaine partie de l'arbre syntaxique.

Il ne faut pas perdre de vue que les règles d'orthographe ne sont pas toujours suffisantes pour définir les accords :

Ex. : l'homme et la femme qui doivent venir ...
l'homme et la femme qui doit l'accompagner ...
une promenade dans les bois qui finit sous la pluie ...
---------- qui sentent bon le printemps ...

Faut-il avoir recours à la sémantique ou "redescendre" au niveau phonétique ?

Il se pose un problème d'équilibre déjà rencontré à d'autres niveaux :

- utiliser des règles très strictes au risque de rejeter la phrase correcte,

- utiliser des règles plus souples au risque d'accepter des phrases incorrectes, et de surcharger les autres niveaux d'analyse.

D'un point de vue instrumental le traitement de l'orthographe est réalisé par un ensemble d'actions "sémantiques", au sens des langages de programmation (3).

ACCELERATION DE L'ANALYSE.

Le programme initial a été modifié de manière à poursuivre l'analyse après le premier découpage correct. Un paramètre donne le nombre de découpages essayés. Ne sont plus considérés comme différents les découpages où figurent dans une même catégorie plusieurs désinences homophones.

Exemple :

les six pag{es / e} mang{ent / es / e} la bonn{es / e} confitur{es / e} dans la cuisin{es / e}

ainsi le traitement a pu être amélioré de manière sensible, par une recherche systématique des mots grammaticalement ou phonétiquement différents (le temps d'analyse est réduit d'un rapport qui varie de 1 à 6 pour les phrases examinées).

De plus, mettant en évidence des découpages différents syntaxiquement, il a permis de mieux situer les faiblesses de mieux situer les faiblesses de la grammaire initiale.

Exemples :

1er exemple : (i) i est le n° du découpage

(1) ces outils servait à la moissons du grains dans les temps anciens

(2) ces outils servait à la moissons du grains dans les temps en sien
ceci est accepté car "sien" est dans la catégorie " "substantif"

(3) ces outils servait à la moisson du grains dans l'étangs anciens

(4) ces outils servait à la moissons du grains dans l'étangs en sien.

2ème exemple :

(1) ces six scies scient ces six cyprès

(2) c'et six scies s'et six scies prêts
ceci est accepté car c' est dans la catégorie "nom propre"
Il y a là en fait 8 découpages par combinaison des diverses possibilités.

3ème exemple :

(1) nous avons fais un jolis voyages

(2) nous avons fais un jolis voies y à j'
"y" est considéré comme "adverbe" et "j" comme "nom propre".

Introduction d'une méthode d'analyse parallèle.

Découpage et analyse syntaxique traités séparément dans le programme initial ont été rendus simultanés.

Le principe général est le suivant :

. recherche de tous les mots pouvant débuter la chaîne de phonèmes ; d'où une liste de niveau 1.

. lancement de l'analyse jusqu'à la reconnaissance syntaxique d'un mot de cette liste ; ce mot devient la tête de liste.

. recherche de tous les mots pouvant suivre la tête de liste de niveau 1 ; d'où une liste de niveau 2.

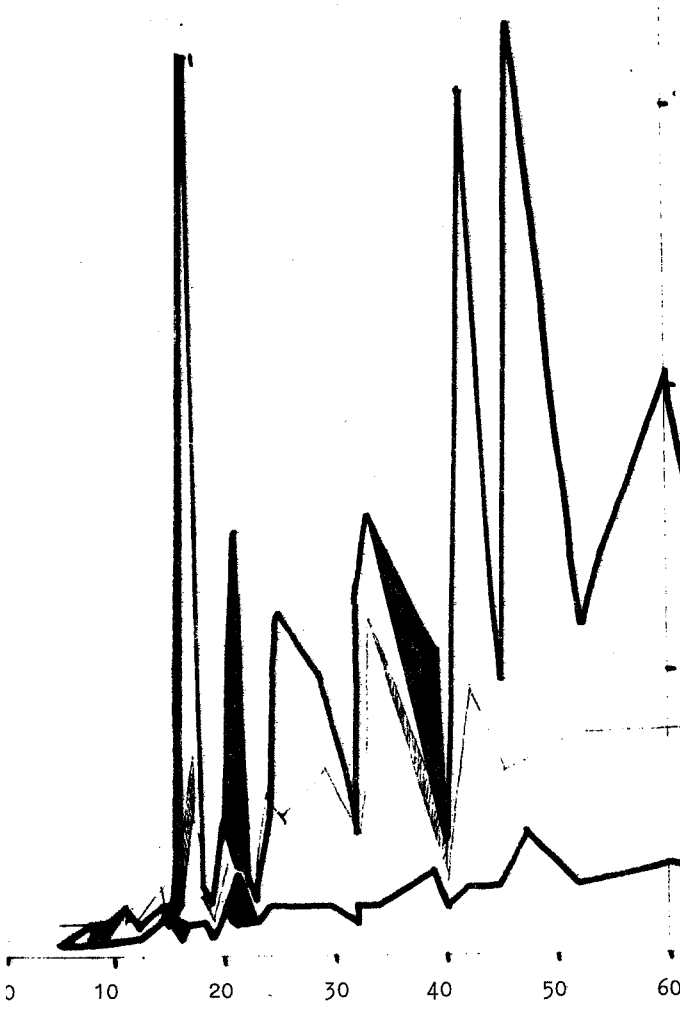
. poursuite de l'analyse pour tenter de reconnaître un mot de la liste de niveau 2 ... etc.

Cette méthode essaie de gagner du temps de découpage et du temps d'analyse en préservant si possible une partie de l'arbre syntaxique après un échec.

L'algorithme assez complexe a conduit à une restructuration des éléments manipulés. On retrouve les mêmes découpages corrects pour chaque chaîne de phonèmes mais ceux-ci n'apparaissent plus dans le même ordre ; en effet dans cet algorithme le premier mot sélectionné dans une liste est celui qui donne le chemin syntaxique le plus direct.

Dans l'algorithme initial le critère de choix reposait uniquement sur la longueur du mot.

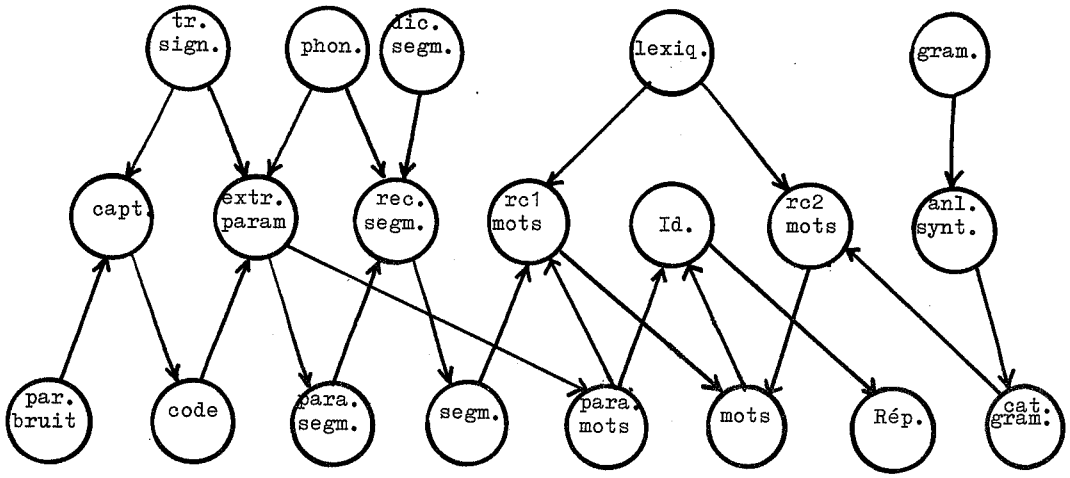
Le temps de traitement est fortement diminué comme le montre le graphique joint qui porte sur 40 phrases de 5 à 62 phonèmes.



Variations des temps de traitement du programme initial et du programme dirigé par la syntaxe.
(sur CII 10070 programmation en FORTRAN IV)

APPLICATION A LA RECONNAISSANCE VOCALE.

Le branchement d'un programme d'analyse syntaxique sur un analyseur phonétique peut se faire de multiples façons. A titre d'exemple nous en donnons une où la liaison logique entre les deux analyses est réduite au minimum : au niveau du choix des mots ...



au point où l'analyse acoustique et phonétique montante rencontre l'analyse syntaxique descendante.

Attention

Le schéma précédent ne représente ni un programme, ni une machine à reconnaître. Il indique simplement les échanges de valeurs que les différents traitements produisent entre objets et données manipulés. Complété de plusieurs critères d'optimisation il sert de donnée source à un générateur de programme ou de machine qui détermine l'enchaînement effectif (séquentiel ou parallèle) des traitements. Nous parlerons ultérieurement de ce générateur qui est une application de la CAO à la reconnaissance des formes.

CONCLUSION PROVISOIRE.

De cette étude fragmentaire il faut à notre avis retenir :

- l'échec des approches statistiques (dont une basée sur une statistique de la longueur des mots, non mentionnée),

- le succès de l'approche déterministe qui impose évidemment des contraintes sévères à la syntaxe.

La séparation entre les deux approches est sans doute due à la taille relativement importante du vocabulaire utilisé.

- le fait que l'analyse syntaxique soit descendante, c'est-à-dire qu' elle contrôle la machine,

- la mise en oeuvre d'un mécanisme parallèle pour atteindre le temps réel et disposer à chaque instant du faisceau des analyses possibles.

Il nous reste à la généraliser en considérant :

- liaisons, élisions, traitement du e muet,

- une transcription objective réelle,

- une grammaire qui permette une meilleure approximation du langage parlé et non plus écrit,

- l'introduction d'une ou plusieurs analyses sémantiques pour lever les ambiguités résiduelles ou prendre le contrôle du reconnaisseur. Ceci peut d'ailleurs nous amener à reformuler le problème de la reconnaissance automatique et abandonner la filière de la transcription intégrale.

Les deux modèles étudiés permettent une estimation de l'efficacité des contraintes syntaxiques.

Le domaine d'emploi du programme présenté dépasse de loin celui de la reconnaissance vocale. Nous envisageons son application à la correction automatique des fautes d'orthographe dans des textes divers (notamment des programmes), au passage automatique d'une sténotypie manuelle à un texte clair, ...

BERDER, le 18 Mai 1973

BIBLIOGRAPHIE

(1) D.R. REDDY, A.E. ROBINSON Phoneme to grapheme translation of english, IEEE Trans. on audio electroacoustics, Vol. AU 16-2, Juillet 1968, p. 240-246.

(2) M. ICHBIAH, H. VALABREGUE, F. LEVERY Analyse syntaxique automatique des phrases françaises, C.R. Congrès AFCET, Septembre 1970, p. 6-15.

(3) KNUTH Semantic of context-free languages, Mathematical systems theory, vol. 2 n° 2.

mɛ ʒœ nɑ̃ lɛ pa vy

mais je ne laid pas vu

ʒɛ̃ rœ gard de partu

je est regardé par tous

vwa tu la mɛ ʒɔ̃ vɛr te ruʒ

vois-tu la maison vers tes roue je

ɛ la yn gʀɑ̃d pœʀ e ɛ lɛ tut blɑ̃ʃ

elle as une grande peur et elle est toute blanche

œ̃ la buʀœ̃ki

un laboureur qui

sœl syʀ la tɛʀ

seul sur la terre

Fig. 1 - Modèle markovien à 33 états

il a rœsy œ̃ ku syr lœ viz aʒ
il a reçu un cou surent le visage

prɔmœne vu dɑ̃ lɛ bwa
promené vous dans les bois

ʒɛ rœgarde partu
je est regardé partout

sœ kɔlɛ tro etrwa
ce collai trop étroit

ilːi sɑ̃ lynɛt
il lis cent lunettes

ilnɛ paa vœgl
il ne est pas aveugle

ʒœnœsɛ pa ɑ̀kɔr ekrirtrɛ bjɛ̃
je ne sait pas encore écrire très bien

lœ ʃa ɛ gnɑ̃
le chat est grand

ʒœ truv sœla ase difisil
je trouve ce la assez dix fils y le

Fig. 2 - Modèle markovien à 5 états

/e si paʒ mâʒ /a bɔn kɔ̃fityr dɑ̃ /a kyizin
les six pages manges la bonne confitures dans la cuisines
art quan su vt art adj su prev art su

/o dœ /a mer ɛ mwɛ̃ frwad
l'eau de la mer es moins froides
art su pren art su ve deg adj

/ɔm kiɛtɛ dɑ̃ /a pjɛr dikilnœ puvɛ detryirsɔ̃otobys avɛksɔ̃ arm
l'hommes qui étaient dans la pièce dits qu'ils ne pouvaient
art su qui ve prev art su vque conj nfg mod

détruire son autobus avec son arme
vt art su prev art su

ʒâfɛrm duzzɛbr dâ /a grâd kaʒ dœfɛr
j'enfermes douze zèbres dans la grande cage de fer
nom vt quan su prev art adj su pren su

/a rivjɛr
la rivière
art su

e /a fɔ̃tɛn kikul dâ /u mɔ̃tap
et la fontaines qui coules dans la montagnes
comm art su qui vi prev art su

Fig. 3 - Modèle déterministe brut

œ̃ tigr ɛ œ̃ gʀɑ̃ ʃa
un tigre est un grand chat

sœsi ɛ puʀ twa
ceci est pour toi

ʒɑ̃ ɛ mɑ̃ʒe kɛlkœfwa
j'en ai mangé quelquefois

...
les six pages mangent la bonne confiture dans la cuisine

la mɛz ɔ̃ vɛʀt a œ̃ ʒɔli twa ʀuʒ kibʀij a lɔʀizɔ̃
la maison verte a un joli toît rouge qui brille à l'horizon

ʒ œ le ɛ vy
je les ai vues

la banan kœ lœ sɑ̃ʒ a pʀiz etɛ tʀo myʀ
la banane que le singe a prise était trop mûre

Fig. 4 - Modèle déterministe avec traitement des accords

iscussion

. BELLISSANT : Je voudrais savoir comment est organisé le dictionnaire, comment
es informations sont rangées à l'intérieur et comment on y accède ?

-Y. GRESSER : Il y a séparation entre radicaux et désinences, dont la longueur
st limitée, et le codage naturel éventuellement comprimé. Ceci est un emprunt
la documentation automatique. Les radicaux sont triés, dans l'ordre, selon le
remier phonème, selon leur longueur, selon l'ordre phonémique. L'accès
e fait à l'aide de deux tables qui contiennent :
le nombre de radicaux commençant par un phonème donné et de longueur donnée,
l'indice de ligne dans le dictionnaire du premier de ces radicaux.

Un programme de tri permet de passer d'une liste de mots sous forme externe
transcriptions phonétique, orthographique, numéro de désinence, attributs) à la forme
nterne (dictionnaire et tables d'accès).

La forme interne des désinences et leur rangement sont analogues.

. ROCHE : Que se passe-t-il lorsqu'il existe un taux d'erreur non nul sur la recon-
aissance des phonèmes ?

-Y. GRESSER : Je n'en sais rien. Dans l'expérience décrite nous avons simplement
herché à voir si une syntaxe générale n'introduisait pas des ambiguïtés qui ne se-
aient pas dues aux erreurs de reconnaissance phonétique, mais uniquement à la
ontinuité de l'élocution.

Le branchement sur la reconnaissance phonétique est un autre problème, que nous
ommes d'ailleurs prêts à aborder. Comme le schéma, indiqué dans mon exposé le
ontre, ce branchement peut se faire très facilement. Il suffit de remplacer dans une
ertaine étape de recherche lexicale un test de coïncidence par un test de ressemblance.

. GUEGUEN : La question a un certain rapport avec celle de Roche. On montre que l'on
eut toujours approcher un automate stochastique par un automate fini de dimension
uffisamment grande. Donc il n'est pas étonnant qu'en prenant l'automate fini vous
yiez de meilleurs résultats. Mais lorsqu'il y aura des erreurs sur l'entrée, la conclusion ris-
que d'être inversée car l'automate fini serait de dimension extrêmement élevée.

J-Y. GRESSER : A quelques détails théoriques près (l'automate utilisé pour l'analyse syntaxique n'est ni fini, ni déterministe) je partage le doute de Guéguen. Il est légitime si l'on ne dispose pas d'une mesure de ressemblance efficace. Il semble que Vives ait récemment trouvé un indice de comparaison entre chaînes de listes de caractères extrêmement performant. Un tel indice devrait permettre de passer de l'entrée idéale à l'entrée réelle sans grande perte. Cela reste à vérifier. L'intérêt d'un modèle stochastique serait peut-être d'accélérer l'analyse. Il reste qu'en tant que modèle de départ, il est inefficace pour résoudre des ambiguïtés relativement simples.

R. CARRE : Il serait intéressant maintenant de faire des erreurs plausibles sur les phonèmes, au départ et de voir progressivement. Si certaines lois peuvent permettre de corriger ces erreurs, l'ensemble pourrait alors être placé par exemple après un détecteur de paramètres.

J-Y. GRESSER : Actuellement, dans l'équipe, Buisson est en train d'élaborer un modèle de reconnaissance, disons "phonétique". Nous l'utiliserons, dès que possible, pour faire des erreurs plausibles.

Il y a un problème important à mentionner, c'est celui de la mesure des erreurs "d'arrivée" sur l'ensemble des émissions de voix manipulées qui est grand et difficile à cerner. Comment définir le taux de reconnaissance d'une phase, comment quantifier la redondance syntaxique ? Nous avons tenté, ailleurs, une approche dont il est difficile de dire, à l'heure actuelle, si elle est bonne. Il n'y a pas de théorie satisfaisante. La présentation des résultats est, de ce fait, limitée à l'énumération de données brutes d'expériences.

F. CARTON : Le phonéticien est assez surpris que le problème des liaisons et des élisions ait été esquivé, parce qu'il aurait l'impression que les jointures seraient une aide plutôt qu'une gêne. Quelle est votre opinion ? Serait-il difficile de les prendre en considération ?

J-Y. GRESSER : Les jointures constituent certainement une aide. A ma connaissance la seule personne qui ait bien étudié liaisons, élisions en français est Schane. Il y a eu d'autres travaux mais qui restent fragmentaires. Le problème actuel est de formaliser le livre de Schane. Une tentative est en cours à Lannion, malheureusement retardée pour des raisons d'enseignement.

G. NOIZET : Le fait qu'il n'y ait pratiquement pas de diminutions du taux de reconnaissance avec l'augmentation du vocabulaire de référence, est très surprenant pour un psycholinguiste....

J-Y. GRESSER : Ce l'est également pour nous.

G. NOIZET : Avant de m'étonner, j'aimerais m'assurer que j'ai bien compris. Pouvez-vous me dire comment le contrôle est organisé pour calculer le taux de reconnaissance en fonction de l'étendue du vocabulaire ?

J-Y. GRESSER : On cherche à reconnaître une liste de mots en les comparant à des prototypes ("template") : chaque mot est confronté à la liste complète des prototypes.

Au mot objet est attaché une signification, si cette signification coïncide avec celle du prototype le plus semblable, ce mot est dit reconnu.

Les réponses multiples proviennent de ce que l'on fait pour chaque mot un classement des prototypes les plus semblables.

Il faut peut-être insister sur le fait que les mots prononcés sont les premiers du dictionnaire et que commençant par /a/ ou /b/, ils diffèrent en cela des autres mots.

J. DREYFUS-GRAF : M. Gresser a parlé de la normalisation en énergie. Est-ce une normalisation au niveau du locuteur, au niveau digital ou analogue ? Par exemple d'une manière analogue on peut compenser des variations de ± 20 dB avec des constantes de temps de l'ordre de 5 à 10 ms.

J-Y. GRESSER : Notre normalisation est assez primitive. Elle est semblable à celle que nous avions employée sur la "Machine de bureau VOCAL", avec les différences entre énergies successives ou simultanées de canal à canal. Je pense qu'il faut normaliser, mais qu'une méthode est suffisante.

J. DREYFUS-GRAF : Vous ne pensez pas qu'en ayant un prétraitement par compensation automatique de l'énergie à $\pm$ 20 dB près, on pourrait faciliter la reconnaissance ?

J-Y. GRESSER : C'est possible, j'ai tout de même peur à la lumière d'autres expériences sur la segmentation phonétique effectuées par Mercier, qu'une normalisation prématurée entraîne une perte d'information : actuellement notre segmentation est presqu'exclusivement basée sur des critères énergétiques et rythmiques.

J. DREYFUS-GRAF : Mais on peut séparer les informations de la normalisation.

J-Y. GRESSER : Je ne sais pas quelle information il faut garder.

J. DREYFUS-GRAF : On pourrait avoir la variation d'énergie globale d'un côté et de l'autre on pourrait normaliser le spectre de fréquences.

J-Y. GRESSER : Je ne pense pas que cela soit si simple. Les interactions entre spectre, durée, énergie, contexte phonétique me paraissent complexes. Le programme de segmentation de Mercier est bâti en 3 ou 4 phases. Chaque phase fait appel à environ 10 paramètres, l'énergie y apparaît plusieurs fois sous forme pertinente et redondante. Dans l'état actuel il nous est difficile d'effectuer une séparation quelconque.

M. CARTIER : Comment ont été choisis les six mots essayés, et les mots de référence ?

J-Y. GRESSER : Les six mots étaient toujours les mêmes, je les ai cités, commençant par la lettre A. Les mots de référence étaient situés au début du dictionnaire.

M. CARTIER : Les résultats ne seraient-ils pas différents si tu faisais un tirage aléatoire pour les mots de référence?

J-Y. GRESSER : Si, c'est évident, mais cela nous coûterait encore plus cher, dans une expérience déjà coûteuse.

Je pense que la liste acheter, adroit, affiche.....n'est pas un vocabulaire optimal. Des résultats obtenus avec d'autres mots ne pourraient que nous surprendre agréablement.

Il est intéressant de noter que c'est la première fois que l'on se pose la question du choix du vocabulaire. Jusqu'ici peu de chercheurs s'en sont préoccupés. Le fait d'employer des vocabulaires de plusieurs centaines de mots sinon plusieurs milliers nous permet de nous interroger sur la dépendance des résultats selon la nature du vocabulaire. Mais quelle est cette nature ? Comment la caractériser ?

M. ROSSI : Il n'est pas étonnant que cela ne se dégrade pas étant donné que les mots commencent tous par la voyelle /a/.

J-Y. GRESSER : C'est tout de même un résultat inattendu en reconnaissance des formes.

M. ROSSI : Mais quand tu étends ton vocabulaire, il y a une forte probabilité pour que les 200 ou 300 mots suivants commencent par une consonne très différente de la voyelle initiale /a/.

J-Y. GRESSER : Bien sûr.

J. GUIBERT : Les résultats sont bons pour les mots commençant par A, le fait qu'il y ait peu de dégradation avec les mots de la seconde tranche qui commencent aussi par B est quand même étonnant.

J-Y. GRESSER : Le fait est que la méthode permet de mettre en évidence ce genre de chose, contrairement à d'autres.

R. DE MORI : J'aimerais poser une question sur la procédure expérimentale. Combien de prototypes avais-tu pour chaque mot ?

J-Y. GRESSER : Un seul.

R. DE MORI : Un seul ? Et comment était-il représenté en mémoire ?

J-Y. GRESSER : Par le codage intégral des "échantillons-vocoder".

R. DE MORI : Y avait-il une normalisation temporelle ?

J-Y. GRESSER : Non.

R. DE MORI : Etcomment ont été faits les essais ?

J- Y. GRESSER : On a enregistré une première liste de 1200 mots prononcés une fois par Mercier. Cette liste a été prise comme référence. D'autres listes plus restreintes ont été enregistrées pour les essais qui ont eu lieu en temps différé.

CODES PHONETIQUES (PHONOCODES) ET REGLES LINGUISTIQUES

Résumé.

Objectifs de la parole naturelle et artificielle. Langues littéraires et techniques. Vocabulaires illimités et restreints : listes ouvertes et fermées. Langages articulés et non articulés. Segmentation humaine et automatique de la chaîne parlée. La loi du maximum de vitesse et du minimum d'effort dans les communications. Apprentissage, redondance et correction d'erreur.

Systèmes logiques de règles phonétiques et linguistiques permettant la définition de codes phonétiques (phonocodes). Choix possibles parmi les classes internationales de phonèmes (voyelles et consonnes) et de silences, avec les symboles graphiques correspondants. Arbre phonémique et distances vectorielles.

Exemples de règles pour la formation et la segmentation de syllabes, de mots ou de phrases représentant des nombres, des opérations ou des instructions.
1) Chaque chiffre élémentaire se compose d'une seule voyelle, pouvant être précédée d'une seule consonne. 2) Chaque fin de voyelle segmente ainsi la chaîne parlée. 3) Une suite de 2 consonnes signifie un mot-instruction, valable jusqu'à l'apparition d'un autre mot-instruction. 4) Chaque opérateur arithmétique ou logique se termine par une consonne suivie d'un silence. 5) Une syllabe est agglutinable si elle ne contient aucune suite de 2 consonnes.

Exemples de phonocodes (basés sur 3 à 12 classes de phonèmes) traduisant un système tel que : binaire, ternaire, quaternaire, octal, décimal, bi-octal, alphabétique, alphanumérique. Vitesses d'élocution : jusqu'à 6 mots élémentaires par seconde. Rangements progressifs et dégressifs. Correspondances entre sons et couleurs.

Optimisation et taux d'erreur : programme statistique pour tester les divers codes phonétiques. Séparation des bruits ou parasites. Tolérances d'articulation. Tests de rimes et d'anti-rimes. Mesures automatiques par appareils de reconnaissance de la parole naturelle ou codée. Redondance et contrôle.

Applications techniques : commandes verbales de téléphones, téléscripteurs, ordinateurs, véhicules, serrures, dispositifs d'alarme, de chiffrage, de machines diverses.

Applications humaines : langages logiques, psychologie de la mémorisation audio-visuelle, assistance au raisonnement et au calcul mental.

Summary

Aims of natural and artificial speech. Literary and technical languages. Unlimited and restricted vocabularies. Articulated and non-articulated languages. Human and automatic segmentation of connected speech. Law of maximum speed and minimum effort. Learning, redundancy and error correction.

Logical systems of phonetic and linguistic rules allowing the definition of phonetic codes (phonocodes). Possible choices among international classes of phonemes (vowels and consonants) and silences, with corresponding graphic symbols. Phonemic tree and vectorial distances.

Examples of rules for forming and segmenting syllables, words and sentences representing numbers, operations or instructions.
1) each elementary digit is composed of a single vowel, which may be preceded by a single consonant. 2) Each vowel-ending serves, therefore, to segment uninterrupted speech. 3) A sequence of 2 consonants signifies an instruction-word, which remains valid until the occurrence of another instruction-word. 4) Each arithmetic or logic operator ends in a consonant followed by a silence. 5) A syllabe is "agglutinative" if it does not contain a sequence of 2 consonants.

Examples of phonocodes (based on 3 to 12 phoneme classes) translating a system which may be : binary, ternary, quaternary, octal, decimal, alphabetic, alpha-numeric. Pronunciation speeds : up to 6 elementary words per second. Progressive and degressive sequences. Relationships between sounds and colours.

Optimization and error-rates : statistical programme for testing various phonetic codes. Separation of noise and interference. Articulation tolerances. Rhyme and anti-rhyme tests. Automatic measurement by apparatus able to recognise natural or artificial speech. Redundancy and control.

Technical applications : for giving verbal instructions to telephones, teleprinters, computers, vehicles, locks and various other apparatus.

Human applications : logical languages, psychology of audio-visual memorization, aids to reasoning and mental calculation.

J. DREYFUS-GRAF
Genève.

Codes phonétiques (phonocodes) et règles linguistiques

par Jean DREYFUS-GRAF, Genève (Suisse)

1. Objectifs de la parole naturelle et artificielle

Une langue naturelle, telle que le français ou l'anglais, a pour objectif d'exprimer et de transmettre la variété illimitée des pensées humaines avec un maximum de vitesse et un minimum d'effort. Grâce aux lois de la combinatoire, elle peut être transcrite à l'aide d'une trentaine d'éléments phonétiques seulement, nommés phonèmes. Avec f phonèmes au total et k phonèmes par mot, on peut concevoir $m_k = f^k$ mots théoriques ou arrangements avec répétition /1/.

Par exemple, selon le tableau de la Fig. 1, si f = 32 phonèmes du total, on peut concevoir 1000 mots à 2 phonèmes, nommés "diphones", ensuite 32 000 "triphones", 1 million de "tétraphones", 32 millions de "pentaphones", 1 milliard "d'hexaphones", 32 milliard "d'heptaphones", 1000 milliards "d'octophones" et ainsi de suite. Mais la puissance de la combinatoire phonémique est encore bien plus formidable. En effet, si un mot est défini comme un enchaînement de phonèmes entre 2 silences, une phrase contenant plusieurs mots prononcés d'une traite devient un "hyper-mot". Ainsi l'enchaînement de 50 phonèmes permettrait la formation d'un milliard de milliards de ces hyper-mots (puisque $32^{50} = 2^{250} \backsimeq 10^{82} \backsimeq (10^9)^9$.

La moitié environ des nombres de mots mentionnés serait pratiquement articulable par la bouche humaine, grâce à une répartition équilibrée des voyelles v et des consonnes c. Mais au cours des millénaires de son développement, l'humanité les a sélectionnés empiriquement, en ne conservant que les arrangements qui semblaient les plus pratiques à prononcer et à mémoriser, de sorte que le dictionnaire d'une langue naturelle ne comporte actuellement que quelque 100 000 mots. Chaque enfant refait un chemin linguistique analogue, en raccourci, pendant les premières années de sa vie /2/.

Le tableau de la Fig. 2 montre une statistique des longueurs de mots dans la langue française, et qui est basé sur 2 textes de culture générale /3/ /4/.

Les voyelles et les consonnes se répartissent environ à égalité (48 % resp. 52 %) en français, tandis que les Anglo-saxons utilisent 2 fois plus de consonnes que de voyelles.

En appliquant les données statistiques de l'étude 2) à un vocabulaire de 25 000 mots différents, par exemple, on obtient la répartition suivante : 12 monophones (voyelles), 688 diphones, 7800 triphones, 4800 tétraphones, 4200 pentaphones, 2800 hexaphones, 2100 heptaphones, 1200 octophones et 1400 "multiphones" de 9 à 14 phonèmes.

On estime le vocabulaire de Shakespeare à 24 000 mots et celui d'un Européen moyen à 12 000 mots. L'enfant utilise 3 600 mots à l'âge de 8 ans, et 9600 mots à 14 ans /5/ .

On constate que les langues naturelles emploient moins de la millionième partie des mots pratiquement articulables qu'on peut concevoir avec 32 phonèmes. Ce gaspillage gigantesque, qui obéit d'ailleurs à une loi générale de la nature, n'est justifié que partiellement par la redondance, destinée à diminuer les taux d'erreurs.

Le cerveau de l'homme alphabétisé est capable de reconnaître rapidement chacun des 32 phonèmes à l'intérieur de chaque mot, et chacun des milliers de mots à l'intérieur des millions de phrases de la chaîne parlée. Par contre la machine ne parvient que très difficilement à opérer ces segmentations, même dans le cas de vocabulaires réduits à une dizaine de mots prononcés par une dizaine de personnes /6/. Le problème numéro un de la reconnaissance de la parole se nomme donc "segmentation". On a proposé divers procédés pour tourner la difficulté. En voici des exemples :

a) Le locuteur ralentit sa prononciation, il accentue chaque phonème et sépare les mots par des silences /7/. b) la machine subit un "apprentissage" préliminaire. Elle enregistre préalablement et plusieurs fois dans sa mémoire chaque mot prononcé par chaque locuteur /8/. c) la machine enregistre du moins une phrase-clef normalisée /12/.

Ces procédés ralentissent inutilement les opérations lorsqu'il s'agit de commander verbalement une machine par une langue numérique, telle que celle du système décimal. En effet, celui-ci est basé sur dix mots, zéro à neuf, seulement, tout en restant capable d'enchaînements illimités.

Il n'est donc pas rationnel de conserver les 32 phonèmes d'une langue naturelle, dont la vocation est littéraire. Nous allons examiner diverses possibilités de construire des langues artificielles conciliant le minimum de phonèmes avec le maximum de vitesse d'articulation et de mémorisation. Fondamentalement, les machines sont logiques, alors que les langues naturelles ne le sont pas. C'est de là que proviennent les difficultés essentielles de leur reconnaissance automatique. Aussi allons-nous proposer des systèmes logiques de règles linguistiques.

2. Hiérarchie des classes internationales de phonèmes

La plupart des langues naturelles de la terre peuvent être transcrites par 93 phonèmes, au total /9/ dont 28 voyelles et 65 consonnes ou semi-voyelles. L'arbre hiérarchique de la Fig. 3 distingue au maximum 12 classes internationales de phonèmes F, dont 6 classes de voyelles V (O,I,A,E,U,Y) et 6 classes de consonnes C(T,S,N,P,D,T) groupées en "plosives et non-plosives". Ces classes présentent des caractères distinctifs généraux indiqués par les flèches de direction "basses-hautes" et qui concernent les compositions dominantes de leurs spectres de fréquence. L'aspect technique de ces dichotomies successives réside dans les analyses de dynamique, de spectres et de durées de phonèmes, déjà décrites dans nos publications antérieures /10-15/. L'aspect phonétique en est résumé par l'arbre de la Fig. 3 : les écartements de ses branches symbolisent les distances vectorielles entre les classes de phonèmes. Celles-ci sont d'autant plus distinctes que leurs numéros hiérarchiques se suivent de plus près. La valeur phonétique centrale de chaque classe de phonèmes est indiquée par une lettre minuscule, son symbole graphique par une majuscule unique de l'alphabet latin. Ainsi la lettre H est le symbole du phonème ʃ = "ch" en français, "sh" en anglais, ou "sch" en allemand. Quelques variantes sont suggérées. Il appartiendra à des statistiques expérimentales de choisir finalement les valeurs phonétiques et les symboles graphiques les plus favorables.

En associant les classes de phonèmes à des couleurs, on constate d'emblée qu'elles sont d'autant plus discernables que leur nombre est plus petit. Ainsi les 6 couleurs "rouge, jaune, vert, bleu, brun, violet" (représentantO,A,I,S,T,N) prêtent bien moins à confusion que les 12 nuances "rouge, orange, jaune, olive, jade, vert, bleu-ciel, cobalt, brun, pourpre, lilas, violet" (représentantO,U,A,E,Y,I,S,H,N,P,D,T). On a donc intérêt à rechercher le minimum de classes de phonèmes compatibles avec le nombre de mots différents à représenter. D'autre part, la vitesse d'articulation sera d'autant plus grande que les mots seront mieux agglutinables, c'est-à-dire enchaînables, grâce à une distribution équilibrée des voyelles V et des consonnes C.

3. Mots élémentaires et non élémentaires

Dans la conception d'un phonocode, nous distinguerons 2 espèces de mots : d'une part, les mots élémentaires, indéfiniment agglutinables, qui expriment les chiffres de base, et d'autre part les mots non élémentaires, segmentés par des silences (plus grands que 2/10 seconde), et qui représentent divers genres d'instruction /12-15/.

Le tableau de la Fig. 4 montre les nombres d'arrangements possibles quand on dispose de 3 à 12 phonèmes au total, et de 1 à 4 phonèmes par mot. La moitié environ de ces arrangements serait pratiquement prononçable.

.../...

Pour que la machine puisse segmenter les mots élémentaires agglutinés il faut lui fournir une règle logique. Par exemple : chaque mot élémentaire est formé par une seule voyelle, pouvant être précédée, mais jamais suivie, par une seule consonne. Dans ce cas, la machine saura que chaque fin de voyelle est un signal de segmentation du mot élémentaire. Disposant de v voyelles et de c consonnes, on peut former $m_e = v+v.c$ mots élémentaires. Inversement, si nous désirons m_e mots élémentaires et que le nombre v des voyelles n'est pas inférieur à celui des consonnes ($v \geqslant c$), la formule suivante fournit le nombre (entier) minimum de phonèmes nécessaires. $f_{min} = 2(m_e + 1/4)^{1/2} - 1$. Par exemple, le système décimal exigera au moins 3 ou 4 voyelles et 3 ou 2 consonnes pour coder les 10 mots élémentaires zéro à neuf. Un système alpha-numérique, comprenant 10 chiffres et 26 lettres, soit 36 mots élémentaires, demandera au moins 6 voyelles et 5 consonnes.

Voyons maintenant quels sont les nombres approximatifs d'arrangements possibles avec les 6 phonèmes nécessaires au système décimal ? D'abord 36 diphonèmes ($m_2 = 6^2$), ensuite 200 triphones ($m_3 = 6^3$), 1200 tétraphones ($m_4 = 6^4$), 7000 pentaphones, 46 000 hexaphones, et ainsi de suite dont la moitié environ serait pratiquement prononçable.

Ainsi, avec 6 phonèmes seulement, le phonocode SOTINA admettrait déjà un nombre de mots très supérieur à celui du vocabulaire usuel d'une langue naturelle, qui ne dépasse guère 12 000 mots, malgré ses 32 phonèmes.

La Fig. 4 résume aussi les capacités de divers autres phonocodes, basés sur 3 à 12 phonèmes, et dont les noms font apparaître la répartition des voyelles et consonnes : TOS, OTI, OSI, SOTI, OTISA, OTISAE, SOTINA, ESOTINA, ESOTINAU, SOTINAPE, ESOTINAHU, SOTINAPERU, YSOTINAPERU, SOTINAPERUDY. En variante, O peut parfois être remplacé par U, I par E (=e), E par E (=ɛ), P ou R ou D par H (=ʃ). etc... De plus, SOTINA peut se prononcer /sotina/ ou /ʃokina/ ou /supena/, etc...

4. Liste des mots élémentaires de quelques phonocodes

Le tableau de la Fig. 5 annexée montre 10 listes de 32 premiers mots élémentaires qui permettent de coder les nombres 0 à 31 et les 26 lettres de l'alphabet pour un système tel que binaire, ternaire, quaternaire, octal, décimal, bi-octal, alphabétique, alpha-numérique. Les étoiles régulières, allant de 4 à 12 branches qui correspondent aux 4 à 12 phonèmes, montrent géométriquement la répartition des voyelles (branches noircies) et des consonnes, concernant les divers phonocodes, depuis SOTI jusqu'à SOTINAPERUDY. Ce dernier, avec ses 12 phonèmes, est capable d'exprimer directement les 36 mots élémentaires d'un système alpha-numérique complet.

Pour les autres codes, dont les nombres de phonèmes sont insuffisants, il faut que 2 mots d'instruction, tels que STI et STA, informent la machine que les mots élémentaires concernent des nombres ou des lettres.

5. Vitesses et durées d'élocution

Voyons maintenant comment se présentent les vitesses et durées d'élocution de phrases phonocodées ? Tandis que la vitesse d'élocution de la parole naturelle ne dépasse guère 3 syllabes par seconde, la phonocode permet de prononcer jusqu'à 6 mots élémentaires par seconde, c'est-à-dire le double. Toutefois la longueur, et par conséquent la durée d'élocution des mots phonocodés sera d'autant plus réduite que le nombre des classes de phonèmes différentes sera plus grand. Le tableau de la Fig. 6 montre les mots "75-Paris" exprimés successivement par divers phonocodes. On constate que la durée d'élocution est 6 secondes avec le système binaire OTI, 3, 1 secondes avec le système décimal SOTINA, et 1,2 seconde avec le système alpha-numérique SOTINAPERUDY. Des études ultérieures devront rechercher les codes les mieux adaptés à divers cas pratiques.

.../...

6. Exemple d'un système logique de règles linguistiques

Après avoir esquissé les règles phonétiques qui sont à la base des phonocodes nous allons les intégrer comme suit dans un système logique et cohérent de règles linguistiques, avec quelques exemples relatifs aux codes SOTINA III, OTISAE I, et ESOTINA I.

Règle logique N° 1, mots élémentaires (indices de segmentation)

Les chiffres 0 à 9 sont traduits par 10 "mots élémentaires", qui sont agglutinables. C'est-à-dire qu'ils se terminent par un "indice logique de segmentation" permettant leur séparation automatique. Cet indice peut être la fin de chaque voyelle. Chaque mot élémentaire est alors constitué par une seule voyelle, pouvant être précédée (mais jamais suivie) par une seule consonne.

Par conséquent chaque nombre (de zéro à l'infini) peut être traduit par une phrase prononcée d'une traite : la segmentation des chiffres constituant un nombre s'opère indépendamment de pauses. La fin d'un nombre complet, ou mot composé , est signalée par un silence plus long que 0,2 seconde, symbolisé par un espace ou un point (.).

Si nous disposons de v voyelles et de c consonnes, soit f = v + c phonèmes, nous pouvons former $m_o = v + c.v$ mots élémentaires. Exemples :

Chiffres	0	(0')	1	2	3	4	5	6	7	8	9	10	100	1 9 7 3
phonocode														
SOTINA	0	(NO)	I	TO	TI	TA	SO	SI	SA	NI	NA	IO	IOO	INASATI
OTISAE	0	(NO)	I	TI	TE	TA	TO	SI	SE	SA	SO	IO	IOO	ISOSETE
ESOTINA	TO	(NO)	TI	TE	TA	SA	SE	SI	SO	NI	NA	TITO	TITOTO	TINASOTA

Règle de logique N° 2, zéros explicites et implicites

Le zéro explicite (0) se distingue du zéro implicite (0') par 2 mots élémentaires différents, tels que 0 et NO, ou TO et NO. Ceci permet d'exprimer des dizaines, centaines, milliers, etc... Exemples :

nombre énoncé usuel	1 5 3 un-cinq-trois	1 5 3 = 100+50+3 cent-cinquante-trois	6 0 9 six-zéro-neuf	6 0 9 = 600+9 six-cent-neuf
phonocode				
SOTINA	ISOTI	INONOSONOTI	SIONA	SINONONA
OTISAE	ITOTE	INONOTONOTE	SIOSO	SINONOSO
ESOTINA	TISETA	TINONOSENOTA	SITONA	SINONONA

Règle logique n° 3, fractions décimales

Le"zéro implicite" (NO) désigne une fraction décimale (zéro virgule) quand il se trouve en position initiale. Ainsi 2/100 = 0,06 = NOOSI (ou NONOSI)

Règle logique N° 4, mot anti-répétitif

Un mot élémentaire "anti-répétitif" évite de répéter le même phonème plus de 2 fois. Ce mot peut être la voyelle A (=avec), soit initiale, soit après une autre voyelle. Ainsi, 1 million = 1 000 000 = 10^6 = un-avec-six-zéros = IASIO. Exemples :

.../...

nombre énoncé	7 000 sept-mille	1.973 mille-neuf-cent septante-trois	10^3 kilo	10^6 méga	10^9 giga	10^{-12} téra	10^{-6} micro
phonocode							
SOTINA	SAATIO	IATINONANONOSANOTI	TINO	SINO	IANAO=NANO	ITONO	NOIASIO=NOSINO
OTISAE	SEATEO	IATENONANONCSENOTE	TENO	SINO	IASOO=SONO.	ITINO.	NOIASIO=NOSINO.
ESOTINA	SOATAO	TIATENOSONONOSONOTA	TANO	SINO	TIANAO=NANO	TITENO	NOTIASIO=NOSINO

Règle logique N° 5, opérateurs arithmétiques ou logiques

Un opérateur arithmétique ou logique est un di- ou triphone se terminant par une consonne ("voyelle + consonne", ou "consonne + voyelle + consonne"). La consonne finale est suivie d'un silence plus long que 0,2 seconde, et symbolisé par un point (.) ou par un espace. Contrairement au mot élémentaire, l'opérateur se termine toujours par une consonne interdisant la segmentation automatique. Exemples :

Symbole	=	<	>	+	x	exp.	-	:	log	(	)	,
Phonocode												
SOTINA	OS.	IS.	AS.	TOS.	TIS.	TAS.	SOS.	SIS.	SAS.	NOS.	NIS.	NAS.
OTISAE	OS.	IS.	AS.	TOS.	TIS.	TAS.	SOS.	SIS.	SAS.	ES.	SES.	TES.
ESOTINA	au choix comme ci-dessus											

Règle logique N° 6, mots - instruction

Un mot-instruction contient 2 consonnes adjacentes (ST, SN, etc...) et 1 voyelle au moins. Il peut se terminer au choix par une consonne ou par une voyelle, suivie d'un silence plus long que 0,2 seconde. Il peut se présenter sous la forme "consonne + consonne + voyelle", ou "voyelle+cons.+cons.+voyelle" etc... Exemples :

énoncé	cardinal	ordinal	alphabet.	stop	start	effacer	no-tél.	go-to	if
phonocode									
SOTINA	STI.	STO.	STA.	STOS.	STAS.	SNIS.	ASNO.	OSTO.	ISTO
OTISAE	STI.	STO.	STE.	STOS.	STES.	STIS	ESTO.	OSTO.	ISTO
ESOTINA	au choix comme ci-dessus								

Ainsi, après "STA" les nombres 1 à 26 signifient les lettres alphabétiques "a" jusqu'à "z", par exemple. Un mot-instruction est valable jusqu'au suivant.

Selon les besoins, on peut adjoindre à ces exemples bien d'autres mots ou règles. Avec le phonocode SOTINA, le nombre des mots peut aller jusqu'à 100 triphones, 600 tétraphones, 3500 pentaphones ou 23 000 hexaphones, c'est-à-dire jusqu'à la moitié aisément prononçable, environ, des arrangements de 6 phonèmes. Concernant l'adjonction de règles, il faut veiller à ce qu'elles s'enchaînent d'une manière strictement logique, car une machine rationnelle ne tolérerait aucune des contradictions qui sont l'apanage des langues naturelles. Une langue peut être considérée comme un système de signaux, dont l'optimisation dépend du but recherché, et qui est soumise à des lois physiques très générales /16/.

.../...

7. Taux d'erreur et optimisation

Un taux d'erreur convenable, quant à l'utilisation d'un phonocode, serait de l'ordre de un pour mille phonèmes, puisqu'une bonne dactylographe parvient à ne faire qu'une faute sur mille frappes. Il s'agit donc d'établir des programmes pour tester "l'intelligibilité" du côté de la machine, et la "tolérance d'articulation" du côté du locuteur. Les tests peuvent être effectués de 3 manières complémentaires : a) par des opérateurs humains, b) par des appareils de reconnaissance automatique cablés (hardware), c) par de tels appareils simulés sur ordinateurs (software).

La séparation correcte des consonnes fricatives (s, ch, f) par rapport aux consonnes plosives (p,t,k) exige un détecteur de pente d'attaque des phonèmes. Ce détecteur peut être soit câblé (hardware), tel que par exemple un régulateur d'amplitude à double-boucle /11-15/ soit simulé sur ordinateur (software). Dans ce dernier cas, la fréquence d'échantillonnage pour les plosives doit être d'au moins 300 Hz.

L'optimisation d'un phonocode dépendra de son champ d'application, et tout spécialement des bruits provenant de l'environnement ou des canaux de transmission. Un craquement peut ressembler à une consonne plosive, et un souffle respiratoire à une consonne fricative. Mais le moment de leur apparition par rapport à une voyelle sera généralement différent. Une analyse complète des énergies, fréquences et temps respectifs dirigera la séparation des parasites par rapport aux phonènes. Afin de déterminer quantitativement les classes de phonèmes et les mots les plus favorables, on pourra utiliser les tests de rimes, mais aussi d'anti-rimes, basés sur les variations individuelles de phonèmes initiaux ou finaux /17/. On peut prévoir des corrections d'erreurs par redondances, confirmations, quittances ou répétitions.

8. Applications techniques : commandes verbales de machines

Les applications techniques des phonocodes sont évidentes, puisque ceux-ci sont proposés dans le but de simplifier les appareils de reconnaissance de la parole. Ils permettront le remplacement progressif des boutons-poussoirs spécialisés par le microphone universel. La parole pourra commander directement diverses machines, telles que commutateurs téléphoniques, ordinateurs, télescripteurs, véhicules, serrures, dispositifs d'alarme, de transmissions secrètes par chiffrage, de machines-outils et autres.

9. Applications humaines : langage parlé (parole) numérique

Mais on peut aussi envisager des applications humaines des phonocodes, car ceux-ci permettront l'introduction de la parole numérique, c'est-à-dire d'un langage parlé logique et universel. Il y aura correspondance cohérente entre les chiffres écrits et parlés de même qu'entre les raisonnements mathématiques correspondants.

Actuellement, cette correspondance n'existe pas. Par exemple, le rangement de chiffres s'effectue par poids progressifs ou dégressifs selon qu'il s'agit d'opérations ithmétiques écrites ou orales : l'addition écrite s'effectue dans l'ordre "unités, dizaines, centaines, etc...", l'addition orale ou mentale se déroule en sens contraire. Les Français et les Anglais disent "vingt-sept" et twenty-seven", les Allemands prononcent dans l'ordre contraire "sieben-und-zwanzig", les Hollandais aussi.

D'autre part, il n'existe aucune représentation logique à la base des mots qui désignent des multiples, tels que "mille=kilo","méga=million", "giga= milliard", "téra=billion". Par contre, en phonocode, tel que SOTINA : "mille" peut s'exprimer par "trois-zéros = TINO" ou par "un-avec-trois-zéros = IATIO" et ainsi de suite.

.../...

Il serait intéressant d'effectuer des programmes de tests, psychologiques cette fois, avec des hommes et surtout avec des enfants, pour déterminer les phonocodes les mieux adaptés à l'assistance audio-visulle du raisonnement et du calcul mental.

Par ailleurs, puisque beaucoup de choses sont caractérisables par des nombres (une montagne par son altitude, un mois de l'année par son numéro, etc...) d'autres correspondances pourront encore s'introduire entre l'oral et l'écrit.

Finalement, les langues naturelles et les phonocodes pourraient se compléter comme des tableaux artistiques et des dessins techniques.

BIBLIOGRAPHIE

/1/ DREYFUS-GRAF, J., Actuateur phonétique (phonacteur) et calculateur, Revue d'Acoustique, Paris, n° 9 - 1970

/2/ DREYFUS-GRAF,J., La Parole humaine et l'informatique (Phonétographe V, vocographe, phonacteur, mélographe, etc...), Automatisme, Dunod, Paris, N° 9 - 1970

/3/ HATON J.P. et LAMOTTE M. Etude statistique des phonèmes et diphonèmes dans le français parlé, Revue d'Acoustique, Paris, n° 16 - 1971

/4/ TUBACH J.P., Etude des contraintes statistiques des groupements phonématiques, C.N.R.S., C.E.T.A., Grenoble, 1970

/5/ KONDRATOV A, Sons et signes, Editions Mir, Moscou 1968

/6/ DREYFUS-GRAF J. La reconnaissance automatique de la parole (Phonétographe IVa), L'Echo des Recherches, C.N.E.T., Paris et Lannion, avril 1971

/7/ DREYFUS-GRAF J. Le Phonétographe (III) : présent et futur, Bull. Techn. PTT, Berne n° 5 - 1971

/8/ GLENN J.W. et HITCHCOCK, M.H., Voice Command System, Electronics, May 10, 1971

/9/ The principles of the International Phonetic Association, Department of Phonetics, University College, London W.C. 1, 1966

/10/ DREYFUS-GRAF J. Machines obéissant à la parole, Proceedings 7th I.C.A., Budapest, 1971, conférence 20-C-15

/11/ DREYFUS-GRAF J. Speech Dynamics and Pitch, Proceedings of the Speech Symposium Szeged, 1971

/12/ DREYFUS-GRAF J. Recognition of Natural and of Artificial Speech (Phonocode) Conference H 9, Reports on the International Conference on Speech Communication and Processing, I.E.E.E.-A.F.C.R.L., Boston, 1972

/13/ DREYFUS-GRAF J. Parole codée (phonocode) : reconnaissance automatique de langages naturels et artificiels, Revue d'Acoustique, Paris, n° 21, 1972

/14/ DREYFUS-GRAF J. Reconnaissance automatique de la parole codée (phonocode), sonore et chuchotée, à paraître dans Revue d'Acoustique, juin 1973

/15/ DREYFUS-GRAF J. Brevets publiés (ou déposés) : SUISSE n°s 535.510, (12.177/71) FRANCE n°s 1.428.460, 71.07332, 72.29499, G.B. N°s 978.303 (23.111/71), (38.725/72), GERMANY n°s 1.206.167. (P 21 09 436.0), (P 22 40 557.8), CANADA n° 770.309, U.S.A. n°s 3.238.301, 3.304,369, (122.612/71), (280.723/72), JAPAN n°s 480.378, (11.566/71), (82.241/72), etc...

/16/ DREYFUS-GRAF J. Physique et Cybernétique (le monde physique comme système de signaux auto-régulés) Helvetica Physica Acta, Vol. 45,1972.

/17/ ROSSI M, Le test de diagnostic par paires minimales, I-IV, Journées d'Etudes d'Aix-en-Provence, 1971, Rapport du Groupe "Communication Parlée", G.A.L.F.

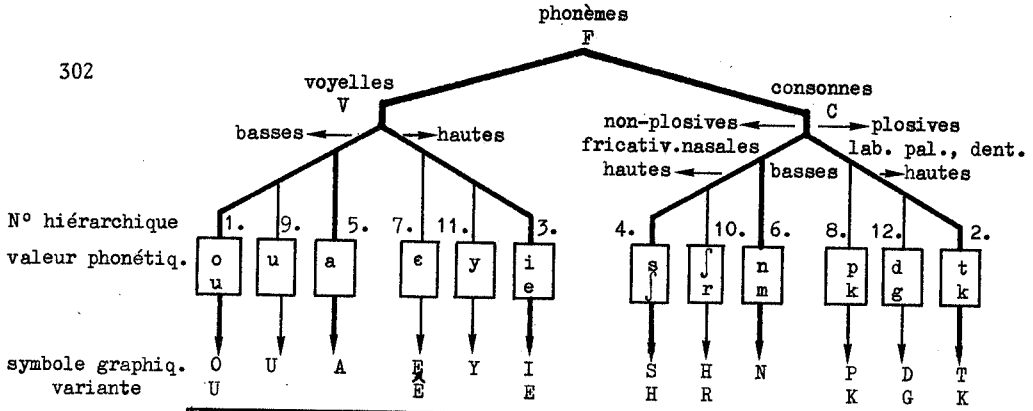
Documentation sur demande à l'auteur : J.A. DREYFUS GRAF, 5, avenue de la Genade 1207 - Genève (Suisse), Tél. (022) 36.34.32

ıonèmes k= ar mot	1	2	3	4	5	6	7	8	50
f^k =	32^1_5 = $2^5_{\simeq}$	32^2_{10} = $2^{10}_{\simeq}$	32^3_{15} = $2^{15}_{\simeq}$	32^4 = $2^{20}_{\simeq}$	32^5 = $2^{25}_{\simeq}$	32^6 = $2^{30}_{\simeq}$	32^7 = $2^{35}_{\simeq}$	32^8 = $2^{40}_{\simeq}$	32^{50} $2^{250}_{\simeq}$
ıots m_k =	32	10^3	32.10^3	10^6	32.10^6	10^9	32.10^9	10^{12}	10^{82}
	mono-phones	di-phones	tri-phones	tétra-phones	penta-phones	hexa-phones	hepta-phones	octo-phones	hyper-mots
rrangements : premier	a	aa	aaa	aaaa	aaaaa	aaaaaa	aaaaaaa	aaaaaaaa	aaa..aaa
deuxième	b	ab	aab	aaab	aaaab	aaaaab	aaaaaab	aaaaaaab	aaa..aab
.....	.	..	...	...	...	...	...	...	...
dernier	Z	ZZ	ZZZ	ZZZZ	ZZZZZ	ZZZZZZ	ZZZZZZZ	ZZZZZZZZ	ZZZ..ZZZ
équence (%) occurence français	19,3	31,5	16,3	9,4	8,4	5,8	4,1	2,4	%

g. 1 Nombre de mots $m_k=f^k$ formables avec f = 32 phonèmes au total, et k = 1,2..8 (ou 50) phonèmes par mot. Mots = arrangements de f éléments K à K avec répétition. Les 32 phonèmes sont symbolisés par les 32 lettres a,b,c,... x, y,z,U,V,W,X,Y,Z. Un mot est un enchaînement de phonèmes entre deux silences (>0,2 seconde). Environ la moitié ($m_k/2$) des mots serait commodément articulable.

honèmes par mot	1	2	3	4	5	6	7	8	9-14	mots différents, ou %=fréquence occurence
épartition dans exte parlé, dont otal phonèmes :	mono-phones (voy.)	di-phones	tri-phones	tétra-phones	penta-phones	hexa - phones	hepta phones	octo-phones	autres multi-phones	
) 50 033 = 100 %	12	615	-	-	-	-	-	-	-	mots différents
	10,73	36,13	16,87	11,61	8,74	6,77	4,21	2,40	2,46	%
)88 377 = 100 %	12	905	7420	-	-	-	-	-	-	mots différents
	19,3	31,5	16,3	9,4	8,4	5,8	4,1	2,4	2,7	%
) estimation our vocabulaire 25 000 mots	12	688	7800	4800	4200	2800	2100	1200	1400	mots différents
	0,05	2,75	31,2	19,2	16,8	11,2	8,4	4,8	5,6	%

g. 2 Répartition statistique des phonèmes dans les mots de la langue française
1) Avec texte de 50 033 phonèmes et alphabet f=33 phonèmes = 15V+18C
longueur moyenne du mot 4,4 phonèmes = 50 033 : 11 443
2) Avec texte de 88 377 phonèmes et alphabet f=35 phonèmes = 15V+17C+3SV
longueur moyenne du mot 3,21 phonèmes
3) Estimation pour un vocabulaire de 25 000 mots = 7800 triphones x 3,21
(mono- et diphones en valeurs absolues, enlevant 19,3+31,5=50,8 % de 50 000 mots).



12 nuances	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
	rouge	orenge	jaune	olive	jade	vert	bleu ciel	bleu cobalt	brun	pour-pre	lilas	viol-let

6 couleurs	1.	2.	3.	4.	5.	6.
	rouge	jaune.	vert	bleu	brun	violet

Fig. 3 Hiérarchie des classes internationales des phonèmes. Arbre phonémique et distances vectorielles. Dichotomies : voyelles/consonnes, basses/hautes, consonnes plosives/non-plosives, fricatives/nasales. Correspondance entre phonèmes et couleurs : 12 nuances et 6 couleurs principales. Voyelles claires consonnes sombres.

v	voyelles	c	consonnes	phonèmes f	mots élémentaires $m_e = v+v.c$	phonocode	applications	mots non-élémentaires: diphones $m_2=f^2$	triphones $m_3=f^3$	tét phor $m_4=f$
1	0,(I)	2	S,T	3	3	TOS,(OTI)	binaire, tern.	9	27	8
2	0, I	1	T,(S)	3	4	OTI,(OSI)	binaire, tern.	9	27	8
2	0,I	2	S,T	4	6	SOTI	tern., quatern.	16	64	25
3	0,I,A	2	S,T	5	9	OTISA	quatern.,oct.	25	125	62
4	0,I,A,E,	2	S,T	6	12	OTISAE	octal, décimal	36	216	129
3	0,I,A	3	S,T,N	6	12	SOTINA	octal, décimal	36	216	129
4	0,I,A,E	3	S,T,N	7	16	ESOTINA	octal, décimal	49	343	240
5	0,I,A,E,U	3	S,T,N	8	20	ESOTINAU	décimal,bi-oct.	64	512	409
4	0,I,A,E	4	S,T,N,P,(H)	8	20	SOTINAPE	décimal,bi-oct.	64	512	409
5	0,I,A,E,U	4	S,T,N,H	9	25	ESOTINAHU	décimal,bi-oct.	81	729	656
5	0,I,A,E,U	5	S,T,N;P,R	10	30	SOTINAPERU	décimal,bi-oct.	100	1000	1000
6	0,I,A,E,U,Y	5	S,T,N,P,R	11	36	YSOTINAPERU	alphabétique	121	1331	1464
6	0,I,A,E,U,Y	6	S,T,N,P,R,D	12	42	SOTINAPERUDY	alphanumétique	144	1728	2093

symbole graphique	0	I(E)	A	E(Ê)	U	Y	S	T	N	P(H)	R(H)	D
valeur phonétique	o(u)	i(e)	a	ɛ	u	y	s(ʃ)	t,k	n,m	p(ʃ)	r(ʃ)	d,g

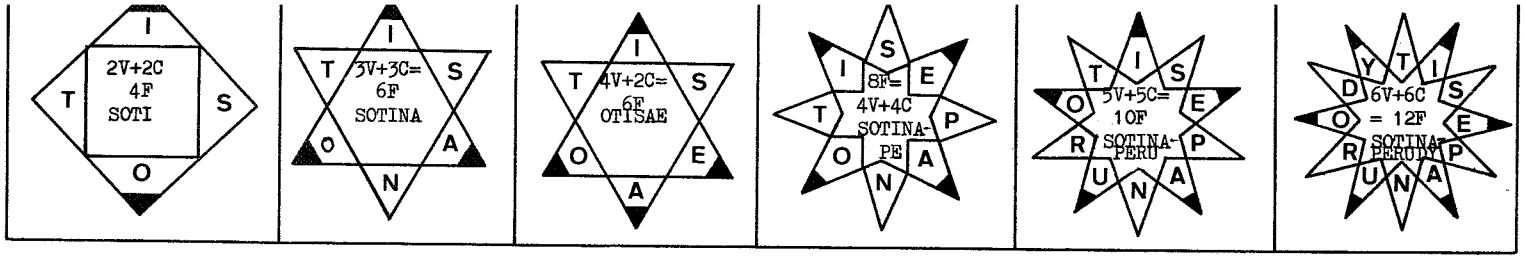
Règles combinatoires de formation.

Mots élémentaires : chacun est formé par une seule voyelle (parmi v), pouvant être précédée par une seule consonne (parmi c), donc mots élémentaires $m_e = v + c.v$

Mots non-élémentaires : autres arrangements (avec répétition). Disposant de f (=v + c) phonèmes au total et de k phonèmes par mot, on peut former $m_k = f^k$ mots. Exemples avec f = 6 : $6^4 = 1296$; $6^5 = 7776$; $6^6 = 46656$; $6^7 = 17.10^6$; $6^8 = 100.10^6$. La moitié environ ($m_k/2$) des arrangements est commodément articulable, et utilisable pour un phonocode. Les mots élémentaires m_e sont inclus dans les diphones $m_2 = f^2$.

Fig. 4. Nombres de mots élémentaires $m_e = v + c.v$ et de mots composés $f_k = f^k$ formables avec v classes de voyelles et c classes de consonnes, c'est-à-dire avec f = c + v classes de phonèmes.

Représentation géométrique



décimal	alphab.	phonocode	3F=OTI,(TOS)	4F = 2V+2C = SOTI		6F = 3V+3C = SOTINA		OTISAE	ESOTINA	ESOTINAU	SOTINAPERU	SOTINAPERUDY
		binaire		ternaire	quaternaire	octal	decimal	décimal	décimal	bi-octal	alphab.	alpha-numérique
0	0	00000	(SO) TO	0000 TO	000 TO	00 NO	00 (N)O	0	TO	00 0	TO	0 TO 0
1	a	00001	(TO) TI	0001 TI	001 TI	01 NI	01 I	I	TI	01 TI	TI	1 TI a NI
2	b	00010	(TOSO) TITO	0002 SO	002 SO	02 TO	02 TO	TI	TE	02 TE	TE	2 TE b NE
3	c	00011	(TOTO) TITI	0010 TITO	003 SI	03 TI	03 TI	TE	TA	03 TU	TU	3 TU c NU
4	d	00100	TITOTO	0011 TITI	010 TITO	04 TA	04 TA	TA	SA	04 TA	TA	4 TA d NA
5	e	00101	TITOTI	0012 TISO	011 TITI	05 SO	05 SO	TO	SE	05 SO	SO	5 SO e NY
6	f	00110	TITITO	0020 SOTO	012 TISO	06 SI	06 SI	SI	SI	06 SI	SI	6 SI f PO
7	g	00111	TITITI	0021 SOTI	013 TISI	07 SA	07 SA	SE	SO	07 SE	SE	7 SE g PI
8	h	01000	TITOTOTO	0022 SOSO	120 TISOTO	10 NINO	08 NI	SA	NI	08 SU	SU	8 SU h PE
9	i	01001	TITOTOTI	0100 TITOTO	121 TISOTI	11 NINI	09 NA	SO	NA	09 SA	SA	9 SA i PU
10	j	01010	TITOTITO	0101 TITOTI	122 TISOSO	12 NITO	10 IO	IO	TITO	0α NO	NO	10 TITO j PA
11	k	01011	TITOTITI	0102 TITOSO	123 TISOSI	13 NITI	11 II	II	TITI	0β NI	NI	11 TITI k PY
12	l	01100	TITITOTO	0110 TITITO	130 TISITO	14 NITA	12 ITO	ITI	TITE	0γ NE	NE	12 TITE l RO
13	m	01101	TITITOTI	0111 TITITI	131 TISITI	15 NISO	13 ITI	ITE	TITA	0δ NU	NU	13 TITU m RI
14	n	01110	TITITITO	0112 TITISO	132 TISISO	16 NISI	14 ITA	ITA	TISA	0ε NA	NA	14 TITA n RE
15	o	01111	TITITITI	0120 TISOTO	133 TISISI	17 NISA	15 ISO	ITO	TISE	0φ TO	PO	15 TISO o RU
16	p	10000	TITOTOTOTO	0121 TISOTI	200 SOTOTO	20 TONO	16 ISI	ISI	TISI	10 TIO	PI	16 TISI p RA
17	q	10001	TITOTOTOTI	0122 TISOSO	201 SOTOTI	21 TONI	17 ISA	ISE	TISO	11 TITI	PE	17 TISE q RY
18	r	10010	TITOTOTITO	0200 SOTOTO	202 SOTOSO	22 TOTO	18 INI	ISA	TINI	12 TITE	PU	18 TISU r DO
19	s	10011	TITOTOTITI	0201 SOTOTI	203 SOTOSI	23 TOTI	19 INA	ISO	TINA	13 TITU	PA	19 TISA s DI
20	t	10100	TITOTITOTO	0202 SOTOSO	210 SOTITO	24 TOTA	20 TOO	TIO	TETO	14 TITA	RO	20 TETO t DE
21	u	10101	TITOTITOTI	0210 SOTITO	211 SOTITI	25 TOSO	21 TOI	TII	TETI	15 TISO	RI	21 TETI u DU
22	v	10110	TITOTITITO	0211 SOTITI	212 SOTISO	26 TOSI	22 TOTO	TITI	TETE	16 TISI	RE	22 TETE v DA
23	w	10111	TITOTITITI	0212 SOTISO	213 SOTISI	27 TOSA	23 TOTI	TITE	TETA	17 TISE	RU	23 TETU w DY
24	x	11000	TITITOTOTO	0220 SOSOTO	320 SISOTO	30 TINO	24 TOTA	TITA	TESA	18 TISU	RA	24 TETA x TY
25	y	11001	TITITOTOTI	0221 SOSOTI	321 SISOTI	31 TINI	25 TOSO	TITO	TESE	19 TISA	U	25 TESO y SY
26	z	11010	TITITOTITO	0222 SOSOSO	322 SISOSO	32 TITO	26 TOSI	TISI	TESI	1α TINO	E	26 TESI z NO
27		11011	TITITOTITI	1000 TITOTOTO	323 SISOSI	33 TITI	27 TOSA	TISE	TESO	1β TINI	A	27 TESE I
28		11100	TITITITOTO	1001 TITOTOTI	330 SISITO	34 TITA	28 TONI	TISA	TENI	1γ TINE	I	28 TESU E
29		11101	TITITITOTI	1002 TITOTOSO	331 SISITI	35 TISO	29 TONA	TISO	TENA	1δ TINU	O	29 TESA U
30		11110	TITITITITO	1010 TITOTITO	332 SISISO	36 TISI	30 TIO	TEO	TATO	1ε TINA		30 TUTO A
31		11111	TITITITITI	1020 TITOSOTO	333 SISISI	37 TISA	31 TII	TEI	TATI	1φ TITO		31 TUTI Y
STI	STA	mot instruction										

Fig. 5 Phonocodes : formation des mots élémentaires (chiffres et lettres) (Règle fondamentale : chaque "mot" comprend 1 voyelle ou 1 consonne + 1 voyelle)

code	mot instr. numéri-que	7 5 7 5	mot instr. alpha-bétiq.	P 16	A 1	R 18	I 9	S 19	durée d'élo cutio sec.
binaire OTI(TOS)	00	1 0 0 1 0 1 1 TITOTOTITOTITI	000	1 0 0 0 0 TITOTOTOTO	1 TI	1 0 0 1 0 TITOTOTITO	1 0 0 1 TITOTOTI	1 0 0 1 1 TITOTOTITI	6
ternaire SOTI	STI	2 2 1 0 SOSOTITO	STO	1 2 1 TISOTI	1 TI	2 0 0 SOTOTO	1 0 0 TITOTO	2 0 1 SOTOTI	4,2
quaternaire SOTI	STI	1 0 2 3 TITOSOSI	STO	2 0 0 SOTOTO	1 TI	2 0 2 SOTOSO	1 2 1 TISOTI	2 0 3 SOTOSI	4,2
octal SOTINA	STI	1 1 3 NINITI	STA	20 TONO	1 NI	22 TOTO	11 NINI	23 TOTI	3,5
décimal SOTINA (ESOTINA)	STI	7 5 SASO	STA	16 ISI	1 TI	18 TINI	9 NA	19 TINA	3,1
décimal OTISAE	STI	7 5 SETO	STA	16 ISI	1 I	18 TITA	9 SO	19 TESE	3,1
alpha-numérique SOTINAPERUDY		7 5 P A R I S SESORANIDOSADI							1,2

Fig. 6 Les mots "75-Paris" selon divers phonocodes, avec durées d'élocution

Discussion

J. GUIBERT : Avez-vous entrepris des tests sur des vocabulaires d'une dizaine de mots avec une dizaine de locuteurs et les taux de reconnaissance sont-ils sensiblement supérieurs à ceux que l'on obtient avec une langue naturelle?

J. DREYFUS-GRAF : Nous n'avons pas encore effectué de tests spécifiques, mais des programmes de tests sont actuellement en préparation. Ils permettront d'établir les tolérances d'articulation à l'émission et les taux d'erreur à la réception, en vue d'optimiser les phonocodes. Cependant, sur la base de nos longues années d'expériences acquises, on peut prévoir des taux de reconnaissance très supérieurs à ceux qui ont été obtenus avec des langues naturelles.

L'amélioration des taux de reconnaissance est déjà une conséquence logique de l'élargissement des classes de phonèmes. Par exemple, il n'y aura plus besoin de se préoccuper de distinguer le /i/ du /e/, ou le /o/ du /u/; puisque /i/ et /e/ tombent dans la même classe "I" et que /o/ et /u/ tombent dans la même classe "O".

MODELISATION DES TRANSMISSIONS PHONEMIQUES

APPLICATION A LA SEGMENTATION DE LA PAROLE

Résumé .

L'article présente l'application d'un filtrage multidimensionnel récurrent à l'identification d'un modèle de transition phonémique. Le système dispose comme entrées de l'évolution temporelle de paramètres phonémiques (une dizaine environ) d'origine quelconque (vocoder à canaux, formants, paramètres du canal vocal...). Il recherche un modèle d'ordre limité (2 à 3 en pratique) caractérisant l'évolution transitoire de ces paramètres et réalisant de plus un filtrage adapté des régions stationnaires. Cette analyse peut être orientée vers la détermination de variables représentant la transition considérée en synthèse de la parole, ou complétant les paramètres phonémiques bruts en reconnaissance, ou vers la segmentation automatique en ligne par détection des régions transitoires.

Summary

A recursive method for the modelisation of phonemic transitions is presented. It is based on a multivariable kalman filtering applied to any given set of parameters. The sought model of low order identifies the behavior of all the parameters jointly and perform an adapted filtration of the stable regions. The method can be designed to segmentate continuous speech by on-line implementation, or to memorize a few significant parameters representing the transitions in the speech synthesis and recognition problems.

G. CARAYANNIS

Ecole Nationale Supérieure des Télécommunications

1. ROLE DES TRANSITIONS PHONEMIQUES DANS LA STRUCTURE D'UNE PHRASE.

L'introduction du phonème comme unité linguistique de base a répondu au besoin d'établir un code pour représenter le message parlé ; analogue au code de représentation du message écrit , l'unité de base pour le message écrit étant le caractère.

Pourtant une grande différence existe entre les deux représentations, le message parlé étant un code analogique , la représentation par des phonèmes n'est qu'une discrétisation de ce code , et par conséquent une approximation. Cette approximation se révèle assez grossière vu le rôle que jouent les transitions phonèmiques dans le message parlé.En effet,ces transitions phonémiques, exprimant l'évolution des paramètres d'un état stable à un autre état stable , sont très informatives , car une infinité de trajectoires entre deux régions stables seraient théoriquement possibles. En outre , un phénomène de voisinage se manifeste souvent, ce qui implique un conditionnement de l'état stable en question par les états voisins. Ainsi, chaque état stable se présente sous un éventail d'aspects différents suivant son voisinage.

On peut regrouper l'information contenue dans une phrase parlée sur differents supports fondamentaux :

- variations d'amplitude du signal vocal (enveloppe énergétique)
- variations du fondamental.
- régions stables d'un phonème .
- régions de transition entre phonèmes.
- durée des états stables ou transitoires.

Les régions stables, dont seule l'étude est, en général , entreprise en détails, ne comportent donc qu'une partie relativement faible de l'information. D'où l'interêt de détecter et de représenter par un formalisme adapté , les zones transitoires en général méconnues de la phrase.

C'est dans ce sens qu'est dirigée cette étude qui propose l'établissement d'un modèle abstrait de l'évolution transitoire de paramètres, d'origine quelconque , supposés représentatifs du signal vocal.

On notera que , malgré l'importance du problème , un nombre très restreint de travaux y a été consacré. La raison peut en être recherchée dans l'introduction relativement récente des méthodes de modelisation dans le traitement du signal vocal. RABINER , cependant , en vue de la synthèse vocale , propose l'ajustement d'une équation differentielle du second ordre avec amortissement critique (1 paramètre libre) pour décrire l'évolution des formants lors d'une transition , mais ce modèle demeure rudimentaire .

2. MODELISATION DES TRANSITIONS .

Soit $\underline{d}_n$ un vecteur représentant les paramètres issus d'une méthode d'analyse quelconque , à l'instant n . Un modèle abstrait caractérisant l'évolution temporelle de $\underline{d}_n$ est constitué par une relation de recurrence d'ordre k :

$$\underline{d}_n + \omega_1 \underline{d}_{n-1} + \ldots + \omega_k \underline{d}_{n-k} = 0 \qquad (1)$$

dont il s'agit de déterminer les coefficients ω_i au mieux . L'équation (1) ne pouvant être idnetiquement vérifiée sur la longueur de l'échantillon, on est amené à concéder une erreur $\underline{e}_n$ dont la matrice de covariance R_n est supposée connue.

$$\begin{cases} \underline{d}_n + \omega_1 \underline{d}_{n-1} + \ldots + k \underline{d}_{n-k} = e_n \\ E(e_n) = 0 \ , \ E(e_i . e_j^T) = 0 \ , \ E(e_n . e_n^{rT}) = R_n \end{cases} \qquad (2)$$

Le critère d'optimalisation retenu est le minimum de variance du vecteur ω_n estimé exprimé par :

$$C = \text{Trace } P_n \ , \quad P_n = E\left[(\omega - \hat{\omega}_n)(\omega - \hat{\omega}_n)^T \right] \qquad (3)$$

La solution du problème est alors fournie par un filtre de Kalman multidimensionnel défini par l'estimateur linéaire récurrent :

$$\hat{\omega}_n = \hat{\omega}_{n-1} + K_n \left(\underline{d}_n - D_n^T \hat{\omega}_{n-1} \right) \qquad (4)$$

avec $D_n^T = \underline{d}_{n-1} , \underline{d}_{n-2} , \cdots , \underline{d}_{n-k}$

dont le coefficient matriciel K_n est ajusté de manière optimale par :

$$\begin{cases} P_n = (I - K_n D_n^T) P_{n-1} \\ K_n = P_{n-1} D_n (R_n + D_n^T P_{n-1} D_n)^{-1} \end{cases} \qquad (5)$$

L'initialisation des équations (5) est réalisée à un instant choisi par:

$$\hat{\omega}_o = 0 \quad \text{et} \quad P_o = \left\{ p_{ij} : \exp - (|i - j| / \theta) \right\}$$

La décroissance systématique de la trace de P_n traduit la convergence et l'apport d'information résultant de la nouvelle mesure disponible $\underline{d}_n$.

3. APPLICATION A LA SEGMENTATION DE LA PAROLE .

Une approche analytique de la parole est nécessaire pour toute étude de quelque ambition en synthèse ou en reconnaissance vocales . Il est, en effet , indispensable quand le traitement d'un vocabulaire important est requis, de décrire et de mémoriser les caractéristiques du signal à partir d'unités de dimension suffisamment restreinte pour ne pas aboutir à un volume de données prohibitif. La détermination et l'enchainement de ces unités devient alors le problème crucial .

Ces unités d'origine acoustique auront avantage , bien qu'à l'évidence ceci ne soit pas indispensable /1/ , à coincider avec des éléments linguistiques classiques pour bénéficier de l'acquis de cette discipline . La nature et la qualité de la segmentation est donc avant tout conditionnée par la pertinence des paramètres extraits du signal. Un paramètrage idéal ferait apparaître sans équivoque des régions de stabilité caractérisant la présence sous-jacente d'une forme acoustique. Un tel paramètrage serait inaccessible ou simplement trop coûteux à mettre en oeuvre .

Il en résulte que les méthodes usuelles de segmentation se contentent de rechercher des zones de stabilité approximative en procédant ou non sur l'ensemble des paramètres /2/ /3/ , à partir de seuils de discrimination ajustables ou non /4/. On a , alors , affaire à un problème de détection de forme .

Une approche fine du problème est constituée par la segmentation après reconnaissance où chaque occurrence d'un vecteur de paramètres se voit appliquer une procédure de reconnaissance complète destinée à vérifier son appartenance à une classe phonémique particulière et à identifier celle-ci . Une telle méthode entraîne un temps de calcul très important . C'est pourquoi , on a le plus souvent recours à la détermination de diverses clefs simples , permettant le classement grossier en catégories phonétiques /5/ /6/ . Il s'agit là d'une reconnaissance élémentaire parfaitement adaptée à l'implantation d'une procédure de reconnaissance hiérarchique , mais qui se doit d'être rigoureusement exacte.

La méthode proposée ici est intermédiaire car elle réalise une segmentation par classification différentielle et par détection des régions transitoires dans l'ensemble des paramètres . Supposant en effet la convergence obtenue , le caractère récurrent de la technique de modelisation permet de surveiller l'évolution des paramètres $\underline{d}_n$ par prédiction :

$$\hat{d}_n = - \sum_{i=1}^{k} \omega_i \, \underline{d}_{n-i} \quad ; \quad \varepsilon_n = d_n - \hat{d}_n \qquad (6)$$

Par ailleurs , le vecteur $\hat{\omega}_n$ étant caractéristique d'une zone temporelle dont les paramètres respectent une certaine cohérence , un critère intrinsèque du type :

$$\gamma_n = (\hat{\omega}_n - \hat{\omega}_{n-1})^T (\hat{\omega}_n - \hat{\omega}_{n-1}) \qquad (7)$$

rend compte du franchissement de la frontière d'une classe sans qu'il soit utile d'identifier celle-ci .

Cette procédure a été mise en oeuvre avec succès sur différents types de paramètres : canaux de vocoder , formants calculés par FFT , paramètres issus de la modelisation du canal vocal (voir /7/). La figure 1 donne un exemple d'application dans ce dernier cas . L'ordre k du système est limité à 3 et un nouveau jeu de paramètres $\underline{d}_n$ est fourni en synchronisme avec les périodes fondamentales pour les sons voisés et toutes les 10ms pour les sons non-voisés . De façon à montrer le pouvoir discriminant de l'algorithme , on a omis de le réinitialiser après chaque détection de transition . La désensibilisation du critère de segmentation , dû à l'intégration par le filtre d'événements de natures différentes ne survient qu'après quelques secondes de parole. Il apparaît cependant que l'exactitude du modèle sur les transitions rapides est conditionnée par le nombre d'échantillons disponibles . On peut considérer qu'une convergence acceptable est obtenue par un nombre d'échantillons supérieur à 3.k.

Il est à noter , d'autre part , qu'une telle procédure est succeptible de compléter les informations issues d'une analyse rudimentaire du signal. Les paramètres du modèle traduisent une loi d'évolution des composantes de $\underline{d}_n$, dont l'existence ne saurait être soupçonnée par le simple examen des courbes temporelles . Les coefficients ω_i constituent , ainsi , des paramètres d'ordre supérieur qui seraient adjoints avec profit aux paramètres bruts précédents.

4. CONCLUSIONS

La modelisation des transitions phonémiques par un filtre de kalman multidimensionnel a donné lieu à une technique de segmentation qui se caractérise par :

- Une application en ligne sur un jeu de paramètres quelconque .
- Une prise en compte globale de l'évolution de tous les paramètres.
- Un critère de segmentation lié au modèle et opérant donc une classification.

Mais l'étude déborde largement ce cadre grâce à des modifications mineures. En effet , l'utilisation du lissage conjointement avec la prédiction (ici seule utilisée) engendrerait un modèle tenant compte à la fois du passé et du futur de la transition . Un tel modèle serait bien adapté à la génération automatique des transitions en synthèse par règles de la parole .

Bien plus, la méthode d'identification adoptée , constitue une approche directe au problème de reconnaissance . L'application de l'algorithme en période d'apprentissage , fournit un modèle de répartition des occurrences sous la forme des caractéristiques d'un bruit multidimensionnel coloré . Cette procédure , homogène avec les traitements précédents , fait l'objet des études actuelles.

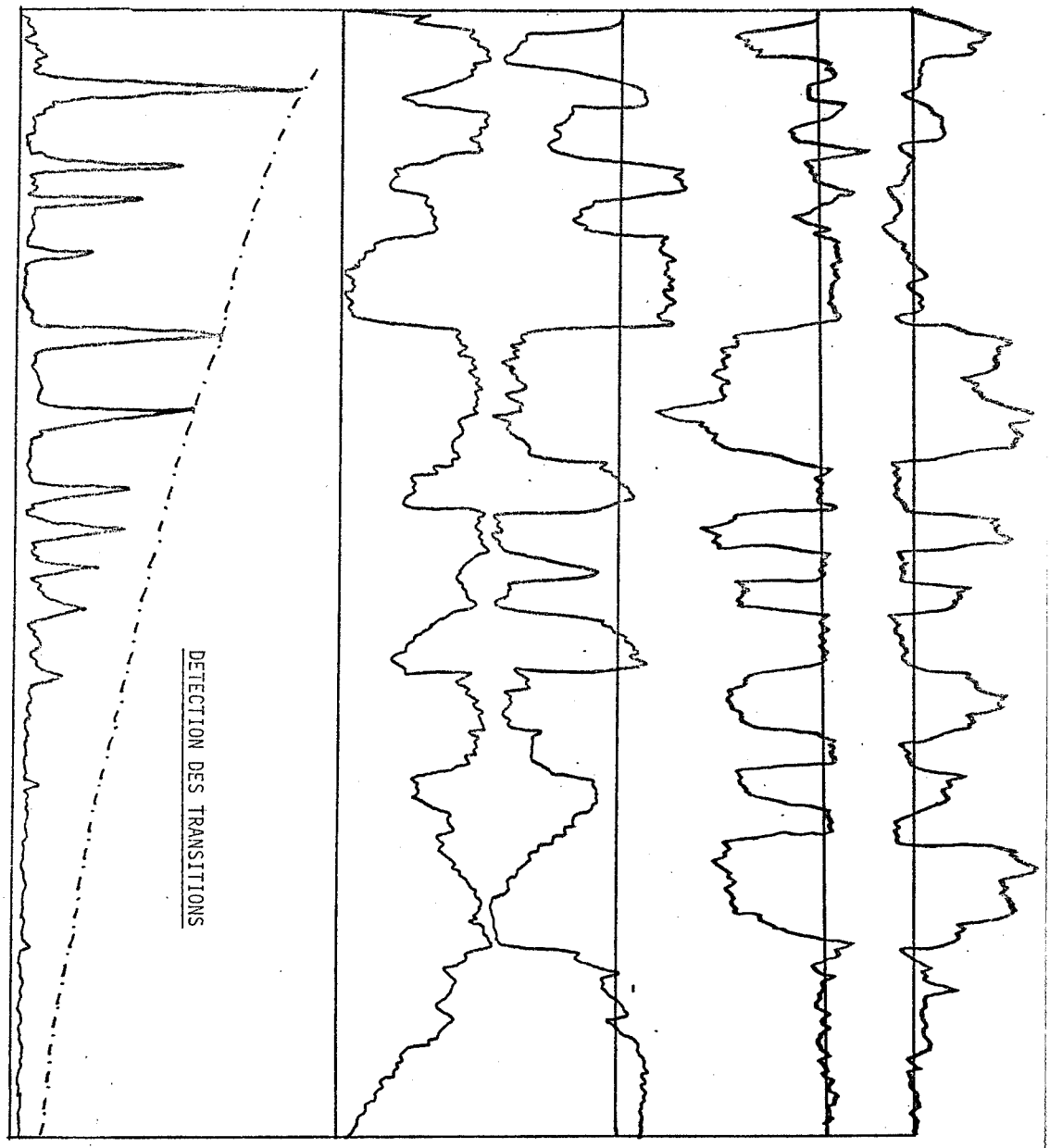


Figure 1 : Détection des transitions SANS réinitialisation de l'algorithme (paramètres issus de /7/)

REFERENCES :

/1/. W.A. LEA : An approach to syntactic recognition without phonemics.
IEEE Trans. on Audio Vol 21 . To be published April 73

/2/. D.R. REDDY : Segmentation of speech sounds
J. Acoust. Soc. Am. , Vol 40 , N° 2 , pp 307 -312 - 1966

/3/. D.R. REDDY , P.J. VICENS : A procedure for segmentation of connected speech .
J. Audio. Eng. Soc. Oct 1968

/4/. S.K. DAS , D.F.STANAT : Segmentation of utterances of a known phrase using linear threshold techniques.
IEEE Trans. on Audio Vol 20 N°2 - pp 142 -150 - 1972

/5/. K.P. LI , G.W. HUGHES , T.B. SNOW : Segment classification in continuous speech.
IEEE Trans. on Audio . Vol 20 N°2 - pp 142 -150 -1972.

/6/. ALINAT : Reconnaissance de phonèmes en temps réel en vue de réaliser des liaisons à faible débit . Application à la stenotypie automatique
Rapport final d'Etudes DRME N° 198/71 - 1972 .

/7/. C. GUEGUEN , G. CARAYANNIS : Analyse de la parole par filtrage optimal.
4ème Journées du groupe " communication parlée du GALF .
Bruxelles 1973.

Discussion

J-P. HATON : Avez-vous fait une étude systématique des différentes façons dont il est possible de réinitialiser votre système, et avez-vous trouvé une méthode optimale ?

G. CARAYANNIS : Par nature, le filtre de Kalman intègre les événements compris entre son initialisation et l'instant courant. Si ces événements sont de nature contradictoire, il en effectue une moyenne ce qui le rend peu sensible à une nouvelle variation des propriétés du signal. Il convient donc de réinitialiser le modèle à chaque détection d'une zone transitoire. Le problème est donc lié à la qualité du critère de détection et à la rapidité de convergence après la réinitialisation qu'il convient donc d'optimiser. Le critère de détection intrinsèque se révèle sensible et l'on aura intérêt à fixer un seuil relativement faible car un défaut de détection est en général reconnu plus grave qu'une segmentation additionnelle. La rapidité de convergence dépend du choix des informations a priori, il faut qu'en tout état de cause le nombre d'échantillons disponible soit suffisant entre deux transitions.

J. GENIN : Quelles sont les performances de la segmentation obtenue par cette méthode ?

G. CARAYANNIS : Cette méthode a surtout été appliquée sur les paramètres issus d'un modèle du canal vocal comme présenté en figure 1. La finesse de la segmentation, dépendant essentiellement de la qualité des paramètres est difficile à évaluer pour d'autres systèmes. Cependant, on peut affirmer que sur les paramètres de A. Maissis, la segmentation est meilleure que celle jusqu'ici pratiquée. Dans le cas des canaux de vocoder, la méthode n'a pu être comparée avec des techniques conventionnelles.

SEGMENTATION ET RECONNAISSANCE PAR COMPARAISON DYNAMIQUE DE LA VOIX PARLEE.

Résumé

Cet article concerne une approche analytique de la reconnaissance de mots en temps réel. Les mots, prononcés dans un micro, sont directement prétraités par un analyseur spectral-codeur et acquis en temps réel par un ordinateur. Une segmentation sur critères acoustiques en "phonèmes" est effectuée pendant l'acquisition et les segments obtenus sont ensuite comparés à des segments-types par une méthode dynamique qui effectue en même temps une normalisation temporelle non linéaire. La méthode de comparaison dynamique a d'abord été testée à la reconnaissance globale, en temps réel, de mots isolés. Différents résultats et conclusions sont donnés.

Summary

This paper concerns an analytical approach to the problem of real time speech recognition.

Words, pronounced through a microphone, are pre-processed by a spectral analyzer-coder, fed into a computer and segmented in real time. The segments are then compared to paradigms by a dynamic matching procedure, which does a non linear temporal normalization of the segments. The method of dynamic matching has been first tested in the global recognition of isolated words. Different results and conclusions are given.

J-P. HATON
Université de Nancy I.

SEGMENTATION ET RECONNAISSANCE PAR COMPARAISON DYNAMIQUE DE LA VOIX PARLEE

*****

J.-P. HATON

I - INTRODUCTION.

L'approche analytique de la reconnaissance de la parole consiste en une segmentation du message en éléments (syllabes, phonèmes, segments minimaux), suivie d'une reconnaissance de ces segments. Cette approche est, certes, plus délicate qu'une approche globale de reconnaissance de mots, par exemple, mais elle est plus générale et beaucoup plus intéressante dans le cas de très grands vocabulaires.

Cet article décrit les deux phases de segmentation et de reconnaissance dans un système de reconnaissance phonémique de mots.

La segmentation, fondée sur des critères purement acoustiques, fournit des "segments minimaux", correspondant aux zones de stabilité relative, ou de transition, du signal de parole.

Les segments obtenus pouvant varier beaucoup en longueur, selon le type d'élocution, ou la position dans le mot, il importe d'effectuer une normalisation temporelle de ces segments. Pour cela, nous utilisons une méthode dérivée de la programmation dynamique, qui permet une normalisation temporelle, automatique, non linéaire, en cours de reconnaissance [1].

La programmation dynamique a été également utilisée, avec des algorithmes différents, pour la reconnaissance globale de mots, en particulier au Japon [2].

II - DESCRIPTION DU CADRE EXPERIMENTAL.

Les applications potentielles de la reconnaissance vocale sont presque toutes liées à la possibilité d'une réponse quasi-instantanée de la machine. Nous nous efforçons de nous placer le plus possible dans ce contexte "temps réel". C'est ainsi que les expériences reportées ici concernent, sauf spécification contraire, la reconnaissance en temps réel, c'est-à-dire à la vitesse d'élocution, de mots prononcés dans un micro de qualité médiocre en salle d'ordinateur (niveau de bruit : 65 db). L'ordinateur utilisé est un T 2000 de la Télémécanique Electrique.

Les mots prononcés sont prétraités à l'aide d'un analyseur spectral à 25 filtres, suivi d'un codeur qui fournit une forme binaire, par codage des pentes de la courbe amplitude-fréquence. Ces formes binaires sont acquises par l'ordinateur à une fréquence de 100 Hz.

III - SEGMENTATION.

Bien que la parole soit de nature essentiellement transitoire, on peut y distinguer des zones de stabilité relative (corps d'une voyelle par exemple), et des zones de transition (fin d'une voyelle, explosion d'une plosive , ...). La méthode de segmentation utilisée ici effectue cette distinction en fournissant des "segments minimaux", souvent assimilables aux phonèmes (mais une plosive, par exemple, sera composée de deux segments, et une voyelle longue pourra donner lieu à plusieurs segmentations).

Elle procède, d'autre part, à une pré-classification des segments en grands types (segments "voyelles" , "fricatifs" ,"plosifs" ...) lorsque cette classification est possible. Il en résulte un gain de temps appréciable au moment de la reconnaissance.

Pour assurer ces deux fonctions de segmentation et de préclassification, nous utilisons deux jeux de paramètres acoustiques, déterminés à partir du spectre :

- trois paramètres principaux reliés au nombre de "1" du sonagramme binaire dans trois zones de fréquences (basses, moyennes et hautes) ;

- des paramètres secondaires tenant compte des variations à l'intérieur même d'une zone de fréquence.

L'efficacité de la méthode a été testée en comparant la segmentation obtenue à une segmentation "idéale", subjectivement effectuée à partir des sonagrammes binaires. On constate que le nombre de segments est de l'ordre de 1,5 à 2 fois le nombre de "phonèmes" retenus. On obtient donc des segmentations multiples, peu gênantes pour le traitement ultérieur de la chaîne phonémique, mais très peu "d'oublis". La figure I donne un exemple typique de cette segmentation.

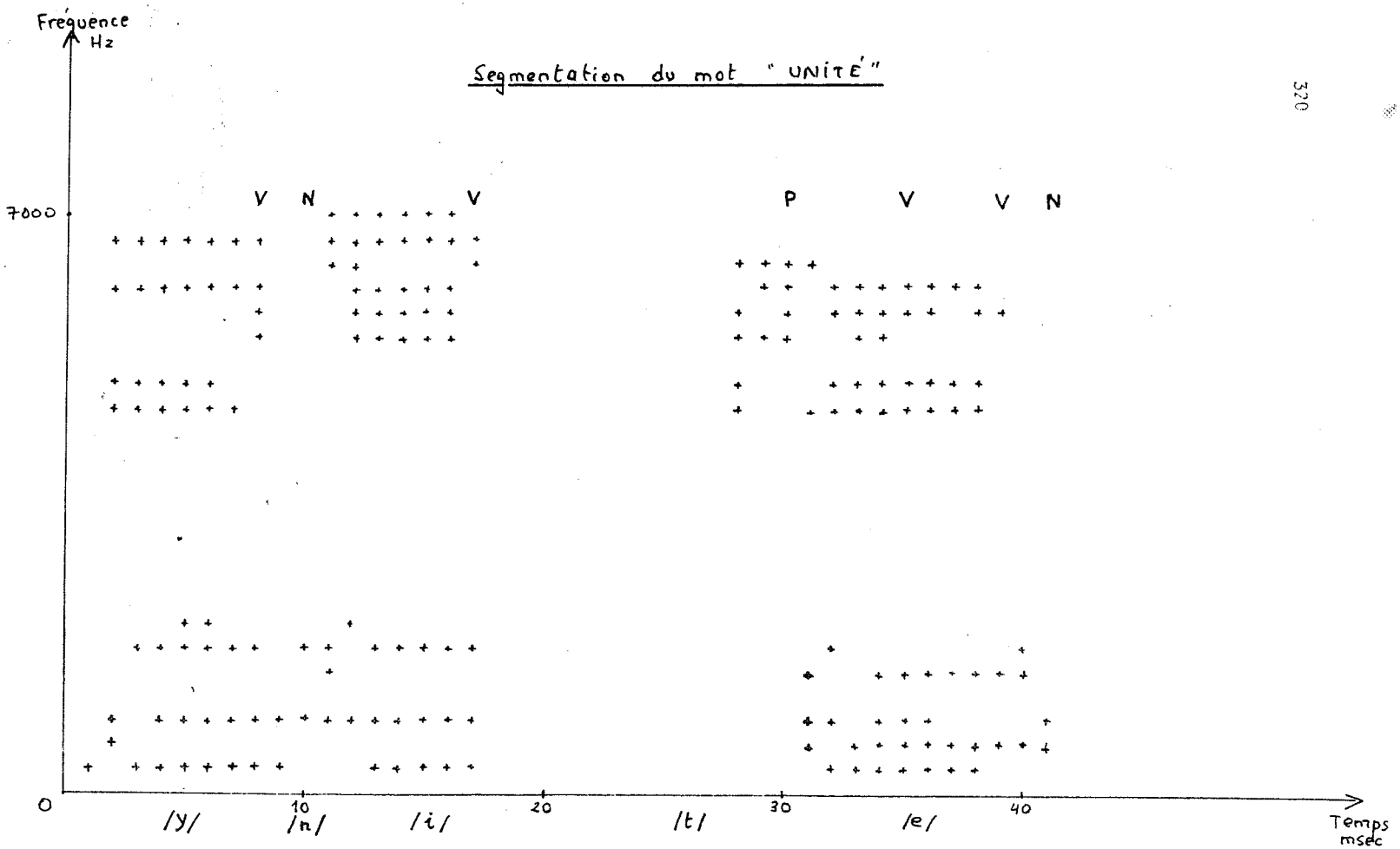
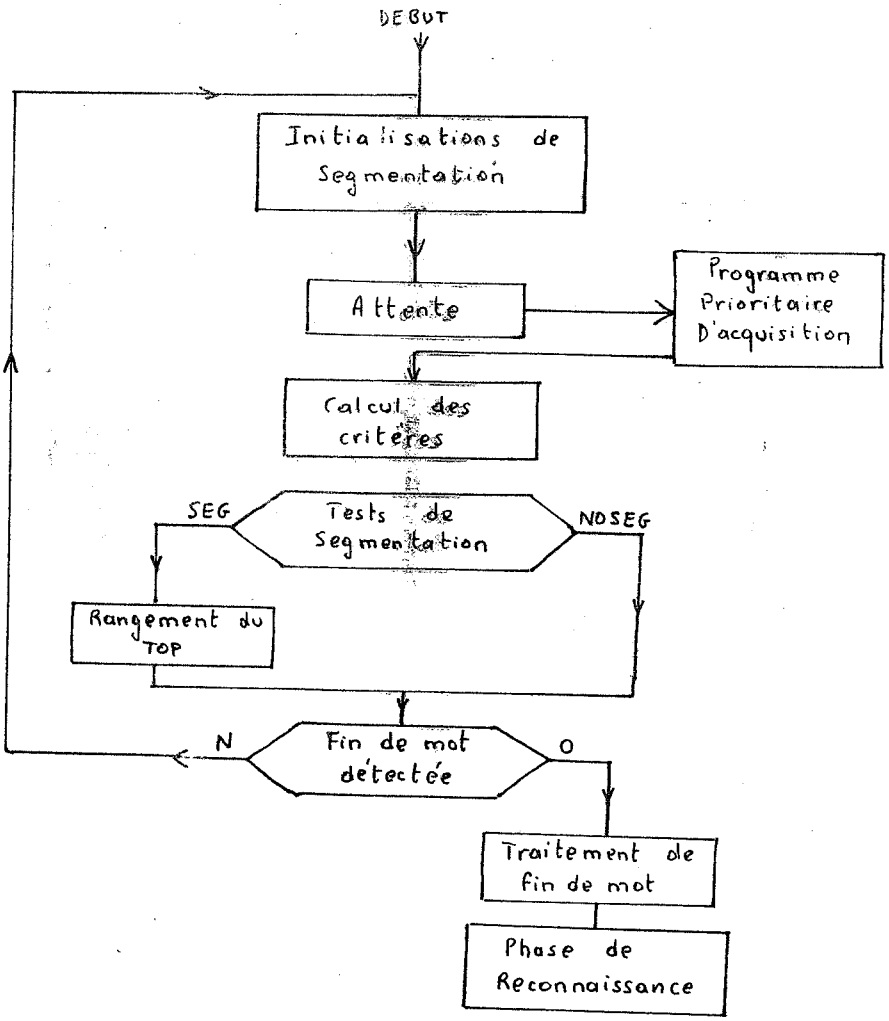

Fréquence
Hz
Segmentation du mot " UNITÉ "
7000
V
N
V
P
V
V
N
0
10
20
30
40
/y/
/n/
/i/
/t/
/e/
Temps
msec

Fig. 1



Segmentation d'un mot en cours d'acquisition

Fig. 2

Le temps de traitement est variable, mais n'excède pas 1,4 ms. Pour gagner du temps de reconnaissance, la segmentation est effectuée en cours de prononciation d'un mot, entre les acquisitions de données (deux acquisitions étant séparées de 10 ms). La figure II schématise le déroulement des opérations de segmentation et de reconnaissance pour la prononciation d'un mot.

IV - RECONNAISSANCE DYNAMIQUE.

Lorsqu'une fin de mot est détectée par le système, on passe alors en phase de reconnaissance pendant laquelle les segments sont classés par comparaison à des segments-types stockés en mémoire. Mais les variations de longueur qui peuvent affecter un même segment sont considérables, aussi est-il nécessaire de procéder à une normalisation temporelle. Plutôt que d'effectuer a priori une normalisation arbitraire, nous utilisons une méthode de comparaison dynamique qui effectue une normalisation non-linéaire en cours de reconnaissance. La méthode dérive du principe d'optimalité de la programmation dynamique, elle permet de trouver la similitude optimale entre deux formes avec le minimum de calculs, au sens d'une métrique donnée.

Pour tester la méthode et en connaître les limites, nous l'avons d'abord appliquée à la reconnaissance globale de mots d'un dictionnaire. L'inconvénient majeur réside dans un temps de calcul assez long. Diverses améliorations ont permis de réduire ce temps, en particulier la microprogrammation du calcul de distance entre deux échantillons et l'utilisation d'une fenêtre temporelle linéaire F, qui permet de limiter la comparaison à une portion seulement d'une forme (fig. 3)

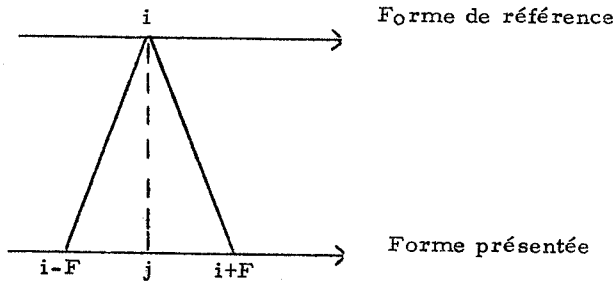


Figure 3.

Dans ces conditions, les divers algorithmes essayés ont donné des résultats satisfaisants. On obtient ainsi de 98 à 100 % de reconnaissance sur un vocabulaire d'une trentaine de mots, pour un locuteur, avec un temps de calcul bien inférieur à 1 s. Pour plusieurs locuteurs, le taux de reconnaissance tombe, pour le même vocabulaire, à 80 % environ, sans aucune normalisation fréquentielle.

Ces performances ne constituent pas la limite du système, car, en particulier, le vocabulaire n'a été limité à 30 mots que pour des raisons de saturation de la mémoire centrale actuelle du T 2000. De toute façon, le but de ces expériences préliminaires n'était que de tester la validité de la méthode, qui s'avère être une des meilleures que nous avons utilisées jusqu'à présent, moyennant quelques précautions.

Cette méthode a été ensuite adaptée à la reconnaissance analytique de segments. Les expériences ont été menées, pour l'instant, en temps différé, à partir de mots préenregistrés sur bande magnétique. Les mots utilisés étant peu nombreux, il faudra attendre le fonctionnement complet en temps réel pour chiffrer exacte ment les résultats obtenus. On a cependant constaté que ces résultats étaient très encourageants, surtout si l'on considère à chaque fois, non seulement le segment reconnu, mais aussi les 2 ou 3 segments les plus probables.

V - CONCLUSION ET PERSPECTIVES.

Les méthodes de normalisation temporelle par programmation dynamique, qui se sont avérées efficaces en reconnaissance globale de mots, sont utilisées ici en vue de la reconnaissance analytique de mots, après segmentation acoustique du signal de parole. Le système complet doit fonctionner bientôt en temps réel, et on sera ainsi ramené au problème de l'optimisation d'une chaîne phonémique et à la recherche lexicale d'un mot.

BIBLIOGRAPHIE.

[1] J-P. HATON "La programmation dynamique : principes et applications à la reconnaissance de la parole" - Laboratoire d'Electricité et d'Automatique - Note interne. février 1973.

[2] SAKOE and CHIBA "A dynamic programming approach to continnons speech recognition" 7th Int. Cong. on Acoustics Budapest 1971.

Discussion

G. ROGER : Pour ceux qui ne connaissent pas les principes de la programmation dynamique, est-ce qu'un des deux orateurs pourrait les rappeler ?

J-P. HATON : On peut vous donner pour commencer quelques références (voir bibliographies).

M. ROSSI : Gresser a parlé tout à l'heure de critères phonétiques de début et de fin de mot. Je serais intéressé par la nature des critères acoustiques trouvés.

J-Y. GRESSER : Les critères trouvés sont purement mathématiques. Début et fin de mots sont les échantillons qui permettent d'atteindre un certain extremum. Mais l'extremum est global, c'est-à-dire qu'il est relatif à une mesure définie sur l'ensemble du mot.

J-S. LIENARD : Il y a deux ans nous avons fait un travail assez voisin de ceux-ci, en utilisant une méthode, qui se ramenait à une programmation dynamique, sur plusieurs locuteurs. Elle nous avait donné des résultats encourageants. On peut réduire le temps de calcul en utilisant une fonction de stationnarité au sens où je l'ai exposé hier, et en ne conservant du message que les instants caractéristiques. Nous avons aussi réussi à ne garder qu'un spectre sur 7 (dans le temps).

D'autre part, nous étions arrivés à une fenêtre optimale, ramenée sur le message lui-même en temps réel, qui était de l'ordre de 250 à 300 ms.

J-P. HATON : C'est à peu près l'ordre de grandeur que l'on trouve, mais je pense qu'il est plus intéressant de permettre à la fenêtre de varier, ce qui n'aggrave pas le problème.

J-Y. GRESSER : Après la segmentation décrite, comment se fait la normalisation des diphonèmes ? De manière élastique ou non ?

J-S. LIENARD : Il s'agit de deux méthodes différentes à partir de la même fonction de stationnarité.

J-Y. GRESSER : Dans la méthode indiquée, procèdes-tu effectivement après la segmentation à des déformations non linéaires ?

J-S. LIENARD : Au moment de la segmentation, lorsqu'on passe de la segmentation à des éléments discrets normalisés, on fait une compression élastique, sinon cela n'aurait pas de sens. Mais c'est une autre méthode que celle indiquée.

G. MERCIER : 1. Quels segments pensez-vous prendre ? Ne pensez-vous pas que les syllabes seraient des segments plus valables que les phonèmes étant donné que la programmation dynamique est d'autant meilleure que les segments sont longs.
2. Combien d'éléments de référence pensez-vous utiliser pour représenter chaque segment au niveau de l'apprentissage et comment pensez-vous procéder pour cet apprentissage (ou cette normalisation) si vous faites un apprentissage ?

J-P. HATON : 1. Les segments que nous utilisons correspondent, comme nous l'avons dit, aux zones de stabilité ou de transition du spectre. Il nous semble que la programmation dynamique peut être intéressante dans ce cas, car, si ces segments sont relativement courts, les variations relatives de longueur qui les affectent sont très importantes, la normalisation temporelle effectuée est alors efficace.
2. Nous avons pour l'instant une quarantaine de segments de référence, mais ce nombre n'est pas limitatif. Nous avons déjà fait remarquer que l'apprentissage est très important dans notre méthode. Il s'agit d'avoir de "bonnes" formes de référence. Pour cela nous pensons utiliser un nombre important de segments d'apprentissage et de dégager, pour chaque classe, un "squelette commun" par une méthode itérative.

APPLICATION DE LA PROGRAMMATION DYNAMIQUE A LA RECONNAISSANCE DE MOTS ++

ésumé

Nous présentons quelques résultats sur la reconnaissance de mots isolés ou tués dans le discours naturel, comme une première évaluation de l'efficacité es méthodes non linéaires de normalisation temporelle.

ımmary

In several experiments we tested the efficiency of a non-linear time ɔrmalization technique using dynamic programming, on the automatic recognition ' words spoken either in isolation or in continuous speech.

J-F. BARS +, J-Y. GRESSER, M. QUERRE
Centre National d'Etudes des Télécommunications

Alors étudiant à la Faculté des Sciences (Rennes)

+ Cette étude a été partiellement soutenue par le CRI (convention 70 105/71 09).

Un même mot prononcé plusieurs fois dans des contextes différents se présente à l'analyse automatique sous des aspects variés. Parmi les déformations les plus "visibles" certaines ont trait à la rapidité et au rythme d'élocution. Nous nous sommes limités aux déformations temporelles.

Le point de départ est le codage d'un mot selon une suite d'échantillons. SLUCKER [1] a proposé une méthode de comparaison entre deux suites qui représentent le même mot ou groupe de sons. Cette méthode est basée sur une recherche, par déformation non élastique, des échantillons qui coïncident le mieux à l'intérieur des mots. L'algorithme de base est emprunté à la programmation dynamique.

La procédure de reconnaissance est une recherche d'identification ("template matching") ou de ressemblance optimale. ZAGORUJKO [2] l'a appliquée avec succès à un vocabulaire de 200 mots différents, où chacune des 4 occurences d'un mot, prononcées par un seul locuteur, sert tour à tour de modèle. Pour diminuer le temps de calcul il ne met en jeu la programmation dynamique qu'au dernier stade de la décision. SAKOE et CHIBA [3] ont présenté un algorithme simplifié qui leur a permis de reconnaître parfaitement les chiffres japonais énoncés de manière continue par un locuteur.

Une expérience préliminaire sur la reconnaissance de chiffres isolés, prononcés par un locuteur, nous a également permis d'atteindre une reconnaissance parfaite. Nous l'avons estimée suffisamment encourageante pour faire des essais approfondis.

Les résultats présentés constituent une première estimation de la méthode. Pour cela nous nous sommes abstenus de compliquer la prise de décision qui se fait en une seule étape.

Les limites de notre codage sont bien connues.

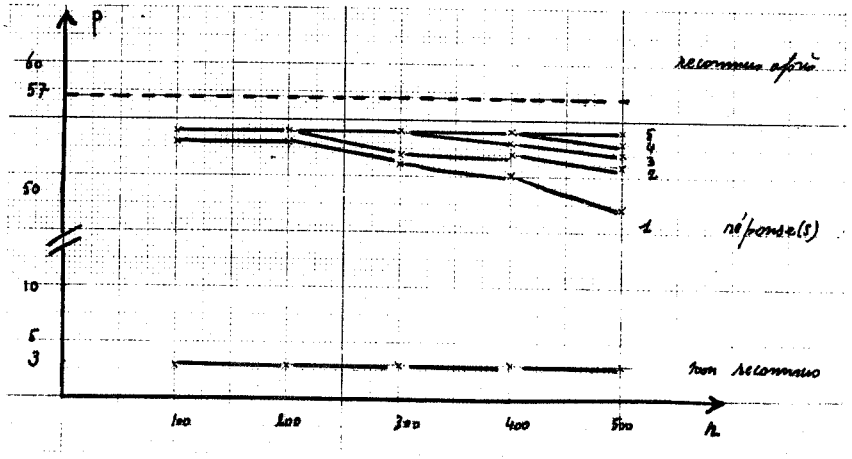
Nous utilisons un vocodeur, à 14 canaux, dont la bande passante est dite téléphonique. Dans chaque canal l'énergie instantanée est repérée selon 16 niveaux.

MOTS ISOLES.

Le vocabulaire de référence était constitué de 1200 mots tirés du premier Larousse en images, vocodés pour un seul locuteur (MERCIER).

Les essais de reconnaissance de p mots parmi n ont été faits en prenant toujours les n premiers mots du vocabulaire.

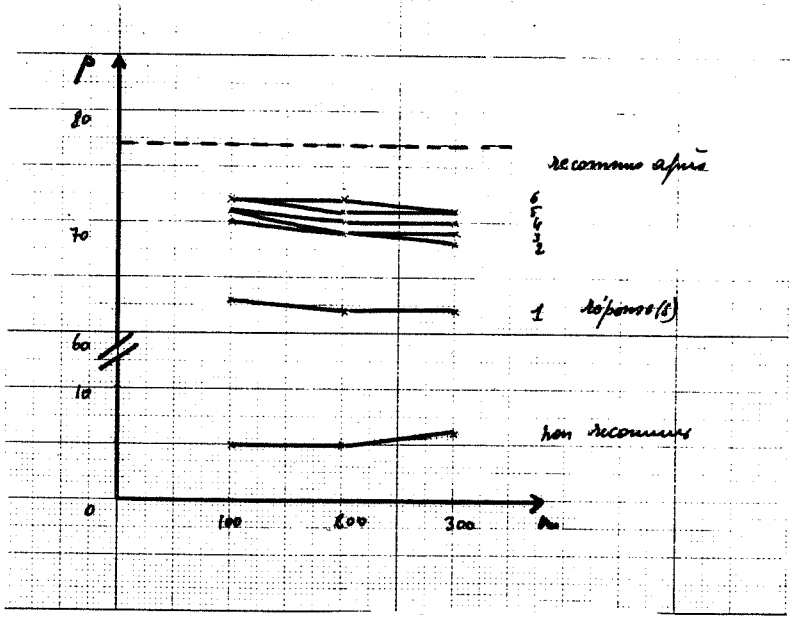
1. - Six mots différents (acheter, adroit, affiche, aider, argent, avion) ont été prononcés chacun 10 fois par le même locuteur. Les résultats sont (après élimination de 3 enregistrements mauvais) donnés par le graphique ci-dessous.



Ils sont d'autant plus satisfaisants que certaines erreurs proviennent de confusion entre mots très voisins comme aimer, aider et deux prononciations du mot acheter.

Un fait remarquable est la faible décroissance des performances selon la taille du vocabulaire considéré, la chute n'étant sensible qu'à partir de 300 mots.

2. - Les 77 premiers mots du dictionnaire ont été prononcés une fois par le même locuteur.



Les résultats sensiblement moins bons que précédemment font apparaître la nécessité d'une décision hiérarchisée. Ils restent néanmoins satisfaisants si le système de reconnaissance comporte un dispositif correcteur d'erreur.

Là encore la stabilité des performances est remarquable.

3. - Un troisième enregistrement des 88 premiers mots a été comparé à 100 mots de l'enregistrement initial.

mots reconnus après						non reconnus
1	2	3	4	6	réponse(s)	
78	6	1	1	1		1

L'utilisation d'un seuil de ressemblance donne des performances légèrement inférieures.

1	2	3	4	6	8	
74	8	1	2		2	1

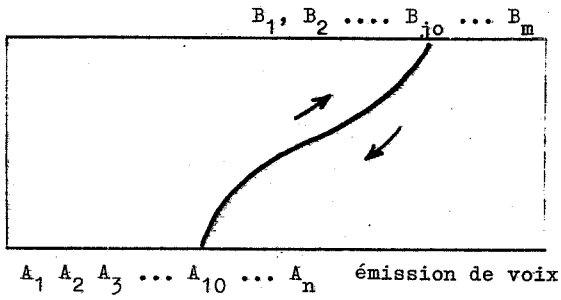
Les performances sont meilleures que celles de l'essai précédent. Cela tient sans doute à la qualité de l'enregistrement et au fait que les mots soient légèrement plus différents.

Je renvoie pour les difficultés d'emploi de la méthode à l'exposé de J-P. HATON. Le temps de calcul proportionnel à la taille du vocabulaire est très long (de plusieurs s à plusieurs 10 s sur CII 10070) ; c'est le prix qu'il semble falloir payer pour des performances que d'autres méthodes ne nous ont jusqu'ici jamais permis d'atteindre.

REPERAGE ET RECONNAISSANCE DE MOTS DANS LA PAROLE CONTINUE.

La reconnaissance des langages à mots clés, ou le repérage d'un mot dans un "bruit" de parole non pertinente est une extension du problème précédent.

Vu le nombre d'échantillons constituant une émission de voix il est impossible d'utiliser l'algorithme tel quel. Un repérage grossier permet de délimiter la zone de calcul. On utilise pour cela une petite partie centrale du mot, qu'il déplace le long de la phrase en recherchant une bonne coïncidence énergétique. 3 échantillons ont permis une localisation correcte dans 90 % des cas examinés.



Les limites du mot ainsi repérées ne sont pas connues. Il existe plusieurs débuts et fins possibles pour lesquels on peut définir une ressemblance S (Aj, Bj) (i = 1, n, j = i, m).

L'expérience a montré qu'il existait un point B_{jo} tel que :

$\forall i \quad S(Ai, Bjo) = mm\ [S(Ai, Bj)\ j = 1, m]]$

et un point A_{io} tel que :

$\forall j \quad S(A_{io}, Bj) = mm\ [S(Ai, Bj)\ \ i = 1, n]$

La ressemblance choisie entre le mot de référence et le mot prononcé est :

S(Aio, Bjo),

et par définition Aio et Bjo sont les limites du mot.

Le calcul est effectué en deux "parcours" où l'on recherche :

$S(Ai, Bjo) = \min\ [S(Ai, Bj)\ \ j = 1, m]$

$S(Aio, Bjo) = \min\ [S(Ai, Bjo)\ \ i = 1, n]$

L'algorithme est adapté à un pas négatif et positif.

Dans la pratique un seuil de ressemblance a été adopté, au delà duquel la recherche reprend.

Les résultats sont encore partiels. Ils portent sur le repérage de 21 mots prononcés par un même locuteur au moins deux fois dont une à l'intérieur des 32 phrases où l'on a cherché à les reconnaître. Ces phrases extraites du "premier Larousse en images" ont été composées à partir de 120 mots différents.

Par la même occasion 3 types de distance entre échantillons ont été essayés. Ceci pour atténuer l'effet de l'intensité d'élocution.

Le temps de calcul initialement long a été divisé par dix grâce à une réécriture des programmes.

Liste des "mots" recherchés

abeille	grand-mère	argent
action	allumé	
adroit	animal	
la flèche	animaux	
six ans	année	remarque : certains mots de la liste sont très voisins. Pour les phrases essayées figuraient également des mots très proches mais ne figurant pas sur la liste (faute de plusieurs occurences) ex : l'affiche la flèche adroit aboie
agent	appelle	
aile	nager	
parents	après	
l'an	arbre	
trois	voici	

résultats

Distance utilisée	Pourcentage de réponse correcte (rejet ou reconnaissance)	Temps de calcul (extrêmes et moyenne) (selon le second programme)		
Différence entre énergie	81,2	0,04 s	0,53 s	2 s
Différence des pentes entre canaux	84,4	0,04	0,56	2,3
Distance élaborée	87,5	0,04	0,69	2,7

CONCLUSION.

Appliquée à la reconnaissance des mots la méthode paraît intéressante mais, à moins de disposer d'une machine extrêmement rapide ou de réduire le vocabulaire à quelques dizaines de mots, il est difficile de l'utiliser telle quelle.

Un dernier mot sur l'influence du codage initial. ZAGORUJKO utilisait un codage grossier à travers cinq filtres. Le vocodeur fournit une analyse plus fine. Il semble toutefois que le principe de reconnaissance ne puisse s'accomoder d'une trop grande finesse qui ferait apparaître des variations du signal, jusqu'ici négligées.

Comment doit-on interpréter la quasi-absence d'influence de la taille du vocabulaire de référence sur les performances ? Sans doute par le fait que les premiers mots sont suffisamment distants des mots suivants. Nous aurions ainsi fait apparaître involontaiement des grappes ("clusters") dans l'espace de séparation.

REFERENCES BIBLIOGRAPHIQUES.

[1] G.S. SLUCKER Nelinejnyj metod analiza recevyh signalov, trudy N.I.I.R., n° 2, 1968, p. 76-82.

[2] V.M. VELIČKO, N.G. ZAGORUJKO Avtomatičkoe Raspoznavanie 200 ustnyh komand, Vyčislitel'nye sistemy, n° 37, 1969, p. 73-76.

[3] H. SAKOE, S. CHIBA (NEC) A dynamic programming approach to continuous speech recognition, 7e congrès international d'acoustique, Budapest, août 1971, CR Vol. 3, p. 65-68.

CONSTANTES ET VARIABLES DANS L'ESTIMATION

DE LA "FREQUENCE" DES PHONEMES EN FRANCAIS

Résumé

. Nous disposons d'une dizaine de listes de "fréquence" des phonèmes du français. S'il est absolument normal que les pourcentages diffèrent d'une liste à l'autre, la comparaison des résultats laisse néanmoins apparaître des différences parfois considérables tant au niveau des chiffres avancés qu'à celui du rang respectif des phonèmes. Or, les niveaux de langue et la petitesse de certains échantillons ne peuvent à eux seuls expliquer ces variations.

L'établissement de ces listes, en effet, n'a pas été sans difficultés; car chaque auteur ne comptabilise pas, en réalité, des éléments semblables, au point qu'à notre avis la cause principale des écarts enregistrés d'une liste à l'autre réside dans la transcription elle-même.

De ce fait, les réelles variations de "fréquence" qui existent dans le discours sont masquées par un manque d'unité au niveau de l'établissement des unités phonologiques. Un certain nombre de résultats, par exemple, donne à penser que le rapport Consonnes/Voyelles est pratiquement égal à 1. Or il n'est pas nécessaire de recourir à de fastidieux calculs pour constater que la proportion des phonèmes consonantiques est nettement supérieure à celle des phonèmes vocaliques en français parlé.

Comme l'a montré G.K. ZIPF, les phonèmes d'une langue ont des "fréquences" relativement stables. Dans les limites importantes offertes par les principales listes dont nous disposons, nous voudrions rappeler les constantes dans la répartition des fréquences uniquement d'après les principaux critères articulatoires.

Summary

We have around ten lists of "frequency" of french phonemes. It is perfectly normal that the percentages differ from one list to another. But the comparison of the results sometimes allows considerable differences, to appear at the level of advanced figures as well as at the respective level of the phonemes. Now, the level of language and the insignificance of certain samples cannot in and of

themselves explain these variations.

In fact, the establishment of these lists was accomplished not without difficulty because each author does not take account of similar elements, to the point, in our opinion, the principal cause of the differences recorded between the lists can be explained by the manner of transcription.

From this fact, the real frequency variations which exist in the speech are masked by a lack of unity at the level of the establishment of phonological unities.

A certain number of results lead us to think that the relationship Consonant/Vowel is practically equal to 1. Now, it is not necessary to resort to fastidious calculations to see that the proportion of consonant phonemes is clearly greater than that of vocal phonemes in spoken french.

As G.K. ZIPF has shown the phonemes of a language have relatively stable frequencies. Within the important limits offered by the principal lists which we have, we would like to recall the constants according to the main articulatory criteria.

Fr. WIOLAND
Institut de Phonétique de Strasbourg.

CONSTANTES ET VARIABLES DANS L'ESTIMATION DE LA "FREQUENCE" DES PHONEMES EN FRANCAIS

G.K. ZIPF en I939 (1), P. CHAVASSE en I948 (2), A. VALIMAN en I956 (3), le Docteur J.Cl. LAFON en I96I (4), Monsieur P. GUIRAUD en I963 (5), P. DELATTRE en I965 (6), Messieurs J.P. HATON et M. LAMOTTE en I97I (7), Monsieur M. HUG en I972 (8) et nous-mêmes (9), entre autres, publièrent des listes de fréquence des phonèmes du français.

On est en droit de se demander dans quelle mesure elles sont représentatives de la "Langue", dont on ne peut d'ailleurs parler qu'en termes de probabilité, car :

1 - Les pourcentages ne sont obtenus qu'à partir de petits échantillons de "Discours" :

(1) G.K. ZIPF et F.M.ROGERS, Phonemes and variphones in four present-day romance languages and classical latin from the viewpoint of dynamic philology, in Archives Néerlandaises de Phonétique expérimentale, t. XV., I939, pp.111-147.

(2) Essai sur la phonétique statistique de la langue française et son application à l'étude de l'intelligibilité d'une conversation, Annales de Télécommunications, t.III, n°1, Janv. I948.

(3) Les bases statistiques de l'antériorité articulatoire du français, Le Français Moderne, 27ème année, n°1, Janv.I959, pp.I02-II0.

(4) Message et Phonétique, P.U.F., I961, pp.I42-I46.

(5) Structure aléatoire de la double articulation, Bulletin de la Société de Linguistique de Paris, t.58, fasc.1, I963, pp.I35-I55.

(6) Comparing the phonetic features of English, German, Spanish and French, Heidelberg, I965, p.97.

(7) Etude statistique des phonèmes et diphonèmes dans le français parlé, Revue d'Acoustique, n°I6, I97I, pp.258-262.

(8) La distribution des phonèmes est-elle aléatoire ? Bulletin de l'Institut de Phonétique de Grenoble, Vol.1, I972, pp.93-I07.

(9) Estimation de la "fréquence" des phonèmes en français parlé, Travaux de l'Institut de Phonétique de Strasbourg, n°4, I972, pp.I77-204. La liste des fréquences a été publiée, en partie, par M. J. CALVET, Etude phonétique des voyelles du Wolof, l'Enseignement du français en Afrique, XIV, I965, p.13.

• ZIPF a utilisé la Chrestomathie de Passy,

• CHAVASSE, des extraits d'ouvrages littéraires des siècles derniers,

• Le Docteur LAFON, une partie des bandes d'enregistrement qui ont servi à l'élaboration du français élémentaire,

• VALDMAN, DELATTRE, HATON et LAMOTTE et nous-même, divers enregistrements,

• HUG, des extraits du Canard Enchainé et des Propos d'Alain.

2 – La longueur des échantillons varie dans un rapport de 1 à 15 : de 5000 phonèmes chez ZIPF à 77.702 pour notre échantillon en passant par IOOOO (GUIRAUD, LAFON, HUG), 26896 (VALDMAN), 30000 (CHAVASSE) et 50033 (HATON/LAMOTTE).

Mais comme les échantillons ne sont évidemment jamais semblables, fussent-ils d'un même niveau de langue, il est tout à fait normal que les pourcentages diffèrent d'un auteur à l'autre. En pratique, nous faisons confiance à la liste que nous avons sous les yeux : J. PEYTARD et E. GENOUVRIER par exemple, qui publient en I970 (1) une liste de fréquence des phonèmes, reprennent la liste publiée par P.R. LEON en I966 (2) qui lui-même reprend celle établie en I96I par J.Cl. LAFON; ou alors nous prenons le temps d'en établir une nouvelle à partir d'échantillons qui nous paraissent plus représentatifs et nous constatons, à notre grande satisfaction ou à notre grand étonnement, des différences parfois importantes avec les résultats antérieurs.

---

(1) Linguistique et Enseignement du français, Larousse, Paris, I970, p.42.

(2) Prononciation du français standard, Didier, Paris, I966.

Une comparaison, dans la mesure ou l'on peut employer ce terme, des résultats publiés, nous a néanmoins semblé utile : Voir le tableau 1.

Quelques tendances générales se dégagent à première vue : certains phonèmes sont toujours plus fréquents que d'autres.

Mais une comparaison plus approfondie laisse perplexe :

- le rapport du pourcentage le plus élevé sur le plus bas - par phonème, atteint plus de 8,3 pour /ə/.

7 pour /ɥ/, 4,4 pour /j/, 4 pour /z/ et /ɛ̃/, 3,5 pour /R/ et /ɔ/, 3 pour /ø/, 2,5 pour /ɡ/, /oe/ et /ɛ/, 2,3 pour /b/ et /ʒ/, 2 pour /œ̃/, /o/, /v/, /ɲ/, /e/, /y/ et /ʃ/, 1,8 pour /l/, ce qui est pour le moins étonnant.

- On retrouve également cette disparité au niveau du rang des fréquences : le phonème donné comme le plus fréquent par les neuf listes se trouve être :

4 x /R/, 3 x /a/ et 2 x /e/.

Dans les autres cas : /R/ occuperait 4 x le 2ème rang
1 x le 17ème rang

/a/ occuperait 1 x le 2ème "
3 x le 3ème "
1 x le 4ème "
1 x le 7ème "

/e/ occuperait 1 x le 2ème "
1 x le 4ème "
3 x le 6ème "
1 x le 7ème "
1 x le 10ème "

Parmi les autres phonèmes les plus fréquents :

/l/ occuperait 1 x le 2ème rang
3 x le 3ème "
3 x le 4ème "
1 x le 5ème "
1 x le 9ème "

/ə/ occuperait 1 x le 2ème rang
1 x le 7ème "
3 x le 8ème "
2 x le 12ème "
1 x le 13ème "
1 x le 26ème "

/ɥ/ occuperait 1 x le 2ème "
1 x le 3ème "
1 x le 4ème "
3 x le 6ème "
1 x le 7ème "
1 x le 8ème "
1 x le 10ème "

Si pour certains phonèmes de fréquence très peu élevée les écarts relevés peugent à la rigueur être considérés comme logiques, il n'est néanmoins pas possible de les attribuer au choix des échantillons : ce n'est pas, en effet, parce qu'il s'agit de français lu, ou récité, ou parlé dans quelque situation particulière que ce soit, que la fréquence d'un /j/ ou d'un /z/, par exemple, peut être quadruplés, pas même celle d'un /ə/. Seule la petitesse des échantillons pourrait expliquer une partie des écarts. Il est évident qu'à partir de très petits échantillons la fréquence et le rang de certains de ces phonèmes peut varier de façon notoire. Mais le plus grand écart enregistré par HUG qui compare 5000 phonèmes des Propos d'Alain et 5000 phonèmes de La Mare au Canard, échantillons de niveau de langue très différents, ne dépasse le rapport de 2 que pour les phonèmes /g/ et /ŋ/, phonèmes peu fréquents : 0,34 contre 0,80 et 0,02 contre 0,14% respectivement (1); dans tous les autres cas, aucune commune mesure avec les écarts qui paraissent sur le tableau 1. Or le plus petit échantillon, celui de ZIPF est réalisé à partir de 5000 phonèmes.

La cause principale des variables enregistrées est d'un autre ordre. Il nous a, en effet, été très difficile d'établir ces listes

(1) ouvr. c. pp. 103 et 105

"comparatives". Nous avons constaté que le nombre des phonèmes retenus variait d'un auteur à l'autre, qu'un même phonème pouvait être comptabilisé sous deux aspects différents ou confondu avec un autre, au point que nous nous demandons si toute comparaison entre les différentes listes est non seulement justifiée mais possible. Car, en définitive, chaque auteur ne transcrit pas de la même façon, ce qui revient à dire que les pourcentages ne correspondent pas en réalité à des éléments semblables. Les écarts, de ce fait, peuvent, en effet, être considérables.

Les variables qui apparaissent sur les listes de fréquence des phonèmes du français sont donc dues, pour la plus grande part, non pas à la nature des échantillons comme on serait en droit de le penser, mais tout simplement à la transcription elle-même. Loin de nous la pensée qu'il n'existe qu'une transcription : le choix est souvent difficile, doit-elle correspondre à une théorie ou à ce qui a été, non pas prononcé par le locuteur, fut-ce avec "l'accent" d'une région de France, mais perçu par l'auditeur, en l'occurence celui qui transcrit ? Or nous savons qu'il y a de fortes chances pour que dans ce cas la même personne ne transcrive pas deux fois de la même façon le même stimulus sonore. S'agit-il d'autre part du français parlé du théâtre, de la radio, de la télévision, d'une conférence, récité en quelque sorte, stéréotypé parfois, ou vraiment spontané soit sous forme de monologue, soit sous forme de dialogue avec toutes les hésitations, répétitions, inhérentes à un tel niveau de langue ? Mais de toute façon, en bonne logique, à quoi bon établir des listes de fréquence si le mode de comptage n'est pas aussi rigoureux que possible ?

De ce fait, les réelles variations de "fréquence" qui existent dans le Discours sont masquées par un manque d'unité des transcriptions au niveau des unités phonologiques.

D'autre part, dans le Discours, comme l'a montré ZIPF, les phonèmes ont des fréquences "stables" à l'intérieur de certaines limites; les écarts sont évidemment plus importants pour certains

phonèmes tels que /ə/, les phonèmes de liaison, ou pour des raisons thématiques.

Or ce n'est pas cette stabilité qui est mise en évidence par ce tableau "comparatif".

Il est aisé de le démontrer en considérant le rapport Consonnes/Voyelles chez les différents auteurs :

Phonèmes

	Consonantiques %	Vocaliques %
Chavasse	50,69	49,31
Lafon	50,9	49,1
Haton/Lamotte	51,96	48,04
Hug	54,54	45,46
Valdman	54,8	45,2
Wioland	56,5	43,5
Delattre	56,8	43,2

Les premiers pourcentages donnent à penser que le nombre des phonèmes consonantiques et vocaliques est pratiquement égal en français parlé. Il n'est pas nécessaire de recourir à de fastidieux calculs pour constater que la proportion des phonèmes consonantiques est nettement supérieure à celle des phonèmes vocaliques. Ce rapport varie évidemment selon le niveau de langue du fait de la prononciation des /ə/. Nous avons constaté en dépouillant 5 émissions radiophoniques de style très différent (1) soit 31.731 phonèmes, que plus le locuteur soigne sa prononciation, plus le pourcentage de phonèmes vocaliques augmente. C'est, en effet, dans l'émission où J. VILAR monologue en

(1) ouvr. c. p.179

a) Constatations :

2 phonèmes	/R/ et /l/	représentent environ	25% (1)	des occurences des phonèmes consonantiques
3 "	+ /S/	"	35%	"
5 "	+/t/, /d/	"	50% (2)	"
7 "	+/p/, /k/	"	66%	"
9 "	+/m/,/n/	"	77%	"
10/20 "	+/v/	"	82,5%	"
13 "		"	90%	"

La première moitié des phonèmes consonantiques représente donc 82,5%

La dernière moitié " " 17,5%

b) Répartition selon les critères articulatoires

Voir le tableau 3. Quelques remarques à propos de ce tableau:

. Répartition selon la participation des cordes vocales :

6 sourdes : 37,35%

14 sonores : 62,65%

Les quatre sourdes /p/,/t/,/k/ et /s/ sont responsables de plus du tiers des occurrences des consonnes : 33,81%.

Si l'on compare les sourdes et les six sonores qui leur correspondent, la règle établie par ZIPF qui veut que dans la plupart des langues les sourdes soient plus fréquentes que les sonores se vérifie :

6 sourdes : 37,35% / 6 sonores orales correspondantes : 20,70%

Mais il semble que cette règle s'applique surtout en français aux occlusives :

---

(1) Cf. VALDMAN, ouvr. c., p.109 et LAFON, ouvr. C. p.142

(2) HATON/LAMOTTE obtiennent ce pourcentage avec les seules quatre premières consonnes, ouvr. c.p. 260.

3 occlusives sourdes : 23,54%

3 occlusives orales et sonores: I0,62% soit plus double de sourdes; ce rapport ne se vérifie pas pour chaque paire d'occlusives ayant un même lieu d'articulation, mais la sourde est toujours plus fréquente que la sonore correspondante.

Il en va tout autrement pour les constrictives :

3 constrictives sourdes : 13,81%
3 constrictives sonores correspondantes: 10,08%

La différence, peu importante, n'est favorable aux sourdes qu'à cause de /s/ : 10,27%, nettement plus fréquent que sa correspondante sonore : 2,65%. Dans les deux autres séries, labio-dentales et post-alvéolaires, ce sont les sonores qui sont les plus fréquentes :

/v/ : 4,78% /ʒ/ : 2,65%
/f/ : 2,48% /ʃ/ : 1,06%

• Répartition selon le mode articulatoire :

9 occlusives : 45,84%
11 constructives : 54,16%

Il est intéressant de noter que la répartition selon ce critère est proportionnelle au nombre d'articulations considérées. Les différences de fréquence ne se produisent qu'à l'intérieur de chacune des deux séries.

• Répartition selon les lieux d'articulation :

Voir le tableau 4.

Les deux lieux d'articulation intrabuccaux les plus antérieurs sont statistiquement les plus importants (1).

S'il est indéniable que le mode du français est plutôt antérieur (2), il n'en demeure pas moins que les articulations consonantiques postérieures sont responsables de 23,7% des occurences; or ce pourcentage n'est pas particulièrement faible,

Les articulations les plus antérieures, les extrabuccales, et les articulations les plus postérieures ont des fréquences d'occurrence pratiquement semblables : 22,64% et 23,70%; mais pour les premières au bénéfice des occlusives (I5,38%), pour les autres, à celui des constrictives (I5,58%), fréquences également très rapprochées.

On retrouve une répartition comparable entre les occlusives et les constrictives intrabuccales les plus antérieures : 22,16% et 23,89%.

Comme on peut le constater sur le tableau 5, ces quatre groupes d'articulations sont d'importance sensiblement égale.

---

(1) Voir à ce sujet ZIPF, ouvr. C.,
W. MANCZAK, Fréquence d'emploi des occlusives labiales, dentales et vélaires, Bulletin de la Société de Linguistique de Paris, t. 54, fasc.1, I959, pp. 208-2I4.

(2) A. VALDMAN, ouvr. c., pp. I02-II0.
P. DELATTRE, Les modes phonétiques du français, Studies in French and Comparative phonetics, Mouton, I966, pp.9-13.

Les articulations intrabuccales antérieures sont donc deux fois plus nombreuses que les articulations intrabuccales postérieures d'une part, et que les articulations extrabuccales d'autre part. Voir le tableau 6.

2 - Phonèmes Vocaliques

Les pourcentages sont exprimés par rapport aux seuls phonèmes vocaliques.

a) Constatations :

1	phonème	/a/	représente environ	17,5%	des occurences des phonèmes vocaliques
2	"	+ /ɛ/ "	"	30%	"
3	"	+ /e/ "	"	43%	"
4	"	+ /i/ "	"	55% (1)	"
7	"	+ /ɔ/, /ɑ̃/, /u/	" +	75%	"
8/16	"		"	81,5%	"
10	"		"	90%	"

La première moitié des phonèmes vocaliques représente donc 81,5%

La dernière moitié " " 18,5%

La similitude des pourcentages avec les phonèmes consonantiques est remarquable.

---

(1) cf. LAFON, ouvr. c., p.142

b) Répartition selon les critères articulatoires :

• Répartition selon les lieux d'articulation :

10 voyelles antérieures (1) : 74,02% soit 3/4 environ
6 voyelles postérieures : 25,98% soit 1/4 environ

• Répartition selon les mouvements des lèvres :

5 voyelles non labialisées : 57,69%
11 voyelles labialisées : 42,31%

La proportion des non-labialisées est très forte, les quatre premières voyelles étant, en effet, les non-labialisées orales.

• Répartition selon la position du voile :

12 voyelles orales : 83,45%
4 voyelles nasales : 16,55%

La proportion des voyelles nasales est faible (2).
La voyelle /ɑ̃/ est responsable de presque la moitié des occurences des voyelles nasales, les deux voyelles postérieures de plus des 3/4.

• Répartition selon l'aperture : Voir le tableau 7

Sont relativement fréquentes dans l'ordre croissant :
- les voyelles de petite aperture
- la voyelle /ə/
- et surtout les voyelles de grande aperture.

Sont relativement peu fréquentes :
- les voyelles d'aperture moyenne
- et surtout les voyelles nasales /ɛ̃/, /œ̃/, /ɔ̃/.

A l'intérieur de ces groupes on peut relever les tendances suivantes :

---

(1) Y compris /ə/, selon les critères articulatoires traditionnels de classement.

(2) Cf. DELATTRE, Nasalité vocalique en français et en anglais, French Review, t. XXXIX, octobre 1965, n°1, pp.103-104.

- Voyelles orales :

A chaque degré d'aperture, la voyelle de la série antérieure non labialisée a toujours une fréquence très nettement supérieure à celle des deux autres séries réunies. La voyelle de la série postérieure est également plus fréquente que celle de la série antérieure labialisée, comme on peut le constater sur le tableau 8.

Les voyelles non labialisées à deux timbres sont trois fois plus fréquentes que celles de la série postérieure, elles-mêmes trois fois plus fréquentes que celles de la série antérieure non labialisée.

Il convient enfin de souligner la fréquence de /ə/ par rapport à celle des voyelles à deux timbres de la série antérieure labialisée /ø/ et /oe/ : 7,82% contre respectivement 1,38% et 1,15%.

- Voyelles nasales :

Ce sont les voyelles postérieures qui sont les plus fréquentes, mais comme pour les voyelles orales, c'est la voyelle de la série antérieure labialisée qui est la moins fréquente. Voir le tableau 8.

Lorsqu'on évoque l'antériorité des articulations vocaliques du français (1), il faut se garder de croire qu'elle se réalise au préjudice des seules voyelles postérieures; comme pour les consonnes, c'est la série intermédiaire qui est la plus touchée comme on peut le voir sur le tableau 9.

---

(1) A. VALDMAN, Les bases statistiques de l'antériorité vocalique du français. French Review, vol. XXXI, octobre 1957, n°1, pp.317-321. P. DELATTRE, ouvr. c., (Studies...), pp.10-11.

CONCLUSION

. Si l'on considère l'ensemble des phonèmes:

- le rapport consonnes/voyelles
- la distribution selon le mode articulatoire
- la distribution selon les lieux d'articulation

sont pratiquement proportionnels au nombre d'articulations intéressées, alors que

- les articulations sourdes sont plus fréquentes, et que
- les articulations nasales et labialisées le sont moins.

Il existe donc une stabilité de répartition des phonèmes au niveau des principaux critères articulatoires malgré les variations très importantes d'occurrences d'un phonème à l'autre.

Mais si l'on considère séparément articulations vocaliques et articulations consonantiques, la distribution selon les lieux d'articulation et selon les degrés d'aperture(voyelles) n'est plus proportionnelle au nombre d'articulations concernées.

. 4 phonèmes: /R/,/a/,/l/,/s/ sont responsables de plus du quart des occurrences: 27,4%

8 phonèmes: +/ɛ/,/e/,/i/,/t/, de la moitié: 49%.

15 phonèmes, des 3/4.

Les 18 premiers phonèmes de 82%.

Les 18 derniers phonèmes de 18%.

La question qui se pose est de savoir pourquoi certains phonèmes, les mêmes dans la plupart des langues, ont un rendement nettement supérieur à d'autres. On constate que les phonèmes les plus fréquents sont également les plus perceptibles. /R/ et /l/, étant seuls dans leur catégorie, ont une vaste latitude articulatoire; /a/est une articulation de base: intensité la plus grande pour le moindre effort articulatoire; /s/ a les bruits dans les zones de fréquence les plus élevées. C'est d'autre part la partie antérieure de la langue, région la plus innervée de tous les organes articulatoires, qui se trouve être "l'organe articulatoire" de 8 des 9 phonèmes les plus fréquents (1).

C'est donc l'économie des mouvements articulatoires qui semble régir la fréquence d'occurrence des phonèmes.

(1) Sans oublier que /r/ fut antérieur en français jusqu'au XVII° siècle.

C'est de propos délibéré que nous nous sommes contenté d'exprimer la fréquence des phonèmes en pourcentages, et nous sommes conscient des risques encourus par toute comparaison de pourcentages. Nous voulions simplement montrer que ni les variations en fonction des échantillons, ni les constantes au niveau articulatoire, n'étaient mises en évidence sur ces différentes listes.

Nous travaillons actuellement avec un ensemble plus important de phonèmes, sur fiches perforées, d'après la méthode statistique (1), en nous intéressant non seulement à la fréquence des phonèmes mais à leur distribution en fonction de leur position dans le groupe rythmique et de leur entourage.

F. WIOLAND

Institut de Phonétique de Strasbourg

---

(1) MONJALLON A. Introduction à la méthode statistique. Vuibert,Paris, 1958.
TORTRAT A. Principes de statistiques mathématiques, Dunot, Paris, 1961.
REEB G. et FUCHS A. Statistiques commentées, Gauthier-Villars, Paris, 1967.
MULLER Ch. Initiation à la statistique linguistique,Larousse, Paris, 1968.

T1 352	ZIPF	CHAVASSE	VALDMAN	LAFON	GUIRAUD	DELATTRE	WIOLAND	HATON/ LAMOTTE	HUG
R	7,44	7,4	7,6	6,9	2,54	8,67	7,58	8,37	8,12
a	5,60	5,3	6,5	8,1	7,42	7,04	8,11	7,50	6,52
l	6,88	6,4	6	6,8	3,8	6,14	5,89	6,71	6,28
s	5,06	5,6	6,4	5,8	6,46	5,06	5,75	5,90	6,02
ɛ	5,88	3	4,5	5,3	6,64	2,83	5,55	4,81	5,92
e	4,98	7,5	8,2	6,5	5,07	8,14	5,28	5,82	4,32
t	4,9	5,4	5,3	4,5	6,34	5,59	5,39	4,93	5,01
i	4,50	5,8	5,9	5,6	6,90	5,23	5,08	6,79	5,26
d	4,54	4,5	4	3,5	3,56	4,18	4,24	4,51	4,94
p	3,96	3,4	3,5	4,3	3,57	4,60	3,88	3,94	3,93
k	3,32	3,8	4,9	4,5	4,85	3,67	3,75	3,90	3,83
m	3,42	3,2	3,8	3,4	2,58	3,46	3,91	2,98	3,14
ə	4,86	5,2 4,9	3,7	4,9	0,90	3,21	3,39	5,42	7,53
ɑ̃	3,52	3,3	3,5	3,3	3,87	3,20	3,21	3,73	3,15
n	3,04	2,4	2,7	2,8	3,22	3,02	3,09	2,63	2,77
v	3,46	2	2	2,4	2,~~22~~	2,57	3	2,46	1,73
u	3,48	2,1	2,4	2,7	2,41	2,70	2,62	2,28	2,03
ɔ̃	1,48	2	2,2	2	3,03	1,62	2,27	2,13	2,43
ɣ	1,46	2,7	1,8	2	2,74	1,98	2,01	2,46	2,25
o	1,44	2,2	1,1	1,7	/	1,10	1,97	1,07	1,40
j	1,92	0,6	1,8	1,5	0,74	1,86	1,76	0,43	1,57
ʒ	1,98	1,2	1,4	1,7	0,99	1,67	1,57	1,23	0,85
z	2	1,5	1,4	0,6	2,37	1,35	1,55	0,75	1,50
f	1,26	1,3	1	1,3	1,45	1,48	1,38	1,44	1,26
ɔ	2,14	0,9	1,9	1,5	3,28	2,13	1,28	2,16	1,96
ɛ̃	0,36	0,9	1,2	1,4	1,12	1,03	1,16	0,99	1,19
b	1,82	1,1	0,8	1,2	1,4	1,31	1,08	1,18	1,15
w	1,14	0,8	0,9	0,9	0,8	1,33	1,03	0,94	0,91
ʃ	0,76	0,5	0,5	0,5	0,39	0,57	0,61	0,56	0,52
g	0,36	0,5	0,4	0,3	0,76	0,65	0,56	0,54	0,57
ø	0,42	0,6	1,3	0,6	0,97	0,72	0,51	/	0,54
œ	0,30	0,5	/	0,3	/	0,76	0,44	0,60	0,38
œ̃	0,56	0,5	0,5	0,5	0,27	0,44	0,54	0,49	0,54
ɥ	0,68	/	0,3	0,7	0,10	0,49	0,37	0,20	0,36
ʏ	0,12	0,1	0,07	0,1	0,15	0,15	0,14	/	0,08
ɑ	0,94	0,6	0,5	0,2	/	0,01	0,05	/	/

PHONEMES CONSONANTIQUES	%	RANG	PHONEMES VOCALIQUES	%	
/R/	7,8	1			
		2	/a/	7,6	
/l/	6,2	3			
/s/	5,8	4			
		5	/ɛ/	5,6	11
		6	/e/	5,4	
/t/	5,3	7	/i/	5,3	
/d/	4,3	9			
/p/	4	10			
/k/	4	10			
/m/	3,6	12			
		13	/ə/	3,4	
		13	/õ/	3,4	
/n/	2,9	15			
/v/	2,7	16			
		17	/u/	2,5	
		18	/ɔ̃/	2,1	
		18	/y/	2,1	
		20	/o/	1,9	3,2
/j/	1,8	21			
/ʒ/	1,5	22			
/z/	1,5	22			
/f/	1,4	24			
		25	/ɔ/	1,3	
		26	/ɛ̃/	1,2	
/b/	1,1	27			
/w/	1	28			
/ʃ/	0,6	29	/ø/	0,6	1,1
/g/	0,6	29			
		32	/œ/	0,5	
		32	/œ̃/	0,5	
/ɥ/	0,3	34			
/ɲ/	0,1	35	/ɑ/	0,1	
	56,5			43,5	

Tableau 3:

354	Lieux d'articulation							Totaux:%
	/p/ 7,06		/t/ 9,42				/k/ 7,06	23,54
	/b/ 1,95		/d/ 7,61				/g/ 1,06	10,62
	/m/ 6,37		/n/ 5,13			/ɲ/ 0,18		11,68
Totaux:%	15,38		22,16			0,18	8,12	45,84
		/f/ 2,48		/s/ 10,27	/ʃ/ 1,06			13,81
		/v/ 4,78		/z/ 2,65	/ʒ/ 2,65			10,08
						/j/ 3,19	/w/ 1,77	
						/ɥ/ 0,53		5,49
				/l/ 10,97			/R/13,81	24,78
Totaux:%		7,26		23,89	3,71	3,72	15,58	54,16

Tableau 4:

Lieux d'articulation	Nombre de phonèmes	%
Bilabial	3	15,38
Labio-dental	2	7,26
Alvéo-dental	3	22,16
Alvéolaire	3	23,89
Post-alvéolaire	2	3,71
Palatal	3	3,90
Vélaire	4	23,70
	20	100

Tableau 5:

Articulations	%	Nombre de phonèmes
Extrabuccales	22,64	5
Occlusives alvéo-dentales	22,16	3
Constrictives alvéolaires	23,89	3
Vélaires	23,70	4
Post-alvéolaires/Palatales	7,61	5

Tableau 6:

	Extrabuccales	Intrabuccales		
		Antérieures	Série interméd.	Postérieures
%	22,64	46,05	7,61	23,70
Rapports	3	6	1	3

Tableau 7:

Nombre de phonèmes	Phonèmes	%
3	/i/ /y/ /u/	22,76
3	/e/ /ø/ /o/	18,16
1	/ə/	7,82
3	/ɛ/ /œ/ /ɔ/	17,01
3	/ɛ̃/ /œ̃/ /ɔ̃/	8,73
3	/a/ /ɑ/ /ɑ̃/	25,52

Tableau 8:

	+ fréquence		-
ORALES	Série antérieure non labialisée	Série postérieure labialisée	Série antérieure labialisée
	/i/ 12,18 /e/ 12,41 /ɛ/ 12,87 /a/ 17,47	/u/ 5,75 /o/ 4,37 /ɔ/ 2,99 /ɑ/ 0,23	/y/ 4,83 /ø/ 1,38 /œ/ 1,15
	e/ɛ 25,28	o/ɔ 7,36	ø/œ 2,53
NASALES	Voyelles postérieures	Voyelle antérieure non labialisée	Voyelle antérieure labialisée
	/ɔ̃/ 4,82 /ɑ̃/ 7,82	/ɛ̃/ 2,76	/œ̃/ 1,15

Tableau 9:

Série	Nombre de phonèmes	%
Antérieure non labialisée	5	57,69
Postérieure labialisée	6	25,98
Antérieure labialisée	4	8,51
/ə/	1	7,82

P. JUBAN : Dans votre tableau n° 7 donnant le rapport de la fréquence des phonèmes vocaliques avec l'aperture, vous avez remarqué que la fréquence était plus forte pour les phonèmes d'aperture plus fermée et d'aperture plus ouverte, et plus faible pour les apertures moyennes. C'est peut-être une interprétation abusive dans la mesure où la neutralisation est possible entre les degrés d'apertures moyennes (e/ɛ, ø/œ, o/ɔ), neutralisation qui est impossible avec les degrés supérieur ou inférieur. On obtient alors, selon vos calculs, une fréquence de 35 % pour les apertures moyennes, donc supérieure aux apertures fermées et ouvertes.

F. WIOLAND : Il est bien évident que les rapports établis n'ont de valeur qu'en fonction des regroupements opérés. Rien n'empêche de considérer trois niveaux d'aperture pour les voyelles orales. Mais s'il est vrai qu'en position inaccentuée la "neutralisation" entre les voyelles d'aperture moyenne peut être relativement fréquente, il en va tout autrement en position accentuée. Comme les dépouillements ont été réalisés à partir d'échantillons de français réellement prononcé et que les cas ambigus ont été peu nombreux, il nous a semblé préférable de considérer 4 degrés d'aperture.

J-Y. GRESSER : Pensez-vous que ces statistiques aient un sens dans la chaîne parlée ?

F. WIOLAND : Le but de notre intervention était justement de montrer, en comparant les résultats publiés, combien il était difficile de leur attribuer une valeur. Nous pensons par contre, que si les échantillons de discours étaient :
- plus importants
- plus diversifiés
- transcrits selon un même principe
un traitement statistique des résultats, avec possibilité de vérification sur l'enregistrement serait certainement révélateur de certaines tendances du français parlé. Il ne s'agirait évidemment pas d'une simple fréquence absolue ou relative, mais de relations établies en fonction de la position des phonèmes
- dans la syllabe
- dans le mot
- par rapport à l'accent
- les uns par rapport aux autres, etc...
Nous y travaillons actuellement.
Nous pensons néanmoins avoir rappelé certaines constantes non négligeables au niveau articulatoire.

J-S. LIENARD : Les différences observées dans les distributions des phonèmes publiées par différents auteurs sont très édifiantes, et proviennent sans doute, comme vous le remarquez, des différentes options adoptées lors de la transcription phonétique. Mais, j'ai été surpris par votre interprétation de la distribution des phonèmes comme un reflet de l'économie des mouvements articulatoires : d'une part, il est bien difficile de parler d'économie, ou de coût, en matière de la physiologie; d'autre part la répartition des phonèmes, éléments statiques, ne donne pas d'indications sur les mouvements articulatoires, qui correspondent aux transitions entre phonèmes. Par contre, je crois que l'étude de la répartition des diphonèmes, déjà abordée par plusieurs chercheurs français, pourrait être extrêmement fructueuse en ce domaine.

F. WIOLAND : Une étude distributionnelle peut être faite sur la base des seules relations structurales existant entre des phonèmes déjà inventoriés, sans références à leurs manifestations phonétiques. Mais il n'en n'est pas moins vrai que l'étude des faits distributionnels, réalisés ici au niveau perceptif, révèle les tendances articulatoires d'une langue. Les propriétés distributionnelles des phonèmes s'appuient sur leurs propriétés articulatoires. Comme vous le dites très justement cette étude trouve son prolongement naturel au niveau des habitudes combinatoires de la langue qui nous intéresse. Puisque l'on constate une économie au niveau du rendement distributionnel, pourquoi refuser ce terme au niveau articulatoire ?

Mme LHOTE : A votre question : " La question qui se pose est de savoir pourquoi certains phonèmes, les mêmes dans la plupart des langues, ont un rendement nettement supérieur à d'autres " soulevée à la fin de l'exposé sur la cause de la grande fréquence (d'emploi) de certains phonèmes que l'on retrouve dans d'autres langues, on pourrait faire remarquer que les sons les plus employés sont, pour les voyelles et les consonnes, des sons aigus. Il n'est certainement pas sans intérêt de remarquer que notre oreille semble éprouver dans la communication parlée une prédilection pour les aigus, en particulier dans l'apprentissage d'une langue étrangère.

En ce qui concerne les variations des résultats selon les auteurs, on peut attirer l'attention sur /z/ dont les pourcentages varient de 2.37 à 0.6; si l'on se rappelle que /z/ est un phonème de liaison très employé dans les pluriels français en langue parlée, le résultat 0.6 laisse à penser .

LA SIMULTANEITE DES INDICES DANS LA PERCEPTION DU VOISEMENT DES OCCLUSIVES

Résumé

En se limitant aux occlusives orales du français, on relève au moins six indices du trait de voisement :

- vibrations laryngées
- durée de la tenue
- durée de la voyelle préconsonantique
- transition de F1
- explosion : durée et caractéristiques spectrales
- durée d'établissement du voisement (VOT)

Dans le cadre de syllabes VC -, nous avons réduit dans des proportions variables la durée de la tenue des trois occlusives sourdes tout en conservant intact l'ensemble des autres éléments acoustiques.

Les résultats d'une tâche d'identification montrent que la durée de la tenue constitue un indice important pour la perception du trait de voisement mais qui ne peut, à lui seul, déterminer de manière non ambiguë le caractère sourd ou voisé de la consonne.

Summary

Confining oneself to the oral stops of French, one would find the correlates of voicing to be the following :

- laryngeal vibrations
- duration of closure
- duration of vowel preceding the consonant
- F1 transition in this vowel
- burst : duration and spectral characteristics
- voice onset time

Varying sections of the closure segment of voiceless consonants in a syllabe (VC) have been segmented without affecting the other acoustical cues.

The results of an identification task show that the duration of the closure segment constitutes an important cue in the perception of the voicing feature.

However, taken alone, this cue cannot determine the voiced or voiceless character of the consonant in an unambiguous manner.

Willy SERNICLAES
Université Libre de Bruxelles

La recherche des corrélats acoustiques a mis en évidence la multiplicité des indices susceptibles d'intervenir dans la perception d'un seul trait phonétique.

Ainsi que l'a souligné FANT (1967), le concept de trait distinctif se retrouve sur le plan perceptif à condition de ne pas l'assimiler à un seul paramètre important.

En se limitant aux occlusives orales du français, WAJSKOP et SWEERTS (1973) ne relèvent pas moins de six indices, supports potentiels de l'opposition entre sourdes et voisées (voir figure 1) :

- présence de vibrations laryngées dans la tenue de la voisée.
- durée de la tenue, plus brève pour la voisée.
- durée de la voyelle préconsonantique, plus longue devant la voisée.
- transitions de F_1 de cette voyelle, plus longues pour la voisée.
- l'explosion est plus brève pour la voisée. Les détentes des occlusives sourdes et voisées présentent également des différences dans leur composition spectrale (qui n'ont pas été reprises dans la figure 1).
- l'établissement du voisement de la voyelle postconsonantique (V.O.T.), plus rapide pour la voisée.

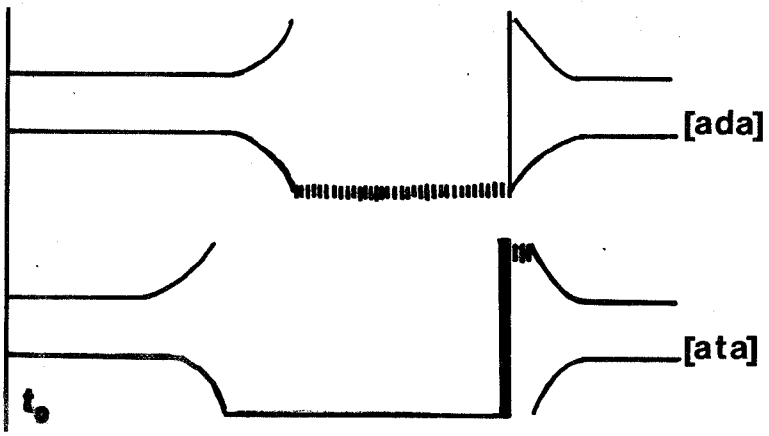


Figure 1 - Schémas des séquences /ata/ et /ada/ illustrant l'évolution des principales caractéristiques spectrales en fonction du temps.

Parmi ces indices, la durée de la tenue et l'explosion ont joué un rôle prépondérant dans les résultats de WAJSKOP et SWEERTS (1973). Leur expérience consistait en une tâche d'identification des 3 consonnes sourdes [p], [t] et [k], en syllabe [a] + consonne dont l'explosion et des portions croissantes de la tenue avaient été excisées. Cependant, il n'a pas été possible d'évaluer l'importance exacte du paramètre temporel, la segmentation étant, dans cette expérience, accompagnée automatiquement de la suppression de l'explosion.

Selon notre hypothèse, la perception du trait de voisement dépendrait de la présence de plusieurs indices, l'importance de chacun d'eux étant fonction du contexte. Dans cette perspective, l'estimation du rôle joué par l'un des indices doit procéder par modifications hautement spécifiques de la substance sonore.

Dans le travail dont nous allons rendre compte, la durée de la tenue des 3 occlusives sourdes a été réduite dans des proportions variables en laissant intact l'ensemble des autres éléments acoustiques de la séquence VC-, y compris l'explosion. En procédant de la sorte, nous désirions évaluer le poids percepti de la durée de la tenue dans un cadre VC- complet.

Une série expérimentale composée uniquement de stimuli V + C sourde a été constituée. Les stimuli ont été obtenus à partir de séquences VCV digitalisées et copiées sur bande ordinateur.

Un programme mis au point par P. JOSPA *, nous a permis de fixer les limites des syllabes, de mesurer avec précision les durées des tenues et d'exciser des segments compris à l'intérieur de la tenue tout en conservant l'explosion. Pour chaque consonne sourde nous avons ainsi obtenu, en plus de la séquence VC- qui présente une tenue complète, quatre stimuli en excisant successivement quatre portions de la tenue équivalentes à 20% de la tenue complète. (voir figure 2).

Nous avons demandé à 19 sujets francophones d'identifier ces consonnes en choisissant l'une des 6 réponses [p], [t], [k], [b], [d], [g]. Une série préparatoire comprenant également des séquences V + C voisée précédait directement la série expérimentale.

---

* Assistant à l'Institut de Phonétique (U.L.B.)

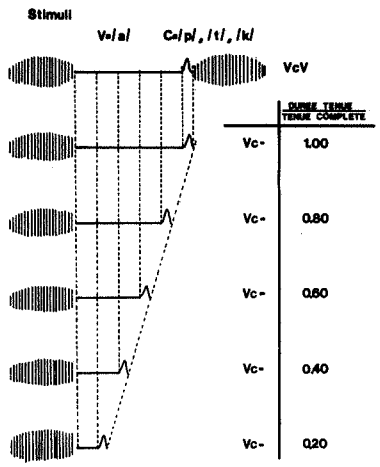


Figure 2 - Caractéristiques des séquences VC- utilisées dans notre expérience.

Les résultats (voir figure 3) ont été testés par un plan d'analyse de la variance à 3 facteurs dont deux à effet fixe ("consonne" et "durée") et un à effet aléatoire ("sujets"). L'influence de la durée de la tenue est significative à .01, et l'examen des écarts partiels entre les résultats obtenus pour chaque durée - par la méthode S de SCHEFFE (1959) - indique que la décroissance du taux de réponses sourdes dépasse le seuil de signification de .01 lorsque le rapport durée de la tenue / tenue complète atteint .4. C'est précisément dans la région d'abscisse située entre .6 et .4 que le rapport durée de la tenue / tenue complète franchit la limite inférieure de la distribution des rapports temporels tenue voisée / tenue sourde mesurés dans un groupe de 12 sujets francophones. (voir figure 3).

D'autre part, les écarts entre les résultats de [k] et ceux des autres consonnes sont significatifs à .01. On remarquera que le taux de réponses sourdes obtenu pour [t] et [p] est déjà très faible lorsque la tenue est complète. Cet effet pourrait trouver son origine dans l'élimination des indices de voisement situés dans la partie postconsonantique de la séquence VCV.

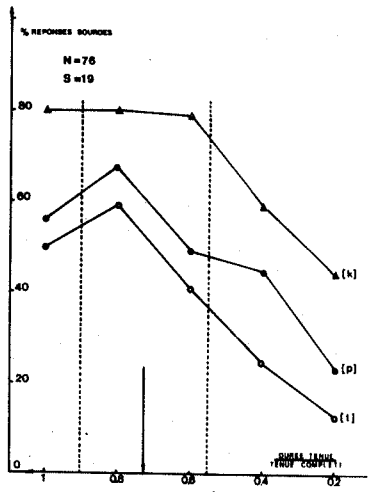


Figure 3 - Résultats de l'expérience d'identification.
En abscisse: durée de la tenue consonantique exprimée en proportion de la tenue complète. Sur le même axe la flèche situe la médiane des rapports correspondants entre tenue voisée et tenue sourde, mesurés dans un groupe de 12 locuteurs francophones. Les lignes interrompues indiquent les limites de la distribution de ces rapports. En ordonnée : pourcentage de réponses sourdes obtenu pour chaque stimulus.

Enfin, l'interaction "consonne - durée" n'est pas significative. On peut donc considérer que la réduction temporelle de la tenue a des répercussions semblables sur la perception du trait de voisement lorsque l'on fait abstraction des écarts entre les résultats globaux des trois consonnes.

Les résultats obtenus dans une expérience antérieure (WAJSKOP et SWEERTS, 1973) montrent que l'influence de la réduction temporelle de la tenue des occlusives sur la perception du voisement se présente de manière assez différente lorsque l'explosion est excisée. Ces résultats ont été replacés dans la figure 4 où nous les avons mis en parallèle avec les données de notre expérience.

En l'absence d'explosion, les stimuli sont constitués de la voyelle /a/ suivie du souffle qui intervient dans la tenue consonantique. Pour /ap-/ cependant, l'occlusion se faisant aux lèvres, le souffle est pratiquement inexistant.

Ceci explique sans doute que, pour cette syllabe, lorsque l'explosion est segmentée la durée de la tenue ne conditionne plus de modification régulière du taux de réponses sourdes, la limite de la tenue étant probablement imperceptible.

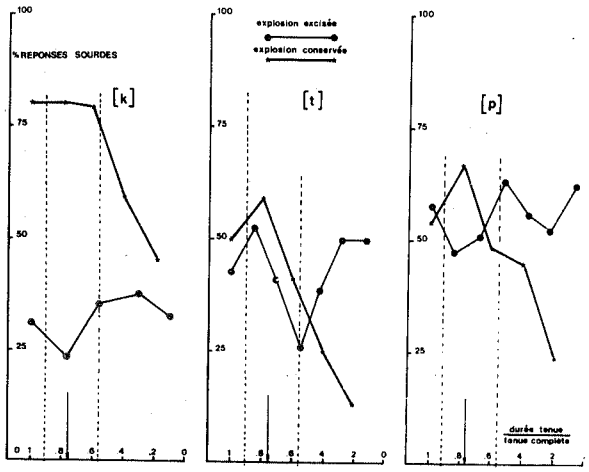


Figure 4 - Mise en correspondance des résultats d'identification obtenus lorsque l'explosion est présente (notre expérience, traits épais) et ceux se rapportant aux séquences VC- dont l'explosion a été excisée (expérience de WAJSKOP et SWEERTS, traits fins).

Par contre pour /at-/ les résultats suivent une évolution semblable à celle qui a été obtenue dans notre expérience - pour autant que le rapport durée de la tenue / tenue complète reste du même ordre que ceux qui ont été mesurés pour les occlusives voisées. Pour des durées plus brèves, et ce malgré la présence de souffle dans la tenue, cet indice deviendrait inopérant lorsque l'explosion est absente.

En ce qui concerne /ak-/, remarquons d'abord que la suppression de l'explosion entraîne une chute importante du taux de réponses sourdes. Pour cette consonne, l'explosion constituerait un indice de voisement aussi important que la durée de la tenue. Notons d'autre part que la durée de la tenue n'apporte plus d'information lorsque l'explosion est excisée : les effets de l'explosion et de la durée de la tenue sur la perception du non-voisement de [k] seraient en interaction négative.

Cette interprétation reste cependant sujette à caution, car nous pourrions également nous trouver en présence d'un phénomène semblable à celui qui a été constaté pour /at-/ lorsque le rapport temporel durée de la tenue / tenue complète devient très faible.

En conclusion, nous retiendrons que la durée de la tenue constitue un indice important pour la perception du trait de voisement des occlusives mais qui ne peut à lui seul, déterminer de manière non ambiguë le caractère sourd ou voisé de la consonne.

D'autres indices pourraient également intervenir, l'importance de chacun dépendant du point d'articulation de la consonne :

- pour [p] et [t], il s'agirait essentiellement des indices contenus dans la partie post-C de la séquence VCV
- pour [k], des indices présents dans l'explosion.

Ces différents indices s'intégreraient dans un schème perceptif dont les caractéristiques dépendraient de la consonne et du contexte considérés.

BIBLIOGRAPHIE.

FANT, G. Sound,Features,and Perception, S.T.L.-Q.P.S.R., 2-3, Stockholm,1967.

OHMAN, S. Perception of Segments of VCCV utterances.
J. Acoust. Soc. Am., 40, 1966, 979-988.

SCHEFFE, H. The Analysis of Variance
New-York, J. Wiley, 1959, 66-72.

WAJSKOP,M. La perception des consonnes intervocaliques.
International of Review of Applied Linguistics, 1973.(à paraître)

WAJSKOP, M. and SWEERTS, J.
Voicing cues in oral stop consonants.
Journal of Phonetics, 1973, 1, 121-130.

Discussion

M.ROSSI : Les sujets devaient-ils choisir entre deux stimuli de type p/b, k/g ou t/d ou entre les six consonnes citées ? Dans cette deuxième hypothèse pensez-vous que les résultats, notamment quand l'explosion est excisée, puissent être affectés par la reconnaissance du lieu d'articulation ?

W. SERNICLAES : Les sujets devaient choisir entre les six consonnes. Dans le cas où l'explosion avait été excisée, on a observé un pourcentage assez élevé d'erreurs de point d'articulation. Ceci ne devrait cependant pas affecter la validité des résultats car des travaux antérieurs (OHMAN (1966); WAJSKOP (1973)) tendent à montrer que la perception du trait de voisement et celle du point d'articulation sont, dans une certaine mesure, indépendantes.

A. NEMETH : Avez-vous étendu vos investigations sur l'étude de la mélodie ? Il est indubitable , en effet, que pendant l'occlusion sonore la mélodie tombe d'une octave comme la mélodie est une fonction continue du temps (on le constate) il y a un glissement de mélodie au début de la voyelle qui suit l'occlusion. Avez-vous constaté que ce glissement rapide est un trait distinctif ou bien il passe inaperçu ?

W. SERNICLAES : Cet indice, qui joue probablement un rôle important, n'a pas encore été étudié.

Il nous semble que de nombreux indices sont susceptibles d'intervenir dans la perception du voisement. Remarquons à ce propos, que la liste que nous avons reprise n'est pas limitative. Elle ne concerne que les indices qui ont fait l'objet d'une investigation antérieure.

Mme LHOTE: Le groupe VCV représente deux syllabes. Vous avez choisi VC dans ce groupe VCV, ce qui suppose une syllabe fermée.

Or, le rôle de l'explosion n'est pas le même dans la syllabe ouverte CV que dans la syllabe fermée VC - en ce qui concerne la perception de la sonorité.

Dans ce test de la durée de la tenue (de l'occlusive) - et de l'explosion - dans la perception de sonorité, vous semblez avoir négligé le rôle de la coupe syllabique.

W. SERNICLAES : Ce problème a effectivement été négligé pour l'instant.

Le choix de VC dans l'ensemble VCV s'inscrit dans le cadre d'une première étape de l'étude de la perception de la consonne. Il nous semble que le rôle de l'explosion doit être considéré en fonction de l'ensemble des indices présents dans la séquence VCV.

Dans cette perspective, supprimer l'explosion tout en conservant l'ensemble des autres éléments de la séquence VCV - y compris ceux situés dans la partie post-consonantique- eut été préférable.

Mais à l'époque où la première expérience a été réalisée (WAJSKOP et SWEERTS, 1973) nous ne disposions pas du programme-ordinateur nécessaire pour effectuer cette opération.

LA PERCEPTION DES PAUSES

Résumé

Cet exposé décrit une partie d'une enquête sur les seuils de perception pour les pauses de la parole. Des travaux précédents ont montré que la durée minimum d'un arrêt de signal nécessaire afin d'être entendu (durée limite) est plus longue pour les pauses entre les unités syntaxiques principales (juncture pauses) que pour les pauses à l'intérieur de ces unités (hesitation pauses). Une première expérience a montré que les durées limites varient comme fonction de la complexité syntaxique à l'emplacement de la pause. Cependant une deuxième expérience a démontré que cette variation ne se produit pas quand la courbe mélodique est effacée en tant qu'indication de la structure syntaxique. Un des buts éventuels de cette recherche sera la construction d'un modèle pour la génération des durées limites.

Summary

This paper describes part of an investigation into perceptual thresholds of the pauses of speech. Previous work has shown that the minimum duration of signal break necessary in order for a pause to be heard (threshold duration) is longer for pauses between major syntactic units (juncture pauses) than for pauses within such units (hesitation pauses). A first experiment showed that threshold duration varies as a function of the syntactic complexity at the pause position. A second experiment, however, demonstrated that this variation does not occur when intonation is removed as a cue to syntactic structure. Long terme aims of this research include the construction of a model for the generation of threshold durations.

A. BUTCHER

Université de Kiel - Institut de Phonétique.

LA PERCEPTION DES PAUSES

Andrew Butcher

0. On peut considérer l'étude de la perception des pauses comme une partie de la question plus générale de la perception de la durée. D'une part, une pause peut être perçue à un endroit où l'analyse instrumentale ne découvre aucun arrêt du signal acoustique, d'autre part, pour l'auditeur un arrêt de plus d'une demi-seconde peut très facilement passer inaperçue. Donc, la durée-limite - la durée au delà de laquelle une pause est perçue - peut montrer une variation entre zéro et plus de 500 msec. Nous étudions à présent, avec des locuteurs et des auditeurs allemands, les paramètres qui déterminent la valeur de ce seuil temporel. Je vais décrire ici des expériences qui examinent deux de ces paramètres - la structure syntaxique et l'intonation - leur relation entre elles et avec la perception des pauses.

1. Les pauses de la parole ont été classifiées de différente façon par chaque investigateur. Un des critères préféré est celui de la structure syntaxique. Maclay et Osgood (1959) et Blankenship et Kay (1964) ont considéré cette question,tous les deux, d'un point de vue distributionnel, en se servant des cadres syntaxiques de Fries pour classifier les mots qui se trouvent après les pauses d'hésitation. Wilkes et Kennedy (1969) ont aussi déduit une relation entre la production des pauses et la variation dans la forme grammaticale, fondée sur le temps nécessaire pour le rappel à court terme des mots consécutifs.

Une des expériences les mieux connues sur la perception des pauses est celle de Boomer et Dittmann (1962) qui ont défini la structure syntaxique selon la courbe

mélodique. Donc ils ont différencié deux degrés de complexité. Leurs pauses se sont trouvées ou à l'intérieur ou à l'extérieur des "phonemic clauses". Leur expérience a démontré une différence de durée limite entre ces deux types. Dans l'analyse d'un corpus spontané une telle dichotomie élémentaire devient contre-intuitive. Il s'agit d'une variation plus étendue à l'intérieur de ces catégories, conforme à la variation dans la complexité de la structure syntaxique aux emplacements dans et entre les phrases phonémiques.

Ruder et Jensen (1972) ont examiné la perception des pauses à cinq niveaux de complexité syntaxique au moyen de leur système pour l'ajustement des pauses, par laquel les sujets ont ajusté la longueur des pauses entre la même paire de mots ayant différentes relations syntaxiques l'un avec l'autre. Cinq niveaux de complexité ont donné cinq durées-limites. Un des nombreux aspects intéressants de cette expérience est que l'on a quantifié la complexité syntaxique au moyen de l'index de la complexité structurale proposé par Chomsky et Miller (1963).

Notre première expérience a combiné des traits des deux dernières études citées. L'expérience originale de Boomer et Dittmann a considéré quatre phrases de différents locuteurs dans un corpus. Chaque phrase a comporté une pause de conjoncture et une pause d'hésitation. Les périodes de silence ont été excisées et un morceau de bande vide d'une longueur entre 100 et 500 msec. a été introduit dans un des emplacements. Chacune des phrases avec un tel silence a été accouplée avec une version de la même phrase sans silence. Ces paires, arrangées au hasard avec un nombre de paires témoins - sans aucune période de silence - ont été présentées à des auditeurs qui devaient constater si les paires semblaient différentes ou identiques. De cette manière, Boomer et Dittmann ont établi un seuil pour les pauses de conjoncture entre 500 et 1000 msec. et pour les pauses d'hésitation à 200 msec. environ.

Notre expérience a éliminé trois variables non désirées. 1. Variation de locuteur 2. Variation de signification de la phrase 3. Variation de longueur de la phrase. On a employé une seule phrase, enregistrée dans une chambre insonorisée, prononcée sans pauses, et on a inséré à l'un des quatre emplacements une période de silence d'une longueur entre 50 et 350 msec. (Des tests préliminaires où l'on a employé des longueurs maximales de 700 msec. ont démontré que l'on peut s'attendre à des seuils plus bas avec cette méthode où les pauses ont été insérées plutôt qu'excisées.)Les sujets étaient 25 lycéens de la première classe. La présentation des phrases et la tâche demandée aux sujets sont restées les mêmes. La phrase employée était:

> Ein Tonbandgerät und einer neuer Plattenspieler stehen auf dem Tisch.

Les emplacements des pauses, leur complexité structurale, et leur contexte phonologique sont montrés par le tableau I. La figure I montre les résultats de 900 (36 x 25) jugements expérimentaux. En supposant que la durée-limite soit celle pour laquelle 75% des sujets ont fait un jugement correct, on voit un écart très évident entre les pauses de conjoncture (1 & 3) et les pauses d'hésitation (2 & 4). Par contre, on voit très clairement que le seuil monte en corrélation étroite avec la complexité syntaxique. Plus la structure à travers un emplacement est complexe, plus la durée nécessaire à la perception d'une pause dans cette emplacement est longue. Ce résultat confirme la constatation de Ruder et Jensen qu'il s'agit de plus de deux seuils dans le discernement des pauses, mais se trouve en contradiction avec eux et avec Martin et Strange (1968a,b) quand il suggère que la longueur des pauses a en effet des conséquences pour le décodeur non moins que pour l'encodeur, respectivement comme indication ou comme fonction de la complexité syntaxique.

2. Néanmoins il est évident que la structure syntaxique n'est pas la seule influence sur la perception des pauses. Le contexte des emplacements des pauses dans notre phrase expérimentale a varié dans au moins trois aspects phonologiques - segmentaire, accentuel, mélodique. Il y avait naturellement aussi des variations non-phonologiqes de fréquence, d'intensité et de durée dans les syllabes environnantes. En considérant d'abord les aspects phonologiques, l'expérience suivante, dans notre série a examiné un facteur auquel on n' a guère pris garde dans les différents travaux sur les pauses, c'est-à-dire le rôle de l'intonation comme indicateur de la structure syntaxique. Nous voulions vérifier si la perception des pauses est influencée par la structure syntaxique implicite elle-même, ou seulement quand cette structure est rendue phonologiquement explicite au moyen de la configuration mélodique.

Donc on a répété notre première expérience en employant une phrase de la même structure syntaxique, prononcée en monotonie (fréquence du fondamental 100 Hz.). Cette fois-ci on a maintenu le contexte immédiat des pauses invariable, avec une fricative sourde en avant et une occlusive sonore en arrière. La phrase employée était:

> Ein Rasthaus bzw. ein modernes Gasthaus befindet sich auf dieser Strecke.

On n'a pas contrôlé le rythme indépendamment de l'intonation. Cependant des oscillogrammes montraient que la variation d'intensité a été diminuée et du point de vue auditif la proéminence des syllabes accentuées a été en pratique neutralisée par l'absence de variation mélodique. La longueur des pauses variait entre 50 et 300 msec. et les sujets étaient 15 lycéens de la première classe. Les résultats des 465 (31 x 15) jugements expérimentaux sont résumés dans la figure 2 et le tableau 2.

On peut voir que toutes les durées-limites se trouvent pour ainsi dire au même niveau - avec un écart maximale de

10 msec. Les pauses de la même longueur ont été perçues dans tous les quatre emplacements syntaxiqes avec la même facilité. Ces résultats suggèrent que ce n'est pas la structure syntaxique elle-même qui influence la perception des pauses, mais la configuration mélodique qui indique cette structure. Une nouvelle expérience complémentaire répétéra le procédé en employant des mots sans signigication avec des configurations mélodiques allemandes surimposées. D'autres paramètres - par exemple dans le contexte segmentaire: sonorité, position, manière, allongement non-phonémique etc. - restent naturellement à examiner, de même que la question importante de l'influence du tempo.

3. Peut-être doit-on ici préciser exactement quelle est la résponse que l'on obtient des sujets dans de telles expériences et ce que l'on entend par 'durée-limite' à ce propos.

Premièrement, ce type d'expérience demande que l'on distingue entre deux stimuli et pour cette raison, c'est une tâche plus facile que de demander l'identification des pauses dans un texte spontané courant. Donc il s'ensuit que les seuils sont plus bas.

Deuxièmement les pauses sont insérées plutôt qu'excisées. Cette méthode entraîne l'absence des indications comme l'allongement non-phonémique des segments précédentes, ce qui peut provoquer une certaine qualité d'anomalie dans les phrases qui ont des pauses. Par contre, quand les pauses sont excisées, l'anomalie réside dans les phrases sans pause - les phrases-témoins y compris. Donc notre méthode de composer les stimuli donne aussi des durées-limites plus courtes que par exemple la méthode de Boomer et Dittmann.

Troisièmement, il s'agit ici d'une limite absolue - un jugement de la présence ou de l'absence d'un phénomène. Des investigations préliminaires ont suggéré qu'il serait aussi intéressant d'obtenir des jugements de normalité ou

de "fluidité" - autrement dit de juger si une pause fait partie de la structure phonologique, ou si elle est une intrusion dans cette structure. Ruder et Jensen (1972) donnent à ces deux types les noms de "fluent pause" et "hesitation pause". J'aimerais mieux employer les termes 'pause' et 'interruption'. Il faut souligner que cette classification n'est pas faite simplement selon l'emplacement de la pause, mais selon la longueur de la pause à un emplacement donné. C'est-à-dire que l'on pourrait avoir affaire à deux durées-limites - l'une au delà de laquelle une pause serait perçue et l'autre au delà de laquelle cette pause serait considérée comme une interruption de l'activité langagière. Evidemment, si, dans un emplacement quelconque, ces deux seuils coïncident, alors n'importe quelle pause dans cette emplacement serait entendue comme une interruption. Cela pourrait arriver par exemple dans l'emplacement 4 dans nos expériences. Nous espérons pouvoir en rendre compte dans un avenir peu éloigné.

4. On a considéré comme un des buts possibles de cette étude la formalisation de la description de la perception des pauses. On arriverait à une telle formalisation par la dérivation des règles pour la génération des durées-limites. Lorsque la pertinence de chaque paramètre sera établie, la valeur de chaque paramètre pertinent, à l'emplacement dans le signal acoustique dont il s'agit, composerait la consommation d'un modèle pour la génération de la durée-limite pour le discernement d'une pause (aussi peut-être d'une interruption) à cet emplacement.

Dans le cas où la valeur de ce premier serait zéro ou négative, alors n'importe quel arrêt du signal dans cet emplacement serait discerné comme une pause. Si le deuxième seuil était aussi zéro ou inférieur à zéro, alors n'importe quel arrêt du signal dans l'emplacement serait entendu comme une interruption. L'intention est de construire de telles

règles peu à peu, au moyen d'expériences du type esquissé ici et de les vérifier statistiquement avec l'aide d'un corpus enregistré. De cette façon,on pourrait améliorer le modèle, pour le rendre plus conforme aux données.

------oOo------

Cette recherche a pu être conduite grâce à l'aide de la Deutsche Forschungsgemeinschaft (DFG).

Institut für Phonetik, Universität Kiel,
2300 KIEL, Olhausenstr. 40-60, RFA.

"LA PERCEPTION DES PAUSES" : BIBLIOGRAPHIE

BLANKENSHIP,J. & KAY,C.(1964) Hesitation Phenomena in English speech: a study of distribution. Word 20 360-372.

BOOMER,D.S. & DITTMANN,A.T.(1962) Hesitation pauses and juncture pauses in speech. Lang Speech 8 148-158.

MACLAY,H. & OSGOOD, C.E. (1959) Hesitation phenomena in spontaneous English speech. Word 15 19-44.

MARTIN,J.G. & STRANGE,W.(1968a) Determinants of hesitations in spontaneous speech. J Exp Psychol 76 474-479.

MARTIN,J.G. & STRANGE,W.(1968b) The perception of hesitation in spontaneous speech. Percept Psychophys 3 427-438.

MILLER,G.A. & CHOMSKY,N.(1963) Finitary models of language users. Dans: Luce,R., Bush,R. & Galanter, E. (Eds.) Handbook of Mathematical Psychology II 419-92.

RUDER,K.F. & JENSEN,P.J. (1972) Fluent and hesitation pauses as a function of syntactic complexity. J Speech Hear Res 15 49-60.

WILKES,A.L. & KENNEDY,R.L.(1969) Relationship between pausing and retrieval latency in sentences of varying grammatical form. J Exp Psych 79 241-245.

FIG.1 POURCENTAGE DE DISCERNEMENTS CORRECTS DE PAUSES DE DURATION EGALE DANS QUATRE EMPLACEMENTS - INTONATION "NORMALE"

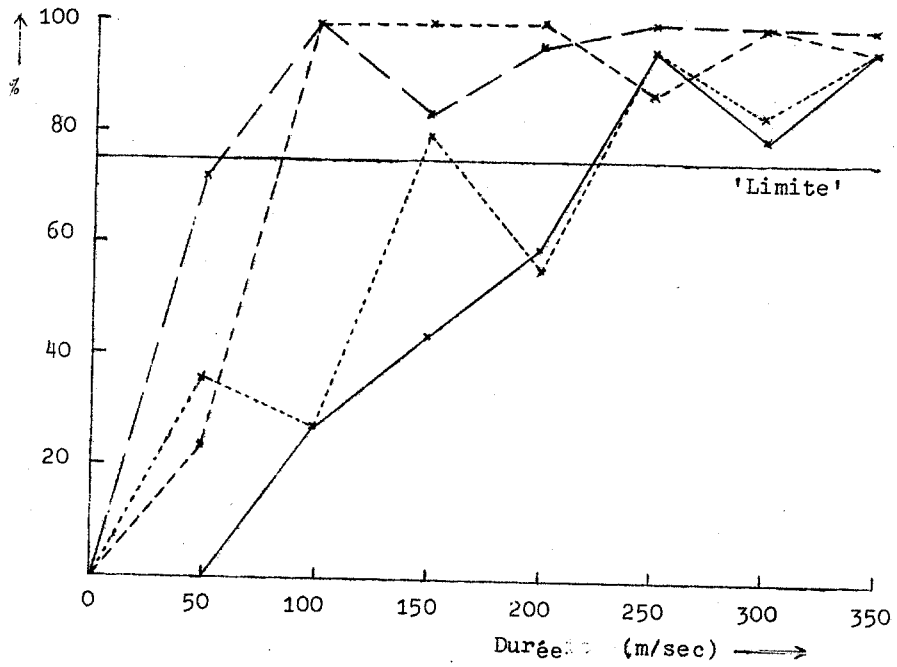


TABLEAU 1 VARIATION CONTEXTUELLE DES PAUSES

Emplacement	Valeur de complexité syntaxique	Contexte segmentaire	Contexte accentuel	Contexte mélodique	Durée limite (m/sec)
—— 1	1,86	P//V	- 2//- -		225
— — 2	1,50	V//P	1 -//1 -		55
..... 3	1,85	V//F	2 -//1 -		150-225
- - - - 4	1,67	F//P	- -//- 1		85

STIMULUS:

Ein Tonbandgerät /1/ und ein neuer /2/ Plattenspieler /3/ stehen auf /4/ dem Tisch.

FIG.2 POURCENTAGE DE DISCERNEMENTS CORRECTS
DE PAUSES DE DURATION EGALE DANS QUATRE EMPLACEMENTS
- INTONATION MONOTONE

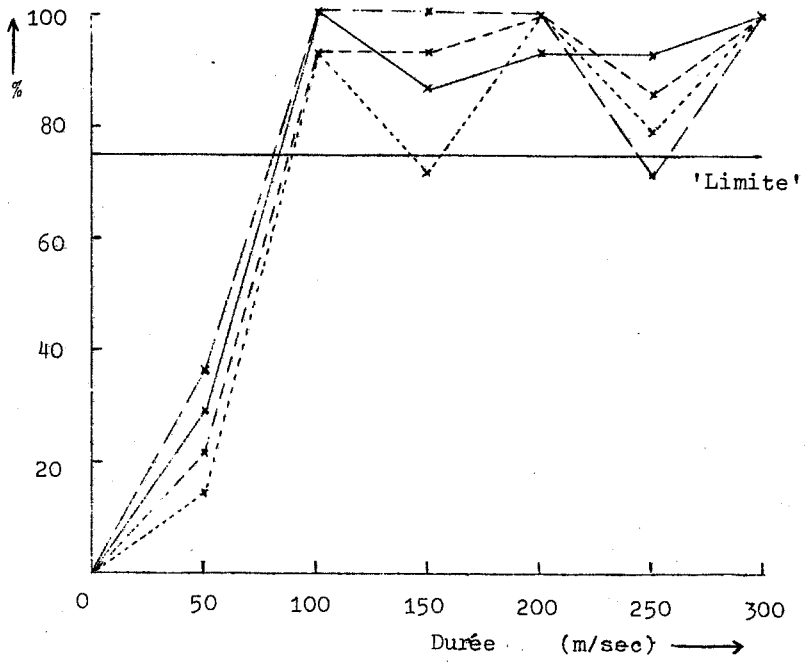


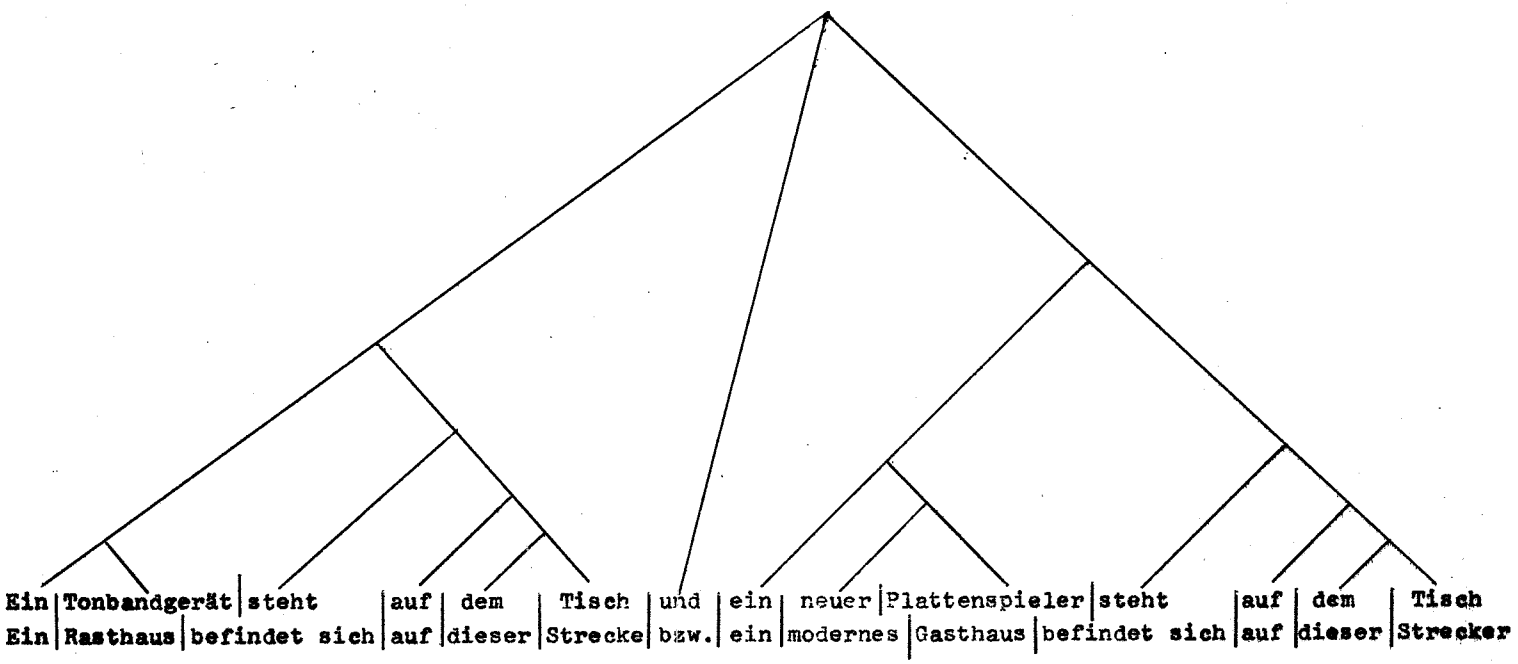
TABLEAU 1 VARIATION CONTEXTUELLE DES PAUSES

Emplacement	Valeur de complexité syntaxique	Contexte segmentaire	Contexte accentuel	Contexte mélodique	Durée limite (m/sec)
——— 1	1,86	F//P	1 2//- 2	—//—	80
— — 2	1,50	F//P	1 -//1 2	—//—	75
····· 3	1,85	F//P	1 2//- 1	—//—	85
- - - - 4	1,67	F//P	- -//1 -	—//—	85

STIMULUS:

Ein Rasthaus /1/ bzw. ein modernes /2/ Gasthaus /3/
befindet sich auf /4/ diese Strecke.

FIG.3 - ANALYSE STRUCTURALE DES PHRASES EXPERIMENTALES



LA PAUSE GRAMMATICALE COMME PHENOMENE ALEATOIRE

DANS LE NEERLANDAIS PARLE.

Résumé

Quelques données statistiques sur le phénomène de la pause grammaticale sont présentées pour un texte lu en néerlandais. On considère les diverses transitions des états parole et pause. La pause grammaticale peut fournir des renseignements sur un moyen de segmentation pour le discours continu en unités plus larges que des mots pris isolément.

Abstract

Some statistical details are presented for the grammatical pause occuring in a prepared spoken dutch text. The grammatical pause seems to be an indication for the segmentation of speech in units greater than words.

Cornelius KAMMINGA
Université Technique de Delft (Hollande).

La pause grammaticale comme phénomène aléatoire dans le néerlandais parlé.

Le phénomène des silences ou les pauses dans le discours continu a obtenu depuis longtemps beaucoup d'attention. D'une part sont traités les aspects techniques d'une conversation téléphonique, d'autre part trouve t'on la plupart des travaux en psycholinguistique dans les publications de Mme Goldman Eisler.
Comme celle-ci le fait remarquer, le phénomène d'un discours continu est caractérisé par "some regularities". Dans une étude faite auparavant pour le discours continu en néerlandais - composé de réponses d'un interview - on a observé une certaine régularité dans les pauses.
Vu la structure complexe de la pause comme l'affirment plusieurs psycholinguistes on a essayé, avec l'aide du concept de l'information sélective de la théorie d'information de Shannon, de voir s'il est possible de segmenter le langage parlé.
On a donc choisi un certain aspect des "unfilled pauses", c'est-à-dire la pause grammaticale. Cet aspect est traité d'une manière statistique.
Selon Goldman Eisler on peut observer dans la parole courante les pauses suivantes:
1) la pause grammaticale
2) la pause non-grammaticale

ad 1) Se fait remarquer avec
- ponctuations naturelles
- conjonctions
- pronoms interrogatifs
- remarques

ad 2) Se fait remarquer avec
- faux départ
- répétitions

En général le discours continu se trouve contenir environ 45% de pauses non-grammaticales et 55% de pauses grammaticales. En lisant à haute voix un texte préparé on rencontre presque sans exception la pause grammaticale.
A cet effet un locuteur entraîné a lu un texte en néerlandais. Ce texte est enregistré et analysé par rapport aux pauses et par rapport à la parole.
De cette façon, on peut étudier sans équivoque le phénomène de la pause grammaticale.
Voici quelques données du texte analysé.

longueur totale	14'4"
durée totale de la pause	4'4'"
durée totale de la parole	10'
nombre des pauses	353
nombre total des mots	2333

Afin de distinguer les silences phonétiques d'avec les vraies pauses on a introduit dans les mesures de la durée des pauses une valeur de seuil de 0,2 sec.
Dans le texte analysé on trouve un pourcentage de 32 pour les pauses.
En regardant la distribution des fréquences (fig. 1) on peut remarquer une différence très nette en la comparant avec la distribution de la parole (fig. 2).

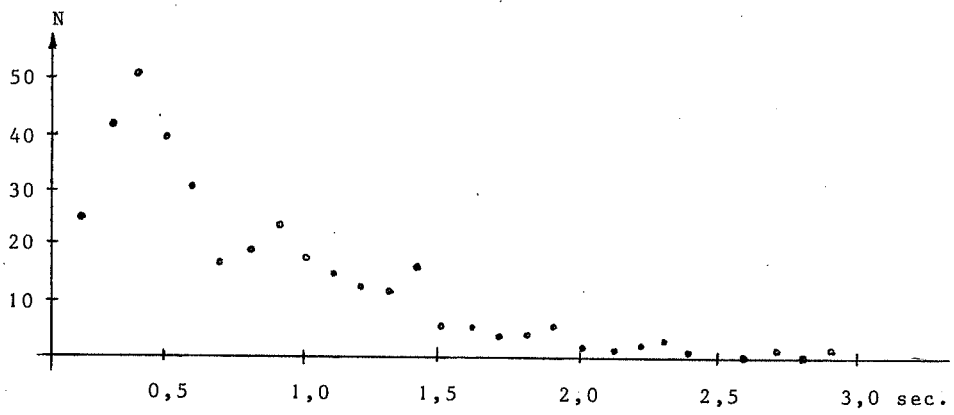


fig. 1

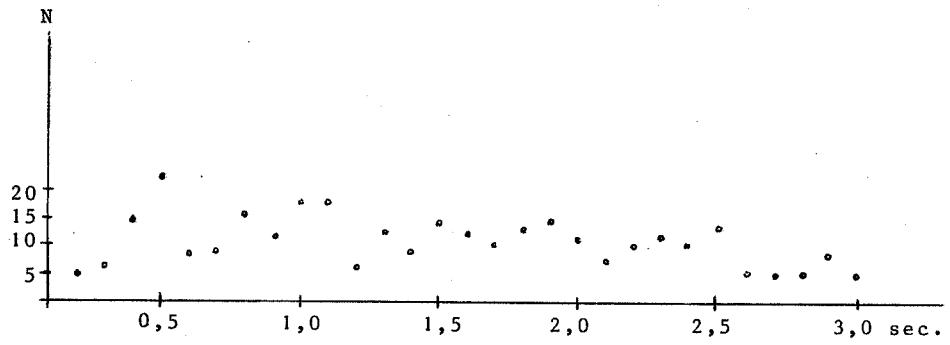


fig. 2

Il est important de vérifier comment la pause grammaticale se situe par rapport à la quantité d'information précédant et suivant une pause. A cet effet, on a introduit, avec l'aide de la statistique, des mots de ce texte: la quantité

$$I(i) = - \log \hat{p}(i) \text{ bit}$$

avec $\hat{p}(i)$ comme fréquence du mot i dans le texte.
Ce calcul de l'information a fourni les donnees suivantes.

Quantité d'information moyenne par mot	8,28 bit
Quantité d'information premier mot après une pause	7,37 bit
Quantité d'information dernier mot avant une pause	9,41 bit

Valeur moyenne de l'information entre 2 pauses 32 bit.

Les donnees indiquées ci-dessus fournissent une conclusion remarquable en cequi concerne l'interruption de la séquence d'information par le phénomène de la pause.
Après la pause la quantité d'information du premier mot atteint une valeur qui est moins que l'information moyenne pendant la

parole, tandis que la quantité d'information du dernier mot précédant la pause se trouve être plus grande que la valeur moyenne.
Une première hypothèse semble indiquer un décodage du message par l'interlocuteur pendant la pause. Mais cette hypothèse est infirmée si l'on calcule la corrélation entre la quantité d'information du dernier mot avant la pause et la durée de la pause. On trouve pour cette corrélation une valeur de 0,12.
Un tel comportement du courant d'information, comme indiqué ci-dessus est pourtant tout en fait contraire à ce qu'on trouve dans la littérature traitant ce sujet.
Outre l'aspect du courant d'information à la base de la quantité d'information par mot on a étudié la structure de Markov du texte considéré.
Pour cela on a conçu le texte comme une chaîne de Markov à deux états. Les résultats qu'on obtient de cette manière pour les transitions des états.

pause → pause
parole → parole
parole → pause

sont indiqué dans les figures 3, 4 et 5.
On peut remarquer que la transition parole → parole montre une distribution relativement uniforme contrairement à la transition pause → pause. La transition parole → pause comme indiqué dans la figure 5 ne fournit que des généralités.
En conclusion on peut affirmer que la pause grammaticale est un indice très net pour la segmentation de la parole en unités plus grandes que les mots.
La recherche continue en ce moment pour des échantillons plus vastes et avec l'aide de dictionnaires de fréquences.
Pour voir si les effets signalés sont typiques de la parole et pour voir s'ils sont liés au néerlandais, les recherches vont s'étendre sur le français, l'italien et l'anglais.

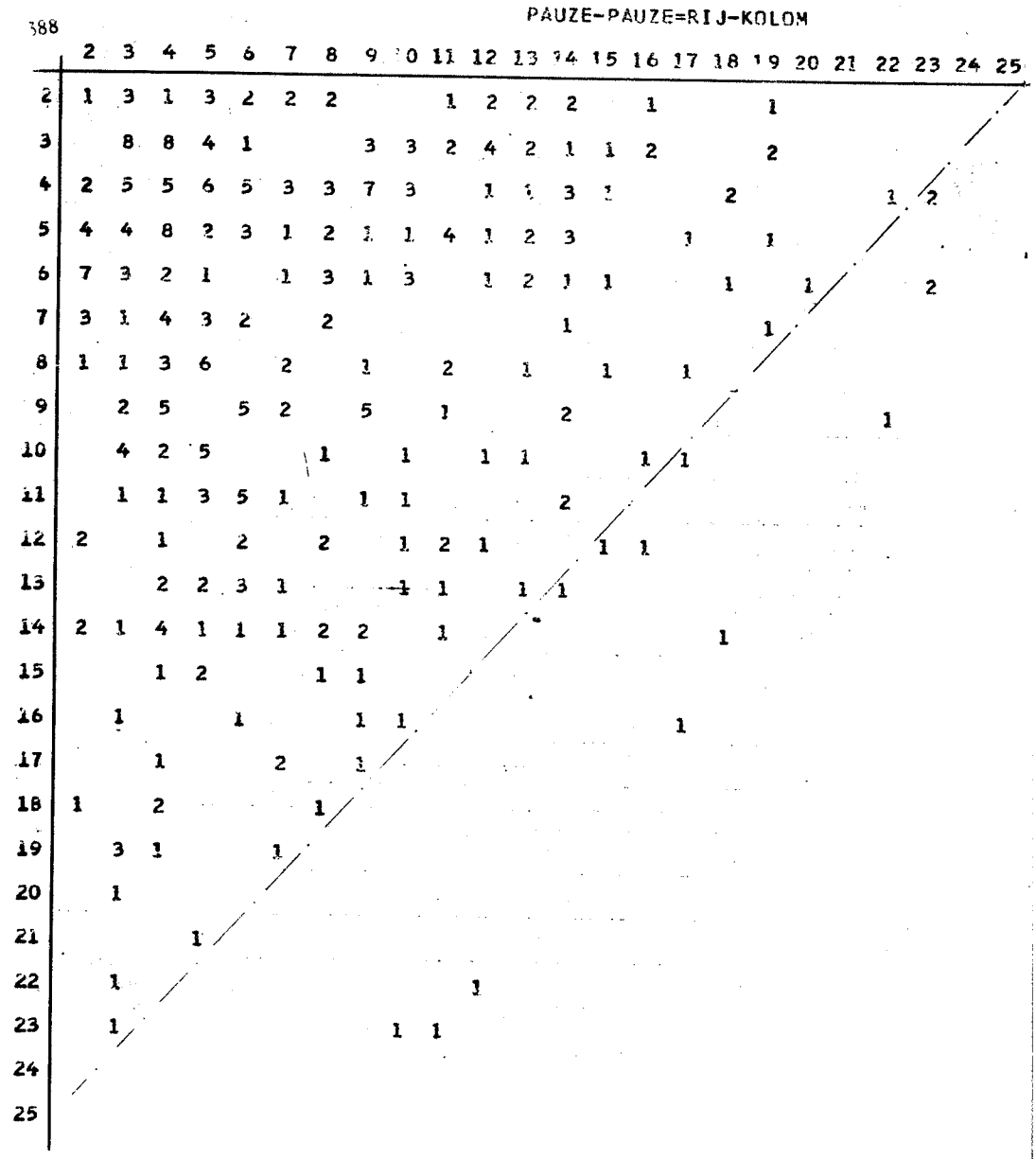

PAUZE-PAUZE=RIJ-KOLOM
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
2 1 3 1 3 2 2 2 1 2 2 2 1 1 1
3 8 8 4 1 3 3 2 4 2 1 1 2 2
4 2 5 5 6 5 3 3 7 3 1 3 2 1 2
5 4 4 8 2 3 1 2 1 1 4 1 2 3 1 1
6 7 3 2 1 1 3 1 3 1 2 1 1 1 1
7 3 1 4 3 2 2 1 1
8 1 1 3 6 2 1 2 1 1 1 1
9 2 5 5 2 5 1 2 1
10 4 2 5 1 1 1 1 1 1
11 1 1 3 5 1 1 1 2
12 2 1 2 2 1 2 1 1 1
13 2 2 3 1 1 1 1 1
14 2 1 4 1 1 1 2 2 1 1
15 1 2 1 1 1
16 1 1 1 1 1
17 1 2 1 1
18 1 2 1
19 3 1 1 1
20 1
21 1
22 1 1
23 1 1 1
24
25

fig. 3

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
2			1	1	1										1									
3				1					1	1				2	1		1							
4				1		1	1		1	2	1				1	1	2	1						
5	1	1		1	1	3	1			1			2				2	1	1	1		1	3	
6	1					1			2					1	2					1				
7				3	1						1	1							1				1	
8		1	1	1	1				1	1				2	1	1	1	1		1		1		
9			1		1		2	1				1				1				1	3	1		
10		1					2		1	1		3	1					2				2		2
11	1		2	2	1			1					1	1			1		1		1		2	1
12			1	1															1					
13		2	2						1	2								1		1			1	1
14							1	3			1						1							1
15				3	1	1				2			1		1			1			2			
16			2						1	2				1		1		1						1
17		1	2						1	1	1											2		
18			1			1			1		1	1	1	1	1	1	1		1				1	
19	1			1		1	1		2						1			3			2			
20				1				1	1					1				1	1					1
21				1					1	1		1										2		
22			1				1		2	1			1			1								2
23	1			2				1				1		1		1			1	1				
24				1	1		1			2		1		1		2								
25			1				2		1						1				2		1	1	2	

fig. 4

SPRAAK-PAUZE=RIJ-KOLOM

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
2	1			2	2																			
3	1		1	1	2		1	1																
4	2	4	1			1	2	2			2	1												
5	1	3	4	4	1	1	1	1	2	1		2	1											
6		1	2					2			1			1	1									
7			2	3			1	1			1				1									
8	1	1	2	2	3	2		2		2	1													
9		2	3	2	1			2					1				1							
10	2	1	1	2	1		1	2	2	1		1	3				1							
11	1	6	1	2	3			1	2			1	1											
12		1		1		2	1																	
13			2		3	1	2	1	1				2				1							
14	1		1	2	1		2							1										
15	2		2	1	1		3	1		3	1	1												
16	2	2	1	2	2	1					1		1		1									
17		1	2		2	3						1						1						
18	1	3	1	1		1		1		1	1		2								1	1		
19		2	5	2	2				1		1	2												
20		2	3						1				1		1	1		1		1				
21	1	2		1				1	1															
22			2	2	1			1	2		1	1												
23	2	2	1	1						2		1		1					1					
24	1	1		1	2		1		1	2						1								
25	2	1	1	1	1	2	1	1		1	1		1		1									

fig. 5

Bibliographie.

1) D.S. Boomer
Hesitation and grammatical encoding
Language and speech
1965, p. 148 - 158

2) A.T. Dittmann - L.G. Llewellyn
The phonemic clause as a unit of speech decoding
Journal of personality and social psychology.
1967, p. 341 - 349

3) F. Goldman Eisler
Psycholinguistics 1968.
Academic Press - New York.

Discussion

A. NEMETH : Vous constatez que le dernier mot avant la pause porte plus d'information que le mot qui suit la pause. Vous dites encore que c'est contraire aux affirmations des psychologues.

A mon avis, la contradiction n'est qu'apparente. Même si le mot est rare (le calcul élémentaire lui attribue donc beaucoup d'information) psychologiquement il est prévisible, car l'esprit humain travaille à un niveau plus élevé que la théorie d'information élémentaire.

C. KAMMINGA : Je suis d'accord avec vous en ce qui concerne le niveau, où travaille l'esprit humain - on a réfuté le modèle gauche/droit dans la psychologie depuis des années parce que la structure de Markov n'existe pas pour les mots seuls - néanmoins on peut utiliser le phénomène de quantité d'information plus élevée comme un indice de prédiction de la pause. Puis, je veux insister sur le fait de segmentation de la parole par la pause, c'est-à-dire dans des unités contenant en moyenne 6,6 mots dans notre texte.

TABLE RONDE SUR LA LINGUISTIQUE, PERCEPTION ET RECONNAISSANCE DE LA PAROLE.

M. ROSSI - Présentation

L'objet de cette table ronde n'est pas tant de confronter les résultats présentés pendant ces Journées que de faire état des problèmes qui se posent dans chaque discipline, afin de faire le point et de développer la collaboration entre ingénieurs, phonéticiens, linguistes et psychologues. Dans la mesure où cet objectif sera atteint, je pense qu'on aura remédié aux insuffisances qui ont été évoquées ce matin.

Les problèmes qui seront soulevés concerneront donc les relations entre Linguistique, Perception et Reconnaissance de la Parole. La discussion sera ensuite ouverte et, si certains problèmes n'ont pas été évoqués, je demanderai à ceux qui veulent en faire état d'intervenir en cours de discussion.

1er thème de discussion

Des psycho-linguistes comme SAVIN ont parlé de la non-réalité perceptive du phonème. Le phonème n'étant pas entendu au sens strict d'un point de vue phonologique. Le phonème selon ces recherches est perçu à un niveau supérieur d'intégration, c'est-à-dire au niveau de la syllabe (résultats qui rejoignent ceux de CHISTOVITCH et de LEHISTE en particulier). Alors se pose le problème suivant :

L'identification des phonèmes isolés est-elle justifiée ?

Je pense que non, si on se réfère à l'expérience suivante : on sait que les transitions apportent beaucoup d'information dans la parole. Si, dans une liste de mots, on supprime les états stables et on ne maintient que les transitions, le taux de reconnaissance ou d'identification est très faible. Il est aussi faible que le taux d'identification que l'on obtient, par exemple, en plaçant bout à bout des voyelles et des consonnes produites isolément.

En revanche, si l'on supprime les transitions et si on maintient les états stables en contexte, la reconnaissance est très bonne (expérience de BOULANGER, Nancy); par conséquent, se pose le problème de la coarticulation qui affecte même les états stables. La coarticulation pénètre entièrement les voyelles ou même les consonnes.

Dans cette perspective, il est donc essentiel de définir des règles phonétiques dans chaque contexte; règles phonétiques contextuelles qui tiendraient compte des contraintes de distribution, qui permettraient un choix entre des unités moins nombreuses et qui pourraient fournir un coefficient contextuel du système phonique. Cela reviendrait , dans une certaine mesure, à étendre les lois concernant la normalisation des voix individuelles à la normalisation des contextes.

Deuxième conséquence de ce fait : c'est l'importance des unités de haut niveau (des unités qui dominent le phonème). Exemple : la syllabe, le mot, l'unité accentuelle et l'unité intonative dans l'identification et la reconnaissance. Ces unités de haut niveau sont marquées ou non marquées phonétiquement.

Pour les unités marquées, il n'y a pas de problème; mais les unités non marquées, comment les faire intervenir ? Peut-être en utilisant des procédures comme celles qu'a présentées GRESSER hier.

Je pense qu'il faudrait faire intervenir pour la reconnaissance de ces unités de haut niveau également la micro-mélodie; d'après les recherches qui sont en cours actuellement (VAISSIERE), il semble y avoir au niveau micro-mélodique une distinction assez nette entre les unités sémantiques pleines et les unités sémantiques vides.

GRESSER a utilisé la syntaxe après le stade de la reconnaissance et on lui a posé le problème de savoir ce qui allait se passer dans le cas d'une erreur de reconnaissance d'un phonème; mais ne pourrait-on pas inverser le problème et utiliser la connaissance syntaxique non pas à la fin de l'étape Reconnaissance, mais au départ pour faciliter la reconnaissance, la segmentation, etc.... Je pense que la connaissance des structures syntaxiques combinées avec les règles phonétiques contextuelles permettraient une reconnaissance beaucoup plus rapide des unités de bas niveau.

Les recherches que j'ai citées (SAVIN , en particulier) posent le problème de l'interaction entre les modèles perceptifs et les modèles de reconnaissance et de classification. Peut-on ignorer dans la reconnaissance de la parole les modèles perceptifs ? Doit-on d'un côté, développer des modèles de classification et de reconnaissance des formes et, d'un autre côté, des modèles perceptifs sans qu'il y ait osmose entre ces deux types de modèles ? N'y aurait-il pas avantage à ce que les premiers s'inspirent des seconds de façon à accélérer les solutions pour la reconnaissance de la parole ? C'est donc là un premier thème de discussion avec les conséquences qui s'imposent.

Le 2ème thème de discussion peut s'énoncer ainsi :

La reconnaissance des formes, la méthodologie de la reconnaissance des formes peut-elle être utilisée en linguistique pour la classification phonologique, la classification syntaxique, etc... Réciproquement, les modèles linguistiques existants ont-ils été ou peuvent-ils être d'une quelconque utilité dans la reconnaissance de la parole ? Je pense que c'est là un problème très intéressant pour les linguistes et les phonéticiens qui veulent savoir quelle est la validité des modèles linguistiques en cours.

Le 3ème thème est relatif aux problèmes de normalisation auxquels ont fait allusion LIENARD et MAISSIS en particulier.

On parle de normalisation des voix individuelles, mais est-ce une énormité de dire que la normalisation ou la reconnaissance des voix individuelles doit précéder les recherches ultérieures sur la reconnaissance. Est-ce que la normalisation des voix individuelles ne permet pas de développer des méthodes qui nous permettront d'avancer grandement dans le domaine de la reconnaissance de la parole ?

R. DE MORI

A propos du premier problème, il me semble que d'après LIBERMAN, on est assez d'accord que la langage n'est pas un simple chiffrage, mais un codage complexe gouverné par une grammaire. Ce n'est pas une simple traduction de paramètres, mais une génération qui est gouvernée par une grammaire qui est inconnue mais dans laquelle les phonèmes d'une syllabe ont une interaction et les paramètres sont dépendants de cette interaction. Je crois que ce point est fondamental pour ceux qui travaillent sur la chaîne parlée (connected speech).

J'ai essayé personnellement de faire la reconnaissance de mots isolés et on peut chercher à résoudre ces problèmes en ignorant ce qui a été dit avant, c'est-à-dire qu'il y a des phénomènes complexes au niveau de la coarticulation; mais quand on veut reconnaître la parole en contexte, on ne peut pas ignorer les travaux des psychologues de la perception.

Alors, se pose le problème très difficile : celui d'isoler de façon automatique certains morceaux du spectre dans lesquels les effets de coarticulation sont particulièrement évidents. Je pense aux exemples d'OHMAN qui a montré comment, dans certaines coarticulations CVC, ces effets sont particulièrement nets.

J-Y GRESSER

Je voudrais rectifier l'impression que j'ai pu donner hier dans mon exposé. En fait, mon schéma au tableau n'était pas du tout un schéma de machine ou de programme de reconnaissance. Les relations qui étaient définies étaient des relations logiques. Disons que c'est un peu une variante du modèle de FANT dans le cas particulier que j'ai exposé et qu'avec le générateur de machine qu'on est en train de faire, on peut effectivement commander la machine par la syntaxe, c'est-à-dire avoir un processus où on ne cherchera au niveau acoustique que les éléments qui seront présents dans le message. En fait, on peut développer l'approche qui a été faite. Il n'est pas nécessaire dans le message de rechercher tous les éléments, c'est-à-dire de décoder tout le continuum acoustique. On peut très bien rechercher les éléments-clés et cela à tous les niveaux : sémantique, syntaxique, lexical, phonétique et phonologique et même à des niveaux inférieurs où on pourrait chercher certains traits particuliers. C'est au niveau phonétique que la notion de trait pertinent et de trait redondant est assez délicate à manier. Les expériences de perception nous suggèrent de chercher d'abord ces traits pertinents. On a également l'impression qu'il ne faut chercher que ceux-là. Il serait également intéressant d'avoir une description de règles phonétiques en contexte; cela reviendrait à préciser la notion de pertinence ou de redondance et ainsi on ne pourrait rechercher que les éléments nécessaires.

G. NOIZET .

Du point de vue de la perception du langage, ce qui apparaît important au psychologue est que le stimulus verbal, segmentable à différents niveaux, est identifiable à différents niveaux également, niveaux du phonème, de la syllabe, du mot, du syntagme, etc.... mais ce qui est surtout important, c'est que le processus de bouclage de l'identification peut se fermer à un niveau sans qu'il soit fermé à un niveau inférieur. Le fait que le processus se ferme à un niveau (par exemple, que le sujet identifie un syntagme) peut lever une ambiguïté à un niveau inférieur.

Nous avons de multiples exemples expérimentaux de ce point de vue là. Par conséquent, il ne faut pas penser que le processus d'identification se ferait selon une sorte d'architecture cartésienne où il faudrait tout identifier au niveau inférieur et puis, ensuite, procéder à des regroupements. A chaque niveau, il y a choix des indices pertinents. A ce point de vue là, M. ROSSI a fait allusion à SAVIN sur la non-

réalité perceptive des phonèmes; c'est une expérience intéressante : son interprétation est assez difficile. Je ne suis pas sûr personnellement de partager les vues de SAVIN et BEVER.

En deux mots, SAVIN et BEVER présentent à des sujets des logatomes et leur demandent de reconnaître soit un phonème de ce logatome, soit le logatome dans lequel se trouve ce phonème. Ils mesurent des temps de réaction et ils constatent que le temps de réaction est plus court pour reconnaître le logatome dans lequel se trouve le phonème que pour reconnaître le phonème. Par conséquent, le processus va plus vite au niveau supérieur qu'au niveau inférieur. Ils ont fait de nombreuses vérifications, suivant la position du phonème (position initiale, médiane, finale); il y a certaines différences mais les résultats vont toujours dans le même sens.

Il y a une façon simple, peut-être même trop simple, d'interpréter ceci et qui consisterait à dire que, quand nous faisons des expériences en psychologie de la perception, les sujets font ce qu'on leur dit de faire. Si on leur présente des logatomes, ils travaillent à ce niveau-là. C'est-à-dire qu'ils essayent de boucler le processus d'identification à ce niveau-là. Si on leur donne des phrases, ils travaillent au niveau des phrases.

Pour conclure, je dirais que je vais tout à fait dans le sens de la remarque de ROSSI et je m'étonne de voir séparés les problèmes de reconnaissance et de perception, parce qu'après tout la perception est bien une reconnaissance et une identification. Pour ma part, si je vois très bien comment les choses peuvent s'opposer du point de vue des modèles, du point de vue épistémologique, il me paraît curieux qu'une distinction de modèles se transforme en une distinction de domaines et c'est pourquoi, en ce qui me concerne, je souhaiterais beaucoup une confrontation des deux aspects.

G. ROGER .

Ce qui me paraît important, quand on fait de la reconnaissance de la parole, c'est de bien connaître les mécanismes intellectuels de la reconnaissance parce qu'à ce moment-là, on pourrait faire de bons systèmes de synthèses. (Par exemple : si l'on savait la classe des sons qui sont toujours perçus comme "a" on pourrait trouver le mécanisme le plus économique pour faire des "a" en synthèse) et réciproquement, en reconnaissance, il serait intéressant de bien connaître les phénomènes de phonation pour connaître dans toute la classe des signaux que l'on aura à reconnaître,

ceux qui sont limités par les mêcanismes de coarticulation.

Donc, je pense qu'on devrait associer davantage la perception à la synthèse et la phonation à la reconnaissance.

G. PERENNOU .

En reconnaissance des formes, on a toujours à analyser les formes que l'on veut reconnaître, c'est-à-dire analyser pourquoi elles sont mises dans une classe.

L'exemple suivant fera comprendre comment la structure du problème peut être liée à la méthode utilisée.

Si l'on veut placer les nombres pairs et les nombres impairs dans un plan et si on veut les séparer par une méthode de seuils placés sur la droite, on se heurtera à de nombreuses difficultés. Si on connaît la nature algébrique du problème, on aura tout de suite le schéma à appliquer.

Donc, au point de vue de la structure du message vocal, l'analyse algébrique grammaticale, relationnelle, est à mon avis fondamentale.

J-J. MASSOT .

Les automates fabriqués sont beaucoup moins puissants et bien moins adaptés aux problèmes de reconnaissance de formes et de reconnaissance de parole en particulier. Il me semble qu'on parle des problèmes de reconnaissance de l'ensemble du système vocal, et dans l'état actuel de la technique, cela semble difficile.

J-Y. GRESSER .

Un automate est aussi bien capable de reconnaître des voyelles qu'un être humain, mais il faut chercher à faire mieux que l'être humain.

J-P. HATON .

Le problème est, à mon avis, une question de méthode à utiliser. GRESSER est le seul à avoir entrevu une solution pour l'instant. Disons qu'il s'agit avant tout de formaliser un modèle dans une machine.

J-Y. GRESSER .

Il y a effectivement des modèles différents. Mais certains problèmes mathématiques se posent toujours, notamment cette approche descendante décrite par ROSSI. Depuis le début de nos travaux, cela fait seulement un an que nous avons un modèle physique ou mathématique satisfaisant.

G. NOIZET .

Les objectfs sont différents. Le psycho-linguiste veut trouver comment le locuteur concret fonctionne.

La simulation nous intéresse dans la mesure où elle est un modèle du comportement du locuteur, mais il est évident aussi que l'on peut fabriquer un automate qui ne simule pas le comportement humain. En ce qui nous concerne, cela nous apporte beaucoup de voir dans quelle mesure des modèles de reconnaissance automatique peuvent être utilisés pour simuler le comportement et quelles sont les limites de cette simulation.

Je n'ai pas voulu opposer un modèle descendant à un modèle ascendant; le modèle de l'identification chez un sujet réel est un modèle en boucle qui fonctionne en parallèle; mais tout ceci reste à démontrer.

P. JUBAN .

Est-ce que les psycho-linguistes considèrent qu'il y a symétrie entre les modèles qui rendent compte de la perception et de la production chez un locuteur concret ? Il semble qu'il y ait un privilège conféré à la perception sur la production.

G. NOIZET .

Je ne pense pas que ce soit le même modèle théorique qui puisse rendre compte de la perception et de la production.

P. JUBAN .

N'est-ce pas privilégier un aspect particulier en faisant référence à la perception par rapport à la reconnaissance automatique de ne tenir compte que du côté perceptif

du locuteur concret comme une sorte de Pré-modèle à fournir à la reconnaissance automatique ?

S. CASTAN .

Puisque dans le Groupe de la Communication Parlée on essaye de faire des travaux en commun, ne pourrait-on pas entreprendre, en utilisant un même échantillon limité de voyelles, un travail où les psycho-linguistes feraient des expériences de reconnaissance par l'homme et d'autre part par la machine en utilisant les algorithmes de GRESSER et d'HATON.

J-P. HATON .

Nous n'avons pas fait de tests de reconnaissance de voyelles jusqu'à présent. Il est certain que dans diverses conditions (de bruit, par exemple), on peut envisager une machine qui fonctionnerait mieux que l'oreille humaine.

Fr. LONCHAMP .

Pour les chercheurs oeuvrant sur la reconnaissance de la parole, le problème est d'éliminer au maximum les ambiguïtés au niveau phonétique. Ils utilisent pour cela tout ce qui est techniquement possible même s'il est certain que ce n'est pas utilisé par l'homme.

En ce qui nous concerne, le modèle parfait serait celui qui rendrait compte des erreurs, qui permettrait de prévoir les erreurs, leur direction et leur grandeur.

A la limite, s'il existait un système simple qui permettrait d'avoir 100 % de reconnaissance, les études de reconnaissance de la parole s'arrêteraient tandis que celles des psycho-phonéticiens continueraient.

G. NOIZET .

Ce qui fait la limite du modèle descendant, c'est qu'il y a nécessairement un prétraitement, ce que NEISSER appelle dans "Cognitive Psychology" un niveau préattentif.

G. PERENNOU .

Des mots artificiels constitués à partir de formants sont-ils aussi bien reconnus que des mots naturels ?

A. MAISSIS.

Au niveau des voyelles synthétiques que nous avons utilisées pour nos expériences, disons qu'elles ont une qualité différente. En ce qui concerne leur perception et pour qu'une voyelle synthétique soit reconnue à 100 % il faut que les positions des formants ne correspondent pas à la réalité. Donc, il faut introduire une distorsion - ceci n'est peut-être pas dû uniquement aux formants mais aussi au rapport d'intensité entre les deux premiers formants, à la largeur de bande.

M. ROSSI .

Je pense que des indices comme l'intensité ou l'intonation jouent également un rôle important.

J-Y. GRESSER .

Comment les psycho-linguistes caractérisent-ils les erreurs des sujets dans les expériences ? Ont-ils des modèles-types de caractérisation de ces erreurs ?

G. NOIZET .

En psycho-linguistique, on utilise effectivement des matrices de confusion. Dans les tests de perception avec l'homme, non seulement il y a des erreurs, mais il peut y avoir une évolution dans les erreurs. Je pense que c'est là une information d'un type particulier. Les erreurs commises ont également pour effet de ralentir le processus perceptif. Le processus est retardé et, quand il y a persistance dans les erreurs, les seuils d'identification sont beaucoup plus élevés. Mais ce sont là des aspects assez spécifiques de la perception humaine.

J-P. HATON.

On peut s'étonner du fait que les méthodes de prédiction d'erreurs utilisées en psycho-linguistique sont finalement des méthodes mathématiques alors qu'on s'attendrait plutôt à une prédiction des erreurs qui provienne du modèle de perception lui-même.

G. NOIZET .

Que représente l'erreur ? C'est un premier bouclage erroné du processus d'identification. La supposition faite en particulier au niveau de la perception des phreses, c'est que l'hypothèse faite conduit à des heuristiques de type syntaxique.

M. ROSSI.

En ce qui concerne le taux de reconnaissance des unités phoniques, est-ce qu'on a tenu compte des valeurs contextuelles qu'elles prennent après telle ou telle consonne, ou encore des contraintes phonologiques, et dans quelle mesure cela peut-il faciliter la reconnaissance ?

J-Y. GRESSER .

Cela a été fait d'un point de vue statistique, mais autrement, je ne pense pas que cela ait été fait.

M. ROSSI .

Il s'agirait de remplacer un modèle probabiliste par un modèle déterministe qui serait à l'image du modèle linguistique.

R. DE MORI .

Personnellement, je travaille sur la recherche d'une grammaire qui décrive ces effets de l'interaction de la coarticulation. Le problème mis en évidence par LIBERMAN est à mon avis le suivant :

- dans la perception, il y a une force complexe de traduction des commandes. Dans la reconnaissance, il faut faire l'opération inverse à l'aide d'une grammaire.

M. ROSSI .

Vous revenez donc à un modèle d'analyse-par-synthèse.

R. DE MORI.

Ce n'est pas la même chose, car le modèle d'analyse-par-synthèse se fonde sur la minimisation de certains traits fonctionnels, tandis que notre modèle cherche à mettre en évidence les productions qui peuvent régler la génération du pattern d'une syllabe à partir du phonème.

Fr. LONCHAMP.

Aux Journées de LENINGRAD en 1966, LIBERMAN a montré que pour avoir des réponses unanimes de reconnaissance à la parole synthétique : exemple : " ϵ " isolé opposé à " ϵ " en contexte, ex. bϵt, les valeurs des formants devaient être légèrement différentes. Donc, il semble bien que le contexte joue un rôle important au niveau de la production.

M. ROSSI.

Il y a deux choses :

D'une part, l'analyse contextuelle qui peut se faire séquentiellement et d'autre part, les contraintes que l'on ne peut analyser que si l'on connaît les unités supérieures (par exemple, le mot ou la syllabe).

En italien, il est facile de reconnaître le mot parce qu'il porte un accent, mais c'est beaucoup plus délicat en français.

M. WAJSKOP.

L'intervention de LONCHAMP repose le problème des niveaux de calibrage dans les processus d'identification. Il serait plus prudent de revenir au modèle (cité par

NOIZET) de type descendant fonctionnant à l'aide de bouclage et d'un système de pré-traitement en parallèle. C'est à l'aide d'un tel système que se résolvent les problèmes d'identification grâce à des problèmes de suppléance.

En un sens, la position de GRESSER est donc raisonnable lorsqu'il affirme qu'à l'heure actuelle la machine est déjà plus puissante que l'homme. Mais cela est dû au fait que la machine fonctionne à des niveaux et dans des contextes pré-déterminés.

S. CASTAN.

La machine a de bonnes performances par rapport à l'homme dans des conditions très particulières pour un seul locuteur. Il n'en est malheureusement pas de même pour plusieurs locuteurs, car on ne connaît pas la loi de variation des paramètres d'un locuteur à un autre.

G. PERENNOU.

A propos du phonème de distorsion dû à la coarticulation, est-ce qu'on observe des phénomènes du genre illusion acoustique ? Par exemple, si le signal évolue vers "i", reconnaîtra-t-on le "i" avant de l'avoir atteint ?

M. ROSSI.

C'est certain; au niveau de l'identification dans la parole, toutes les unités dans le contexte vont subir le même glissement.

R. DE MORI.

Nous nous posons actuellement le problème de savoir quelle est la méthode d'analyse à employer pour faire de la reconnaissance. Aux Etats-Unis, certains chercheurs sont convaincus que pour reconnaître le langage continu, il faut utiliser des systèmes plus complexes que les bancs de filtres traditionnels et employer des techniques sophistiquées pour détecter les formants, faire des modèles du conduit vocal, etc....

Quelqu'un a-t-il essayé d'utiliser des méthodes numériques, c'est-à-dire le cepstrum ou l'analyse simultanée de la mélodie.

Cl. GUEGUEN.

Nous avons utilisé un modèle mathématique de la membrane basilaire, c'est-à-dire un système qui était régi par des équations ou dérivées partielles.

Ce système donne des paramètres qui sont d'origine spectrale si on se place en un point de la membrane basilaire; il met en relief des phénomènes de propagation le long de la membrane basilaire, c'est-à-dire que pratiquement on étire le signal et on l'analyse.

M. CARTIER.

La question est de savoir sur quels paramètres il faut travailler, car au départ, il faut bien comprimer les données de la parole. Ensuite, on fait des tests pour voir si la compression est pertinente ou pas. C'est une méthode possible, mais elle peut être dangereuse car, comme le disait WAJSKOP, avec des sons artificiels, on peut faire comprendre un peu n'importe quoi.

M. ROSSI.

A ce sujet, DELATTRE au cours d'une discussion animée avec FISCHER-JORGENSEN, faisait remarquer que les résultats obtenus lors de la synthèse étaient valables uniquement dans le cadre de ces expériences de synthèse, mais n'étaient pas applicables à la parole naturelle.

2ème thème de la discussion .

M. ROSSI.

Abordons le deuxième thème : l'utilité des méthodes de classification en linguistique ou inversément l'utilisation des modèles linguistiques dans des modèles de classification et de reconnaissance de la parole.

Cl. GUEGUEN.

Je serais personnellement tenté par l'utilisation des méthodes de classification automatique. Les raisons en sont nombreuses, d'abord parce qu'on les utilise depuis quelque temps déjà (exemple, l'analyse factorielle) et on utilise également la méthode

de DIDAY pour essayer d'optimiser une mesure de distance.

Au point de vue de l'amélioration des performances, tout l'aspect didactique et l'aspect recherche peut être gouverné par des approches du type classification automatique.

G. PERENNOU.

Je suis d'accord avec GUEGUEN dans la mesure où des méthodes ont chaque fois des résultats positifs mais lorsque les résultats sont négatifs, on ne peut rien en conclure. Je veux dire par là qu'une analyse factorielle avec baisse de dimensions n'est pas complète et tout ce qu'on a laissé de côté avec l'analyse factorielle est en fait peut-être la partie discriminante. En cela, la méthode ne donne aucun critère de jugement et c'est un inconvévient.

Cl. GUEGUEN.

L'analyse factorielle permet de savoir quels poids sont placés sur les différents axes, donc d'avoir une idée de l'approximation qui a été faite.

J-Y. GRESSER.

Mon impression est que quand on procède à une analyse factorielle, on obtient des facteurs que l'on ne sait pas interpréter.

W. SERNICLAES.

L'analyse des correspondances peut fournir une représentation simple des relations contenues dans certains types de données. Dans le cas d'une tâche d'identification, par exemple, comportant un grand nombre de stimuli et de réponses, cette méthode permet de saisir des aspects essentiels des résultats.

G. MERCIER.

En reconnaissance de la parole, on peut utiliser l'analyse factorielle à plusieurs niveaux.

D'abord, on peut segmenter la parole en syllabes ou bien en phonèmes (exemples : voyelles, consonnes, plosives, fricatives, etc.). A chaque niveau de la hiérarchie on peut utiliser certains facteurs. Ces facteurs peuvent être déterminés par les méthodes mathématiques de l'analyse factorielle ou par les méthodes de reconnaissance de formants, etc...

A chaque instant, donc, on réduit l'information pour faire une division, mais en gardant toujours un maximum d'information. Certains facteurs doivent être conservés mais, par contre, d'autres doivent être éliminés.

3ème thème de la discussion.

M. ROSSI - RAPPEL DU THEME :

La normalisation et la reconnaissance des voix individuelles doit-elle précéder les recherches sur la reconnaissance ?

J-Y. GRESSER.

Avec CARTIER, nous travaillons actuellement sur le phonétographe qui est une machine sensée réaliser une adaptation acoustique indépendamment du locuteur. Nous avons fait l'expérience de la reconnaissance d'un vocabulaire limité et avons procédé à une normalisation au niveau lexical. En d'autres termes, quand un mot est analysé par le phonétographe, il en résulte une séquence de phonèmes et pour améliorer les performances, nous avons cherché à adapter aux locuteurs, c'est-à-dire à faire en sorte que le système soit optimisé pour un nombre important de locuteurs. Malheureusement, notre nombre de locuteurs était limité à 15. Dans l'état actuel de nos recherches, il nous faut par mot du vocabulaire un représentant par locuteur, ce qui veut dire que si on a 100 locuteurs, il faut 100 représentants par mot, etc... Je pense qu'une adaptation au niveau lexical avec une reconnaissance acoustique relativement grossière est tout à fait inadéquate et qu'il faut plutôt lui substituer une adaptation au niveau acoustique. Cependant, l'inconvénient de la méthode de MAISSIS me semble être un temps relativement long à la normalisation.

A. MAISSIS .

Je ne pense pas que le temps soit excessivement long (environ 150 usec, par phonème).

Il faut, en effet, normaliser les sept paramètres du phonème par rapport au coefficient de chaque classe. Comme il y a 22 classes, cela revient à faire 22 normalisations, et donc à calculer 22 distances par rapport au centre de gravité de toutes les classes. Tout cela en 150 usec. Je ne pense pas que cela soit très long. A mon avis, on ne peut pas dire que chaque locuteur a ses propres caractéristiques mais plutôt qu'il y a des groupes de locuteurs.

M. ROSSI .

Faites-vous la normalisation pour un phonème entre les individus ou en fonction du système de phonèmes chez un individu ?

A. MAISSIS .

Dans chaque phrase, il y a plusieurs occurences de chaque phonème ("a" par exemple). Ce phonème est prononcé par plusieurs locuteurs et dans différents contextes. Ces diverses occurences représentent de nombreux cas possibles de prononciation du phonème "a". Dans tous les paramètres qui représentent ce phonème, on peut découvrir une relation qui n'est pas linéaire.

C'est en ce sens que mon approche s'écarte de l'analyse factorielle du fait que la loi que l'on découvre entre les paramètres est non-linéaire. C'est par le coefficient de cette relation qu'on peut par la suite normaliser les paramètres.

Suite à l'intervention de M. ROSSI qui se demande dans quelle mesure il est possible d'étudier la nomalisation au niveau des syllabes, je précise qu'à mon avis, l'application de cette idée serait très difficile, puisqu'elle exigerait l'examen de très nombreux cas particuliers. Par contre au niveau de paramétrisation du signal vocal, il est possible de concevoir deux jeux de paramètres; le premier donnera un information concernant les caractéristiques générales du locuteur, et permettra alors de normaliser le second. Autrement dit, la mesure de certains paramètres dits de normalisation permettra de rapporter les paramètres principaux à des conditions de locution standard, rendant ainsi possible le calcul d'une distance phonémique significative. Mon travail sur le sujet illustre cette idée et les résultats obtenus ne peuvent qu'encourager cette direction.

Cl. GUEGUEN .

Je voudrais préciser la finalité de la normalisation. L'idée essentielle est d'utiliser l'identification des locuteurs de manière à adapter de façon automatique les algorithmes de reconnaissance. C'est là une autre contribution de la classification à la reconnaissance des formes. Dans le cas de la reconnaissance du locuteur, le problème est d'enlever toute la partie significative de son message, et dans le cas de la reconnaissance de la parole, d'enlever toute la partie spécifique au locuteur. Ce sont deux problèmes qui doivent être traités en parallèle et notre but était d'utiliser l'identification du locuteur pour ajuster automatiquement les coefficients dans les algorithmes de reconnaissance. En d'autres termes, on ajuste les coefficients de la normalisation en fonction de la classe à laquelle le locuteur appartient.

R. CARRE .

Je pense que les paramètres normalisés devraient résulter d'une analyse par la synthèse appliquée à un modèle de fonctionnement de type articulatoire. Les paramètres détectés à l'analyse seraient les paramètres de commande élaborés au niveau du cerveau.

L. SANTERRE .

Je considère qu'il est impossible de dégager une performance-type au niveau de la programmation neurologique. Si l'on peut espérer dégager une compétence de la langue, il est impossible d'en faire autant de la performance du discours. Chacun parle différemment le matin et le soir. Il faut normaliser au niveau de la syllabe, du débit, du rythme, de l'accentuation, de l'intensité. Il est vain de normaliser un phonème car il varie trop en raison de la coarticulation du débit, du rythme, etc...

A. MAISSIS .

Du point de vue conceptuel, il semble que l'utilisation d'un modèle de conduit vocal soit une très bonne idée, mais finalement je pense que le problème de la normalisation doit être discuté au niveau des paramètres, c'est-à-dire au niveau des données que l'on peut tirer du signal lui-même.

L'idée qu'une étude conjointe des caractéristiques du canal vocal et de son système

de contrôle permettrait de concevoir une méthode de normalisation de la parole me paraît pour l'instant pratiquement inabordable. Le fait que les méthodes de modélisation du canal vocal utilisant une fonction de transfert abstraite ont eu un plus grand succès que les méthodes utilisant un modèle réaliste, reflète l'état actuel de nos connaissances sur la parole et appuient l'idée présentée plus haut selon laquelle c'est au niveau du traitement du signal qu'il faut, en premier lieu, concentrer nos efforts pour normaliser la parole.

CONCLUSION

De cette table ronde, on retire l'impression d'une discussion serrée autour du premier thème.

Les points de vue présentés montrent que la collaboration avec les psychologues- inexistante jusqu'à ce jour - est indispensable et doit se développer. La confrontation des modèles de reconnaissance par l'homme et par les automates peut aboutir à des résultats nouveaux.

Sur le deuxième thème, la discussion a pratiquement tourné court. Il sera nécessaire d'y revenir lors des prochaines Journées, car il intéresse la collaboration essentielle entre les linguistes et ceux qui travaillent sur la reconnaissance automatique de la parole.

De la discussion sur le troisième thème enfin, on retiendra deux idées essentielles :

- utiliser l'identification des locuteurs de manière à adapter de façon automatique les algorithmes de reconnaissance (GUEGUEN);

- identifier les paramètres de commande articulatoire, paramètres qui ne dépendent pas du locuteur et élimineront le problème de coarticulation (CARRE).

La perspective d'une application de ces deux idées, dans deux directions divergentes, est également riche de promesses.

M. ROSSI (Aix).

COMMUNICATION DE D. ROSTOLLAND .

La relation entre voix criée et voix amplifiée.

D. ROSTOLLAND.

Dans nos expériences, nous avons essayé de comparer l'intelligibilité de la voix criée et de la voix parlée amplifiée par haut-parleur. Cette étude est venue de considérations à la fois théoriques et pratiques, c'est-à-dire qu'il existe des courbes qui permettent de prédire l'intelligibilité de la parole dans le bruit et qui ont été tracées en effectuant une translation de la parole normale en ajoutant 6 db pour passer à un niveau fort, 6 db à un niveau très fort, etc...

Nous avons donc pensé que la distorsion introduite dans la voix criée devait en modifier l'intelligibilité.

Nous avons utilisé comme matériel verbal les listes disyllabiques de FOURNIER. En ce qui concerne les résultats, nous avons trouvé que l'ordre de masquage successif de 3 types de voix était le suivant :

- on commence par masquer la voix criée, ensuite la voix chuchotée et enfin la voix parlée.

Nous avons travaillé à 3 niveaux différents :

- 55 db correspondait au niveau de la voix chuchotée
- 80 db " " parlée
- 100 db " " criée

Dans tous les cas, on a trouvé que la voix parlée amplifiée est de loin mieux comprise que les deux autres types de voix.

Nous avons également comparé les positions formantiques entre les différents types afin de voir comment le triangle vocalique se déplace.

En général, on peut dire qu'en voix criée :

- le F1 a tendance à s'élever
- le F2 est à peu près stable
- le F3 a tendance à diminuer.

QUESTION : Avez-vous travaillé sur l'intensité des formants ?

D. ROSTOLLAND .

On a l'impression que les formants en voix criée ont tendance à se rapprocher d'un certain niveau et, en ce sens, on peut parler d'un nivellement des trois formants sur le plan de l'intensité.

Fr. LONCHAMP.

Pour la voix criée, il y a une forte élévation de Fo, avez-vous contrôlé ce paramètre au niveau du masquage ?

Quel était le type de test que vous avez utilisé ?

D. ROSTOLLAND.

Les sujets devaient répéter les mots et on comptait le pourcentage des mots bien reconnus.

Fr. LONCHAMP.

On sait que l'onde de pression glottique change de forme, ceci voulant dire que la distribution de l'énergie aura une importance sur les fréquences. A mon avis, c'est ce qui modifie la hauteur des formants et cela pourrait expliquer les différences que vous constatez dans votre expérience.

D. ROSTOLLAND .

Sur les sonagrammes, on se rend compte qu'il y a moins d'informations en voix criée parce que les formes spectrales sont moins bien définies, donc il est normal que la reconnaissance soit moins facile qu'en voix parlée.

Fr. LONCHAMP .

Il me paraîtrait intéressant de faire une expérience sur les trois types de voix avec la même valeur Fo afin d'avoir une comparaison plus juste au niveau des performances de reconnaissance.

TABLE RONDE SUR LA SYNTHESE PAR REGLES

J. VAISSIERE : La synthèse par règles consiste en la conversion automatique d'une suite de symboles représentant un texte sous sa forme écrite en une onde sonore continue qui représente une des images acoustiques possible de ce texte. Les programmes de synthèse consistent essentiellement à transformer la suite de symboles en une série de paramètres variables dans le temps, capables de commander un synthétiseur. Ces paramètres sont de deux natures : d'une part, les paramètres acoustiques qui permettent de synthétiser les phonèmes ou allophones successifs représentant le texte (valeur des formants à chaque instant t, ou valeur de l'énergie dans les 15 ou 20 canaux d'un synthétiseur à canaux par exemple), et d'autre part les paramètres prosodiques, tels que la durée et la hauteur à affecter à chaque phonème.

J. GENIN : Il semble que vous restreignez les possibilités de synthèse par règles aux seuls synthétiseurs à canaux et à formants, alors que l'on peut l'étendre à toutes sortes de synthétiseurs, depuis un simple haut-parleur attaqué par un codeur jusqu'au simulateur du conduit vocal. Un des avantages serait par exemple l'utilisation de paramètres physiques (pression, tension des cordes vocales) pour la génération des paramètres prosodiques.

E. LHOTE : Les recherches faites sur la synthèse par règles n'apportent-elles pas une contribution aux recherches faites sur la redondance du langage au niveau acoustique?

J. VAISSIERE : La synthèse par règles est justement rendue possible par le caractère redondant de la parole, et elle essaie d'exploiter au maximum cette redondance afin de réduire le plus possible le taux de données nécessaire au codage de la voix. Elle contribue donc directement aux recherches faites dans ce sens.

E. LHOTE : A-t-on un élément redondant au niveau du spectre de repos? Ne peut-on pas en synthèse supprimer le spectre de repos et garder seulement les spectres de transition afin de tester l'information apportée par les spectres de repos?

J. VAISSIERE : Redondance et continuité dans le temps sont 2 phénomènes qui vont de pair. Dans ce sens, les spectres de repos, correspondant aux zones relativement les plus stables de l'onde, c'est-à-dire aux zones dans lesquelles les paramètres évoluent

le moins rapidement, sont plus redondants que les spectres de transition où la situation est inverse. Les occlusives, telles que "p, t, k, b, d ou g" ne possèdent pas de partie stable et une description de leurs transitions avec les phonèmes adjacents en fonction du temps est nécessaire et suffisante à leur synthèse. Mais les spectres de repos contiennent l'essentiel de l'information sur la nature de la plupart des autres phonèmes et ils ne peuvent être supprimés.

J. SAP : Pensez-vous qu'il soit absolument nécessaire d'engendrer automatiquement la fréquence du fondamental pour créer une parole intelligible?

J. VAISSIERE : De même que nous pouvons parler à pitch constant et être cependant compris, une parole synthétique à fréquence fondamentale fixe peut être parfaitement intelligible , mais elle ne pourra être "naturelle". La décision d'insertion d'un programme de génération de fréquence du fondamental et de la précision apportée par ce programme dépend de la qualité désirée; ce programme peut se limiter à engendrer une macromélodie plus ou moins grossière (nous entendons par macromélodie la courbe générale du fondamental, exempte des détails et des perturbations dus à la nature des phonèmes) ou insérer également la micromélodie.

J.N. CONTENSOU : Est-ce que la micromélodie et la mélodie sont deux problèmes bien distincts ou se recouvrent-ils?

J. VAISSIERE : Ce sont deux problèmes bien distincts. La micromélodie est due à des phénomènes de coarticulation, alors que la ligne générale de mélodie est liée à la structure syntaxique de la phrase, à sa sémantique, à la place des accents, etc...
On peut parler d'une voix monocorde et ceci dans n'importe quelle langue, les variations dues à la coarticulation des phonèmes (micromélodie) se conservent intégralement. Une occlusion partielle (comme pour la consonne "v") ou totale (comme pour la consonne "b") crée obligatoirement des différences de pression dans le conduit vocal, ce qui provoque des perturbations de la ligne générale du fondamental. En fait, les effets de micromélodie se superposent à la ligne générale de macromélodie et la résultante est la courbe du fondamental que nous obtenons en analysant la voix.

J. GENIN : La micromélodie des consonnes va toujours dans le sens d'un fléchissement des valeurs du fondamental.

J.N. CONTENSOU : La mélodie prend-elle parfois une importance telle que la micromélodie soit masquée?

J. VAISSIERE : La micromélodie des consonnes apparaît plus nettement lorsqu'elle se superpose à une mélodie générale ascendante. Les effets de micromélodie sont très nettement diminués dans le cas d'une chute rapide de la mélodie, pour des mots prononcés à un rythme rapide. A un rythme moyen ou lent,la micromélodie n'est jamais masquée, quelle que soit la mélodie.

J. SAP : Quel est le rôle joué par la micromélodie dans la perception du phonème lui-même?

J. VAISSIERE : L'insertion de la micromélodie des occlusives et des fricatives voisées améliore l'intelligibilité de ces phonèmes. En général, la micromélodie contribue au naturel de la voix.

J. GENIN : Nous avons parlé de la micromélodie des consonnes. Et les voyelles?

J. VAISSIERE : Alors qu'une occlusive et une nasale ont une micromélodie différente, les 15 voyelles françaises ont même micromélodie lorsqu'elles occupent même position dans un groupe de mots. Elle présentent beaucoup plus de stabilité que les voyelles américaines (qui peuvent être plus ou moins diphtonguées et dont la forme dépend aussi de la place du stress dans le mot). La micromélodie finale et initiale des voyelles dépend de la nature des phonèmes environnants : après une occlusive sonore telle que "b", la fréquence du fondamental croît au début de la voyelle, alors qu'elle décroît après une occlusive sourde.

Mais la forme de la voyelle est essentiellement déterminée par sa position dans le schéma général de macromélodie, qu'elle suit dans son ensemble. Par ce biais, la forme de chaque voyelle est étroitement liée à la place des accents de groupe dans la phrase.

J.N. CONTENSOU : Vous faites intervenir accent et micromélodie. Y a-t-il des règles de corrélation entre la micromélodie et l'intensité?

J. VAISSIERE : L'impression subjective d'accent est due en français à des variations de durée, et non d'intensité, les syllabes recevant l'accent étant plus longues que les autres. Il n'y a donc pas de rapport étroit entre accent et intensité. Par contre, la micromélodie

des consonnes est fortement liée à l'intensité : la valeur minimale de fréquence du fondamental correspond au minimum d'énergie au cours de l'élocution des occlusives et des fricatives voisées, et ces deux minima correspondent au moment précis de l'occlusion.

E. RUSCONI : De quel ordre est l'importance de la micromélodie?

J. VAISSIERE : Son importance varie avec le locuteur. La déviation est environ de 20 à 30% de la valeur de la voyelle précédente.

R. DESCOUT : Etant donné que la micromélodie vient principalement de la coarticulation, donc de quelque chose de naturel, on devrait pouvoir trouver un modèle physique l'approchant que l'on puisse appliquer dans tous les cas puisque ce phénomène semble indépendant de tout le reste.

J. GENIN : Nous pouvons renvoyer l'assistance à la lecture du dernier article de K. Ishizaka et J.L. Flanagan : "Synthesis of voiced sounds from a two-mass model of the vocal cords" B.S.T.J. Vol. 51. N° 6, Juillet-Août 1972. Ces articles (ainsi que bien d'autres peut-être moins convaincants) montrent la relation qui existe entre la pression intraglottique et la fréquence des vibrations des cordes vocales. Partant de là, s'il y a variation de cette pression il y a automatiquement variation du fondamental dans un sens et avec une amplitude théoriquement prévisibles. Ce qui reste à établir est l'amplitude. On peut espérer dériver par un modèle de ce genre les écarts de micromélodie à partir des écarts de pression de la désadaptation du conduit vocal et des considérations de ce genre là...

A-t-on actuellement des idées sur les valeurs quantitatives à donner à la mélodie?

J. VAISSIERE : Le respect de la forme générale de la courbe du fondamental est plus important que les valeurs. Ceci dit, pour une utilisation rationnelle des bits de quantification en fréquence, il y a intérêt à adapter la plage de variation en fréquence afin de ne pas réduire la dynamique en s'encombrant de niveaux jamais atteints par une seule voix de synthèse. Pour une synthèse représentant une voix masculine, 120 hz de plage me semblent suffire de 80 à 200 hz.

J. GENIN : Quelle est la signification de ce paramètre d'amplitude?

J. VAISSIERE : L'amplitude est fonction du locuteur, et du rythme avec lequel il parle. Ainsi par exemple, une relative pourra avoir une plage réduite et s'opposer au reste

de la phrase par un rythme plus rapide. Les séquences porteuses d'information sont prononcées plus lentement.

J. GENIN : Il y aurait donc un paramètre, de nature sémantique peut-être, qu'il faudrait tenir en compte pour les débattements des écarts de mélodie?

J. VAISSIERE : En synthèse par règles, il ne faut peut-être pas s'encombrer de détails concernant la sémantique de la phrase. On choisit alors un rythme a priori et une amplitude de variations du fondamental correcte. Comme nous ne pouvons avoir accès automatiquement à la sémantique de la phrase, et qu'une simple analyse syntaxique est insuffisante à déceler les zones de la phrase porteuses de l'essentiel de l'information, nous nous contenterons de synthétiser tous les groupes de mots de la phrase avec le même rythme.

J.S. LIENARD : Que pensez-vous de l'importance relative de la mélodie en français et en anglais. Personnellement, je pense qu'elle a moins de valeur phonétique en français qu'en anglais. Une phrase prononcée en anglais sans mélodie est beaucoup moins compréhensible qu'une phrase française prononcée dans les mêmes conditions. Nous faisons naturellement la distinction entre valeur phonétique et valeur expressive.

Mme LHOTE : Le langage scientifique français est en effet assez peu marqué et il est simple du point de vue prosodique. Mais le langage français non scientifique a une intonation beaucoup plus riche et plus variée que son équivalent anglais.

J. VAISSIERE : Si le rythme semble jouer un rôle comparable dans les deux langues en question et si l'intensité joue certainement un rôle plus important en anglais qu'en français (où toutes les voyelles sont plus tendues), il est délicat de comparer le rôle respectif de la mélodie en français et en anglais. Le rapport très précis entre la fréquence du fondamental et la position du stress en anglais n'est pas encore, à ma connaissance, clairement déterminé (disons qu'il y a plusieurs écoles...). En conséquence de quoi, il est difficile d'évaluer la valeur distinctive des évolutions du pitch, dans l'une et l'autre langue.

J.N. CONTENSOU : Le calcul des transitions entre les formants pose encore à l'heure actuelle un problème très difficile à résoudre. Par contre, les problèmes posés par la prosodie semblent déjà plus accessibles. Si on renonce à une synthèse complète par

règles (qui sous-entend le calcul des transitions entre les phonèmes) à cause des difficultés posées par les règles de coarticulation et si on adopte la solution d'une synthèse faite à partir de mots préenregistrés, le vocabulaire sera par le même coup limité. Il est probable que le nombre de phrases synthétisables sera lui aussi limité. S'il est suffisamment limité, il est plus économique de mettre en mémoire les occurrences de prosodie de ces phrases, plutôt que de générer la prosodie de ces phrases par un système de règles plus ou moins complexes et qui permettent d'engendrer la mélodie de phrases quelconques.

A part le cas de la génération de nombres, a-t-on déjà l'exemple d'applications en synthèse par mots où il soit plus rentable de mémoriser les règles de prosodie plutôt que les occurrences de prosodie?

J. VAISSIERE : En synthèse de nombres à partir d'éléments préenregistrés, une solution mixte est en effet en général adoptée : mémorisation d'une ou plusieurs occurrences de mélodie pour un seul élément et règles de synthèse pour assurer la continuité de la courbe du fondamental entre les éléments mis en séquence (Rabiner, Schafer et Flanagan: Computer Synthesis of Speech by Concatenation of Formant-Coded Words. The Bell System Technical Journal. Vol. 80. 1971) ou pour modifier les valeurs absolues de l'occurrence de prosodie mise en mémoire en fonction de la position de l'élément dans le nombre (Genin : An Audio Response Unit for Telephone Needs. 1972. Conference on Speech Communication and Processing).

Dans le cas des synthèses actuelles de phrases (qui ne sont pas des nombres) faites à partir d'un vocabulaire de base, les mots sont mis en mémoire avec un certain contour du fondamental, correspondant à celui avec lequel le locuteur a prononcé les mots lors de l'enregistrement. Les phrases sont formées par simple concaténation des éléments de base et il en résulte un défaut de mélodie. Les difficultés d'insertion de la micromélodie font qu'il est sans doute peu rentable de créer la ligne mélodique ab nihilo : on peut garder en mémoire les informations concernant la micromélodie de chaque mot et générer seulement la macromélodie en fonction de la position de chaque mot dans la phrase. De même la mise en mémoire de deux occurrences de mélodie (par exemple une mélodie montante et une mélodie descendante pour le même mot) peut être envisagée : le choix entre l'une des deux occurrences est alors décidé par une analyse syntaxique sommaire de la phrase.

J. GENIN : Il ne me semble pas qu'enregistrer plusieurs occurrences de mélodie pour un seul mot soit une solution rentable. Nous avons fait au CNET une étude sur la

synthèse des nombres à l'aide d'un vocodeur à canaux. Chaque mot a été enregistré dans un contexte isolé, le plus recto-tono possible de façon à avoir dans les données du vocodeur uniquement la micromélodie. On a constaté (et cette constatation est peut-être due à la faible qualité du vocodeur) qu'il suffisait de corriger cette ligne de base par des variations linéaires du pitch : on détermine en fonction de la place du mot dans le nombre la valeur à partir de laquelle il faut démarrer la fréquence du fondamental et la pente que l'on choisit parmi les 2ou 4 possibles. Il n'est également pas difficile de modifier la durée de ces éléments (par exemple par un basculement de l'horloge de 10 msec à 20 msec durant la synthèse de certaines zones du nombre, dans le vocodeur à canaux). Des règles telles que celles que je viens de décrire sont d'autant plus nécessaires dans le cas de synthèse de mots ne représentant pas des nombres, où il est encore plus difficilement concevable d'avoir à mémoriser plusieurs occurrences de mélodie.

J. GUIBERT : Une voix naturelle n'est jamais parfaitement périodique au niveau même des cordes vocales et le degré d'apériodicité varie d'un locuteur à l'autre : il est particulièrement important chez les sujets anormaux. Les articles lus à ce sujet remontent à quelques années (P. Lieberman : Perturbations in Vocal Pitch. J.A.S.A. Vol. 33. p. 597) et du même auteur : Some Acoustic Measures of the Fundamental Periodicity of Normal and Pathological Larynges. J.A.S.A. Vol. 35, p. 344). J'aimerais savoir s'il y a des études récentes faites dans ce domaine et si les résultats sont utilisés en synthèse?

J. VAISSIERE : Selon B.L. Cardoza et R.J. Ritsma (Conference on Speech Communication and Processing. Cambridge 1967; The Perception of Imperfect Periodicity), le son n'est plus perçu comme voisé si les variations excèdent quelque 10% de la valeur de la période, et il est perçu comme parfaitement voisé si les variations sont inférieures à quelques %. Peut-être cette apériodicité contribue-t-elle au naturel de la voix, mais il ne me semble pas qu'elle soit utilisée en synthèse.

J. GENIN : Un défaut des voix de synthèse vient justement de ce que la voix est parfaitement périodique et il est par contre prouvé qu'ajouter du bruit aux sons voisés améliore nettement l'intelligibilité et le naturel de la parole. Une transition d'amplitude par exemple (cf : Thèse de R. Carré) provoque un élargissement des raies, une meilleure localisation de l'énergie dans l'espace, et par là, une meilleure définition de la fonction de transfert du canal vocal.

J. VAISSIERE : Le fait de rajouter ou non du bruit durant la synthèse des voyelles me semble d'importance secondaire, bien qu'il soit prouvé que cela améliore effectivement

le naturel de la voix (S. Maeda et O. Fujimura : Factors of the Glottal Wave that contribute to the Naturalness of Speech. Logopedics Phoniatrics. Annual Bulletin N° 2. Tokyo 1968). Il est plus important de générer une courbe du fondamental qui ait des pentes "naturelles".

J. SAP : J'ai entendu dire que la perception de l'intonation était liée à la position des formants non seulement dans le cas de la voix chuchotée (Meyer-Eppler : Realisation of Prosodic Features in Wihspered Speech. J.A.S.A. 29, n° 1. 1957), mais aussi dans la voix naturelle.

J. GENIN : Il y a une explication biologique à ce phénomène. Pour augmenter la fréquence du fondamental, on augmente la pression de l'air dans les poumons, ce qui provoque un mouvement vers le haut des organes subglottaux. Cette élévation entraîne une diminution de la longueur du conduit vocal, ce qui influence la hauteur des formants.

J.S. LIENARD : Il y a en fait deux sortes de perception de hauteur : une perception de la hauteur du spectre de raies et une perception de la dominance d'énergie en fréquence. On peut ainsi attribuer une certaine hauteur à un bruit.

J. GENIN : Le problème essentiel de la synthèse par règles reste celui de la définition des paramètres acoustiques ou articulatoires pour la synthèse des phonèmes et de leurs transitions avec les phonèmes adjacents.

E. RUSCONI : Nous essayons d'établir un tableau équivalent à celui qu'a donné Mattingly pour l'anglais (J. Holmes et I. Mattingly : Speech Synthesis by Rules. Language and Speech n° 7. 1964) pour la langue italienne.

(Le groupe de M. Rusconi travaille depuis 1964 sur la parole, à l'institut de Polytechnique et d'Electrotechnique de Turin dont le Directeur est le Professeur R. Sartori).

J. VAISSIERE : Un tel tableau n'existe pas pour la langue française et il reste à établir pour une synthèse par règles sur synthétiseur à formants. Des règles ont été établies pour un synthétiseur à canaux par A. Nemeth.

En conclusion, beaucoup de recherches restent à effectuer en ce qui concerne la langue française : paramètres formantiques pour une synthèse à formant, paramètres articulatoires pour une simulation du conduit vocal. Les outils sont déjà définis, mais

les règles pour les commander le plus adroitement possible font défaut, du moins en partie. On sait déjà que la voix obtenue par synthèse par règles peut être de bonne qualité (l'expérience la plus récente que je connaisse est celle de D. Klatt au M.I.T. ACOUSTIC Acoustic Theory of Terminal Analog Speech Synthesis. Conference on Speech Communication and Processing. Boston 1972). Et les applications de cette synthèse dépendront sans doute de plus en plus de la qualité de la voix de synthèse obtenue, le prix des mémoires allant en diminuant.

Texte établi par J. VAISSIERE
(M.I.T.)

LISTE DES PARTICIPANTS

M. ABRY, C.	Université de Grenoble
Mme ADRIAEN, M.	Université de Toronto
M. ALEGRIA, J.	Université Libre de Bruxelle, Lab. Psychologie Expérimentale
M. APELDOORN, N.	Université Libre de Bruxelles, Institut de Phonétique
M. BAETENS BEARDSMORE	Université Libre de Bruxelles, Institut de Phonétique
M. BARIAUX, D.	U.L.B. - Faculté des Sciences Appliquées
M. BAUDRY, M.	C.E.A. Saclay
M. BEECKMANS, R.	U.L.B. - Institut de Phonétique
M. BELLISSANT, C.	IMAG, Grenoble
M. BERTELSON, P.	U.L.B., Lab. de Psychologie Expérimentale
M. BOE, L-J.	Université de Grenoble
M. BOSQUET	U.L.B. - Faculté des Sciences Appliquées
M. BOTHOREL	Université de Strasbourg
Mme BOUARD, D.	CNET, Lannion
M. BOURGENOT	THOMSON - Paris
M. BUISSON, L.	CNET, Lannion
M. BUTCHER, A.	Institut für Phonetik. Université de Kiel
M. CAELEN, J.	I.U.T. Toulouse
M. CARAYANNIS, G.	ENST - Paris
M. CARRE, R.	ENSERG. Grenoble
M. CARTIER, M.	CNET. Lannion
M. CARTON, F.	Institut de Phonétique, Nancy.
M. CASTAN, S.	I.U.T. Toulouse
M. CHEVALIER, H.	C.E.A. Grenoble
M. CHEVRIE, C.	INSERM. Paris
Mme CHEZE, M.	Institut de Phonétique, Université de Paris
M. COLIN, B.	THOMSON, Paris
M. COLLIER, R.	Katolieke Universiteit van Leuven
M. CONTENSOU, J-N.	IRIA. Rocquencourt
M. CONTINI, M.	Université de Grenoble
M. COSTERMANS, J.	Université Catholique de Louvain
M. DEBROCK, M.	Katolieke Universiteit van Leuven
Mme DE KONINCK, A-M.	U.L.B., Laboratoire de Psychologie Expérimentale
M. DEMAN, P.	THOMSON - Paris
M. DE MORI, R.	Politecnico, Turin
M. DESCHAMPS, R.	Radio Télévision Belge

M. DESCOUT, R.	CNET - Lannion
M. DE VRIENDT Séra	U.L.B. / V.U.B.
M. DIDAY, E.	LABORIA - IRIA - Paris
M. DISSOUBRY, R.	Paris
M. DOURS, D.	I.U.T. Toulouse
M. DREYFUS-GRAF	Ing. EPFZ - Genève
M. DUPEYRAT, B.	C.E.A. Saclay
M. EL-MALLAWANY	CNET - Lannion
M. ESTEBAN, D.	I.B.M. - La Gaude
M. FACCA, R.	I.U.T. - Toulouse
M. FORET, J.	Lab. Phys. Travail - Paris
Melle GALVAGNY, M-H.	Université de Mons
M. GARDERET, P.	C.E.A. Grenoble
M. GENIN, J.	CNET - Lannion
M. GRAILLOT, P.	CNET - Lannion
Melle GRIJPMA, D.	U.L.B. - Institut de Phonétique
M. GRESSER, J-Y.	CNET - Lannion
M. GUEDJ, R.	Thomson - Orsay
M. GUEGUEN, C.	ENST - Paris
M. GUIBERT, J.	CGE - Marcoussis
M. HATON, J-P.	Universite de Nancy
M. HENNEBERT, D. Dr	Hôpital Universitaire St-Pierre - Bruxelles
M. HENNEBERT, P. Dr	Bruxelles
M. HOLENDER, D.	U.L.B. - Lab. Psychologie Exp.
M. HOUGARDY, J.	U.L.B. - Faculté des Sciences Appliquées
M. JACOBY, A.	U.L.B.- Psychologie-Pédagogie
M. JOSPA, P.	U.L.B. Institut de Phonétique
M. JUBAN, P.	U.E.R. - Rennes
M. KAMMINGA, C.	Université de Delft
M. KENNY-LEVICK, S.	U.L.B. - Institut de Phonétique
Mme KONOPCZYNSKI, G.	Université de Besançon
M. KOSTER, J-P.	Institut de Phonétique- Hambourg
M. LAMB, J-M.	U.L.B. - Institut de Phonétique
M. LANDERCY, A.	Université de Mons

M. LAURENT, G.	S.L.E. - CITEREL - Lannion
M. LEBART	CREDOC - Paris
M. LEBRUN, Y.	U.L.B.
Melle LECOMPERE T.	Université de Besançon
Mme LHOTE, E.	Université de Besançon
M. LEROY M	Pro-Recteur de l'U.L.B.
M. LIENARD, J-S.	CNRS - Orsay
M. LONGCHAMP, Fr.	Université de Nancy II
M. LOOSE, P.	Anvers
M. MAISSIS, A.	ENST - Paris
M. MASSOT, J-J.	C.G.E. - Marcoussis
M. MEHLER, J.	CNRS - Paris
M. MERCIER, G.	CNET - Lannion
Mme METTAS, O.	CNRS - Paris
M. MEZZALAMA M.	Université de Turin
M. MLOUKA, M.	LIMSI - Orsay
M. MORAIS, J.	U.L.B. - Lab. Psychologie Exp.
M. MOST, R.	U.L.B. - Institut de Phonétique
M. NEMETH, A.	I.B.M. - La Gaude.
M. NOIZET, G.	Université de Provence
M. PAU, L-F.	IMSOR - Copenhague
M. PECKELS, J-P.	I.B.M. La Gaude
M. PERENNOU, G.	I.U.T. - Toulouse
M. PERROT, J.	Université de Paris III
Mme PETIT, A-M.	Université de Provence
M. PINEL, J.	Thomson - Paris
M. PYNTE, J.	Université de Provence
M. PYO, J.	Université de Besançon
M. QUERRE, M.	CNET - Lannion
Melle RADEAU, M.	U.L.B. - Lab. de Psychologie Exp.

M. RENARD, G.	LIMSI. Orsay
M. RENARD, R.	Université de Mons
M. RENKIN, A.	Université Libre de Bruxelles. Institut de Phonétique
M. ROCHE, C.	Lab. Central de l'Armement. Arcueil
M. ROGER, G.	C.G.E. Marcoussis
M. ROINSOL, G.	Institut National de Jeunes Sourds. Paris
M. RONGY, J.	Telecontrol. Bruxelles
M. ROSSI, M.	Université de Provence. Aix
M. ROSTOLLAND, D.	Lab. Physiologie du Travail. Paris
M. ROUX, M.	ISUP - Paris
M. RUSCONI, E.	Université de Turin
M. RUWET, N.	Université de Paris III
M. SAP, J.	Laboratoire de Marcoussis
M. SANTERRE, L.	Université de Montréal
M. SCHEELINGS, F.	U.L.B. Institut de Phonétique
M. SERNICLAES, W.	U.L.B. - Institut de Phonétique
Mme SIMON, P.	Université de Strasbourg
M. TEIL, D.	LIMSI - Orsay
M. TESTON, B.	Université de Provence
M. THEYSKENS, L.	U.L.B. - Institut de Phonétique
M. TILKOV, D.	Université de Sofia
Mme TISSEYRE, Fr	U.L.B. - Lab. Psychologie Exp.
Melle VAISSIERE, J.	CGE - Marcoussis
M. VIVES, R.	CNET - Lannion
M. VANDERHAEGEN, C.	U.L.B. - Lab. Psychologie Exp.
Mme VAN HOUT A.	Hôpital St Pierre- Neurolinguistique
Mme VEILLON, Fr.	Université de Grenoble
M. WAJSKOP, M.	U.L.B. - Institut de Phonétique
M. WIOLAND, F.	Université de Strasbourg.