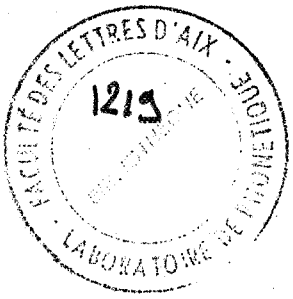
GROUPE *COMMUNICATION PARLEE* DU

GROUPEMENT des
ACOUSTICIENS de
LANGUE
FRANÇAISE



JOURNEES D'ETUDES
SUR LA PAROLE
GRENOBLE

o o o
o o
o

19 - 20 FEVRIER 1970

GROUPEMENT des ACOUSTICIENS de LANGUE FRANÇAISE
Groupe " Communication parlée "

J O U R N E E S D ' E T U D E S S U R L A P A R O L E

19 - 20 Février 1970

G R E N O B L E

Exposés et Discussions sur quelques problèmes de perception, synthétiseurs, synthèse par règles et reconnaissance de la parole.

Les Journées d'Etudes sur la Parole qui se sont déroulées les 19 - 20 février 1970 à l'Ecole Nationale Supérieure d'Electronique et de Radioélectricité de Grenoble ont été organisées avec la participation du Comité spécialisé : Intelligence artificielle et Reconnaissance des Formes de l'A.F.C.E.T. pour répondre principalement à deux préoccupations :
- .réunir des chercheurs travaillant en France dans le domaine de la parole,
- .aborder quelques thèmes précis, chacun des thèmes faisant l'objet d'un exposé de synthèse suivi de discussions.

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

.S O M M A I R E.

LA PERCEPTION DE LA PAROLE : ORIENTATIONS ET PERSPECTIVES (*)

par

M. W A J S K O P

Institut de Phonétique - Université libre - Bruxelles

Une distinction commode permet de classer rapidement la plupart des travaux menés dans le domaine de la perception de la parole suivant un axe méthodologique qui sépare les recherches s'effectuant dans le cadre d'un *modèle actif* de celles que l'on peut regrouper sous la rubrique du *modèle passif* .

C'est au physicien anglais MACKAY (1952) que l'on doit la meilleure formulation de ces deux modèles. Hypothèse centrale de son modèle actif : un acte de duplication qui génère une copie du stimulus à reconnaître. Cette opération exige de l'organisme qu'il dispose d'un nombre fini de commandes agissant sur les effecteurs du processus de copie. " *The elementary acts of replication define the basic vocabulary in terms of which the artefact describes its own experience.* ... And since our artefact uses as its elementary symbols the elementary acts of replication, no problems of *identification* arise. Complex concepts are represented by complexes of symbolic (internal) acts of representation. " (p. 114).

C'est ce texte qui fournira aux théories dites actives une infrastructure psycho-philosophique confortable.

1. *La THEORIE MOTRICE DE LA PAROLE* a pris son point de départ dans les Laboratoires Haskins auxquels nous sommes redevables d'excellents travaux sur les indices acoustiques de la parole et auxquels s'attachent les noms de F.COOPER, A.LIBERMAN et celui du regretté P.DELATTRE. Ces chercheurs ont été frappés par la multiplicité et la simultanéité des indices qui "marquent" une seule caractéristique phonétique. Les investigations qu'ils entreprennent entre 1952 et 1962 sur la discrimination et la reconnaissance des unités phonologiques les conduisent à la formule suivante (A.LIBERMAN, 1957) : " Articulatory movements and their sensory effects mediate between the acoustic stimulus and the event we call perception. " (p. 122).

(*) Texte résumé de la conférence faite le 19 février 1970 .

La netteté de cette affirmation est cependant tempérée par la notion que la référence aux mouvements articulatoires ne s'exprime pas au niveau périphérique. En 1962, l'hypothèse sera renforcée : " In any event, there is evidence from perceptual studies that speech sounds are perceived by reference to the articulatory movements that produce them, and that this articulatory reference is important for the distinctiveness of speech as perceived. " (p. 6).

Pour A.LIBERMAN, cette hypothèse se fonde d'abord sur la relation bi-univoque qu'il discerne entre les commandes neuromotrices et la perception du phonème et ensuite sur le fait qu'il est impossible de retrouver une telle relation entre la perception du phonème et le signal acoustique.

Cette conclusion, comme celle énoncée au 18ème Congrès de Psychologie (Moscou, 1966), fait suite à une série de travaux :
- recherches psychophonétiques sur la perception catégorielle des consonnes, la perception "continue" des voyelles et des éléments prosodiques, l'ambiguïté des indices acoustiques et le recouvrement spatio-temporel que l'on observe entre les indices et les phonèmes ;
- analyses électromyographiques mettant en relief la stabilité de l'image neuro-musculaire des sons de la parole ;
- étude des phénomènes associés à la latéralisation de la parole.

Pour A.LIBERMAN et ses associés, la perception phonématique de la parole reste un postulat de base, l'encodage syllabique permet de réduire le nombre d'unités discrètes qui doit être perçu par unité de temps et de contourner l'obstacle que constitue la résolution temporelle de l'oreille : " ... 30 sounds per second would overreach the temporal resolving power of the ear : discrete acoustic events at this rate would merge into an unanalyzable buzz... " (A.LIBERMAN, 1967).

2. Quant au *MODELE D'ANALYSE-PAR-SYNTHESE* proposé par STEVENS (1960) et HALLE-STEVENS (1962), il dérive plus directement de la pensée de MACKAY et des travaux sur la synthèse artificielle de la parole effectués dans le Laboratoire d'Electronique du M.I.T. Au coeur du système fonctionne une boucle de rétroaction active où l'onde sonore se transforme en une unité spectrale grâce à une comparaison constante entre le spectre fourni à l'entrée et sa copie générée à partir de "règles" stockées en mémoire. L'ordre dans lequel seront essayées les différentes descriptions spectrales dépendra de trois facteurs :
   - les données fournies par l'analyse préliminaire du signal,
   - les données provenant des informations spectrales déjà déposées dans la mémoire,
   - les données provenant des essais en cours.

Des descriptions articulatoires aux phonèmes, de ceux-ci aux séquences phonémiques, puis aux mots et aux phrases, l'ensemble hiérarchisé des synthétiseurs décode successivement les informations sonores qui lui sont transmises.

Ces deux théories, pour séduisantes qu'elles soient, ne sont cependant pas à l'abri de critiques sévères (*). Nous ne résumerons que les plus importantes :

(*) En ce qui concerne la théorie motrice de la parole, on consultera surtout G.FANT (1962), H.LANE (1965 et 1970), MORTON et BROADBENT (1967). Le modèle de STEVENS-HALLE est examiné par G.FANT (1966 et 1967). Voir également la réponse de A.LIBERMAN aux critiques de H.LANE dans Status Report on Speech Research 19/20, july-december 1969, pp. 139-188, et l'attitude favorable adoptée par NEISSER dans *Cognitive Psychology* (1967). Les distinctions entre le modèle de STEVENS et celui de LIBERMAN ont été mises en relief par ce dernier (1968).

- Rien n'autorise à inférer de la richesse luxuriante de l'image spectrographique et de son apparente ambiguïté une quelconque infériorité hiérarchique du décodage acoustique. Ces éléments ainsi que la prégnance de la tradition articulatoire expliquent peut-être la fascination qu'exerce l'apparente simplicité de la description motrice.
  En fait, compte tenu des relations entre les configurations articulatoires et acoustiques, il est possible de retrouver des formes phonétiques qui s'accordent parfaitement avec les discontinuités de la production vocale.

- Une analyse périphérique ayant atteint ce stade où elle permet un appariement neuro-moteur (A.LIBERMAN) ou neuro-sensoriel (K.STEVENS) doit-elle nécessairement passer par le détour d'une création duplicative que rien jusqu'ici n'a prouvé de façon irréfutable ?

- A lui seul, l'apprentissage linguistique suffirait à expliquer le caractère catégoriel des réponses fournies par les sujets. En dehors des conditions naturelles de décodage, l'auditeur s'attache aux gradations qualitatives et délaisse les fonctions d'identification. Les résultats obtenus par STEVENS sur des voyelles isolées et non-isolées renforcent ces conclusions mais ne peuvent démontrer que le processus de décodage soit de nature active.

- Les deux modèles (*) se fondent sur une représentation hiérarchisée de la langue. Tous deux accordent une importance excessive au phonème et à une conception purement phonémique et séquentialisée du codage et du décodage de la parole. La nostalgie de l'invariance et la tendance à plaquer les unités disctètes révélées par l'analyse linguistique sur la substance physique qu'est le discours réalisé semblent méconnaître les principes mêmes sur lesquels s'appuie l'analyse linguistique. C'est une chose que d'admettre à titre d'hypothèse de travail une structure linguistique hiérarchisée qui part du trait distinctif pour aboutir à la phrase ; c'en est une autre que de postuler un décodage qui doit d'abord et de manière absolue récupérer le phonème (**).

3. Que l'on admette ou non la validité des modèles actifs, un certain nombre de problèmes se posent. La *PSYCHOLOGIE DE LA PERCEPTION* s'attache à résoudre ces questions : les mécanismes d'apprentissage et leur détermination expérimentale,
   la taille et les dimensions des unités perceptives,
   le ou les mécanismes de décodage périphérique,
   les mémoires à court et à long terme.
   La plupart des systèmes élaborés par les chercheurs comportent des analyseurs périphériques suivis de mémoires où sont emmagasinées soit les règles qui permettent d'évaluer les stimuli et leurs états probabilistes, soit les unités de décision à critères ou à seuils variables selon l'information transmise par les stimuli et les mémoires. Le "mode passif" des diverses explications proposées permet de relier entre eux les résultats obtenus à des niveaux différents et pour des modalités sensorielles différentes, isolées ou combinées.

   A titre d'exemples, trois groupes de problèmes sont examinés :

(*) Pour K.STEVENS, la version 1960 est seule concernée ici.

(**) Les controverses de ces six dernières années doivent beaucoup à la distinction proposée par CHOMSKY entre la "compétence" et la "performance" . Mais rien ne permet d'affirmer jusqu'ici que les processus de cette dernière adoptent les cadres de la première. Quant au phonème, son importance doit beaucoup à la structure graphique de nos langues. LÜDTKE (1969), dans un travail récent, démontre que le phonème est une entité fictive découlant de l'histoire de l'écriture et de la structure des langues sémitiques.

LA TAILLE DES UNITES PERCEPTIVES

Les expériences de C.CHERRY (1953) (dimensions apparentes des regroupements perceptifs et rôle joué par l'attention dans les mécanismes de la reconnaissance), la réfutation opposée par D. BROADBENT (1958) et la démonstration finale de HUGGINS (1964), indiquent que l'empan du traitement perceptif ne dépend pas de la vitesse du déroulement temporel. Ces expériences ont mis en valeur le rôle joué par la syllabe, unité cognitive plus importante que le phonème. Mais, un examen attentif des résultats obtenus par HUGGINS fait apparaître que la taille de segments perceptifs n'est pas toujours identique et, par conséquent, que la syllabe, pas plus que le phonème, n'est l'unité fondamentale de la perception. Les études sur le mot, la phrase et leur réalité psychologique sont ensuite passés en revue. D'après GARRETT, BEVER et FODOR (1966), les constituants syntaxiques majeurs sont responsables du processus de segmentation perceptive. De plus, le décodage de la phrase serait un processus actif au cours duquel l'auditeur fournit l'analyse structurale de la phrase au lieu de répondre passivement aux indices acoustiques qui la marquent. Ces conclusions sont quelque peu hâtives si l'on tient compte du montage de ces diverses expériences et des résultats obtenus par d'autres chercheurs (LADEFOGED - BROADBENT, 1960).

L'EFFET DE CONTEXTE

L'importance du contexte a été démontrée à tous les niveaux des processus perceptifs. STEVENS l'incorpore dans son modèle d'analyse-synthèse (utilisation des données fournies par les spectres analysés antérieurement). Si donc la perception est conçue comme un procès cognitif, l'effet de contexte en fait intégralement partie (*). On le retrouve dans le microdomaine des séquences syllabiques où les transitions formantiques interviennent dans la reconnaissance vocalique. Si les fréquences des formants sont les corrélats acoustiques de la voyelle, comment pouvons-nous identifier cette dernière alors que ses fréquences caractéristiques ne sont à peu près jamais atteintes dans le déroulement normal de la parole ? L'explication pourrait résider dans l'audition antérieure de segments plus longs et plus lents tandis que la direction et la vitesse des transitions (dans un contexte CVC) joueraient le rôle de contexte à court terme (B.LINDBLOM et STUDDERT-KENNEDY, 1967).

Dans une expérience récente (SERNICLAES-WAJSKOP, n.p.), on observe un effet inattendu du contexte lorsqu'on demande à des sujets d'identifier des vocoïdes isolés présentés à des durées et à des fondamentales variables. Le groupe de sujets soumis à un entraînement préalable avec des stimuli longs (80 ms), obtient des scores de reconnaissance significativement plus élevés que celui soumis à un entraînement identique avec des stimuli brefs (8 ms).

L'ATTENTION ET LA MEMOIRE

Ces phénomènes nous intéressent également dans la mesure où ils nous fournissent des indications .sur l'existence ou la non-existence d'un filtrage périphérique des données sensorielles,
.sur la durée de l'écho qui s'établit dans la mémoire immédiate.

Que les modèles proposés soient actifs ou passifs, ils suggèrent tous la présence d'analyseurs primaires et d'une mémoire-tampon. Le rôle des analyseurs sera plus important dans un modèle passif que dans un modèle actif où le résultat, assez grossier, de l'analyse n'aura d'autre but que d'orienter la stratégie de consultation des règles internes.

Quant à la mémoire à court terme, il faut bien supposer qu'elle existe puisque le son est un phénomène intrinsèquement temporel. Nous avons déjà vu que les unités perceptives de la parole avaient des dimensions variables ; même les plus courtes ont une durée finie et la détection de leurs traits distinctifs nécessite certainement la mise en oeuvre d'une fraction de temps réel. Si ces traits ont un rôle

(*) Cf. l'expérience de MILLER, HEISE et LICHTEN (1951) sur l'intelligibilité et l'importance des dimensions des ensembles de choix.

à jouer dans la reconnaissance, l'information auditive doit nécessairement être préservée sous une forme non discrétisée. Cette mémoire échoïque, comme l'appelle NEISSER (1967), doit avoir un "grain" (taille de l'unité) plus fin que celui de la segmentation qu'elle permet de réaliser et une durée assez longue, tant pour autoriser des comparaisons que pour structurer le message parlé où souvent le contexte nécessaire à l'identification d'un segment est postérieur à ce dernier. Le décrément de l'écho dépend de la nature des stimuli, de la difficulté de la tâche et de l'attitude du sujet. Si l'écho peut se maintenir pendant 10 secondes (ERIKSEN et JOHNSON, 1964), il apparaît également que le phénomène de l'attention sélective peut empêcher l'entrée de certains stimuli. Les expériences en écoute dichotique et en shadowing, très nombreuses depuis 1953, indiquent que les sujets sont capables de s'accorder sur certaines données et d'en négliger d'autres, bien que, dans certains cas d'emploi de stimuli très familiers, on observe une trace mémorielle (TREISMAN, 1960), ce qui indiquerait que les analyseurs de traits distinctifs auraient fonctionné malgré ce système de filtrage par réjection.

4. Le modèle d'analyse par *FILTRAGE SELECTIF* présenté par BROADBENT dès 1958 est un premier exemple de modèle passif. Corrigé (1964) à la suite des travaux de MORAY (1959) et de TREISMAN (1960), il admet l'analyse de tous les signaux présents à l'entrée du système, la sélection ne prenant place qu'après cette opération ; l'information non sélectionnée est seulement atténuée : le filtre est modulé en amplitude. Le concept d'atténuation retrouve celui du seuil. Le modèle proposé par MORTON et BROADBENT (1964-67) a l'avantage d'englober et d'expliquer à la fois les conduites de prédiction (situations de production) et les situations de reconnaissance (visuelle et auditive). Dans ces deux situations, le comportement est au fond le même : il s'agit de fournir un stimulus verbal.

Le coeur du modèle de MORTON-BROADBENT est constitué par les "logogènes" , unités de décision qui correspondent à des mots (le terme est pris dans un sens très large). Quand le degré d'excitation d'un logogène excède un niveau critique, il se déclenche et un mot donné devient disponible. Ou, plus exactement, sa représentation en séquence motrice appropriée est stockée dans une mémoire immédiate. Le mot apparaîtra, qu'il y ait ou non information sensorielle. La présence d'une information sensorielle n'a d'autre effet que de fournir des indices et d'augmenter le degré d'excitation dans certains logogènes.

La présence d'un contexte verbal ou non verbal activera de manière différentielle certains logogènes reliés directement à d'autres ou à des unités d'ordre supérieur : les idéogènes. Les logogènes fonctionnant comme des unités de détection de signal, la présence du contexte aura pour effet de réduire la densité d'information sensorielle nécessaire à leur déclenchement. Ceci explique que des mots à haute probabilité pourront être reconnus malgré un mauvais rapport signal/bruit ou avec un temps d'exposition plus réduit. Si le mot entendu se produit hors contexte, le système devra extraire davantage de "Cues" . C'est dans un cas de ce genre que le modèle passif se rélèle plus adéquat qu'un modèle actif qui serait obligé de générer un nombre très élevé de mots ou d'avoir recours à un feedback par trop raffiné. C'est ce que MILLER constatait déjà en 1962 lorsqu'il démontrait que la réussite d'un modèle actif était sous la stricte dépendance de la première estimation (guess) et que, si celle-ci était trop éloignée du message, l'auditeur devenait incapable de suivre le rythme de la conversation.

5. Les données récentes de la neurophysiologie sur l'existence ou non d'analyseurs périphériques sont ensuite examinées. Si, pour la vision, les travaux de HUEBEL et WIESEL ont éclairé le problème, il n'en est pas de même pour l'audition. Les chercheurs n'ont pu mettre en évidence l'association univoque d'une cellule et d'un son d'une fréquence déterminée, par contre ils ont constaté l'excitation de combinaisons singulières, uniques, d'éléments adjacents qui permettent de conserver une image dynamique de la fréquence sous sa forme spatiale et fonctionnelle (I.C.WHITFIELD, 1969).

Ces résultats ainsi que les discussions qui précèdent doivent ramener notre attention sur l'importance, prônée par G.FANT (1967), d'une analyse exhaustive des patterns auditifs. Le noeud de sa conception réside dans sa théorie des traits distinctifs. Il les considère comme des unités discrètes récurrentes du message linguistique qui possèdent des corrélats sur le plan articulatoire, acoustique et perceptif. " A distinctive feature is thus a unit of the message ensemble rather than a property of the signal ensemble. " (G.FANT, 1967, p. 4) (*).

Les chiffres de retard (100 à 150 ms) obtenus par L.A.CHISTOVITCH au cours de ses expériences de répétition rapide de segments syllabiques ne constituent pas, aux yeux de G.FANT, un argument majeur en faveur d'une théorie motrice qui fait de l'articulation le véritable mécanisme du processus d'identification. Ces retards et les corrections continues qui les accompagnent peuvent s'expliquer d'abord par un décodage sensoriel opérant à l'aide de traits distinctifs et ensuite par une traduction en schèmes moteurs innervant les organes de la parole (G.FANT, 1964-67, p. 116) (**).

L'importance excessive accordée au locus et à des indices isolés a fait perdre de vue l'association intime des aspects spatiaux et temporels observés au cours des conduites d'identification. Comme le note G.FANT, ne serait-il pas préférable de définir le trait "palatal" des occlusives comme une valeur intégrée de différents indices formant une combinaison spectrale particulière (G.FANT, 1964-67, p. 124). Cette attitude qui continue celle des recherches de l'immédiat après-guerre enrichies de tous les apports qui ont suivi, si elle n'est pas très spectaculaire est cependant très respectueuse de la réalité auditive (***). Les unités globalistes du dictionnaire phonétique mis au point par l'équipe du Laboratoire d'Acoustique Musicale de Paris dirigé par M.LEIPP représentent, d'un point de vue très différent, une approche analogue.

6. La recherche sur les *DIMENSIONS ET LES ESPACES PERCEPTIFS* est une autre voie d'approche. Bien que trop peu souvent abordée, elle apparaît dès à présent comme très fructueuse. Des élèves-poètes de l'Abbé ROUSSELOT (R.de SOUZA et A.SPIRE) aux travaux de M.CHASTAING (1964), elle se poursuit par les recherches de Marguerite DURAND (1955), L.N.SALOMON (1958), E.FISCHER-JØRGENSEN (1968), G.HANSON (1967) et S.ERTEL (1969) (****).

(*) Voir les remarques de N.RUWET (1963) sur l'ambiguïté du terme "trait distinctif" et sa signification dans l'oeuvre de R.JAKOBSON (pp. 14-15).

(**) Faut-il ajouter que le shadowing pratiqué par CHISTOVITCH et ses collaborateurs représente une situation de laboratoire très spécifique dont il serait imprudent de généraliser les résultats ?

(***) Il ne nous est pas permis ici de développer ce point de vue que G.FANT a défendu à plusieurs reprises (1967, 1969). Il est très proche de celui proposé par B.MALMBERG (1970) et exige une stricte séparation des niveaux afin d'éviter toute interférence avec les descripteurs métalinguistiques.

(****) Cf. également les travaux de l'Institute for Perception RVO-TNO à Soesterberg (Pays-Bas) ; entre autres : L.C.W.POLS, VAN DER KAMP et R.PLOMP " Perceptual and physical space of vowel sounds " *JASA*, 46, 2 (1969), 458-467. La recherche méthodologique de K.WILSON et S.SAPORTA sera consultée avec profit : " Linguistic organisation " in *Psycholinguistics. A survey of the theory and research problems* (Osgood and Sebeok, ed.) Indiana University Press (1965)

La plupart de ces chercheurs tentent de reconstituer un espace perceptif :
- en partant de comparaisons directes avec les couleurs,
- en demandant aux sujets d'ordonner les stimuli (en général des voyelles isolées) sur des échelles subjectives de différents types,
- en opérant avec des triades afin de juger des différences ou des similitudes entre stimuli.

Les résultats de FISCHER-JØRGENSEN indiquent bien l'existence d'un parallélisme entre les voyelles et les couleurs du moins sur l'axe de la brillance. Sa manière de procéder a l'avantage de ne pas faire appel à des notions linguistiques ; inconvénient concomitant : il est plus difficile d'ordonner les faits. L'utilisation d'échelles multidimensionnelles et l'analyse factorielle ont permis à G.HANSON de mettre en évidence trois dimensions psycho-auditives liées aux paramètres physiques des voyelles. Ceci l'a conduit à adopter une interprétation en termes de traits distinctifs. Dans une telle expérience, on doit naturellement présupposer que les degrés de similitude peuvent être traduits en distances dans un espace perceptif euclidien. Cette dernière supposition devra bien entendu être testée ; de même, il faudra vérifier qu'une investigation de ce genre n'induit pas les sujets à évaluer les sons en fonction de leur valeur symbolique plutôt qu'en fonction de leur valeur linguistique. Mais enfin, tenter d'évaluer les voyelles en termes de " percept " plutôt qu'en termes de paramètres physiques ouvre à la recherche une direction très intéressante. Il est curieux de constater qu'une tentative de reconnaissance automatique des voyelles se basant sur une analyse en traits distinctifs (J.F.HEMDAL et G.W.HUGHES) aboutit à des résultats à peu près semblables compte tenu des différences entre le suédois et l'américain.

Enfin, ces dernières années ont vu se multiplier des recherches sur l'évaluation du poids perceptif de segments acoustiques extraits de la chaîne parlée (S.ÖHMAN, 1966 ; A.COHEN, 1964) ; sur les seuils de détection et d'identification de stimuli verbaux en fonction de différentes distorsions (durée, dimensions du vocabulaire, fréquence fondamentale, filtrage sélectif).

Ces différents essais, s'ils représentent des situations très particulières qu'aucune théorie générale ne permet à présent d'unifier, jettent cependant des lueurs sur le comportement des sujets vis-à-vis de bruits, qui sont aussi des porteurs virtuels d'information linguistique.

De l'ensemble des travaux qui viennent d'être aussi schématiquement présentés se dégage une tendance assez nette : relier les niveaux à l'aide de faisceaux de corrélats. La notion d'indice intégré a repris une vigueur nouvelle dans la formulation qu'en donne G.FANT (1967) : " In an integrated view based on all parameters of importance for a distinction the distinctive feature or rather its speech wave correlate can be conceived of as a vector perpendicular to the hypersurface constituting the multi-dimensional boundary... The main direction of this vector is the sole remaining attribute of the feature if a common denominator of all possible contexts is to be expressed as was the ambition of JAKOBSON, FANT and HALLE (1952) " (1967, p. 9).

Bien qu'exprimé de façon plus complexe, c'est bien le principe de l'invariance, réévalué et rénové, qui reparaît.

Il est permis d'espérer que, sous cette forme, il réunira dans un langage commun des chercheurs agissant dans des directions différentes et tel est bien le but que se propose la rencontre qui s'est ouverte aujourd'hui.

REFERENCES BIBLIOGRAPHIQUES

D. BROADBENT
"Perception and Communication" - Pergamon Press, New-York (1958)

D. BROADBENT - M. GREGORY
"Stimulus set and response set : The alternation of attention"
in *Quarterly Journal of experimental Psychology*, 16 (1964), 309-312

M. CHASTAING
"Nouvelles recherches sur le symbolisme des voyelles"
in *Journal de Psychologie normale et pathologique*, 61 (1964), 75-88

E.C. CHERRY
"Some experiments on the Recognition of Speech, with one and with two Ears"
in *J. Acoust. Soc. Amer.*, 25 (1953), 975-979

L.A. CHISTOVITCH - J.A. KLAAS - I.I. KUZMIN
"The process of Speech Sound Discrimination"
in *Voprosy psikhologii*, 6 (1962), 26-39 - cité par G. FANT (1967)

L.A. CHISTOVITCH - V.A. KOZHEVNIKOV
"Speech : Articulation and Perception", 223-230
J.P.R.S., U.S. Department of Commerce (1965)

A. COHEN - J. 't HART
"Gating Techniques as an Aid in Speech Analysis"
in *Language and Speech*, 7 (1964), 22-39

M. DURAND
"Du rôle de l'auditeur dans la formation des sons du langage"
in *Journal de Psychologie normale et pathologique*, 52 (1955), 347-355

C.W. ERIKSEN - H.J. JOHNSON
"Storage and Decay Characterictics of non-attended auditory stimuli"
in *J. ex. Psychology*, 68 (1964) 28-36

S. ERTEL
"Psychophonetik" - Verlag für Psychologie, Göttingen (1969)

G. FANT
- "Descriptive Analysis of the Acoustic Aspects of Speech"
  in *Logos*, 5 (1962), 3-17
- Comments to Paper D.3 "A Motor Theory of Speech Perception"
  in *Proceedings of the Speech Communication Seminar* (Stockholm, aug.29 - sept.1, 1962), 3 - S.T.L., R.I.T., Stockholm (1963)
- "Auditory Patterns of Speech"
  in *Models for the Perception of Speech and Visual Forms* (Weiant Wathen-Dunn, ed.) - A.F.C.R.L. nov.11-14, 1964 - M.I.T.Press, Cambridge (1967)
- "The nature of Distinctive Features"
  in *S.T.L. - Q.P.S.R.*, 4 (1966), 1-14
- "Sound, Features and Perception"
  in *S.T.L. - Q.P.S.R.*, 2/3 (1967), 1-14
- "Distinctive Features and Phonetic Dimensions"
  in *S.T.L. - Q.P.S.R.*, 2/3 (1969), 1-18

E. FISCHER - JØRGENSEN
"Perceptual Dimensions of Vowels"
in *Z. F. Ph., Sprachwiss. u. Kommunik.*, 21 (1968), 94-98

M. HALLE - K.N. STEVENS
"Speech Recognition : A Model and a Program for Research"
in *I.R.E. Trans. Inform. Theory*, IT.8 (1962), 155-159

G. HANSON
"Dimensions in Speech Sound Perception. An Experimental Study of Vowel Perception" - *Ericsson Technics*, 23, 1 (1967)

J.F. HEMDAL - G.W. HUGHES
"A Feature Based Computer Recognition Program for the Modeling of Vowel Perception" - in *"Models for the Perception of Speech and Visual Form"* (Weiant Wathen-Dunn, ed.) - M.I.T.Press, Cambridge (1967), 440-453

D.H. HUEBEL
"The Visual Cortex of the Brain"
in *Scientific American*, 209 (1963), 54-62

A.W.F. HUGGINS
"Distorsion of the Temporal Pattern of Speech : Interruption and Alternation"
in *J. Acoust. Soc. Amer.*, 36 (1964), 1055-1064

M. GARRETT - T. BEVER - J. FODOR
"The Active Use of Grammar in Speech Perception"
in *Perception and Psychophysics*, 1 (1966), 30-32

P. LADEFOGED - D. BROADBENT
"Perception of Sequence in Auditory Events"
in *Quarterly Jl of Experiment. Phychology*, 12 (1960), 162-170

H. LANE
"The Motor Theory of Speech Perception : A Critical Review"
in *Psychological Review*, 72 (1965), 275-309

H. LANE
"Production et Perception de la Parole : Rapports et Différences"
in *Nouvelles Perspectives en Phonétique*
Conférences et travaux de l'Institut de Phonétique, n° 1, Bruxelles, Presses Universitaires de Bruxelles (sous presse)

A.M. LIBERMAN
"Some results of Research in Speech Perception"
in *J. Acoust. Soc. Amer.*, 29 (1957), 117-123

A.M. LIBERMAN - F.S. COOPER - K.S. HARRIS - P.F. MAC NEILAGE
"A Motor Theory of Speech Perception"
in *Proc. of the Speech Communic. Seminar* (G. FANT, ed. Stockholm, aug.29 sept.1, 1962), 2 Paper D.3 - S.T.L., R.I.T., Stockholm (1963)

A.M. LIBERMAN - F.S. COOPER - D.P. SHANKWEILLER - M. STUDDERT-KENNEDY
"Perception of the Speech Code"
in *Psychological Review*, 74 (1967), 431-461

A.M. LIBERMAN - F.S. COOPER - M. STUDDERT-KENNEDY - D.P. SHANKWEILLER
"On the Efficiency of Speech Sounds"
in *Z. F. Phon., Sprachwiss. u. Kommunikation.*, 21 (1968), 21-32
(Congrès de Moscou - 1966)

B. LINDBLOM - M. STUDDERT-KENNEDY
"Estimating Short-Term Context-Dependence of Formant Pattern Perception Results" - in *S.T.L. - Q.P.S.R.*, 1 (1967), 21-24

H. LÜDTKE
"Die Alphabetschrift un das Problem der Lautsegmentierung"
in *Phonetica*, 20 (1969), 147-176

D.M. MAC KAY
"Mindlike Behaviour in Artefacts"
in *The British Jl for the Philosophy of Science*, 2 (1952), 105-121

B. MALMBERG
"Changements de perspectives en Phonétique"
in *Nouvelles perspectives en Phonétique*
Conférences et travaux de l'Institut de Phonétique, n° 1 - Bruxelles
Presses Universitaires de Bruxelles (sous presse)

G.A. MILLER - F.A. HEISE - W. LICHTEN
"The Intelligibility of Speech as a Function of the Context of the Test Materials" - in *Jl of Experimental Psychology*, 41 (1951), 329-335

G.A. MILLER
"Decision Units in the Perception of Speech"
in *I.R.E. Trans. Inform. Theory*, IT.8 (1962), 81-83

N. MORAY
"Attention in Dichotic listening : Affective Cues and the Influence of Instructions"
in *Quarterly Journal of Experimental Psychology*, 11 (1959), 56-60

J. MORTON - D. BROADBENT
"Passive versus Active Recognition Models, or Is Your Homunculus Really Necessary ?"
in *Models for the Perception of Speech and Visual Form*
(Weiant Wathen-Dunn, ed.) - M.I.T.Press, Cambridge (1967), 103-110

U. NEISSER
"Cognitive Psychology"
Appleton-Century-Crofts, New-York (1967)

S. ÖHMAN
"Perception of Segments of VCCV Utterances"
in *J. Acoust. Soc. Amer.*, 40 (1966), 979-988

N. RUWET
Préface à *"Essais de Linguistique générale de R. JAKOBSON"*
Ed. de Minuit, Paris (1963), 7-21

L.N. SALOMON
"Semantic Approach to the Perception of Complex Sounds"
in *J. Acoust. Soc. Amer.*, 30 (1958), 421-427

W. SERNICLAES - M. WAJSKOP
"Identification de voyelles en fonction de la fondamentale"
in *Rapport d'Activités n° 4 de l'Institut de Phonétique* (sous presse)

K.N. STEVENS
"Toward a Model for Speech Recognition"
in *J. Acoust. Soc. Amer.*, 32 (1960), 47-55

E. TREISMAN
"Contextual Cues in Selective Listening"
in *Quarterly Jl of Experimental Psychology*, 12 (1960), 242-248

I.C. WHITFIELD
"Sensory Processes" - in *"Encyclopaedia of Linguistics, Information and Control"*
(A.R. Meetham - R.A. Hudson, ed.) - Pergamon Press, Londres (1969), 525-535.

° ° °

D I S C U S S I O N

après l'exposé de M. WAJSKOP

Monsieur R I S S E T :

A propos de la théorie motrice de la perception de la parole, P. DENES qui est favorable à cette théorie, a cherché à la confirmer expérimentalement. Il a fait une expérience d'apprentissage de la parole transposée en fréquence (de 180-4500 Hz à 50-1600 Hz, par un procédé semblable à celui de PIMONOV) par deux groupes de sujets adultes : les sujets du premier groupe ne pouvaient qu'entendre, ceux du second groupe devaient répéter les mots ; leur parole subissait la même transposition avant de parvenir à leurs oreilles. Les résultats n'ont pas été significativement différents pour les deux groupes de sujets ; l'expérience n'a donc pas apporté d'argument en faveur de la théorie. Cependant, il reste possible que les adultes n'utilisent pas les mêmes processus d'apprentissage de la parole que les jeunes enfants.

Il est exact que les résultats de P. DENES (1965) ne confirment pas la théorie des Laboratoires HASKINS. Il faut cependant noter que l'allure des deux courbes divergeait de manière considérable. Des problèmes techniques (le rôle de l'auto-écoute) interviennent peut-être pour expliquer ces différences.

Il est possible que les adultes n'utilisent pas les mêmes processus d'apprentissage de la parole que les enfants. Pour les premiers (en regard de l'expérience de P. DENES), on peut ajouter qu'il s'agissait de l'apprentissage d'une deuxième langue.

Monsieur G R E S S E R :

Y a-t-il eu des expériences faites sur les illusions sonores comme sur les illusions visuelles ?

Les illusions sonores ont été moins étudiées que les illusions visuelles. Seules, les hallucinations sonores chez des malades atteints de schizophrénie ont fait l'objet de quelques études ; cf. :

L.N. GOULD "Auditory Hallucinations and subvocal speech : Objective Study in a case of schizophrenia"
in Jl nerv. ment. Disorders, 199 (1949), 418-427

L.N. GOULD "Verbal Hallucinations as Automatic Speech"
in Amer. Jl Psychiatry, 107 (1950), 110-119

McGUIGAN "Covert oral behavior and Auditory Hallucinations"
in Psychophysiology, 2 (1966), 73-80.

On peut également citer les recherches sur " l'effet de transformation verbale " provoqué par la répétition continue d'un même segment verbal ; cf. :

WARREN - GREGORY "An Auditory Analogue of the Visual Reversible Figure"
in Amer. Jl Psychology, 71 (1958), 612-613.

Quelle est l'importance de l'écriture dans la perception de la parole ?

D'une manière générale, le rôle de l'écriture dans la perception de la parole paraît secondaire par rapport au rôle de la parole elle-même. L'information stockée en mémoire est avant tout de nature auditive. Les résultats des travaux de CONRAD indiquent que les éléments visuels (les lettres) sont recodés en représentations auditives. Ce recodage explique que les confusions commises sur les lettres se regroupent selon des similitudes sonores et non selon leur ressemblance visuelle (cf. les recherches importantes de WICKELGREN).

Les articulations syntactiques du discours sont-elles repérables par des phénomènes acoustiques ?

Il est certain que les articulations syntaxiques majeures du discours sont marquées par des indices acoustiques. Il suffit de citer les pauses, les variations de l'intonation, de l'accent et du rythme, les variations de l'intensité vocale. Quant aux articulations mineures, elles sont certainement moins bien repérables sur le seul plan acoustique.

Monsieur T U B A C H :

Les théories sur la perception de la parole ne mettent pas, à mon avis, suffisamment l'accent sur l'aspect prévision, anticipation dans la perception de la parole ; qu'en pensez-vous ?

Il est permis de croire, au contraire, que de très nombreux travaux ont été consacrés à cet aspect de la perception. Malheureusement, ils concernent chacun une facette particulière du problème : mémoire verbale à court terme, empan de la reconnaissance visuelle, coordination main-oeil, redondance du message écrit (en liaison avec la théorie de l'information). Par contre, peu de recherches se sont attachées à cet aspect dans le cadre du message parlé.

Les groupes syntaxiques qui interviennent (peut-être) dans la perception de la parole sont-ils à rapprocher, ou à identifier, aux "groupes rythmiques" ?

Dans les recherches auxquelles il est fait allusion (GARRETT, BEVER et FODOR), les auteurs distinguent soigneusement entre groupes délimités par une analyse syntaxique de type transformationnel et tout ensemble marqué par des critères acoustiques ou retrouvé grâce à une analyse de la substance physique du discours. De toute manière, il s'agit de séparer ici très nettement deux plans distincts.

Monsieur C O M B A S T E T :

Lorsque la voix du sujet est filtrée par un dispositif sélectif, puis renvoyée sur l'oreille, le timbre de la voix est changé. Le conditionnement a une certaine durée : c'est l'effet TOMATIS. Voyez-vous une relation entre ce mécanisme de feed back de la phonation vers l'audition et l'utilisation de la référence articulatoire proposée par la théorie de la perception ?

A.M. LIBERMAN a soigneusement écarté de la Théorie Motrice de la Parole la rétroaction audio-phonatoire. Il s'agit en effet d'un mécanisme périphérique d'importance mineure en regard des processus d'identification.

Monsieur L E I P P :

Vous avez dit : "la constante de temps de l'oreille est de 30 ms. Les avis sur ce point sont extrêmement partagés : on parle de 50 ms, 30 ms, 1 ms, ... Selon que c'est un électronicien, un physicien, un psychologue, il semble que le mot "constante de temps" ait des significations très différentes. Si vous acceptez ce mot comme synonyme de "pouvoir séparateur temporel de l'audition", notre expérience avec les musiciens montre avec évidence que les auditeurs entraînés perçoivent des phénomènes d'une durée voisine de la milliseconde...

De toutes façons, cette constante varie à l'extrême avec les individus ; en particulier, la chaîne ossiculaire semble jouer un rôle déterminant car il s'agit d'un système complexe de leviers rigides couplés entre eux de façon plus ou moins lâche. Les deux muscles qui pilotent le système et accommodent l'impédance mécanique du système sur l'intensité du phénomène acoustique interviennent aussi dans l'inertie, l'amortissement du système ossiculaire. La modification de raideur du système ossiculaire se fait normalement par voie réflexe, mais certains individus semblent pouvoir agir volontairement sur le système. Il est alors évident que la constante de temps varie pour un même individu selon le contexte acoustique et l'entraînement. Entre individus différents, les différences peuvent être énormes. Je pense donc qu'il est nécessaire de définir clairement au préalable de quoi on parle ; si j'ai bien compris, vous appelez constante de temps le seuil de fusion entre phénomènes voisins ; ainsi, quand vous dites que cette constante de temps est de 30 ms, je pense que cela signifie que, selon vous, deux phénomènes voisins de moins de 30 ms fusionnent perceptivement en un bloc unique.

D'autre part, vous avez comparé la constante de temps de l'oeil (25 images seconde pour obtenir au cinéma la fusion entre images successives) avec celle de l'oreille. Je pense qu'il n'y a aucune commune mesure ; la rémanence des cellules du capteur optique de l'oeil est probablement beaucoup plus longue que celle de l'oreille. Ce qui compte, dans ce dernier cas, est l'inscription du phénomène sur la mémoire instantanée (à court terme) ; pour des personnes dont l'oreille moyenne (tympan, système ossiculaire, fenêtre ovale) est très amortie, la milliseconde semble être une moyenne courante, du moins pour les gens qui sont entraînés à l'audition, les musiciens en particulier.

Dans cet ordre d'idées, vous dites que "descendre à 5 ms est du gaspillage". Je ne le pense pas ; en parole, des phénomènes de durée encore plus faible jouent un rôle considérable dans l'intelligibilité, en particulier en ce qui concerne les plosives.

Le texte d'A.M. LIBERMAN, cité en français, au cours de la conférence s'exprime ainsi : " ... 30 sounds per second would overreach the temporal resolving power of the car : discrete acoustic events at this rate would merge into an unanalyzable buzz, ... ". Encore, dans cet article, LIBERMAN entend-il par là des unités du langage émises séparément. Il ne s'agit donc pas d'une constante de temps de 30 ms. Il

est malheureusement exact que les chiffres cités à propos de la constante de temps de l'oreille varient d'un auteur à l'autre selon leur notion du phénomène.

Le seuil de fusion temporelle peut descendre dans certains cas au-dessous de la milliseconde (MILLER et TAYLOR, 1948). Le seuil de discontinuité est plus élevé et dépendra en particulier de l'intensité. PIERON (1955, pp. 396-397) cite le chiffre de 10 ms pour l'ouïe.

Un son très bref, pourvu d'une énergie suffisante, sera détecté. C'est dans le cadre de travaux sur la segmentation temporelle de voyelles qu'il a été dit que descendre au-dessous de 4-5 ms serait du gaspillage. Une durée aussi réduite ne représente pratiquement plus aucun intérêt pour les problèmes linguistiques. Par ailleurs, à d'aussi brèves durées, les constantes de temps d'ouverture et de fermeture de la porte électronique modifient cette durée. Si la constante de temps de la porte ne l'affecte pas (front raide), elle crée des transitoires qui rendent inutile une telle recherche.

Il est exact que des phénomènes encore plus courts existent dans la parole. La détente d'une occlusive s'étend de 8 à 40 ms selon sa nature ; la partie fricative qui la suit peut être plus longue ou non selon les langues. Il est par conséquent difficile d'admettre que des segments inférieurs à 5 ms puissent jouer un rôle dans l'intelligibilité au sens strict de ce terme.

Par ailleurs, l'inscription d'un "phénomène sonore sur la mémoire instantanée (à court terme)" est une notion psychologique se situant à un tout autre plan et à un tout autre niveau que celle de "constante de temps".

En réponse à une question générale posée par R. CARRE sur :

LA SEGMENTATION TEMPORELLE : INFLUENCE DE LA FORME DE PRELEVEMENT SUR LE SPECTRE ET LA PERCEPTION DES SONS SEGMENTES

par A. LANDERCY, Institut de Phonétique
Université Libre de Bruxelles

(Cet exposé résume une suite de travaux qui ont été décrits en détail dans les publications reprises dans la bibliographie.)

Il est intéressant de pouvoir prélever une série d'échantillons déterminés au sein de la chaîne parlée afin de vérifier s'il est possible d'y retrouver certains éléments catégoriels. Cependant, lorsqu'un signal continu est segmenté dans le temps, la fonction de prélèvement crée un spectre parasitaire. Celui-ci doit influencer le moins possible le spectre du signal original mais il faut cependant s'assurer que la fonction de prélèvement maintient la signification physique du signal segmenté. Certaines fonctions de prélèvement, la gaussienne notamment, introduisent un minimum de transitoires mais leur emploi n'était cependant pas recommandé dans notre situation. En effet, si l'on examine la forme d'enveloppe d'un son naturel (3) (une voyelle par exemple) on constate que pendant un certain temps l'amplitude croît ; elle reste ensuite plus ou moins constante pendant sa période de régime et enfin elle décroît plus ou moins lentement. La gaussienne ne reflète pas cette réalité physique. Pour atteindre celle-ci, un appareillage composé de deux segmentateurs électroniques et de leurs organes de commande a été réalisé (1). Cet ensemble nous permet d'extraire en n'importe quel endroit d'une séquence sonore un segment de longueur prédéterminée. La fonction de prélèvement f(t) agit avec des fronts d'ouverture et de fermeture en forme d'arcs exponentiels à constante de temps réglable. La forme d'enveloppe est ainsi très proche de celle correspondant à l'émission vocale naturelle (3).

Pour connaître l'effet de la perturbation spectrale introduite par ce prélèvement, nous l'avons comparée avec celle causée par un prélèvement rectangulaire (2) et trapézoïdal (3). Dans le cas d'un son sinusoïdal, la forme générale du spectre est la même pour ces trois fonctions (fig. 1). Cependant :

a) le premier rebondissement spectral est toujours plus faible pour le prélèvement exponentiel

b) la largeur de la bande centrale est toujours inférieure pour ce même prélèvement

c) la quantité d'énergie spectrale répartie dans la bande centrale y est toujours supérieure.

A titre d'exemple (3), si la constante de temps d'ouverture et de fermeture est choisie de manière telle que 90 % de l'énergie soit transmise au moment où commence la décroissance de la fonction d'enveloppe (exponentielle et trapézoïdale), le rapport du premier au second maximum $M_1/M_2 = 77$ pour le prélèvement exponentiel tandis que $M_1/M_2 = 41$ pour le trapézoïdal. Pour le prélèvement rectangulaire, $M_1/M_2 = 22$. Le pourcentage d'énergie spectrale compris dans la bande centrale varie de 90 % pour le prélèvement rectangulaire à 98 % pour le prélèvement exponentiel.

L'évaluation de l'effet perceptif de tels prélèvements a été étudiée théoriquement, instrumentalement et expérimentalement (3,4).

Grâce aux données de Zwicker (5) sur les bandes critiques et les effets de masque, nous avons construit les spectres de sonie spécifique de différents sons. Par exemple, si l'on présente un 200 Hz à une intensité subjective de X sones/Bark (fig. 2) et que l'on présente en même temps un 500 Hz, celui-ci doit au moins avoir une intensité subjective de Y sones/Bark pour être perçu par l'oreille. En d'autres termes, tout son dont le spectre de sonie se situe à l'intérieur du spectre de sonie du 200 Hz n'est pas perçu par l'oreille.

Si les rebondissements spectraux des sons segmentés sont situés à l'intérieur de la courbe de sonie spécifique de la composante centrale, en principe ils ne doivent pas être perçus par l'oreille. Nous avons déterminé de la sorte que pour les prélèvements exponentiels de sons sinusoïdaux de fréquence f , dans les conditions d'énergie rappelées ci-dessus, et pour des temps de segmentation $t_o \geqslant I/f$, les rebondissements du spectre vers les hautes fréquences seront masqués par la courbe de sonie spécifique de la composante centrale. Dans ces conditions, seules les perturbations spectrales vers les basses fréquences doivent intervenir au niveau de la perception.

Ces prévisions théoriques ont été vérifiées instrumentalement par l'emploi de l'analyseur de sonie Hewlett Packard 8051A. Nous avons constaté qu'effectivement, sur toute la gamme des fréquences audibles et pour des prélèvements exponentiels avec $t_o \geqslant I/f$, les rebondissements vers les hautes fréquences n'apparaissement pas sur le spectre de sonie ; par contre le prélèvement rectangulaire avec $t_o = I/f$, ils étaient apparents.

Il restait donc à vérifier expérimentalement (4) ces prévisions et à démontrer que la perturbation spectrale des basses fréquences est en partie responsable de l'altération subjective de la pureté du son. Juger de la pureté d'un son et surtout comparer, de ce point de vue, deux sons qui sont très proches l'un de l'autre, est cependant une tâche difficile pour des sujets naïfs. Une expérience préliminaire a été effectuée afin d'obtenir un éventail de qualificatifs qui recouvrent le champ notionnel du concept physique de pureté d'un son.

Si l'altération subjective de la pureté du son est due en partie à la perturbation spectrale des basses fréquences, un moyen apparemment simple de mettre ce fait en évidence est de masquer cette perturbation par un son secondaire ; dans ces conditions, le son devra paraître plus pur.

Figure 1

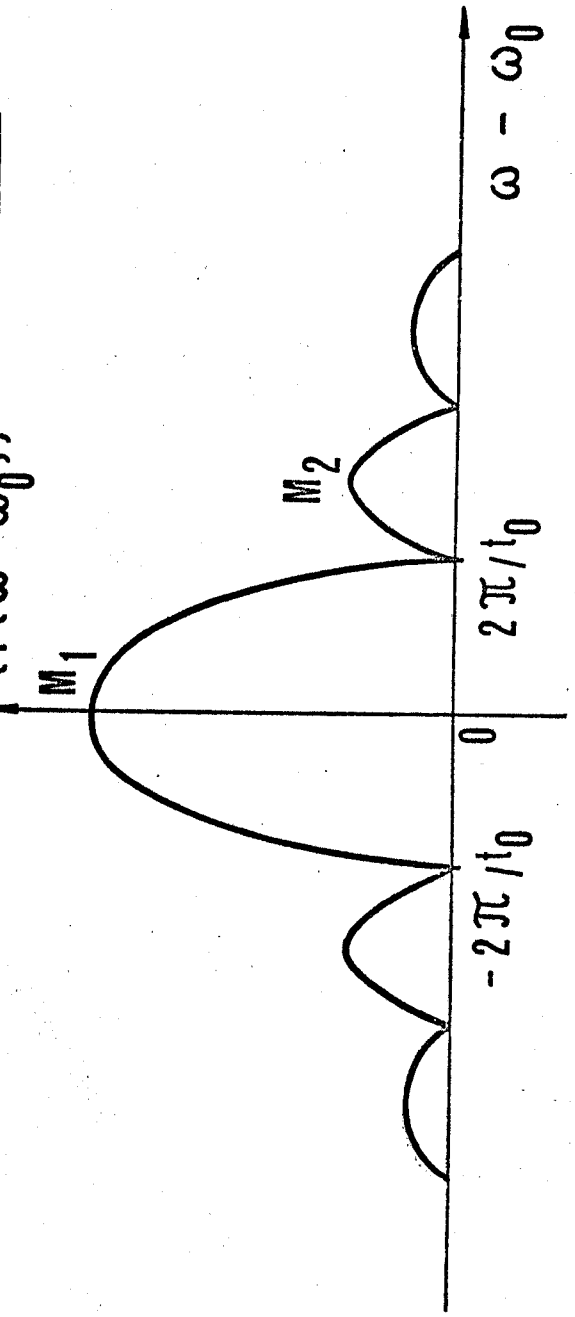


Figure 2

Nous avons employé comme son primaire un 1200 Hz segmenté à 4 ms et comme son secondaire masquant, soit un bruit blanc filtré passe-bas, soit ce même bruit blanc mélangé à un 330 Hz carré (la fréquence 330 Hz ayant été choisie parce le troisième harmonique se situe dans la bande critique précédant celle de 1200 Hz). On présentait le son primaire seul et le son primaire primaire noyé dans le bruit masquant. Le sujet indiquait celui qui lui semblait le plus pur.

Sur 111 observations, 75, soit 68 %, indiquant une sensation de pureté accrue en présence du bruit masquant tandis que 19, soit 17, 1 % ne notent pas de différence entre les deux sons primaires. Ces résultats montrent à suffisance l'importance de la perturbation spectrale vers les basses fréquences dans le cas considéré.

Nous pouvons supposer que pour de telles segmentations de sons complexes (une voyelle par exemple) le spectre de sonie ne sera que faiblement perturbé. En fait, même dans le cas de [a] , voyelle compacte par excellence, le "loudness analyser" montre que les régions formantiques ne sont que légèrement altérées. Ceci explique probablement les scores de reconnaissance élevés obtenus avec des voyelles présentées à de très courtes durées (6) (7).

BIBLIOGRAPHIE

(1) A. LANDERCY, G. SYLIN, M. WAJSKOP
"Etude et réalisation d'un segmentateur électronique et de son organe de commande"
La Revue d'Acoustique II, 5, 31-36 (1959) (France).

(2) A. LANDERCY, M. WAJSKOP
"La segmentation temporelle : forme de prélèvement et perturbation spectrale"
Report of the 6th International Congress on Acoustics - Tokyo - B 67-70 (1968).

(3) A. LANDERCY
"Temporal Segmentation : Influence of the Envelope Function on Perception"
J. of Speech and Hearing Res. (sous presse).

(4) A. LANDERCY
"Relation entre la fonction de prélèvement et la perception des sons segmentés"
La Revue d'Acoustique (sous presse)

(5) E. ZWICKER
"Ein Verfahren zur Berechung der Lautstärke"
Acustica 10, 304-308 (1960)

(6) A. LANDERCY, G. SYLIN, M. WAJSKOP
"La segmentation temporelle. Identification des voyelles françaises en fonction de leur durée"
Rapport d'activités n° 1. Institut de Phonétique -Université libre de Bruxelles, 99-125 (1966).

(7) M. WAJSKOP
"Identification de voyelles en fonction de leur durée"
Report of the 6th International Congress of Phonetic Sciences, Prag, 1967 (sous presse).

Questions de Monsieur RISSET : A propos de la façon d'introduire le minimum de transitoires par un découpage, est-il vraiment indiqué de ne considérer qu'un seul spectre pour toute la durée du stimulus ? Le spectre d'une sinusoïde indéfinie sera bouleversé par tout découpage ayant pour effet de la limiter dans le temps, n'est-il pas plus significatif d'étudier l'effet du découpage sur les spectres à court terme qui correspondent davantage à ce qui est perçu par l'oreille ?

Réponse : C'est pour des segments très brefs que les considérations énoncées ci-dessus présentent quelqu'intérêt. En effet, dès 1947, Doughty et Garner (J. Exp. Psychol. 37, 351-365) avaient mentionné l'importance de la distribution spectrale pour la détermination de la hauteur tonale d'un son pur segmenté. La largeur de la bande centrale dépendant directement de la durée, ils avaient supposé que l'oreille répondait comme si elle était simulée par une largeur de bande d'énergie instanée. Si cette bande était suffisamment étroite, certaines caractéristiques de hauteur pouvaient dès lors apparaître. Il est bien évident que plus la durée de la segmentation augmente plus le spectre global du son tend vers une raie et les effets de la segmentation tendent à devenir négligeables. Cependant, même pour des temps de prélèvement de plusieurs centaines de millisecondes, une enveloppe rectangulaire introduit en début et fin de segmentation un "clic" nettement audible, clic qui disparaît complètement avec une enveloppe exponentielle. Il serait bien sûr intéressant pour des segmentations plus longues d'étudier l'effet du découpage sur le spectre à court terme et en fait pour une durée se rapprochant de la constante de temps de l'oreille encore que les avis divergent sur la valeur de cette constante.

Question de Monsieur ROSSI : Nos recherches en cours semblent montrer que le seuil de glissando est plus fin pour les voyelles segmentées avec front raide que pour celles qui le sont avec front exponentiel. Cette conclusion vous semble-t-elle confirmer vos études sur ce sujet ?

Réponse : La segmentation à fronts raides donne directement la forme temporelle réelle du son à segmenter mais elle perturbe fortement le spectre de ce signal. Par contre, la segmentation à front exponentiel module le signal temporel et perturbe beaucoup moins son spectre. Dans les études que vous menez, votre intérêt se porte essentiellement, je crois, sur la détermination d'un seuil temporel. Il apparaît donc à première vue tout-à-fait normal que dans ce cas le seuil semble plus fin pour des prélèvements rectangulaires. Cependant, si vous effectuez des tests auditifs, la segmentation rectangulaire fait apparaître un clic qui doit gêner les auditeurs. Les fronts exponentiels vous permettront de l'éviter mais vous devrez tenir compte de l'allongement temporel des stimuli dû à la fermeture exponentielle. Une extrapolation des données dans ce sens doit conduire, à mon avis, à des résultats coïncidant avec ceux obtenus par segmentation avec fronts raides.

Question de Monsieur TUBACH : Avez-vous pensé à utiliser pour la segmentation, les fenêtres de Hanning et de Hamming, qui réduisent, elles aussi, la hauteur des lobes latéraux ?

Réponse : Bien que ces fenêtres réduisent les lobes latéraux, leur forme nous a semblé par trop différente de la forme générale d'enveloppe des sons vocaux naturels. Le choix de la forme exponentielle est essentiellement dû au fait qu'elle semble se rapprocher de l'attaque vocale. On pourrait peut-être dire, à la limite, que les perturbations spectrales introduites par ce genre de prélèvement sont du même type que celles qui apparaissent lors de l'attaque des sons voisés.

LES APPAREILS DE SYNTHESE ET LEURS APPLICATIONS

par

R. CARRÉ et J. PAILLÉ

Ecole Nationale Supérieure d'Electronique et de Radioélectricité, Grenoble

I. INTRODUCTION

Il est malaisé de proposer une étude critique des appareils de synthèse et surtout de porter un jugement sur tel ou tel principe, sur telle ou telle réalisation.

Il faut, simultanément, tenir compte de nombreux facteurs : complexité de la réalisation, sécurité de fonctionnement, intelligibilité de la parole de synthèse, paramètres de commande, possibilités d'améliorations, tout ceci étant lié à chacune des applications particulières auxquelles sont destinés les appareils de synthèse.

Si nous pouvons donner des caractéristiques techniques, par contre, il est difficile de parler de l'intelligibilité et encore plus de la qualité de la parole de synthèse. Aussi, pour ces points particuliers, nous nous bornerons à rappeler les appréciations de spécialistes réputés.

Dans un premier temps, nous donnerons brièvement le principe des appareils de synthèse typiques puis leurs caractéristiques. Nous aborderons ensuite le problème de la simulation de la source vocale, une bonne simulation étant indispensable pour l'obtention d'une parole de synthèse de qualité. Enfin, nous développerons quelques applications et utilisations des synthétiseurs.

En conclusion, nous tenterons de dégager les tendances et les perspectives en matière de travaux dans le domaine de la synthèse de la parole.

II. LES SYNTHETISEURS DE PAROLE

Etudions tout d'abord les principes de synthétiseurs types que nous classerons en trois catégories selon une analogie plus ou moins grande avec notre appareil vocal.

Les appareils de la première catégorie peuvent être schématisés par une source dont on modifie le spectre artificiellement. Le vocoder à canaux est un appareil de ce type. Nous pouvons aussi classer le pattern playback dans cette catégorie.

Dans les appareils de la seconde catégorie, on simule le processus de modulation du spectre de notre source vocale.

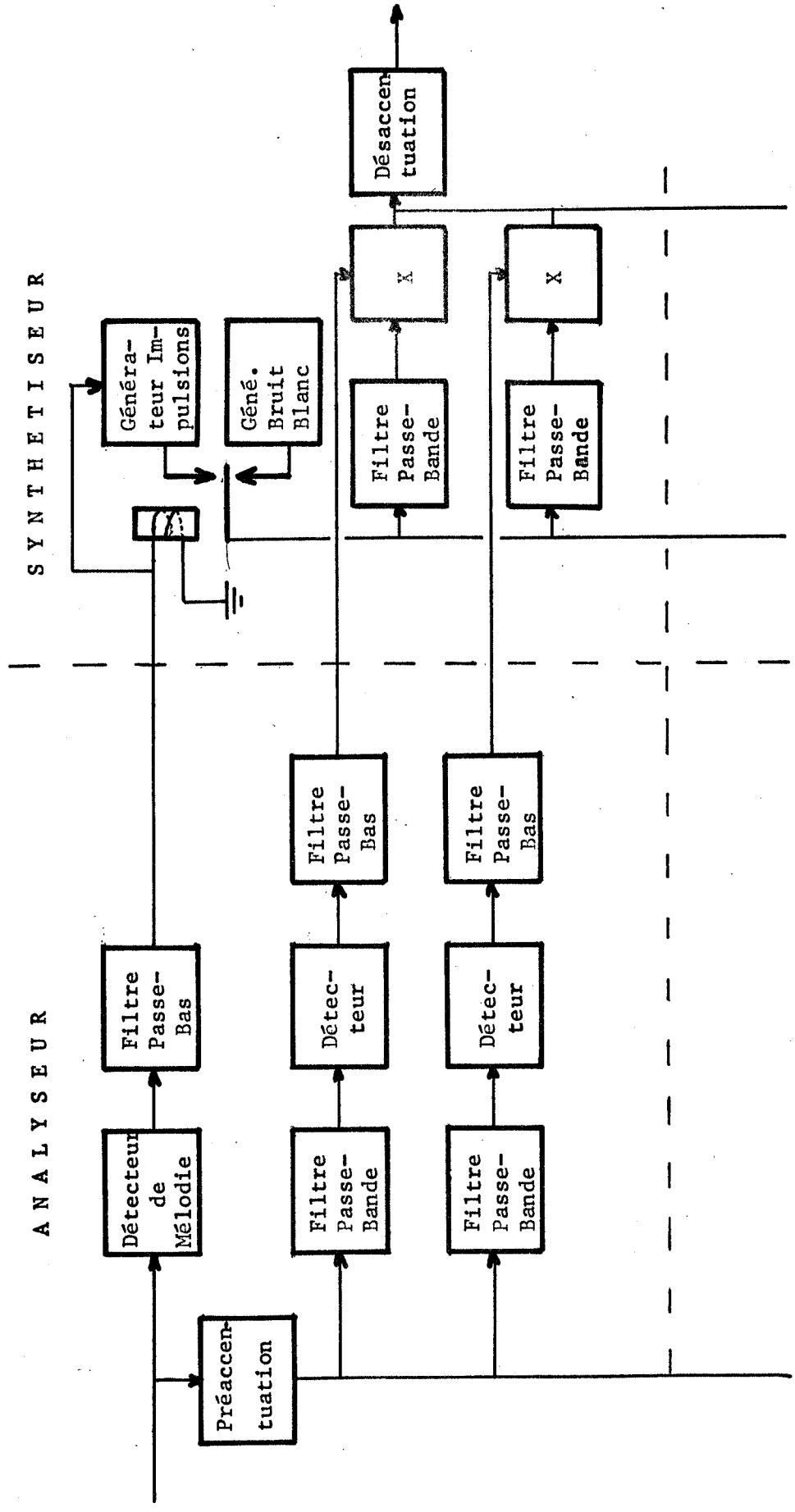

ANALYSEUR
SYNTHETISEUR
Détecteur de Mélodie
Préaccentuation
Filtre Passe-Bas
Filtre Passe-Bande
Détecteur
Filtre Passe-Bas
Filtre Passe-Bande
Détecteur
Filtre Passe-Bas
Générateur Impulsions
Géné. Bruit Blanc
Filtre Passe-Bande
X
Filtre Passe-Bande
X
Désaccentuation

FIGURE 1

Une source d'impulsions attaque des circuits résonnants dont la fonction de transfert globale est équivalente à celle du conduit vocal. Le synthétiseur à formants est un appareil de ce type.

Dans la troisième catégorie, on simule dans sa forme et ses dimensions le conduit vocal.

A. Appareils du type vocoder à canaux

Parmi les appareils du premier type, le plus important est le vocoder à canaux.

Construit au départ dans un but purement technique de compression de bande en télécommunications, il comprend presque toujours un analyseur et un synthétiseur.

L'ensemble d'analyse assure deux fonctions :

- l'élaboration de signaux images de la fréquence fondamentale,
- l'élaboration de signaux images de données relatives au spectre de la parole.

A la synthèse, ces signaux sont utilisés pour la commande de circuits fonctionnant selon un processus inverse de l'analyse.

- FIGURE 1 -

Dans le vocoder schématisé figure 1 et construit en 1939 par DUDLEY (1), le signal image de la parole est appliqué, après amplification de ses composantes à fréquences élevées aux entrées de 10 filtres passe-bande dont les fréquences centrales sont choisies pour couvrir un domaine fréquentiel allant de 0 à 3 000 Hz.

L'écart entre fréquences centrales de deux filtres consécutifs est de 300 Hz. Les signaux de sortie, de niveaux comparables sont alors détectés puis filtrés à l'aide de filtres passe-bas dont la fréquence de coupure est de 25 Hz.

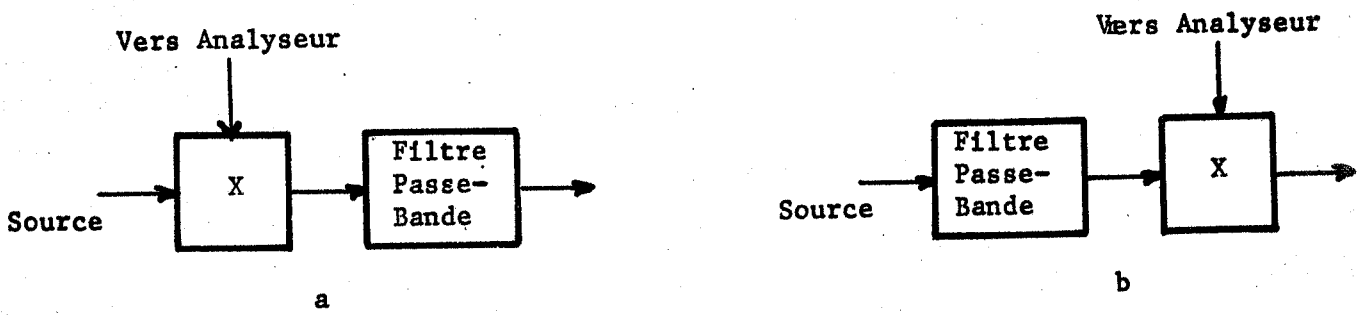
L'amplitude de la tension délivrée par le détecteur de mélodie est proportionnelle à la fréquence de la composante fondamentale du signal de parole.

Cette tension peut être filtrée au moyen d'un filtre passe-bas de fréquence de coupure 25 Hz.

A la réception, une source d'impulsions dont la fréquence de répétition est commandée par la tension issue du détecteur de mélodie, attaque 10 filtres montés en parallèle, identiques à ceux de l'analyseur. Le niveau de sortie de chacun des 10 filtres est alors commandé par la tension élaborée par la voie d'analyse correspondante.

Le signal obtenu après mélange des différentes composantes est représentatif de la parole synthétisée.

En cas d'absence d'excitation vocale, les filtres de synthèse sont automatiquement connectés à une source de bruit pour reconstitution des sons non vocaux.



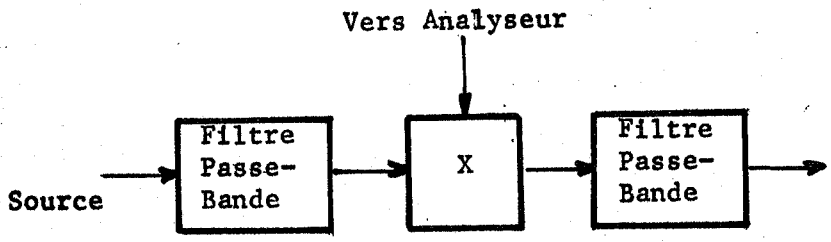
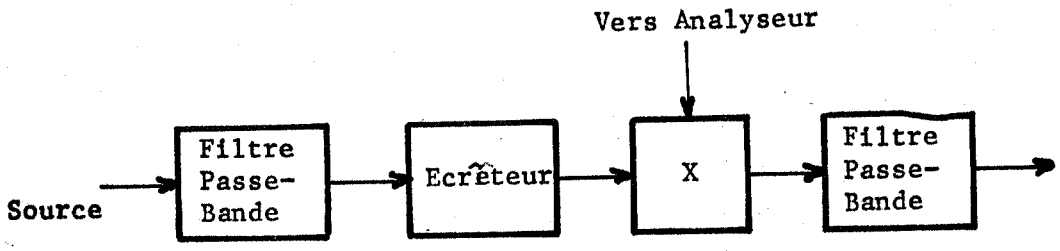


FIGURE 2



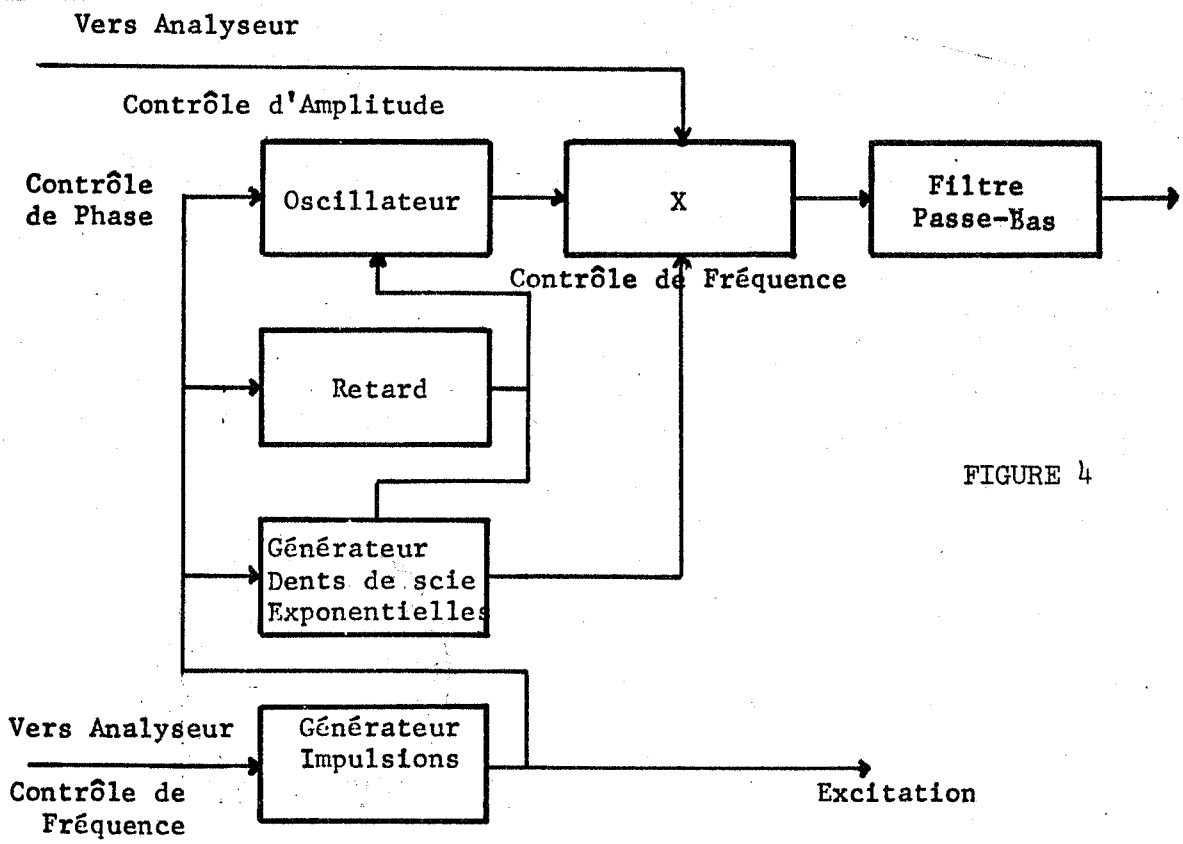


FIGURE 4

L'intelligibilité de la parole de synthèse obtenue au moyen de cet appareil reste tout juste acceptable et on constate une dégradation de la qualité ainsi qu'une sonorité métallique désagréable ; ceci est dû, en particulier, à la reconstitution insuffisamment précise du spectre original, au passage mal défini d'un son vocal à un son non-vocal, à une transmission peu fidèle de la mélodie.

Pour remédier à tous ces défauts, des travaux très importants ont été effectués récemment et, en particulier, au Lincoln Laboratory par GOLD et son équipe (2) -travaux sur les filtres passe-bande : étude de la forme des filtres, de la bande passante, des retards apportés par ces filtres, de leur nombre .

Pour améliorer la qualité de la parole, différentes structures fonctionnelles de synthèse représentées figure 2 ont été essayés.

- FIGURE 2 - (page 24)

Le schéma 2b nécessite un modulateur plus élaboré que celui de la figure 2a.

Le schéma 2c donne un son plus agréable.

Le schéma 2d apporte un nivellement du spectre de la source d'où résulte une diminution de la distorsion spectrale provoquée par les perturbations du signal de source et une amélioration notable de la qualité de la synthèse.

Le V. E. V.

Le schéma 2d avait déjà été proposé vers 1962 par DAVID et al (3) dans le Voice Excited Vocoder.

Le vocoder classique délivrant une parole peu naturelle, métallique, ce défaut pourrait provenir d'une reproduction peu fidèle de la fréquence d'excitation vocale laquelle est souvent perturbée dans la parole naturelle. Aussi a-t-on cherché à transmettre intégralement, au détriment de la compression de bande, une plage de fréquence déterminée (250-940 Hz dans l'appareil de DAVID) Un circuit régénérateur permet, à la synthèse, de créer, à partir des composantes du signal original transmises, des composantes nouvelles et de reconstituer tout le spectre couvert par les filtres de synthèse.

Pour que le spectre du signal appliqué à l'entrée des filtres de synthèse avant modulation soit d'amplitude constante, DAVID avait exploité le circuit schématisé figure 2d (voir figure 3).

- FIGURE 3 - (page 26)

Il semble bien, en fin de compte, que l'amélioration en qualité de cet appareil découle autant de la réduction de la distorsion spectrale que de l'amélioration du signal de source.

Le synthétiseur à réponse vocale

Un canal de synthèse dont les caractéristiques sont intermédiaires entre celles du montage 2a et celle du montage 2d a été proposé par SMITH (4)(figure 4).

- FIGURE 4 - (page 24)

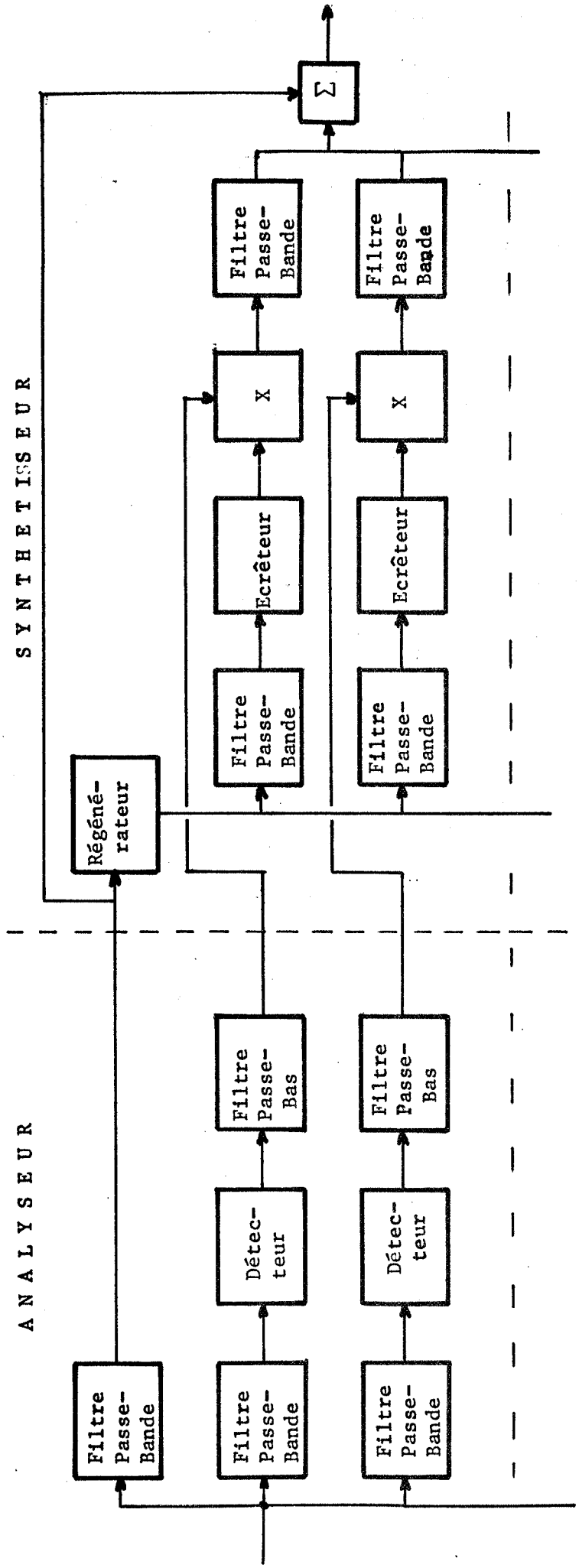

ANALYSEUR
SYNTHETISSEUR
Filtre Passe-Bande
Filtre Passe-Bande
Filtre Passe-Bande
Détec-teur
Détec-teur
Filtre Passe-Bas
Filtre Passe-Bas
Régéné-rateur
Filtre Passe-Bande
Filtre Passe-Bande
Ecrêteur
Ecrêteur
X
X
Filtre Passe-Bande
Filtre Passe-Bande
Σ

FIGURE 3

Ce montage est sensible aux variations d'amplitude à l'intérieur d'une période. Si la constante de temps du signal exponentiel est faible, on a l'équivalent du schéma 2a, si elle est grande, on a l'équivalent du schéma 2d. Dans ce schéma, il n'est pas nécessaire de disposer de filtres passe-bande. La réalisation en est donc simplifiée.

Le vocoder à autocorrélation

Dans le vocoder d'autocorrélation proposé par SCHROEDER (5) (figure 5), on élabore la fonction d'autocorrélation du signal, fonction qui est d'ailleurs directement liée au spectre et on transmet les tensions images des coefficients calculés ainsi que de la fréquence fondamentale.

Les résultats obtenus avec cet appareil sont du même ordre que ceux obtenus au moyen d'un vocoder classique.

- FIGURE 5 - (page 28)

Le vocoder conservant la phase

En vue d'une amélioration de la qualité et après avoir constaté, sur le plan de la perception, l'importance de la phase relative des composantes ou plus exactement de ses variations, on a cherché à transmettre des informations sur ce paramètre informations qui ne pouvaient être transmises avec le vocoder classique.

Au Lincoln Laboratory (6), on ne transmet pas ces informations mais on restitue d'une manière très sommaire des variations de phase à l'aide d'un organe à minimum de phase comportant quatre circuits de formants de fréquences de résonances égales à 600, 1 500, 2 200 et 3 000 Hz. Cet organe est intercalé, entre la source d'impulsions et les circuits de nivellement du spectre des canaux de synthèse. Les variations de phase apportées par ce système imitent grossièrement les variations de phase d'un synthétiseur à formants et améliorent la qualité de la parole produite.

Une expérience plus précise réalisée, antérieurement à la précédente, par GOLD (7) et où les fréquences des circuits de formants variaient selon le spectre de la parole originale avait montré que la parole de synthèse devenait, selon lui, naturelle.

D'autres systèmes ont été proposés qui conservent la phase tel l'analyseur-synthétiseur simulé par FLANAGAN et GOLDEN (8) sur ordinateur. Dans cet appareil, les signaux sont représentés par leurs spectres d'amplitude et de phase ; on ne détecte pas la mélodie ainsi que le passage d'un son vocal à un son non vocal. La réalisation d'un tel ensemble est difficile et la réduction de bande faible (rapport 2). La qualité de la parole résultante est bien meilleure que celle des vocoders classiques.

O'NEILL (9) propose un autre système où le signal de parole est représenté sous forme d'une série de signaux sinusoïdaux amortis exponentiellement.

Là encore, la parole est de haute qualité, mais il faut 10 000 bits/s pour assurer la transmission.

OPPENHEIM (10), dans une simulation sur ordinateur, transmet des informations sur le cepstrum.

Ses expériences confirment l'hypothèse qu'une phase identique à celle du signal original est préférable en synthèse. Mais la capacité de la ligne de transmission est de l'ordre de 8 000 bits/seconde.

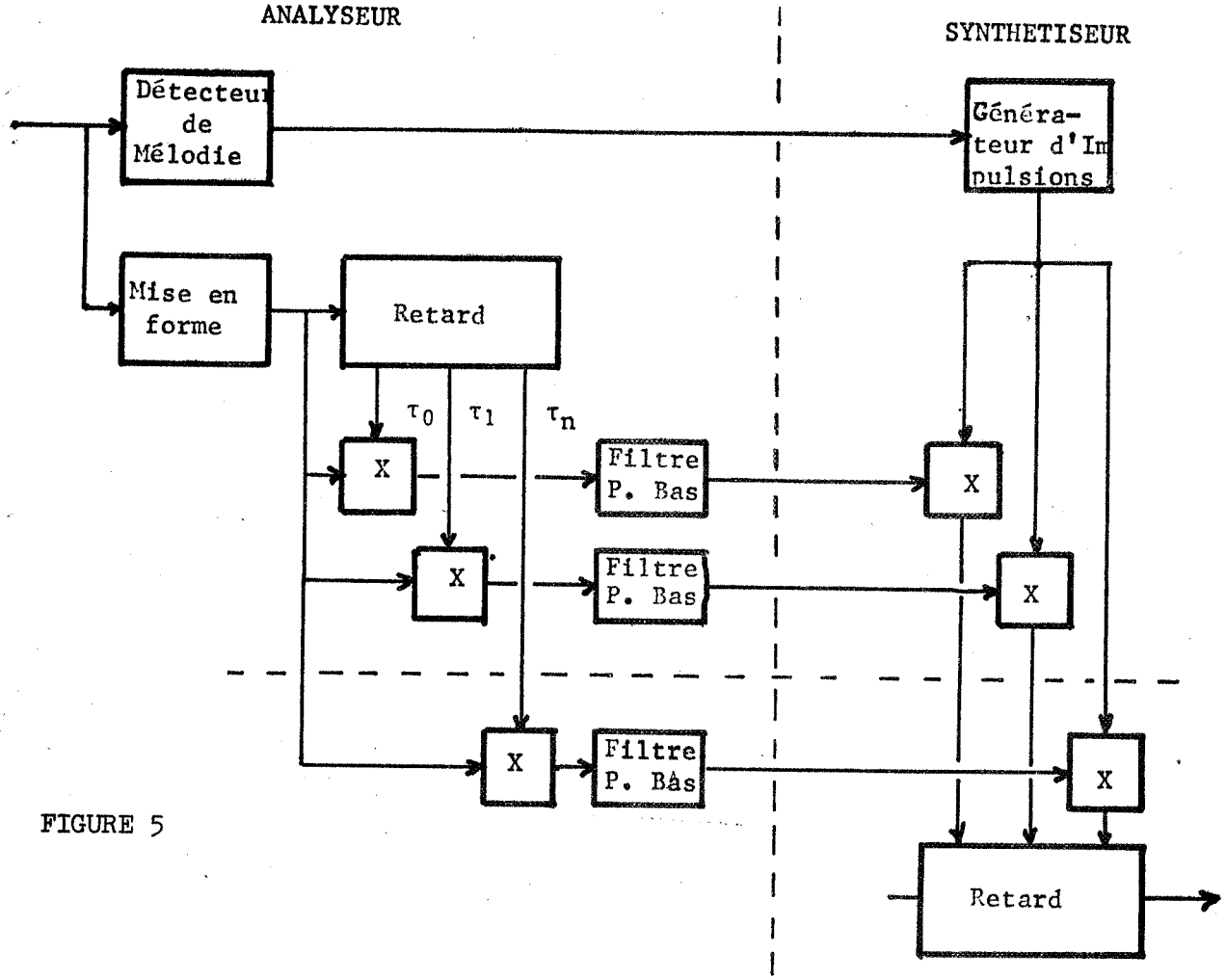


FIGURE 5

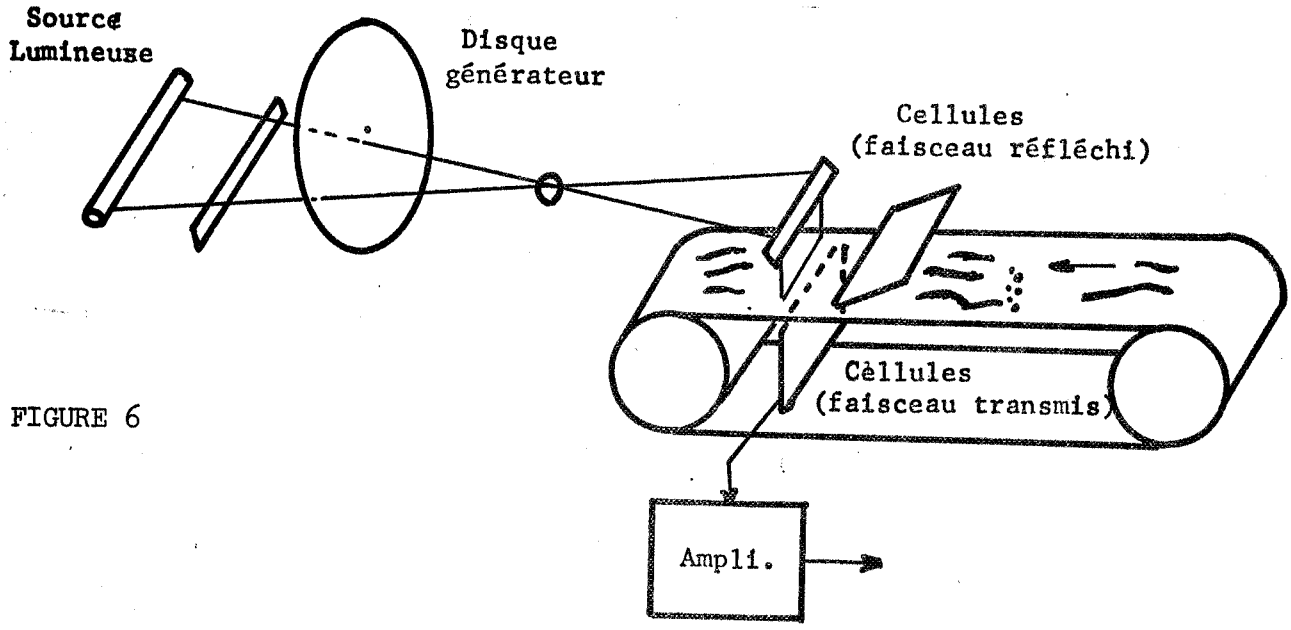


FIGURE 6

Le Pattern Play-Back

Mis au point aux laboratoires Haskins (11), il est à l'origine de nombreux travaux sur la production de la parole.

- FIGURE 6 - (page 28)

La source est un disque générant 50 sinusoïdes de fréquences multiples d'un fondamental à 120 Hz. Les modulations sont réalisées par des dessins tracés sur une feuille de papier transparent, dessins correspondant au spectre simplifié de la parole.

L'appareil de Monsieur LEIPP (12) se rapproche du pattern playback mais la source est ici constituée d'un ensemble d'oscillateurs électroniques.

B. Les synthétiseurs à formants

Dans leur principe même, les appareils que nous venons de voir ne font pas intervenir l'origine de la modulation du spectre de la source vocale, c'est-à-dire, le comportement du conduit vocal et les contraintes qu'il impose.

Dans un synthétiseur à formants, la source d'impulsions attaque des circuits résonnants dont la fonction de transfert globale est équivalente à celle du conduit vocal. Cette fonction de transfert comporte des pôles qui correspondent aux fréquences de résonance des cavités en couplage du conduit vocal et donc aux fréquences des formants.

Un synthétiseur à formants est constitué de plusieurs circuits résonnants chacun d'eux correspondant à un formant particulier. Ce sont des circuits du second ordre dont la bande passante est approximativement constante dans la plage de fonctionnement, leur facteur de surtension varie entre 5 et 30.

Pour effectuer la synthèse d'une séquence de parole il suffit de commander d'une façon adéquate la fréquence de résonance de chacun des circuits à l'image de l'évolution de la fréquence du formant correspondant.

- FIGURE 7 - (page 30)

La figure 7 représente le schéma du synthétiseur OVE II (13) construit par FANT à Stockholm.

Il est constitué de trois canaux.

L'un pour la synthèse des sons vocaux.

Un deuxième pour la synthèse des sons non vocaux et un troisième pour la nasalité.

En fin de compte, onze tensions sont nécessaires pour commander le synthétiseur lesquelles correspondent :

- pour la production des sons vocaux
  - . à la fréquence d'excitation de la source d'impulsions
  - . à l'amplitude vocale
  - . aux fréquences des trois premiers formants
- pour la production des sons non vocaux
  - . à l'amplitude de la source de bruit
  - . aux fréquences de deux formants de bruit et d'une antirésonance de bruit
  - . à l'amplitude du bruit dans les formants vocaux.

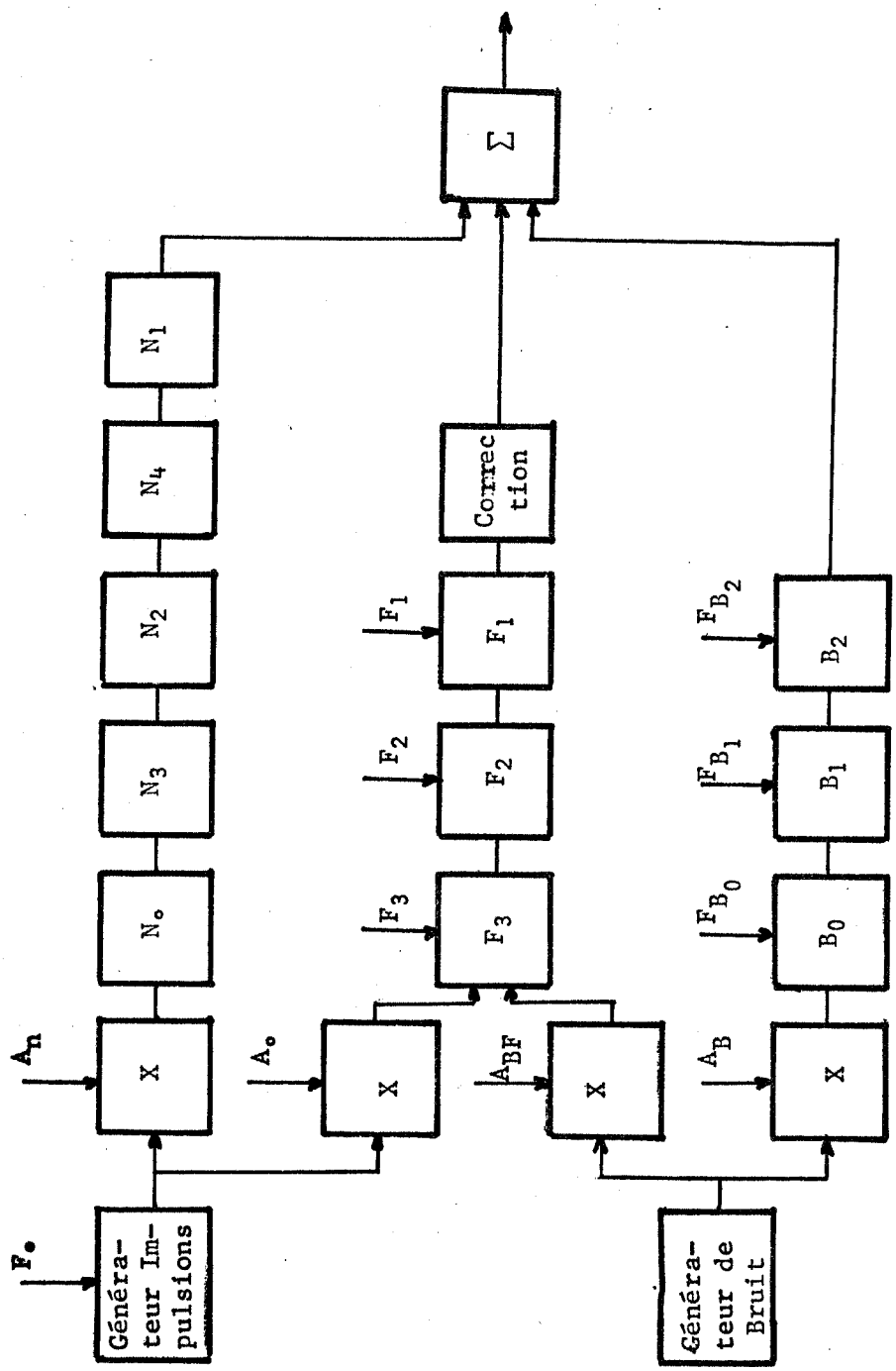

Σ
N1
N4
N2
N3
No
Correc tion
F1
F1
F2
F2
F3
F3
FB2
B2
FB1
B1
FB0
B0
An
Ao
ABF
AB
X
X
X
X
F•
Généra- teur Im- pulsions
Généra- teur de Bruit

FIGURE 7

- pour la nasalité
  . à l'amplitude de la source vocale dans les cavités nasales

D'autres types de synthétiseurs à formants ont été expérimentés. Citons le synthétiseur à structure parallèle où les circuits de formant sont disposés en parallèle. Dans ce cas, il est nécessaire de commander l'amplitude relative des formants.

Cette structure permet d'effectuer de meilleures synthèses de sons non vocaux.

Il est certain que de nombreuses améliorations peuvent encore être apportées à ces appareils : par exemple, on n'a pas encore tenu compte des intéractions source-conduit vocal car on a admis que la source est d'impédance interne très grande par rapport à celle du conduit, ce qui est faux dans certains cas en particulier lorsque les fréquences de formants sont basses.

De même, on n'a pas tenu compte du couplage variable entre le conduit vocal et l'ensemble trachée-poumons lors de l'ouverture des cordes vocales.

C. Simulateur du conduit vocal

La conception du synthétiseur à formants est fondé sur la réalisation d'un ensemble de circuits ayant une fonction de transfert la plus proche possible de celle de notre organe vocal, cette fonction de transfert étant déduite d'observations sur le spectre du signal de parole. Cette structure ne rend pas compte, en particulier, de la nature articulatoire de la parole. Pour cette dernière raison, la simulation de l'appareil vocal dans sa forme et ses dimensions et non plus seulement dans le cadre restreint de sa fonction de transfert nous paraît très intéressante.

Selon la réalisation de STEVENS et al (14), nous découpons le conduit vocal en tranches, centimètre par centimètre, par exemple, à partir des cordes vocales : la longueur d'un conduit étant de l'ordre de 17 cm chez un homme, on obtient ainsi 17 éléments tubulaires de 1 cm de long et de section variable entre 0 et 20 cm^2 environ.

Chacun de ces tubes peut être simulé électriquement par une cellule de la forme suivante :

- FIGURE 8 - (page 32)

avec $L = \frac{\rho \ell}{S}$ et $C = \frac{\ell S}{\rho c^2}$

ρ : densité de l'air, c : vitesse du son.

- FIGURE 9 - (page 32)

Présenté sous la forme schématisée figure 9, l'appareil ne synthétise que des sons soutenus du type voyelle, ou bien des voyelles nasales par adjonction d'un canal en parallèle sur le premier et simulant le conduit nasal, ou bien des sons non vocaux par adjonction d'une source de bruit en série à un endroit déterminé du conduit.

Pour effectuer la synthèse de phrases, ROSEN (15) et plus récemment la firme HITACHI, ont construit un appareil à fonctionnement dynamique.

Les paramètres de commande sont alors des paramètres articulatoires.

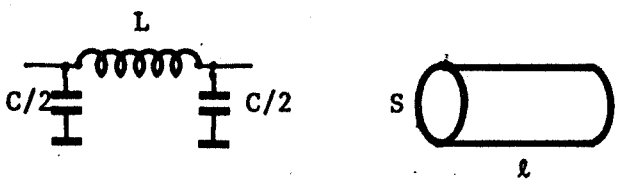


FIGURE 8

FIGURE 9

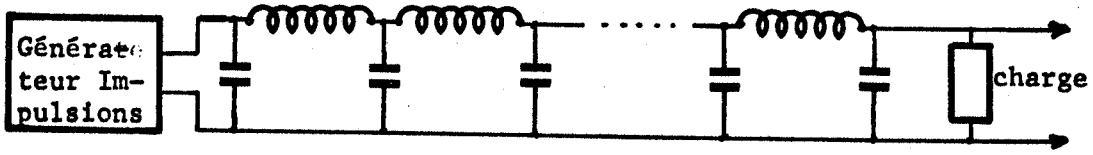
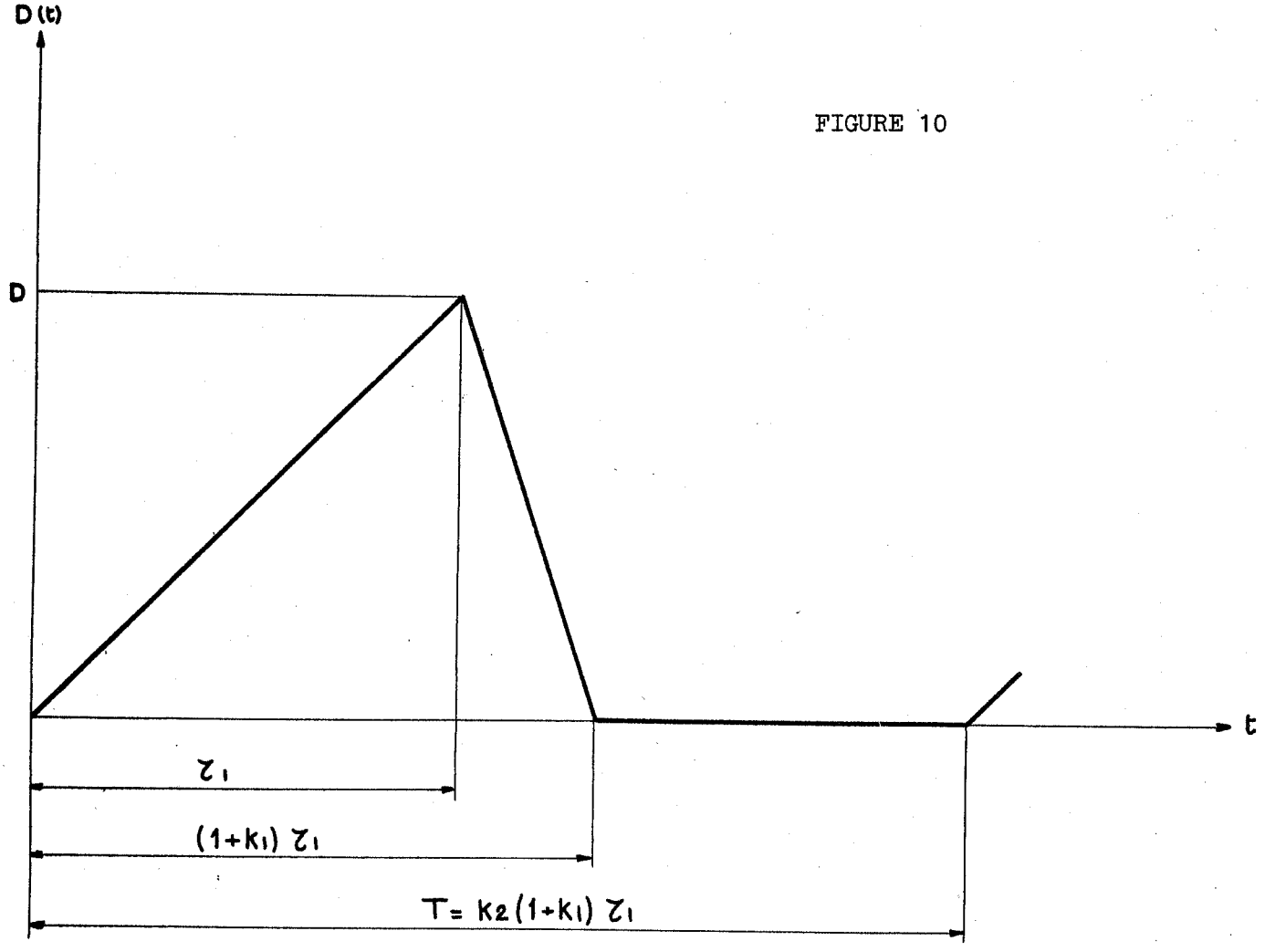


FIGURE 10



III. CARACTERISTIQUES DES DIVERS TYPES D'APPAREILS

Après ce bref rappel du principe des appareils de synthèse les plus représentatifs, essayons de comparer quelques unes de leurs caractéristiques.

Difficultés de construction

Le vocoder à canaux paraît maintenant au point. Il demande à la synthèse au minimum 16 à 20 filtres passe-bande, et autant de modulateurs. C'est actuellement un appareil qui peut être exploité industriellement.

Il semble qu'un effort pourrait être tenté pour construire un vocoder à autocorrélation en utilisant les techniques digitales. En employant de nouveaux circuits presque totalement intégrés, il est certainement possible de construire un appareil de dimensions très réduites.

Le pattern playback optique est certainement l'appareil le plus simple à réaliser, celui à fonctionnement électronique fait appel à des circuits très simples.

Jusqu'alors, le synthétiseur à formants a présenté de grosses difficultés de réalisation. En effet, la commande automatique des fréquences des circuits résonnants est compliquée . Aussi, les réalisations sont ou bien complexes et encombrantes ou bien peu précises. Nous venons de mettre au point un circuit résonnant ne présentant pas ces défauts et qui nous permet de disposer d'un synthétiseur à formants pouvant être commandé par des signaux analogiques ou digitaux (16).

En ce qui concerne le simulateur du conduit vocal, les réalisations effectuées jusqu'à maintenant sont très complexes ou du domaine de la simulation sur ordinateur.

Paramètres de commande - Organes de commande

Actuellement, seul le vocoder avec 15 à 20 paramètres, qui correspondent à la fréquence de la source et au spectre, peut fonctionner correctement en temps réel, c'est-à-dire que l'analyseur peut délivrer en permanence, avec le débit de la parole, les paramètres de commande, ceci sans difficultés particulières sauf peut-être en ce qui concerne la détection du fondamental, mais ce problème existe pour tous les autres types d'appareils.

Différents procédés ont été expérimentés pour détecter en temps réel la dizaine de paramètres nécessaires à la commande d'un synthétiseur à formants et, en particulier, les fréquences des formants.

Le calcul du cepstrum (17) permet de déterminer ces fréquences mais ce calcul est long et sa précision reste à prouver surtout lorsque la fréquence fondamentale est élevée.

Jusqu'à maintenant, les travaux restent du domaine de l'expérimentation et aucune solution satisfaisante n'est encore en vue.

Il est possible que, dans le cas des voix de femmes en particulier, le signal physique (amplitude et phase) soit insuffisant pour la détection des formants et que seul le contexte permette d'effectuer cette détection.

Aussi, l'organe de commande est-il soit un lecteur de courbes soit un ordinateur. Il faut noter que les paramètres utilisés pour le synthétiseur à formants et le pattern playback ont été et sont encore actuellement très étudiés en particulier par les phonéticiens. Leur importance dans des études de perception a été montrée. Toute une somme de travaux effectués à partir du sonagraphe peut donc être exploitée.

Pour commander le simulateur du conduit vocal, les difficultés sont plus grandes encore. Tout d'abord, il faut préciser les paramètres : ouverture de la bouche, position de la langue, etc... puis il faut pouvoir détecter ces paramètres aisément.

Certains tels KADOKAWA (18) ou MERMELSTEIN (19) partent des fréquences des formants en supposant qu'ils sont bien détectés, d'autres à partir des Rayons X, d'autres à partir de recherches électromyographiques. Tous ces travaux sont encore du domaine de la recherche fondamentale.

Débit d'informations

On peut dire qu'en moyenne, le débit d'informations est de 2 400 bits/s pour commander un vocoder classique (20) et de 1 000 à 1 500 bits pour le vocoder à formants (13), mais ces comparaisons ne sont pas significatives si on ne tient pas compte de l'intelligibilité et de la qualité de la parole de synthèse.

Qualité de la parole obtenue

Le vocoder délivre une parole intelligible, mais qui garde en général, un aspect artificiel (sauf par exemple, le vocoder du LINCOLN Laboratory mais ce dernier demande une commande de 4 800 bits/s (6)).

Lorsque la commande d'un synthétiseur à formants a été bien étudiée, l'intelligibilité est bonne (peut-être moins bonne que celle du vocoder d'après FANT (21)) et la parole paraît plus naturelle. C'est tout au moins l'opinion de SCHROEDER (24), de RABINER (23) et de STRONG (22). De même, d'après RABINER (25), le simulateur du conduit vocal a démontré ses possibilités pour générer une parole de haute qualité.

Ces opinions rappelées, il faut dire qu'il est très difficile d'effectuer des comparaisons globales sur le plan de l'intelligibilité et de la qualité. Il faudrait faire des tests dans de mêmes conditions. Les problèmes psychoacoustiques sont ici très nombreux. De toutes façons, on ne peut compter sur le jugement de l'opérateur.

Ce dernier entend ce qu'il s'attend à entendre ; il est habitué à la parole de synthèse et aux défauts de son appareil.

IV. SIMULATION DE LA SOURCE VOCALE

On sait que la perception de la parole est liée au spectre instantanée de celle-ci. Ce spectre est défini à partir des propriétés des excitations et du canal de transmission constitué par le conduit vocal. Si la théorie classique de la production de la parole décrit correctement la transmission des excitations par le conduit vocal, il reste à définir les propriétés spectrales des sources et plus précisément, dans ce propos, celles de la source vocale.

On connaît maintenant d'une manière assez exacte la forme de l'onde de débit issu des cordes vocales. En tout état de cause, des considérations simples permettent déjà de représenter, sous forme idéalisée, une première approximation de ce signal (figure 10)

- FIGURE 10 - (page 32)

En première approximation, le signal débit D(t) est un signal triangulaire d'amplitude D et de période T. Les autres paramètres, définissant ce signal sont τ_1, k_1, k_2 ; leur signification est explicitée par la figure 10.

L'analyse de ce signal va nous permettre de dégager quelques résultats quant aux propriétés spectrales de la source vocale. Nous nous intéresserons en particulier à l'évolution du spectre lorsque la forme et la période du signal varient. Pour cette analyse, nous allons considérer la transformée de Laplace de ce signal. Cette transformée s'écrit :

$$\mathbf{L}\big(D(t)\big) = \frac{D}{\tau_1}\frac{1}{s^2}\left(1 - (1 + \frac{1}{k_1})e^{-s\tau_1} + \frac{1}{k_1}\,e^{-s(1+k_1)\tau_1}\right)\frac{1}{1 - e^{-sk_2(1+k_1)\tau_1}}$$

où s est l'opérateur de Laplace.

Cette fonction est un produit de la forme :

$$L\big(D(t)\big) = F_o(s)\ .\ \frac{1}{1 - e^{-sT}}$$

où $F_o(s)$ est la transformée de Laplace de la fonction D(t) limitée à un cycle d'une période de durée T (le signal s'annulant d'ailleurs avant la fin de la période) et où $1/(1-e^{-sT})$ correspond à la répétition de ce cycle depuis le temps zéro jusqu'à l'infini.

On peut montrer qu'à un facteur 1/T près, les coefficients de la décomposition en série de Fourier, mise sous forme complexe, du signal D(t) sont donnés par $F_o(jn\omega)$ d'où, si l'on représente le module de ces coefficients en fonction de la pulsation, on obtient des raies et l'enveloppe de ces raies est aussi la représentation de $|F_o(j\omega)|$ que l'on déduit de $F_o(s)$ quand on fait $s = j\omega$.

La connaissance de la transformée de Laplace $F_o(s)$ du signal dans la période permet donc d'atteindre le spectre du signal global. Elle permet, en particulier dans notre cas, d'étudier l'influence des paramètres k_1 et k_2 sur l'allure du spectre dont les zéros sont les particularités principales.

Ces zéros sont les racines de l'équation :

$$1 - (1 + \frac{1}{k_1})e^{-s\tau_1} + \frac{1}{k_1}\,e^{-s(1+k_1)\tau_1} = 0 \qquad \{1\}$$

Trois cas sont alors à envisager selon que l'on considère τ_1 constant, le temps où le signal prend des valeurs non nulles dans la période $\tau = (1 + k_1)\tau_1$ constante, ou la période $T = k_2(1 + k_1)\tau_1$ constante.

1° Si τ_1 est maintenu constant, les zéros sont les racines de l'équation {1} et sont indépendants du paramètre k_2 qui intervient seulement dans le facteur multiplicatif 1/T. Il est en effet bien évident que toutes choses égales par ailleurs, plus la durée des valeurs nulles du signal dans la période est grande plus l'énergie contenue dans le signal est faible. Une variation de la période se traduira alors par une modulation d'amplitude du spectre et bien entendu un déplacement des raies.

Quant à la variation de k_1, elle entraîne le déplacement des zéros et une modulation d'amplitude du spectre.

w : largeur variable de la glotte

ℓ : longueur de la glotte

d : épaisseur de la glotte

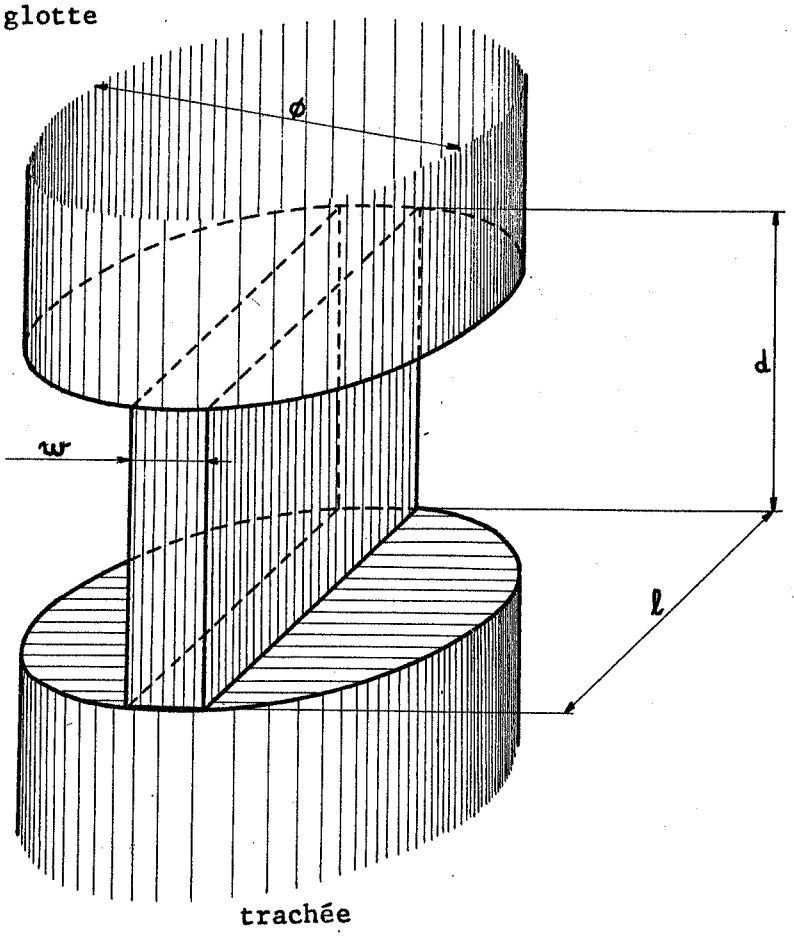


FIGURE 11

2° Si $\tau = (1 + k_1)\tau_1$ est maintenu constant, les zéros sont les racines de l'équation :

$$1 - (1 + \frac{1}{k_1})e^{-s\frac{\tau}{1 + k_1}} + \frac{1}{k_1} e^{-s\tau} = 0$$

et de la même façon indépendants du paramètre k_2 dont la variation entraîne une modulation d'amplitude du spectre. La variation de k_1 entraîne un déplacement des zéros et une modulation d'amplitude du spectre du fait que pour τ = constante, la variation de k_1 entraîne celle de τ_1.

3° Si $T = k_2(1 + k_1)\tau_1$ est maintenu constant les zéros sont les racines de l'équation :

$$1 - (1 + \frac{1}{k_1})e^{-s\frac{T}{k_2(1 + k_1)}} + \frac{1}{k_1} e^{-s\frac{T}{k_2}} = 0$$

et dépendent à la fois de k_1 et de k_2.

Le signal image du débit réel ne présénte pas un graphe aussi simple que celui examiné ci-avant mais on peut dire qu'en première approximation, les résultats précédents sont valables et permettent de voir ce qui se passe lors de la variation de la mélodie ou de la modification de la forme du signal. MATHEWS,MILLER et DAVID (26) ont montré, en particulier, l'existence des zéros pour tout signal pour lequel la dérivée seconde existe et est à variation bornée dans l'intervalle de temps $]0,(1 + k_1)\tau_1[$.

FLANAGAN (27) a étudié en détail le deuxième cas pour lequel $\tau = (1 + k_1)\tau_1$ est constant et qui conduit à une réalisation simple. En effet, dans ce cas, si k_1 est fixé, l'enveloppe du spectre a toujours même allure si l'on fait varier k_2 c'est-à-dire la période et une situation semblable se rencontre si le signal analogue du débit issu du modèle est la réponse impulsionnelle d'un filtre linéaire dont la fonction de transfert est peu différente de la transformée de Laplace évoquée ci-dessus.

Il est bien évident que ce cas n'est qu'un cas particulier et que, lors des régimes transitoires où l'amplitude et la forme du signal varient, il faut considérer le cas général pour lequel D, τ_1, k_1 et k_2 varient.

Ces caractéristiques spectrales ne pourront être restituées que si la source vocale utilisée dans les synthétiseurs est un analogue parfait de la source vocale naturelle.

Quelles sont alors les propriétés physiques de la source vocale naturelle ? Comment a-t-on envisagé le problème et quelles sont nos connaissances actuelles ?

Schématiquement, le larynx humain se présente comme l'indique la figure ci-dessous :

- FIGURE 11 -

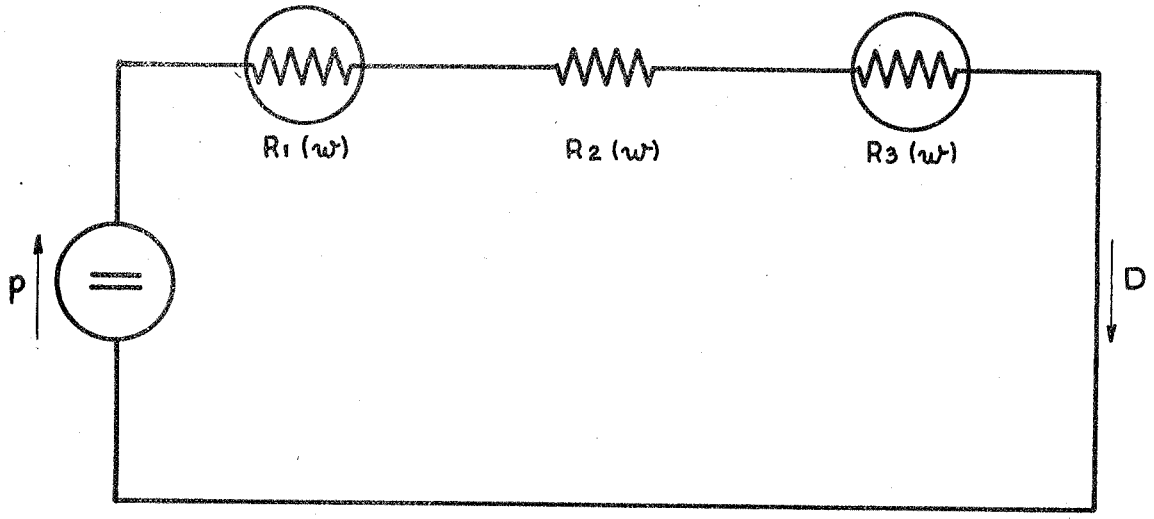


FIGURE 12

On peut, à partir de ce modèle, mener des calculs conduisant à l'expression de la résistance statique présentée par le larynx à un écoulement d'air. L'expression obtenue devra être pondérée par des coefficients tenant compte du fait que la structure naturelle du larynx n'est pas aussi simple.

VAN DEN BERG, DOORNENBAL et ZANTEMA (28) ont donné l'expression de la résistance statique que l'on peut réécrire sous la forme d'une relation entre la chute de pression p en fonction du débit D traversant le larynx.

On a ainsi :

$$\phi = \frac{1,37\ \rho D^2}{2(\ell w)^2} + \frac{12\ \mu dD}{\ell w^3} - \frac{0,5\ \rho D^2}{2(\ell w)^2}$$

où ρ = densité de l'air
μ = coefficient de viscosité cinématique de l'air.

Cette expression est composée de trois termes dont l'un est proportionnel au débit d'air D et les deux autres au carré du débit.

Le terme proportionnel au débit correspond à la chute de pression dans l'intervalle glottique et rend compte de l'application de la loi de Stokes à l'écoulement de l'air dans cet intervalle.

Le terme proportionnel au carré du débit et affecté du signe plus correspond à la chute de pression dans le jonction trachée-glotte et rend compte de l'application de la loi de Bernouilli à l'écoulement de l'air dans cette jonction. Le signe plus provient du fait que l'aire de la trachée est plus grande que celle de la glotte.

Le terme proportionnel au carré du débit et affecté du signe moins correspond à la chute de pression dans la fonction glotte-conduit vocal et rend compte aussi de l'application de la loi de Bernouilli à l'écoulement d'air dans cette jonction. Le signe moins provient alors du fait que l'aire de la glotte est plus faible que celle du conduit vocal.

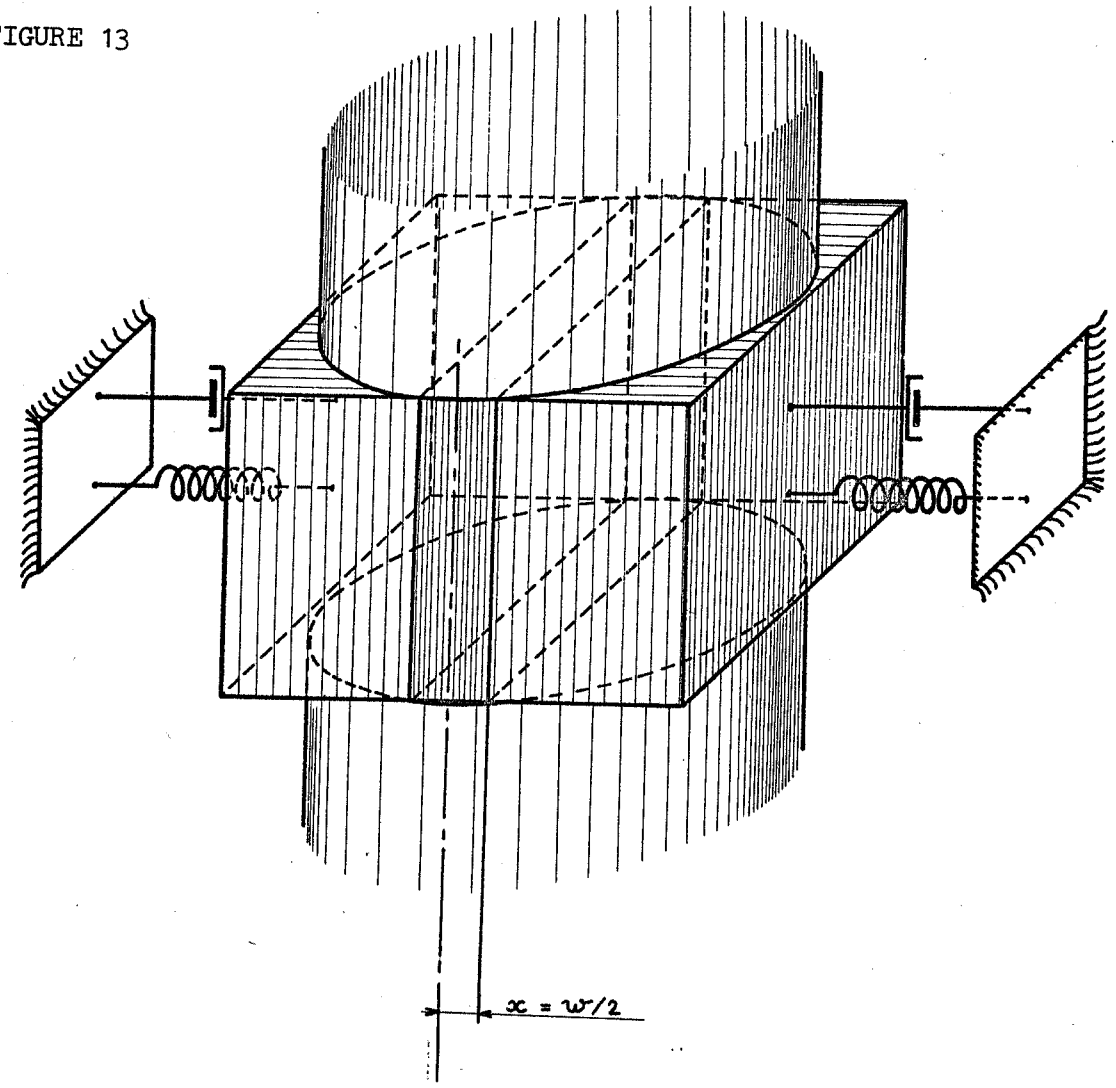
Les coefficients 1,35 et - 0,5 intervenant dans ces deux derniers termes sont des termes de pondération expérimentaux.

Dans la mesure où les poumons se comportent comme un générateur de pression p et où l'impédance présentée par le conduit vocal côté glotte est négligeable devant l'impédance présentée par le larynx, conditions que l'on peut discuter par ailleurs, l'analogue électrique du larynx en régime d'écoulement statique est schématisé ci-dessous :

- FIGURE 12 -

Dans cette représentation, les termes résistifs R_1, R_2, R_3, fonction de la largeur w de la fente glottique correspondent respectivement et dans l'ordre aux termes de l'équation de VAN DEN BERG. On notera que R_1 et R_3 sont des résistances non linéaires de valeur proportionnelle au débit qui les traverse et qu'en outre R_3 est négative. C'est la présence de cette résistance négative associée à des éléments réactifs qui permet l'apparition d'un régime oscillatoire. Les éléments réactifs intervenant dans le régime dynamique sont de deux natures.

FIGURE 13



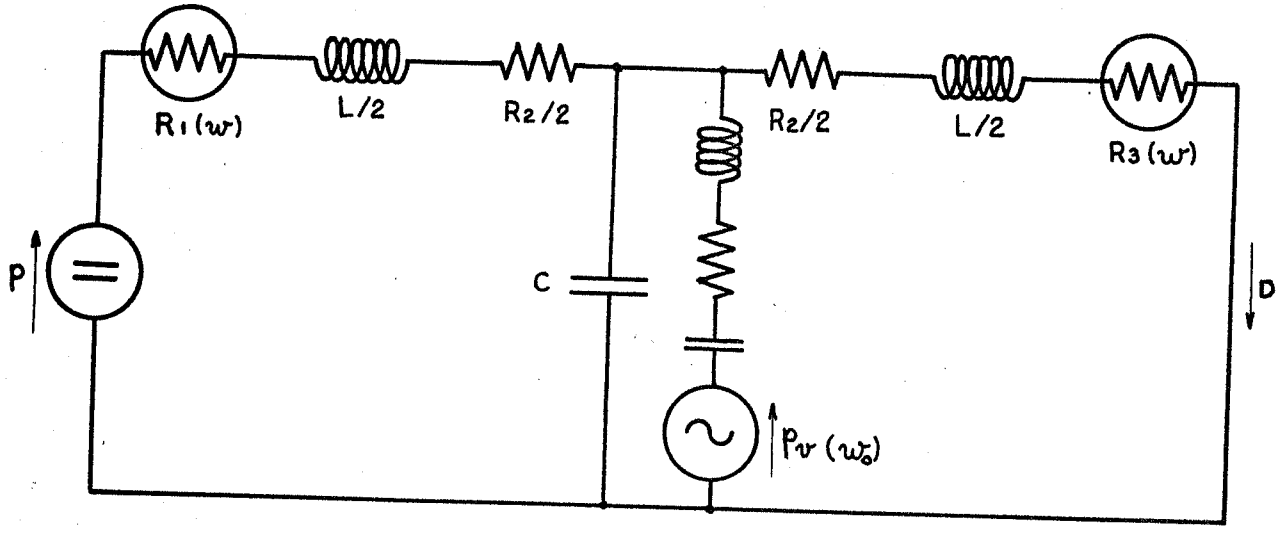


FIGURE 14

En premier lieu, il faut tenir compte des effets liés à l'inertie de l'air en mouvement dans la glotte et de sa compressibilité. Ces effets sont appelés effets hydrodynamiques. On incorporera donc au schéma précédent un élément de self induction et un élément capacitif (nous verrons ceci sur le schéma définitif).

En deuxième lieu, on doit tenir compte d'éléments mécaniques que sont les cordes vocales que l'on peut représenter, en première approximation, comme un système du 2ème ordre. La figure suivante représente alors le système physique complet tel que l'ont proposé FLANAGAN et LANDGRAFT (29) par exemple.

- FIGURE 13 -

Du fait de la symétrie du modèle, seul le mouvement d'une corde peut être pris en considération et le mouvement de l'oscillateur mécanique est décrit par l'équation classique :

$$m \frac{d^2x}{dt^2} + f \frac{dx}{dt} + rx = f(t)$$

où m est la masse de la corde vocale, f le frottement visqueux, r la raideur du ressort, x le déplacement d'une corde soit w/2 et f(t) la force agissant sur la corde vocale. Cette force f(t) est prise égale, dans le modèle présenté, à la moyenne des pressions à l'entrée et à la sortie du larynx, multipliée par la surface des cordes soit :

$$f(t) = \frac{1}{2}(p_1 + p_2) \times \ell d$$

avec

$$p_1 = p - 1{,}37 \frac{\rho D^2}{2(\ell w)^2}$$

et

$$p_2 = - 0{,}5 \frac{\rho D^2}{2(\ell w)^2}$$

D'une autre façon, on peut tenir compte du mouvement des cordes vocales, considérées toujours sous la forme d'un système du 2ème ordre et traduire ceci sous forme d'un circuit électrique équivalent. Si D_v est le débit généré par le mouvement des cordes vocales, on a :

$$D_v = 2 \frac{dx}{dt} \times \ell d$$

Le schéma électrique analogue du larynx devient alors ceci :

- FIGURE 14 -

On écrit à partir de ce schéma le système d'équations liant les différentes grandeurs. Ces équations sont fonction de la largeur w de la glotte et à chaque instant w doit avoir une valeur telle que le système soit compatible. C'est dans ce sens que l'on écrit un programme et que la résolution du problème se fait à l'ordinateur en donnant une valeur initiale à w et en procédant par approximations et itérations successives. C'est une procédure classique due à HAMMING et qui a été utilisée par FLANAGAN, LANDGRAFT (29), CRYSTAL (30).

Une autre méthode consisterait à éliminer entre les équations w de telle sorte à obtenir une équation du type D fonction des éléments connus du système et des pressions aux différents points et de traduire cette équation en un système électrique convenable à constante localisée ceci devant conduire à des réalisations simples.

Quels résultats peut-on attendre d'un tel modèle et quels sont alors les problèmes à résoudre ?

On obtient par ce procédé un graphe du débit D correspondant à celui que l'on peut déduire des mesures in vivo. Ce modèle est donc satisfaisant restriction faite de la phase d'accolement pour laquelle on est arrivé à faire des hypothèses supplémentaires. On introduit alors des conditions aux limites particulières : contact visqueux des cordes vocales pendant la phase d'accolement par exemple. En conclusion, nous pouvons dire que nous avons actuellement la possibilité de concevoir et d'élaborer des sources vocales fonctionnant à l'image de la source vocale naturelle et qui apportent une amélioration de la qualité de la parole synthétisée. Cependant, je vais évoquer rapidement le principal problème qui, à mon avis, se pose pour ces modèles et qui, est celui de la commande. En effet, la commande de la source naturelle est complexe comme chacun le sait ou peut l'entrevoir et les paramètres caractéristiques relatifs à cette commande n'apparaissent pas explicitement dans l'onde sonore finale émise. Il y a donc là un problème d'analyse et d'extraction de paramètres qui peut donner matière à travaux ultérieurs.

V. APPLICATIONS DES SYNTHETISEURS DE PAROLE

Les synthétiseurs de parole sont des instruments particulièrement intéressants car ils nous offrent la possibilité de vérifier la validité de nos hypothèses sur le fonctionnement de notre appareil vocal. Ils permettent une analyse plus précise du signal de parole.

Les paramètres peuvent être systématiquement étudiés. Enfin, les sons de synthèse peuvent être utilisés pour tester l'oreille.

Ceci concerne des études à caractère fondamental sur la parole. La grande majorité de ces études peut être ou a été entreprise à l'aide du pattern playback, du synthétiseur à formants ou du simulateur du conduit vocal.

Dans le domaine des applications, nous évoquerons l'utilisation des synthétiseurs dans le domaine des télécommunications et comme sortie parlée d'ordinateur.

A. Utilisation des synthétiseurs pour vérifier nos hypothèses sur le fonctionnement de notre appareil vocal

La théorie de FANT (26) sur la production de la parole, théorie portant sur l'étude des cavités en couplage de conduit vocal, a été vérifiée à l'aide de synthétiseurs à formants.

Dans le cas de production des sons non vocaux, la théorie de FANT (31) et l'analyse par HEINZ (32) soulignaient la présence d'antirésonances.

L'utilisation d'un circuit antirésonnant dans les synthétiseurs à formants a bien amélioré la qualité des sons non vocaux.

Mais ce qui est applicable à la vérification des théories générales l'est aussi à l'étude de problèmes plus limités tels ceux posés par la formation de certains sons, par exemple.

Les travaux de DELATTRE, COOPER et LIBERMAN (33) sur les consonnes sonores et sourdes ont pu être effectués grâce à de nombreuses expériences de synthèse.

Citons aussi les travaux de FUJIMURA en 1968 (34) sur l'apériodicité au cours de la phonation, travaux effectués à l'aide d'un synthétiseur.

Ces travaux et de très nombreux autres qu'il est impossible de citer ici ont permis de préciser certaines lois, certaines contraintes particulières sur les transitions et en fin de compte de proposer des modèles standards qui permettront d'effectuer des synthèses par règles.

B. Utilisation des synthétiseurs pour préciser l'analyse

Ce procédé d'analyse, proposé en 1961 par BELL, FUJISAKI, HEINZ, STEVENS et HOUSE (35) est appelé par les américains "analysis by synthesis" : l'analyse par la synthèse.

Il est parfois difficile de mesurer la valeur d'un paramètre, un formant par exemple.

C'est souvent le cas pour l'étude des voix de femmes dont le spectre est caractérisé par un petit nombre de composantes d'où ne se dégage pas une image suffisamment précise de la structure formantique.

On fait donc une première mesure entachée d'erreur et on effectue une synthèse. Une comparaison est alors faite soit à l'aide de l'oreille, soit et c'est beaucoup mieux au niveau du spectre. On arrive, par corrections successives, à une bonne identité.

Mais supposons qu'il persiste une différence. Cette différence peut nous faire découvrir des traits intéressants sur la source vocale, par exemple, traits que nous chercherons à interpréter.

Si un phénomène est révélé, il pourra être, à son tour, simulé et intégré au synthétiseur.

Ainsi, l'analyse puis, en conséquence, la synthèse peuvent être améliorées.

Le processus d'analyse par la synthèse est à retenir puisque selon certains parmi lesquels LIBERMAN (36), un processus analogue se déroulerait au niveau de notre ensemble de production et de perception vocale.

C. Utilisation des synthétiseurs pour étudier le rôle et l'importance des paramètres

Il n'est plus question, comme précédemment, de réaliser les synthèses les plus exactes possibles mais, ici, on modifie et on constate.

Quel est, par exemple, le rôle du troisième formant ?

Quelle doit être la fréquence minimale d'échantillonnage de tel paramètre, quelle est la quantification qu'il faut adopter pour coder chacun des paramètres ?

Quel est le rôle de la mélodie dans la reconnaissance des consonnes ? Voir par exemple des travaux de CHISTOVITCH en 1969 (37).

Quel est le rôle des transitions des formants dans la reconnaissance des voyelles ? Voir les travaux de LINDBLOM et STUDDERT-KENNEDY en 1967 (38).

D'autres travaux ont été entrepris pour étudier le rôle de telle ou telle cavité à l'aide d'un simulateur vocal ; citons en particulier les travaux de FANT (31).

De la même manière des recherches devront être entreprises sur les paramètres de l'articulation à partir des premiers travaux de STEVENS et HOUSE (39).

D. Utilisation des synthétiseurs pour effectuer des tests de perception

Les sons de synthèse étant très rigoureusement connus et pouvant être calibrés en quelque sorte, des tests de perception pourront être définis pour étudier le comportement de notre organe de perception.

De nombreux travaux ont été réalisés par FLANAGAN (27) sur les différences juste perceptibles. Ainsi des modifications sont apportées sur tel ou tel paramètres

On cherche à savoir à partir de quelle durée ou à partir de quelle intensité une voyelle est identifiée.

Quel est le rôle de la mélodie et des quatrième et cinquième formants dans la perception des voyelles ? Voir les travaux de FUJISAKI (40).

Comment se déplace notre système de perception à l'écoute de sons isolés et à l'aide de mots références ? Voir les travaux de FOURCIN (41).

On pourrait encore citer de nombreux autres travaux qu'il serait intéressant de développer.

Pour en terminer avec ce chapitre, il faut bien noter le caractère artificiel du classement proposé ici. C'est ainsi, par exemple, que lorsque l'on étudie l'importance des paramètres, il faut aussi aborder les problèmes de perception.

E. Applications des synthétiseurs en télécommunications et comme sortie parlée d'ordinateurs

Seul, actuellement, le vocoder peut être utilisé en télécommunications bien que son rôle ait pu évoluer ces dernières années. Il y a quelques années seulement, on visait avant tout à réaliser une compression de bande ; actuellement, on cherche à réaliser des transmissions sûres en employant des techniques digitales. Un débit de 2 400 bits/seconde peut être acheminé par une voie téléphonique normale.

Les problèmes de capacité de ligne ne se posent plus avec la même acuité grâce au développement de nouveaux procédés de transmission à grande capacité d'information (utilisation de satellites, faisceaux hertziens, rayon laser etc)

Quant aux autres types de synthétiseurs, il n'est pas question pour le moment de les utiliser dans un fonctionnement en temps réel, les problèmes d'analyse étant difficiles à résoudre.

Une autre application directement exploitable actuellement consiste en l'utilisation du synthétiseur comme organe de sortie parlée d'ordinateur. Des résultats intéressants ont été obtenus tels ceux d'I.B.M. avec son unité à réponse vocale (42).

Tous les types de synthétiseurs peuvent convenir à cette application, mais on cherchera à employer le minimum de mémoire dans l'ordinateur d'où l'intérêt d'utiliser des appareils commandés au moyen d'un minimum d'informations.

Dans ce domaine, le développement des travaux de "synthèse par règles" permet une réduction encore plus importante de l'information traitée.

VI. CONCLUSIONS

En résumé, quelles sont les applications immédiates des synthétiseurs ?

Quelles sont les recherches qui doivent être développées dans un proche avenir ?

Nous avons déjà évoqué des applications :

- Utilisations des vocoders à canaux en télécommunications avec transmission digitale et possibilité de codage à caractère secret.
- Utilisations des synthétiseurs comme organes de sortie parlée d'ordinateur avec mise en place progressive de règles de synthèse permettant de limiter les opérations de traitement et de réduire l'occupation des mémoires.

Nous pensons aussi que, dans les années à venir, les synthétiseurs deviendront des instruments courants dans tous les laboratoires effectuant des recherches sur la parole.

Une progression des études fondamentales sur la production et la perception de la parole nous paraît devoir découler d'une perfection toujours plus grande des synthétiseurs lesquels doivent pouvoir simuler au mieux l'appareil vocal, chaque progrès dans l'interprétation des processus permettant d'ailleurs une amélioration des appareils.

Dès maintenant, par exemple, le synthétiseur à formants et le simulateur vocal peuvent être perfectionnés : introduction de l'interaction conduit-source pour le synthétiseur à formants, production automatique de bruit pour le simulateur vocal.

Un travail important doit aussi être effectué dans le domaine de la simulation de la source vocale, travail qui profitera à l'amélioration de tout type de synthétiseur. Il semble, de ce côté, que l'on soit proche d'une solution acceptable.

Dans le cadre des recherches fondamentales au moyen de synthétiseurs, il paraît important de bien préciser le rôle de la phase dans les phénomènes transitoires.

La synthèse du français doit être systématiquement étudiée ce qui nécessitera une analyse acoustique complète.

Cette étude serait, en outre, d'un grand secours pour le développement des travaux sur la synthèse par règles.

Par ailleurs, des études à l'aide d'un simulateur du conduit vocal doivent être effectuées. Les problèmes de synthèse par règles au moyen de cet appareil semblent devoir être plus simples à résoudre : on peut alors aborder directement les problèmes des commandes de l'articulation par le cerveau. Une meilleure connaissance de ces phénomènes jouerait un rôle appréciable dans la connaissance globale de l'ensemble du mécanisme de production et de perception de la parole dans la mesure où les deux phénomènes paraissent très liés, très dépendants l'un de l'autre.

Enfin, il nous paraît important de mettre au point ou d'améliorer quelques tests de perception afin de pouvoir comparer l'intelligibilité et la qualité de la parole produite par les divers types de synthétiseurs.

VII. BIBLIOGRAPHIE

(1) H. DUDLEY
The vocoder
Bell Lab. Record, 17, 122-126, (1939)

(2) B. GOLD, C. M. RADER
The channel vocoder
I. E. E. E. Transactions on Audio and Electroacoustics, AU15, 148-161, (1967)

(3) E. E. DAVID, M. R. SCHROEDER, B. F. LOGAN, A. J. PRESTIGIACOMO
New applications of voice-excitation to vocoders
Proc. Speech Com. Seminar - Stockholm (1962)

(4) C. P. SMITH
Vocal Response Synthesizer
J. A. S. A., 37, 170-171, (1967)

(5) M. R. SCHROEDER, B. F. LOGAN, A. J. PRESTIGIACOMO
New methods for speech analysis-synthesis and bandwidth compression
Proc. Speech Communication Seminar - Stockholm (1962)

(6) J. TIERNEY, J. N. HARRIS
Channel vocoder
I. E. E. E. Com. Conv. Boulder (1965)

(7) B. GOLD
Experiment with speechlike phase in a spectrally flattened pitch-excited channel vocoder
J. A. S. A., 36, 1892-1894 (1964)

(8) J. L. FLANAGAN, R. M. GOLDEN
Phase vocoder
The Bell System Technical Journal, 45, 1493-1509 (1966)

(9) L. A. O'NEILL
The Representation of Continuous speech with a periodically sampled orthogonal basis
I. E. E. E. Transactions on Audio and Electroacoustics, AU17, 14-21, (1969)

(10) A. V. OPPENHEIM
Speech Analysis-synthesis system based on homomorphic filtering
J. A. S. A., 45, 458-465, (1969)

(11) F. S. COOPER
Spectrum analysis
J. A. S. A., 22, 761-762, (1950)

(12) J. SAPALY
Principe de l'appareillage électrooptique icophone II
Revue d'Acoustique, 3-4, 270-272, (1968)

(13) G. FANT, J. MARTONY, U. RENGMAN, A. RISBERG
OVE II synthesis strategy
Proc. of the Speech Communication Seminar - Stockholm (1962)

(14) K. N. STEVENS, S. KASOWSKI G. FANT
An electrical analog of the vocal tract
J. A. S. A., 25, 734-742, (1953)

(15) G. ROSEN
Dynamic analog speech synthesizer
J. A. S. A., 30, 201-209, (1958)

(16) R. CARRE, J. P. BEAUVIALA, J. PAILLE
Formant circuit for speech synthesizers
A paraître dans la revue I. E. E. E. Transactions on Audio and Electro-acoustics

(17) M. R. SCHROEDER, A. M. NOLL
Recent studies in speech research at Bell Telephone Laboratories
5e I. C. A., Paper A21,- Liège (1965)

(18) Y. KADOKAWA, J. SUZUKI
Simple calculation of the vocal tract configuration from three formant frequencies
Journal of the Radio Research Laboratories, 15, 147-164, (1968)

(19) P. MERMELSTEIN
Determination of the vocal-tract shape from measured formant frequencies
J. A. S. A., 41, 1283-1294, (1967)

(20) G. FERRIEU
Particularité d'un vocoder "téléphonique"
Revue d'Acoustique, 3-4, 181-182, (1968)

(21) G. FANT
Keynote address presented at the 1967 Conference on Speech Communication and Processing
I. E. E. E. Transactions on Audio and Electroacoustics, AU16, 3-5, (1968)

(22) W. J. STRONG
Machine-aided formant determination for speech synthesis
J. A. S. A., 41, 1434-1442, (1967)

(23) L. R. RABINER
Digital-formant synthesizer for speech-synthesis studies
J. A. S. A., 43, 822-828, (1968)

(24) M. R. SCHROEDER
Vocoders : Analysis, Synthesis of speech
Proc. of the I.E.E.E. 54, 720-734 (1967).

(25) L. R. RABINER
A model for synthesizing speech by rule
I. E. E. E. Transactions on Audio and Electroacoustics, AU17, 7-13, (1969)

(26) M. V. MATHEWS, J. E. MILLER,et E. E. DAVID Jr.
"Pitch synchronous Analysis of Voiced Sounds"
J. A. S. A., 33, 179-186, (1961)

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

(28) J. W. Van den BERG, J. T. ZANTEMA et P. DOORNENBAL
On the Air Resistance and the Bernouilli Effect on the Human Larynx
J. A. S. A., 29, 626-631, (1967)

(29) J. L. FLANAGAN et L. L. LANDGRAFT
Self-Oscillating Source for Vocal-Tract Synthesizers
I. E. E. E. Transactions on Audio and Electroacoustics, AU16, n° 1, (1968)

(30) H.T. CRYSTAL
A Model of Laryngeal Activity during Phonation
M. I. T. Quarterly Progress Report, 78, July 15, (1965)

(31) G. FANT
Acoustic Theory of Speech production
's - Gravenhage : Mouton (1960

(32) J. M. HEINZ, K. N. STEVENS
On the properties of voiceless fricative consonants
J. A. S. A., 33, 589-596, (1961)

(33) P. C. DELATTRE, A. M. LIBERMAN, F. S. COOPER
Acoustic loci and transitional cues for consonants
J. A. S. A., 27, 769-773, (1955)

(34) O. FUJIMURA
An approximation to voice aperiodicity
I. E. E. E. Transactions on Audio and Electroacoustics, AU16, 68-72, (1968)

(35) C. G. BELL, H. FUJISAKI, J. M. HEINZ, K. N. STEVENS, A. S. HOUSE
Reduction of speech spectra by analysis-by-synthesis technics
J. A. S. A., 33, 1725-1736, (1961)

(36) A. M. LIBERMAN, F. S. COOPER, K. S. HARRIS, P. F. Mac NEILAGE
A motor theory of speech perception
Proc. Speech Comm. Seminar - Stockholm, (1962)

(37) L. A. CHISTOVITCH
Variation of the fundamental voice pitch as a discriminatory cue for consonants
Soviet Physics-Acoustics, 14, 372-378, (1969)

(38) B. E. F. LINDBLOM, M. STUDDERT-KENNEDY
On the rôle of formant transitions in vowel recognition
J. A. S. A., 42, 830-843, (1967)

(39) K. N. STEVENS, A. S. HOUSE
Development of a quantitative description of vowel articulation
J. A. S. A., 27, 484-493, (1955)

(40) H. FUJISAKI
The roles of pitch and higher formants in the perception of vowels
I. E. E. E. Transactions on Audio and Electroacoustics, AU16, 73-77, (1968)

(41) A. J. FOURCIN
Speech source inference
I. E. E. E. Transactions on Audio and Electroacoustics, AU16, 65-67, (1968)

(42) R. H. BURON
Generation of a 1 000 words vocabulary for a pulse-excited vocoder operating as an audio response unit
I. E. E. E. Transactions on Audio and Electroacoustics, AU16, 21-25, (1968)

VIII. DISCUSSION

Monsieur RISSET :

Question : Les synthèses de haute qualité réalisées par HOLMES prouvent qu'il est possible de synthétiser une parole de haute qualité à l'aide de la partie synthèse d'un vocoder à formants. Existe-t-il un semblable "théorème d'existence" pour les vocoders à canaux, à savoir des exemples de parole synthétique de haute qualité -même copiés sur mesure à partir d'une phrase réelle - utilisant la partie synthèse d'un vocoder à canaux ?

Réponse : La partie synthétiseur d'un vocoder à canaux est commandée par un ordinateur dans l'unité à Réponse Vocale construite à la Gaude par IBM. Les paramètres de commande sont déterminés préalablement à l'aide de la partie analyseur mais des retouches sont apportées pour améliorer la parole de synthèse. Malgré ces améliorations, il me semble, mais c'est une opinion tout à fait personnelle, que la synthèse obtenue est de qualité inférieure à celle obtenue au moyen d'un synthétiseur à formants.

Monsieur RISSET

A propos de la simulation du conduit vocal : aux Bell Telephone Laboratories, MERMELSTEIN, RUIZ et ATAL utilisent, outre les données fournies par les rayons X, les mesures électromyographiques, la fréquence des formants, des données supplémentaires obtenues à l'aide d'ultrasons et aussi par des mesures d'impédance acoustique. La réalisation sur ordinateur de ce procédé de synthèse demande peu de multiplications par échantillon, ce qui permet de concevoir un système commercial de synthèse de parole n'utilisant pas de synthétiseur analogique, mais un simple convertisseur numérique-analogique.

Monsieur LIENARD

Question : La fin de votre exposé laisse à penser qu'une reconnaissance basée, non sur le message acoustique proprement dit, mais sur la recherche des mouvements articulatoires, permettrait peut-être de mieux définir les invariants. Mais, les études que mène notre laboratoire sur la phonation montrent à l'évidence que les invariants sont dans les formes du message acoustique et non dans la manière d'articuler du locuteur. Les muscles importants du système phonatoire sont en très grand nombre (au moins une cinquantaine) : leur synergie varie d'un individu à l'autre, comme la conformation anatomique des organes phonatoires. Des locuteurs différents, pour produire le même message acoustique, utilisent en général des mouvements articulatoires très différents, selon leurs possibilités propres : seul compte le résultat acoustique.

Réponse : Au niveau des muscles mêmes de l'articulation, la commande n'est peut-être pas la plus simple (nombre de muscles, conformation différente), mais on peut espérer trouver des invariants à un niveau plus élevé sans aller jusqu'à l'origine de l'information. Il me semble que, de toutes façons, on retrouvera plus sûrement au niveau de ces commandes "des formes" particulières significatives.

Monsieur LEIPP

Question : Dans ce même ordre d'idées, on peut citer le cas du mainate, un oiseau parleur qui peut arriver à reproduire de la parole avec une qualité remarquable. Il possède pourtant un appareil phonatoire complètement différent, en dimensions et en conformation, de l'appareil phonatoire humain. Lorsque l'oiseau a été bien éduqué, il est très difficile à un auditeur non prévenu de penser que cette parole n'est pas d'origine humaine, et l'étude des sonagrammes confirme cette ressemblance.

A propos de la différence entre voix d'homme et voix de femmes, je pense qu'il n'est pas utile de rechercher un paramètre qui serait particulier à l'une ou l'autre. Il suffit de considérer que, statistiquement, la femme parle une octave plus haut que l'homme ; dans ces conditions, si on veut bien admettre que comprendre un mot, c'est percevoir une forme, une "gestalt" sur un diagramme fréquence-temps, cette forme est quantifiée d'autant plus grossièrement en fréquence que la voix est plus aigüe ; nous avons fait toute une série de recherches sur ce point à propos des chanteurs et nous avons publié quelques uns de nos résultats dans les comptes rendus des journée d'étude du Festival International du Son (1969).

Ce n'est certainement pas le fait que l'appareil phonatoire de l'homme ne soit pas homothétique de celui de la femme qui soit en cause. Il est facile de montrer qu'avec des machines à parler" non homothétiques, il est possible de fabriquer des signaux acoustiques tout à fait similaires. En effet, chaque cavité, chaque organe dispose, tant chez l'homme que chez la femme, d'un assez large degré de liberté. Il suffit alors de se rappeler que la fréquence de résonance d'un résonateur est directement proportionnelle à la section du col du résonateur et inversement proportionnel à la racine carrée du volume de la cavité et de l'ajutage. On peut donc fabriquer un formant d'une certaine hauteur avec deux résonateurs de dimensions tout à fait différentes ; un petit résonateur avec une petite section d'ouverture peut donner un formant de hauteur identique à celui que produit un grand résonateur avec des ouvertures plus grandes ; comme le pilotage musculaire entre les diverses cavités et ouvertures est plus ou moins autonome, la différence entre voix d'homme et de femme ne relève donc pas de problèmes d'homothétie. Je rajoute ceci : deux résonateurs de dimensions et de sections différentes peuvent donner des formants de même fréquence moyenne identiques ; mais il est bien évident que leurs intensités relatives, toutes choses égales, peuvent être différentes. Les voix de femme n'ont rien de particulier sinon que la quantification des formants par les spectres à raies délivrés par les cordes vocales est environ deux fois plus grossière pour les voix féminines, d'où, statistiquement, la moins grande netteté des "formes" acoustiques de la voix féminine. Mais il faut bien insister que cette "règle" n'est valable que toutes choses égales, intensité en particulier. Dans certains bruits de fond particuliers, les voix de femmes peuvent "percer" beaucoup mieux que les voix d'hommes; c'est une question d'émergence d'une forme sur un fond, et c'est chaque fois un cas particulier.

Réponse : On peut obtenir une même fréquence de formant avec des configurations différentes, mais c'est beaucoup plus difficile si on veut que les trois premières fréquences de résonances, par exemple, de deux séries différentes de résonateurs soient les mêmes.

Dans le cas des voix de femmes, effectivement, la fréquence fondamentale est, en moyenne, double de celle de l'homme mais, de plus, les fréquences des formants sont de 5 à 10% plus élevées que celles des formants pour les voix d'hommes.

Monsieur LEIPP

Réponse à Monsieur CARRE au sujet des Playbacks : l'Icophone n'a, du point de vue principe de fonctionnement aucun rapport avec le Playback de HASKINS ; la seule similitude réside dans le fait qu'ils relisent tous deux des diagrammes fréquence-temps. Le Playback est basé sur la rotation d'une roue phonique où un faisceau lumineux est modulé par fentes qui sont en ordre harmonique. Dans l'Icophone, chaque cellule photoélectrique déclanche un générateur de sinusoïdes qu'on peut accorder de façon autonome à une fréquence quelconque, et dont on peut régler les intensités relatives à volonté. Le système est donc beaucoup plus souple, de plus l'Icophone répond à des signaux de l'ordre d'une milliseconde, alors que la machine de HASKINS, au dire même de DELATTRE qui nous a rendu visite, répondait pour quelques 50 ms, ce qui était un inconvénient énorme pour simuler les plosives et autres phénomènes brefs. A part cela, l'Icophone n'est qu'un générateur de sinusoïdes mélangeables : c'est tout ce qu'il y a de simple et facile à réaliser. Le modèle qui permettra de moduler la hauteur et de rajouter l'intonation est en cours de construction ; l'agrément de la voix synthétique sera considérablement augmenté, comme nous l'ont montré les essais préalables avec montage sur table.

Monsieur ROSSI

Question : La définition formantique pour une voix de femme ou une voix d'enfant est très mauvaise à cause de la pauvreté des composants dans le spectre de raies. Pourtant la communication linguistique est possible ; bien mieux des voyelles isolées, produites par des enfants, sont facilement reconnues. N'y-a-t-il pas par conséquent d'autre clefs qui permettraient l'identification et la discrimination des voyelles réalisées par les femmes et les enfants ? Je pense en particulier aux caractères spécifiques des voyelles = hauteur spécifique, intensité spécifique, durée spécifique. Il est, par exemple, bien connu que certains sourds qui n'entendent pas les fréquences correspondant aux formants, arrivent à discriminer (a) et (i) par référence à la fréquence fondamentale spécifique de ces deux voyelles.

N'y aurait-il pas au niveau de la source des éléments qui aideraient à cette discrimination ?

Réponse : C'est très possible de par l'interaction conduit vocal-source vocale, mais il faudrait connaître les conditions d'expérimentation.

Monsieur LEIPP

Pour simuler la source vocale, je crois qu'il faut être prudent en ce qui concerne les circuits équivalents... Je ne suis pas le premier à insister sur ce point, puisque BOUASSE avait déjà exprimé ses réticences ; du moins lorsqu'il s'agit de simuler des phénomènes d'écoulements de fluides, comme c'est le cas ici. Un écoulement aérien à travers une série de résonateurs couplés, de forme et de sections d'ouverture variables dans le temps est un phénomène compliqué, dont la théorie est loin d'être au point ! Les simplifications qu'on est amené à faire, on risque beaucoup de perdre l'essentiel du phénomène qu'on se propose d'étudier.

LA SYNTHESE PAR REGLES DE LA PAROLE

par

A. NEMETH, Ingénieur à IBM France, La Gaude.

Monsieur le Président,
Mesdames,
Messieurs,

Je voudrais vous parler de la synthèse par règles de la parole. La synthèse, comme vous le savez, c'est la production de la parole à partir de l'écriture. Par écriture on entend habituellement un ensemble de symboles représentant les sons d'une langue. Mais cela n'est pas obligatoire. Le point de départ pourrait être un dessin, ou tout simplement, l'écriture conventionnelle.

Les symboles de départ sont travaillés par des règles appropriées. Il en résulte une suite de données qui sont exploitables par une machine qui produit la parole.

On a toujours besoin d'une machine qui produit le son. Les règles ne sont que des idées. Pour juger de leur valeur, il faut juger par leur résultat.

Il faut donc trois choses pour produire la parole synthétique. D'abord il faut écrire les symboles représentant la parole. C'est une activité essentiellement humaine. Puis, il faut appliquer les règles. L'homme peut exécuter cette phase, mais il est plus pratique de la confier à une machine. Et, finalement, la troisième phase, la production de la parole, se fait toujours par machine.

Je voudrais vous entretenir uniquement au sujet des règles. Mais fatalement mon exposé va déborder sur la première et la troisième phase. Les règles sont en principe indépendantes de l'écriture et de la machine mais il faut assurer la transition entre les trois phases de la synthèse de la parole.

Mais avant d'entrer dans les détails des règles, on peut se poser la question : à quoi sert la synthèse ? Sur la diapositive que vous voyez en ce moment, sont résumées les réponses. D'abord, il y a un but proprement scientifique. A notre avis, la synthèse est le meilleur moyen de prouver la nécessité des résultats obtenus par l'analyse de la voix humaine. D'autre part, surtout quand on travaille dans l'industrie, il est légitime d'avoir une arrière-pensée d'application. Par exemple, la lecture par téléphone d'un fichier écrit, le langage crypté pour communiquer entre les sous-marins, transmettre la parole par les faibles signaux lors des communications spatiales, etc...

Bien entendu, nous ne sommes pas les premiers à avoir songé à la synthèse de la parole. On peut enregistrer des phrases entières et les rejouer aux bons moments. Un exemple classique en est l'horloge parlante. Le nombre de phrases est infini, mais la qualité pourrait être parfaite.

On peut enregistrer les mots d'un dictionnaire. Le nombre d'enregistrements est toujours très grand, de l'ordre de 100 000, la qualité des mots, s'ils sont examinés isolemment, est impeccable, mais la prosodie des mots juxtaposés est tellement mauvaise que l'on conçoit mal être en présence d'une phrase. Néanmoins, à l'heure actuelle, à notre connaissance, les applications commerciales sont toutes basées sur ce principe.

Il y a une solution plus élégante, c'est construire les mots à partir des syllabes. Comme une syllabe est composée de plusieurs sons, nous appelons cette approche : méthode moléculaire. Il faut citer dans cette catégorie : les digrammes, les dyades, les dyphones.

On peut pousser plus loin cette idée, et construire la parole à partir des sons élémentaires : méthode atomique.

Cette méthode a été choisie par les japonais, par MIT, et notamment par nous à La Gaude. Dans notre jargon, nous appelons spectre le son élémentaire qui n'est rien d'autre que l'image des cavités buccales à un instant donné.

Pour être complet, on pourrait citer la méthode de noyau. La nécessité de modulation par impulsions codées permet la reproduction impeccable de la parole. Un seul élément, la présence ou l'absence d'une impulsion, suffirait pour synthétiser la parole. Malheureusement, nous connaissons mal encore la parole pour pouvoir calculer cette impulsion.

Je viens de dire que nous avons choisi la solution atomique. Ce choix implique le principe du vocodeur à canaux.

Ce principe est vieux. Il y a une trentaine d'années déjà, Monsieur DUDLEY a démontré que l'on peut reproduire une parole compréhensible en excitant un jeu de filtres soit à une cadence régulière, soit aléatoirement. Les voyelles et les consonnes sonores demandent une excitation régulière, par contre, les consonnes sourdes sont imitées par les réponses des filtres excités à des intervalles très rapprochés, mais irréguliers.

Sur cette diapositive, vous voyez l'image d'un son vu à travers -j'ose dire- un vocodeur. La moitié gauche représente une voyelle. L'enveloppe correspond à la configuration de la bouche à un instant donné, configuration transposée, bien entendu, dans le domaine de la fréquence. Les bosses représentent les résonnances des cavités buccales. Comme vous le savez, ces résonnances sont appelées formants.

Les raies verticales représentent la mélodie de la voix. Les raies sont équidistantes et l'inverse de la distance indique le temps qui s'écoule entre deux vibrations des cordes vocales.

La partie droite de l'image représente un son fricatif comme les "S", "F", ou "CH", etc... La structure des raies disparaît et les formants sont moins nettement localisés.

Cette diapositive représente la voix à un instant donné. Plus exactement, dans un laps de temps très court, quelques 20 ms.

Or, la parole est une succession rapide de sons instantanés . La difficulté de la synthèse ne réside pas tellement dans la production d'un son isolé, mais plutôt dans l'ordonnancement de la succession rapide de ces sons.

Cette succession dépend de beaucoup de choses, comme vous le voyez sur cette diapositive.

Tout d'abord, il faut isoler les phrases et leur donner une mélodie appropriée. C'est le problème coriace de la prosodie.

Puis, il faut transcrire l'écriture conventionnelle en notation phonétique. (Vous voyez, je commence à déborder sur la première phase, la phase de l'écriture qui est en dehors des règles). Nous appelons orthoépie cette phase préparatoire. Nous le faisons par machine, car la transcription phonétique d'un long texte est un travail laborieux. D'autant plus qu'il faut indiquer, non seulement les sons, mais, en plus, leur durée.

Les sons, plus exactement, les spectres, sont de deux catégories : les spectres de repos et les spectres de transition. Comme leur nom l'indique, les spectres de transition assurent la liaison entre les sons adjacents tandis que les spectres de repos rendent le milieu du son.

Quant à la mélodie, nous postulons qu'elle ne varie jamais brusquement à l'intérieur d'un mot. Pourtant, dans la parole naturelle, on trouve parfois une rupture de mélodie. Mais, à notre avis, ce n'est qu'un défaut de langage, et aucune information n'est véhiculée.

Par contre, plusieurs sons, notamment les consonnes sonores, provoquent une incursion caractéristique de la mélodie qui facilite la perception du son. A ce phénomène, nous avons donné le nom de "micro-mélodie".

Les détails de chapitres sont illustrés sur les projections ci-après. Commençons par la prosodie. La prosodie, je le rappelle, c'est l'évolution générale de la mélodie pendant la prononciation d'une phrase.

Nous allons voir que la transcription phonétique ne dépend pratiquement que de la succession des lettres. Par contre, la prosodie ne dépend pas des lettres mais de la signification du mot, et surtout du rôle du mot dans la phrase.

C'est pourquoi nous estimons que le problème de la prosodie sort du domaine de l'Ingénieur. Aussi, nous sommes-nous adressés à l'Université de Grenoble, plus exactement au Centre de Traduction Automatique. Il nous semble, en effet, que l'outil forgé pour la traduction, c'est-à-dire, pour l'analyse syntaxique, convient également pour résoudre le problème épineux de la prosodie.

Le problème d'orthoépie est beaucoup plus simple. Malgré les apparences, la prononciation française, ou si vous préférez, l'orthographe française, est régulière. Il est vrai que les règles sont nombreuses. On peut donc confier la transcription à une machine.

Le programme d'orthoépie calcule non seulement la nature du son, mais en plus, la durée, comme le montre par exemple cette projection. L'unité de longueur est la "case" qui vaut environ 40 ms.

A chaque son correspond un spectre. Comme en français, il y a à peu près 36 sons, nous avons 36 spectres, dits de repos, pour les rendre.

A ce stade, le programme remplit chaque cas avec le spectre du repos qui correspond au son.

Mais, la parole dans laquelle les sons se relayent abruptement, avec discontinuité, est à peine compréhensible. Il faut absolument soigner les transitions.

En effet, il est indubitable que l'oreille, ou plutôt, le cerveau, ne perçoit pas les sons les uns après les autres, mais plusieurs sons à la fois. La façon dont un son meurt présage déjà la nature du son suivant. De même que la naissance d'un son permet de reconnaître à posteriori, le son précédent.

L'interaction entre les sons dépend de leur nature. Par exemple, les sons alvéolaires ont une influence énorme sur les voyelles nasales qui suivent. On peut l'expliquer par le mouvement rapide de la langue. Par contre, un son labial n'affecte que peu le son qui suit : la langue bouge à peine.

Ainsi, il faut deux spectres de transition au début d'une voyelle nasale qui est précédée par une consonne alvéolaire. Dans d'autres cas, un seul spectre de transition suffit.

Le nombre des spectres de transition est faible : 90. Avec les spectres de repos, cela fait en tout 126 spectres.

La raison en est, comme on le voit sur la projection suivante, que la configuration de la bouche est la même en allant d'un son à l'autre, qu'en allant d'un troisième son à un quatrième. Exemple : Balle et Mère.

Evidemment, strictement parlant, le nombre des positions est infini. Mais, on peut quantiser les positions. Le son produit change, il est vrai, par petit escalier, mais on ne l'entend pas, car l'oreille, grâce à sa constante de temps lisse les transitions résiduelles.

C'est ainsi qu'avec une bonne centaine de spectres on peut décrire toutes les positions de la bouche.

La projection suivante illustre la recherche des spectres de transition.

La machine a, dans sa mémoire matricielle, les spectres de repos. Sur cette projection les sons sont notés par l'écriture phonétique conventionnelle.

La recherche de l'interaction entre deux sons.
Dans la plupart des cas, il n'y a pas d'interaction car les deux sons ne sont jamais voisins. Par exemple, le son "K" ne suit jamais le son "P". Reprenons l'exemple des sons du mot "MERE". La voyelle est influencée par la consonne initiale "M". Cette influence se manifeste par le nombre des transitions, la durée des spectres, et la micromélodie éventuelle.

La projection suivante résume nos idées sur la mélodie. Nous postulons que la mélodie est une fonction continue du temps à l'intérieur du mot phonique. Le codage de la mélodie devient facile : 15 segments suffisent pour rendre n'importe quelle prosodie. Il ne faut plus coder la mélodie point par point, mais segment par segment. L'économie en taux d'information devient considérable.

Le segment est linéaire, comme l'indique la projection suivante, chaque segment est caractérisé par la pente et par la durée. La jonction entre les segments se fait automatiquement par programme. Si l'on veut déplacer en bloc la courbe entière, il suffit de décaler le premier point. Le premier point seul est codé par 7 bits. Finalement, la mélodie est codée par 50 bits et non pas par un ou deux mille bits par seconde.

La continuité de la mélodie subsiste même aux environs de la micromélodie. Simplement, la pente croît brusquement et la fonction reste continue.

L'exemple de la projection précédente indique que la coarticulation des sons "B" et "L" se manifeste par une allure d'escalier de la mélodie. Plusieurs consonnes peuvent avoir la même allure micromélodique. Comme l'oreille est très sensible à la mélodie, la perception des consonnes sonores est facilitée par le genre de leur micromélodie.

Pour faciliter le calcul de la micromélodie initiale nous avons recencé les configurations consonantiques de la langue française avant la première voyelle du mot. Il y en a une cinquantaine, donc très peu. On peut les insérer dans un programme.

L'assemblage des règles permet l'application indiquée sur la projection suivante, qui est l'unité à réponse verbale à 200 bits par seconde.

On a tout intérêt à réduire l'information pour diminuer la charge de l'ordinateur. De préférence, il faudrait que l'ordinateur réponde à plusieurs centaines d'appels téléphoniques simultanés sans y consacrer plus d'un dixième de son temps.

Dans cette application les spectres ne sont pas codés dans l'ordinateur, mais à l'extérieur. L'ordinateur n'a que l'adresse des spectres. Ainsi, on code l'adresse avec 7 bits au lieu de 45. De même, l'adresse des segments mélodiques est codée par 4 bits au lieu d'une cinquantaine.

Grâce à cette réduction, le nombre de lignes téléphoniques desservies par l'ordinateur est au moins décuplé sans augmentation de charge.

Sur la projection suivante, vous voyez l'alimentation de l'ordinateur en écriture. C'est déjà presque l'écriture conventionnelle.

L'entorse aux règles de l'orthographe est voulue. La lettre "E" à accent aigu est rendue par la gémination de la lettre "EE".

La ponctuation est un peu bizarre. Des chiffres et des caractères spéciaux sont associés en guide de ponctuation. Les mots sont souvent séparés par deux cases non-lettres. C'est ainsi que nous communiquons les ordres à la machine concernant la prononciation désirée. Par exemple, si deux mots ne sont séparés que par un seul espacement, les deux mots seront prononcés d'emblée, sans discontinuité mélodique.

Cette ponctuation abondante est encore nécessaire car la machine ne sait pas encore quels sont les mots à prononcer ensemble, où il faut faire les liaisons, quels sont les mots importants de la phrase.

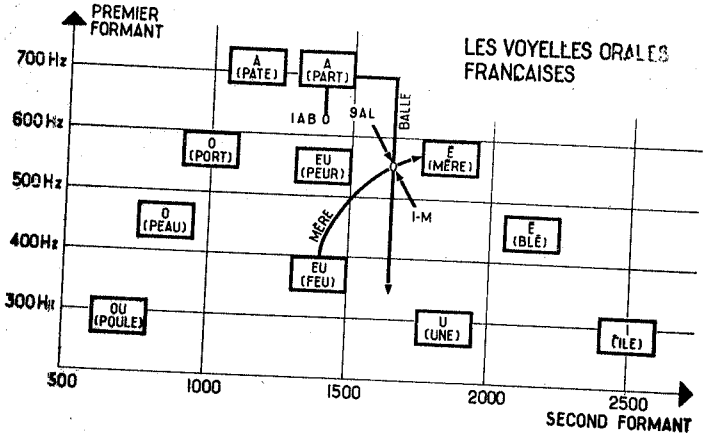
Ce n'est pas la faute de la machine, ni des programmeurs. Il faudrait inclure dans le programme l'analyse syntaxique automatique.

Pour mesurer l'état actuel de l'art, nous devons avouer, et vous allez le constater vous-mêmes bientôt en écoutant la fable : "La cigale et la fourmi", que la qualité obtenue n'atteint pas encore la qualité du téléphone.

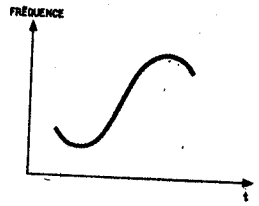
Deux problèmes sont encore à résoudre comme vous le voyez sur la dernière projection. Il y a quelques minutes, je vous ai parlé de la prosodie : il faut la rendre automatique.

Pour le premier problème, il faut remonter trente ans en arrière. Depuis que les vocodeurs existent on admet que la parole est, ou bien engendrée par les cordes vocales, ou bien par la turbulence de l'air sortant de la bouche. Pour simplifier la machine, on admet que les deux phénomènes ne coïncident jamais. Or, c'est une simplification exagérée car l'expérience démontre que la vibration des cordes vocales est presque toujours accompagnée par une friction de l'air dans la bouche.

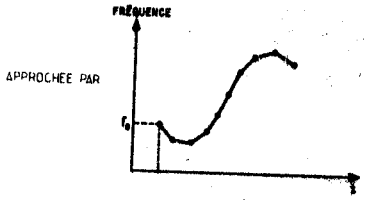
En guise de conclusion, Mesdames et Messieurs, je me permets de me transformer en prophète. Je suis convaincu que le problème de la synthèse de la parole sera résolu par notre génération. Il ne faut pas attendre que nos enfants le fassent. Mais, je le souligne, un chercheur isolé n'y arrivera jamais. Le problème est trop vaste, on a besoin de tout le monde : des linguistes, des phonéticiens, des mathématiciens, des statisticiens, des techniciens. C'est pourquoi, je vous convie à unir nos efforts et en unissant nos efforts, nous allons vaincre.

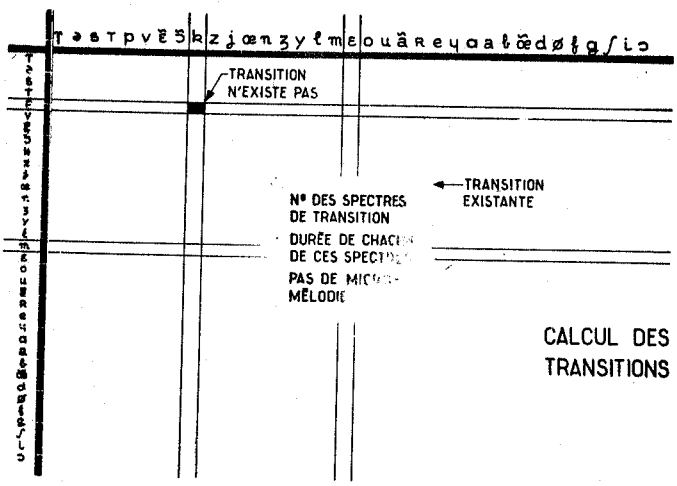


ÉVOLUTION DE LA MÉLODIE

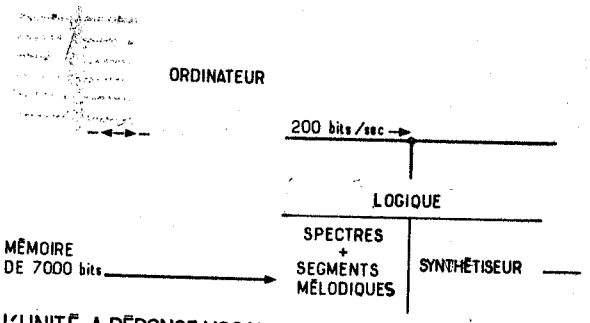


APPROXIMATION DE LA MÉLODIE





CALCUL DES TRANSITIONS



L'UNITÉ A RÉPONSE VOCALE A 200 bits/sec DE PAROLE

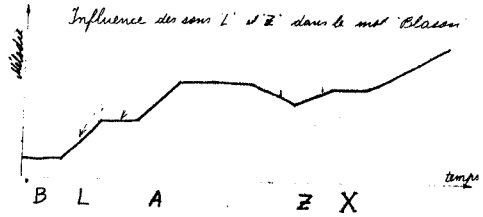
LA 4CIGALE(
LA CIGALE(
SE TROUVAT3•FORT •DEEPOURVUE(•
PAS UN1-SEUL $PETIT1/MORCEAU1•DE MOUCHE6
ELLALA32CRIIER62FAMINE(
SA 4VOISINE.
QUEL QUES 4GRAINES6•POUR 2SUBZISTER(2
JE VOUS3 PAIERAI((
AVANT L'OUT(
INTEERET6=ET 4PRINCIPAL=•
C'EST LA6•SON12MOINDRE6•DEEFAUT.•
DITELLE,5A CETTE16EMPRUNTEUSE(
JE 4CHANTAIS-(
VOUS CHANTIEZ+2
EE BIEN.

ET LA FOURMI=
AYANT CHANTEE TOUT L'ETEE(5
QUAND LA BISE64FUT VENUE=2
OU 4DE 6VERMISSEAU(5
CHEZ LA14FOURMI(/
LA PRIIANT6)DE LUI PREETER(2
JUSQU'A LA1 SAISON6•NOUVELLE.
LUI DITELLE.4
FOI D'ANIMAL.•
LA 5FOURMI6/N'EST PAS84PREETEUSE(
QUE -FAISIEZ VOUS(/AU TEMS CHAUD(2
NUITEE JOUR61/A 6TOUT VENANT(
NE VOUS DEEPLAISE(6
J'EN SUIS1•FORTAISE-&
DANSEZ7-MAINTENANT(4

VI. SEGMENTS MELODIQUES

LA MELODI EST CONTINUE A L'INTERIEUR DU MOT. ELLE EST ROMPUE ENTRE LES GROUPES. 15 SEGMENTS SUFFISENT POUR TOUTE PROSODIE.

VII. MICROMELODIE.



CONCLUSIONS

QUALITE DE VOCODEUR < QUALITE DE TELEPHONE

DEUX PROBLEMES A RESOUDRE :

1. DOSAGE DE LA FRICTION
2. PROSODIE

LA SYNTHESE. POURQUOI ?

POUR ETUDIER LA VOIX, SA PERCEPTION, SA **MELODIE, LES SONS FRICATIFS.**

POUR APPLICATION INDUSTRIELLE :
STOCKAGE DE L'INFORMATION ECRITE = 50 BITS / SEC,
STOCKAGE DE L'INFORMATION PARLEE = 2.000 - 100.000 BITS / SEC.

APPLICATION MILITAIRE : MESSAGE CRYPTE, COMMUNICATION SOUS-MARINE.

APPLICATION SPATIALE : SIGNAUX FAIBLES.

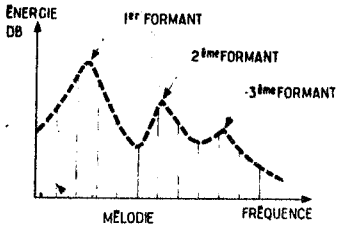
POUR AIDER LA RECONNAISSANCE : PARAMETRES PERTINENTS.

HISTORIQUE DES APPROCHES :

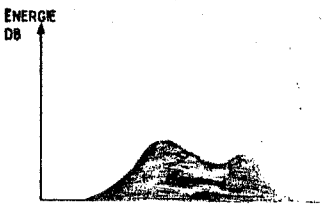
PHRASES :	HORLOGE PARLANTE
MOTS :	LES INSTALLATIONS ACTUELLES
MOLECULES :	SYLLABE, DIGRAMME, DYADE, DYPHONE
ATOMES :	HASKINS, BELL, JAPON, MIT, LA GAUDE
NOYAU :	AUCUNE REALISATION CONNUE

NOMBRE D'ELEMENTS / DIFFICULTES = RELATION INVERSE

PHRASES :	INFINI :	QUALITE PARFAITE
MOTS :	10.000 - 100.000 :	MANQUE DE PROSODIE
MOLECULES :	1000 :	MELODIE DIFFICILE
ATOMES :	100 :	JONCTION A SOIGNER
NOYAU :	1 :	DIFFICULTES INSURMONTABLES ACTUELLEMENT.



SPECTRE D'UNE VOYELLE



SPECTRE D'UNE CONSONNE SOURDE

SYNTHESE DE LA PAROLE A 200 BITS PAR SEC.

I PROSODIE.

II. ORTHOEPIE.

III. RYTHME.

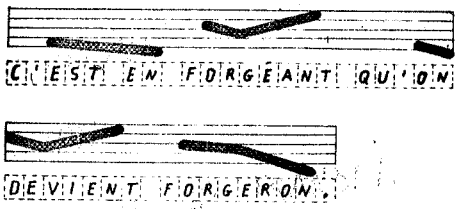
IV. SPECTRES DE TRANSIT.

V. SPECTRES DE REPOS

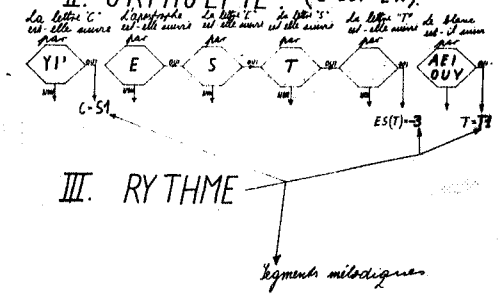
VI. SEGMENTS MELODIQUES.

VII. MICROMELODIE.

I PROSODIE



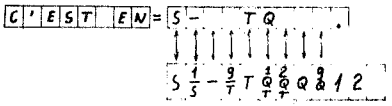
II. ORTHOEPIE (C'EST EN).



III. RYTHME

Segments mélodiques

IV. SPECTRES DE TRANSITION



V. SPECTRES DE REPOS.

SURTOUT AU MILIEU DES VOYELLES LONGUES. CEDENT LA PLACE AUX SPECTRES DE TRANSITIONS.

LE MEME SPECTRE PEUT ETRE DE REPOS OU DE TRANSITION.

576421000135435 $ 1-K 10;

REPONSES AUX QUESTIONS POSEES PAR DES AUDITEURS

Monsieur RISSET :

"Comment l'orthoépie fait-elle la distinction entre les deux "est" des phrases : "C'est en forgeant qu'on devient forgeron" et "Le soleil se lève à l'est" ?".

C'est un cas typique où l'orthoépie est impuissante. Elle est basée sur la succession des lettres et elle ignore la signification des mots. Pour sortir de l'impasse, on instruit l'orthoépie pour le mot le plus fréquent, et pour le mot le moins fréquent, on fait une entorse à l'orthographe. Par exemple, la direction "est" libellée "ESTH". La lettre "H" en français ne se prononce pas, mais sa présence empêche la lettre "T" d'être la dernière et, par conséquent, elle sera prononcée.

D'ailleurs, l'othoépie est en dehors des règles de la synthèse. Nous avons développé, pour diminuer notre labeur et pour faciliter la première phase, celle de l'écriture.

Pour l'orthoépie, il n'y a pas de faute d'orthographe, ni de mot exceptionnel. Lors de l'échafaudage des règles de prononciation, on est souverain. La prononciation de n'importe quel mot est possible à deux conditions : que l'on ait assez de mémoires de machine pour contenir les règles, que l'on arrive à garder dans sa propre mémoire des exceptions et les entorses faites à l'orthographe.

Monsieur CARRE :

"Pourquoi utilise-t-on un vocodeur ? Un vocodeur à formants aurait permis une réduction supplémentaire de l'information ?

La raison est historique. Notre unité à réponse verbale est basée sur le principe des vocodeurs à canaux. La préparation du vocabulaire demandait trop de corrections manuelles. En automatisant de plus en plus les corrections, nous étions amenés -nolens volens à faire la synthèse par règles.

Résumé à des réponses à plusieurs questions au sujet des défaillances du programme d'orthoépie.

On ne peut demander à un programme plus qu'à un être humain. L'orthoépie n'examine pas le contexte. De même, la prononciation d'un mot isolé reste ambigüe, même pour un être humain. Le but de l'orthoépie, c'est soulager le labeur de la première phase, et non pas d'en faire un programme spectaculaire. Le but est largement atteint, nous en sommes très contents. D'autant plus que les exceptions sont vraiment exceptionnelles et facilement contournées.

Réponse à Monsieur GUIBERT, concernant la taille des mémoires nécessaires.

J'ignore leur taille exacte. Toutefois, elle est trop grande pour les mettre entièrement dans la mémoire de l'ordinateur 1 401. Le programme est découpé en sept ou huit morceaux, il est gardé sur bande appelée "moniteur", on appelle les morceaux, les uns après les autres. La règle d'orthoépie demande 7 600 bits. La machine n'admet pas que les règles demandent plus que 8 000 bits. C'est une des raisons, d'ailleurs, qui fait que nous aimons les entorses à l'orthographe.

Une autre question : "Pourquoi le monde des spectres est-il limité à 126" ?

Parce que c'est un chiffre rond pour un ordinateur : $2^7 - 2$. Si l'on voulait soigner davantage les transitions et introduire quelques spectres en plus, il faudrait réserver dans la mémoire $2^8 - 2 = 254$ positions, même si on ne s'en sert pas.

Réponse à la question de Monsieur LEIPP au sujet de la constante de temps de l'oreille.

Nous estimons que la constante de temps de l'oreille est de 30 ms. Bien entendu, on peut parfaitement percevoir des signaux qui sont beaucoup plus courts. Il importe peu de savoir, si la perception des signaux courts dure aussi peu de temps que le signal même, ou bien si elle est étalée sur une trentaine de millisecondes.

Pour nous, la constante de temps, c'est la mémoire de l'oreille. Si le signal se répète, même vaguement, deux ou plusieurs fois pendant ce laps de temps, l'oreille entend un signal mélodieux. Si elle n'y découvre aucune périodicité, le signal est perçu comme un bruit. A ce propos nous avons fait une expérience sur un son "S" digitalisé. On a fait perforé sur cartes les données. En recopiant quelques cartes du milieu du paquet, on n'a pas obtenu le son "S" original, mais un son mélodieux. On peut en conclure que trente fois dix millisecondes de bruit devient trois cent millisecondes de musique. Dix millisecondes, c'est nettement plus court que la constante de temps de l'oreille, la périodicité est perçue. La périodicité de la voix humaine est aussi de l'ordre de 5 à 10 ms. Visiblement, l'oreille est faite, surtout pour entendre la voix humaine.

Réponse à la question de Monsieur DREYFUS-GRAAF

L'importance attribuée aux transitions n'est pas en contradiction avec l'observation que certaines phrases restent compréhensibles même si on les joue à l'envers sur le magnétophone. Le début et la fin d'une voyelle sont presque les mêmes, si la consonne qui précède et celle qui suit cette voyelle sont de même nature.

Réponse à une autre question.

Les règles de la synthèse ont été construites à partir de mots isolés. Elles en portent encore les séquelles : le rythme laisse à désirer. Ce sont surtout les mots longs de la phrase qui sont lentement prononcés.

RECONNAISSANCE DE MOTS ET DE LANGAGES PARLES

par

J.Y. GRESSER, C.N.E.T., LANNION

Cet exposé se compose de deux parties. La première est une courte revue des machines ou programmes à reconnaître les message parlés qui existent actuellement dans quelques pays des deux mondes. Je me suis limité aux systèmes entièrement automatiques fonctionnant en temps réel ou avec un temps de réponse assez court. Parmi les autres systèmes certains mettent en oeuvre des traitements si longs qu'il n'est pas envisageable de les essayer dans les conditions réelles de la reconnaissance vocale, d'autres laissent une certaine part à l'intervention humaine qui, si minime soit-elle, vient fausser la règle de l'automaticité complète.

Pour intéressantes qu'elles soient les solutions qu'ils proposent ne peuvent être que de bonnes mises sur la voie. Ce sont à mon avis les premiers systèmes qui permettent de jalonner les progrès vers la reconnaissance automatique du discours parlé.

La seconde partie est un inventaire des problèmes posés par la reconnaissance automatique. Il s'ordonne selon la structure habituelle des machines à reconnaître : chaîne linéaire composée d'un capteur, d'un extracteur de paramètre et d'un étage de décision. Cette structure, qui sera aussi discutée, apparaîtra par la suite bien plus comme une commodité d'exposition que comme le plan type d'un dispositif. Elle est utile dans la pratique car sa modularité permet d'insérer à un endroit ou l'autre, une "boite noire" simulant tel ou tel mécanisme (partiel) de la perception et de transposer ainsi les théories issues de la biologie, la psychologie ou d'ailleurs.

Pour finir, j'essayerai d'établir des points de communication entre les expériences en cours, dans les différentes équipes de langue française : normalisation de la présentation des résultats, rassemblement d'un fond commun de données vocales à des fins d'analyse de la parole ou d'essai des machines à reconnaître.

I. D'AILLEURS ET D'ICI

Si l'on juge de l'importance accordée à un sujet par le nombre de publications qui s'y rapportent, assurément les Etats-Unis et le Japon viennent en tête par le nombre des équipes qui semblent intéressées à la reconnaissance de la parole. Mais il ne faut pas mésestimer les travaux effectués en Union Soviétique qui, pour moins diffusés qu'ils soient, n'en ont pas moins donné lieu à des expériences impressionnantes. Il y a jusqu'ici en Europe peu de réalisations concrètes. Cela marque-t-il un certain retard ou un manque d'intérêt ?

. ETATS-UNIS

Les premières études sur la reconnaissance automatique des chiffres parlés, eurent lieu vers 1951. Depuis, après avoir bénéficié de l'intérêt général porté à la reconnaissance des formes, dans les années 60, les recherches sur la parole ont peu progressé. Alors que la reconnaissance automatique des caractères imprimés est en voie de commercialisation, la reconnaissance vocale est bien loin d'en être au même point. L'une des raisons de ce retard est certainement la petitesse relative des équipes, mais la difficulté théorique est l'obstacle fondamental. Reconnaître la parole est pour le moins aussi difficile que reconnaître l'écriture manuscrite, et l'on sait que ce problème ne trouvera pas de si tôt une solution.

Les estimations de la Rand Corporation pour l'entrée dans la pratique du dialogue parlé avec un calculateur oscillent entre un optimiste 1975 et un prudent 2000.

Depuis un certain temps déjà la reconnaissance d'un vocabulaire limité à quelques dizaines de mots prononcés par un très petit nombre de locuteurs est considérée comme chose acquise et de peu d'intérêt... commercial. Aussi les grands laboratoires de recherche appliquée ont-ils abandonné le sujet aux universités et aux instituts qui leur sont reliés. Là on peut se replier sur les études fondamentales qui semblent actuellement nécessaires aux progrès désirés : extension du vocabulaire traité aux deux ou trois mille mots de la langue courante, traitement du discours naturel et généralisation à un grand nombre de locuteurs.

Les deux pôles géographiques d'intérêt sont Palo Alto (Cal.) et Cambridge (Mass.). A Palo Alto les études sont menées par l'équipe de R. Reddy dans le cadre du Projet d'Intelligence Artificielle de STANFORD dirigé par J. Mc Carty. Reddy est probablement un des seuls à avoir abordé de front le problème de la parole continue (42) (43a) (43b). La thèse de VICENS (66) est une synthèse pratique des résultats obtenus jusqu'aux premiers mois de 1969. Il est intéressant de citer ceux-ci - ils ont été obtenus sur la machine à deux unités de traitement PDP6, PDP-10 du projet (le PDP-6 traitant les modules temps réel du programme), machine prolongée d'un appareil de prétraitement analogique (en gros un dispositif de comptage des passages par zéro dans quatre domaines de fréquence) connecté aux dispositifs usuels d'entrées -. Vicens a obtenu :

- "98% d'identification correcte pour une liste de 54 mots, effectuée en 2 à 3 s par mot, après 4 passages, pour un locuteur ;

- "85 à 90 % d'identification correcte pour une liste de 54 mots, enregistrée par 10 locuteurs, identification effectuée en 9 à 12 secondes par mot, après 9 passages ;

- "97 % d'identification correcte pour une liste de 70 mots français, prononcée par un seul locuteur, identification effectuée en 2 à 3 secondes par mots ;

-"92 % d'identification correcte pour une liste de 561 mots ou courtes phrases, prononcées par un seul locuteur, identifications effectuées en 10 à 16 secondes après 3 passages.

La reconnaissance automatique de phrases d'un langage défini par une syntaxe simple (langage de commande des actions d'un robot, et petit calculateur de "bureau").(*)

A Cambridge des recherches sont menées au M.I.T. (RLE, LINCOLN laboratories) et à l'Université HARVARD. Elles portent sur l'analyse, la syntaxe est la reconnaissance de la parole (4) (15b) (56) (57). Depuis un certain temps K.N. STEVENS a entrepris dans le cadre de Bolt, Beranek et Newman une étude systématique des propriétés acoustiques des sons anglais, en vue de la reconnaissance automatique. BOBROW a présenté il y a un an (FJCC, 1968) un programme de reconnaissance de 100 mots ou phrases simples pour 30 locuteurs (après apprentissage), prononcés dans de bonnes conditions d'enregistrement.

Il y a eu d'autres recherches (64) (34) mais elles ne semblent pas à la pointe du progrès. Il est peut être intéressant de signaler dès maintenant le contrat de la RCA avec la poste américaine sur une étude qui doit aboutir à l'utilisation de la reconnaissance vocale pour le tri postal, d'ici quelques années (46).

. JAPON

En 1961, la NEC (Nippon Electric Corporation) présentait avec l'Université de KYOTO une "machine à écrire phonétique". C'était une machine assez imposante qui devait effectuer de manière analogique prétraitement, classification, analyse, segmentation, reconnaissance et écriture de la parole continue en temps réel (47 b,a). Cette expérience semble avoir été abandonnée (bien que l'appareil figure toujours au catalogue des produits et recherches de la NEC). Un grand nombre d'études partielles ont été effectuées sur les éléments susceptibles d'être utilisés dans un système de reconnaissance. Notamment, la perception et la reconnaissance des voyelles, isolées ou non, a été approfondie. Il n'y a pourtant eu jusqu'ici qu'assez peu d'expériences synthétiques (7) (18) (25) (61) (68).

. UNION SOVIETIQUE

Les Russes ont beaucoup moins publié sur leurs expériences pratiques de reconnaissance vocale que sur leurs travaux théoriques de reconnaissance des formes (41)(60). La diffusion des publications est un problème linguistique et politique. Toutefois, on montre volontiers aux visiteurs du centre de calcul de l'Académie des Sciences d'U.R.S.S., à Moscou, comment un calculateur universel BESM-3M est capable d'effectuer certains calculs et transferts d'informations commandés à la voix (67). Le programme reconnaît en temps réel un mot parmi 40, prononcé, semble-t-il, par un assez grand nombre de locuteurs. On trouvera d'intéressantes références dans un article de FALTER et OTTEN qui date malheureusement de 1967 (11 b).

. EUROPE

Des recherches sont menées sur la parole en Suède (12) (13). Au Danemark BECKER (46) pour illustrer une méthode pratique de conception des

---

(*) CASTAN conteste la valeur des résultats obtenus par VICENS sur la plus longue liste de mots : d'une part ces mots sont bien différentiés, d'autre part apprentissage et reconnaissance ont été effectués sur la même liste où chaque mot ne figurait que 5 fois.

appareils à reconnaître les formes, a construit une petite machine capable de reconnaître dix chiffres prononcés par une seule personne. Malheureusement un tel appareil n'est pas adaptable et BECKER ne précise pas à partir de quel point il faut recommencer l'étude pour un autre locuteur.

Les Allemands montent des expériences de reconnaissance vocale, tantôt, dans un but pratique (65), tantôt pour simuler ou vérifier certaines théories physioacoustiques (70). Pour avoir un bilan des recherches actuelles il est recommandé de consulter le compte-rendu du Congrès de Cybernétique (BERLIN, Avril 1970).

Actuellement aucune des réalisations concrètes n'a pu être essayée suffisamment longtemps pour que l'on puisse discuter des résultats obtenus.

J'ignore si les études entreprises en Italie ont dépassé la reconnaissance des voyelles (15a).

Pour les études menées en France, ou en Belgique, je renvoie aux notes rédigées par chacune des équipes participant à ce séminaire.

. AMERIQUE LATINE

On commence à aborder sérieusement à l'Université de BUENOS AIRES (Prof. MAZZARO) la synthèse et aussi la reconnaissance de l'espagnol parlé.

BILAN PROVISOIRE

Les résultats les plus spectaculaires sont assurément ceux des Russes et des Américains. Il est difficile de connaître l'état d'avancement réel des travaux en U.R.S.S. et l'intérêt que l'on y porte. Les témoignages et les écrits sont trop fragmentaires pour que l'on puisse juger du degré de généralité du programme de VISSOTSKY. Cette critique devrait d'ailleurs en saine pratique expérimentale s'adresser à tous les travaux de reconnaissance vocale. Souvent l'échec, pourtant instructif, de certaines expériences a été masqué par l'imprécision dans la présentation des résultats. On peut espérer que la collaboration franco-soviétique en informatique viendra nous apporter quelques lumières.

Les Japonais, après que certaines équipes se soient lancées dans l'aventure de la machine à écrire phonétique, en espérant aboutir rapidement, procèdent de façon méthodique et très complète à l'étude critique de quelques idées clés avant de les appliquer à la reconnaissance automatique dans les conditions réelles. Y aura-t-il une machine à reconnaître le japonais à l'Exposition Universelle d'OSAKA ?

En France, faute de motivations, on paraît hésiter à dépasser le stade des expériences préliminaires. Il y a aussi d'autres raisons qui sont d'ordre financier - les études de reconnaissance coûtent très cher et leur rentabilité n'est pas évidente - et d'ordre humain - les équipes sont nouvelles, très réduites et ont encore peu songé à unir certains de leurs efforts.

En 1969, l'équipe de REDDY était pratiquement la seule à avoir présenté un système évolué de reconnaissance de la parole, évolué quant à l'ampleur du vocabulaire traité, évolué quant au traitement de la parole continue (certes limitée à de courtes phrases), dans un contexte réel d'utilisation (le projet HAND-EYE-EAR, ou la petite machine pédagogique DESCAL).

Dans cette revue, je n'ai pas abordé, les études générales sur la reconnaissance des formes et les machines. Ce n'était pas le sujet de l'exposé. Pourtant certaines sont aussi importantes pour les progrès de la reconnaissance automatique d'un langage parlé que celles qui prennent la parole pour objet principal.

II. PROBLEMES ET SOLUTIONS ACTUELS

J'aurais pu aborder l'étude critique des machines à reconnaître par une synthèse des théories de la perception auditive chez l'homme ou les animaux et montrer comment une machine doit s'y conformer. Mais ces théories sont fragmentaires et il serait vain pour un profane de tenter ce que les spécialistes n'ont pas réussi jusqu'ici : donner un modèle de la perception, un modèle utilisable, c'est-à-dire un algorithme. Il se peut que cet algorithme n'existe pas. M. EDEN, après avoir longtemps piétiné sur la reconnaissance de l'anglais manuscrit , prétend qu'aucune machine, si perfectionnée soit-elle, ne sera à l'abri d'erreurs qui sembleraient grossières pour un être humain. C'est probable, mais une telle machine sera capable d'autres performances qui compenseront ces défaillances.

Rien ne prouve que l'on doive juger la reconnaissance automatique selon les mêmes critères que la perception humaine (ou l'idée que l'on s'en fait). Les recherches sur la reconnaissance des formes visuelles ont longtemps stagné faute d'être placées dans un contexte correct, où la reconnaissance des caractères imprimés a certes une place importante mais qui n'est plus exclusive. Il serait dommage de recommencer les mêmes erreurs en reconnaissance vocale. Le problème est peut être résolu sans que nous le sachions.

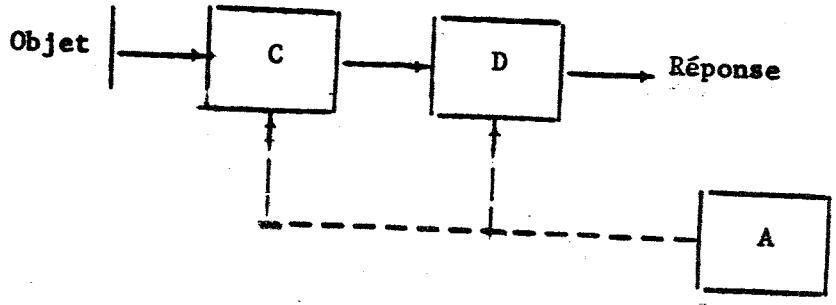
Une machine à reconnaître la parole c'est, à la suite l'un de l'autre, un micro ou un magnétophone, un capteur, un étage de décision (figure 1). Lorsqu'on lui présente un objet, elle peut fournir différents types de réponse : donner un nom à l'objet ou déceler sa présence (s'il appartient à une certaine classe) ou encore déclencher une action. Parfois la réponse aide au réglage des paramètres de décision (et aussi du capteur) c'est l'apprentissage.

- FIGURE 1 -

A ce schéma très simple, je préfère en substituer un plus explicite (figure 2), et aussi plus classique. Il n'est pas exclu qu'un seul appareil puisse réaliser plusieurs des fonctions décrites : celle du capteur (ou codeur), le prétraitement (lissage ou cadrage du code obtenu), l'extraction de paramètres (tentative pour réduire le nombre d'éléments manipulés, la "quantité d'information" au sens des informaticiens, c'est-à-dire projeter ou transformer l'espace des formes), la décision (séparation dans l'espace des formes par une méthode globale, statistique, ou par un cheminement le long d'un graphe).

Dans un tel schéma, qui n'exclut pas à chaque phase la mise en oeuvre de traitements parallèles, les fonctions sont abordées séquentiellement dans l'ordre où je les ai citées. Considérons par exemple la décision : ce peut-être une séparation dans R^n ou un automate déterministe ou autre chose. La parole y arrive comme une séquence de symboles. C'est cette séquence qu'il s'agit d'analyser mais les symboles mêmes qui la composent ne sont pas connus avec une certitude absolue, leur valeur n'a été affectée qu'avec une certaine marge d'erreur. Il faut pouvoir leur attribuer une valeur différente pour obtenir une réponse convenable : -

FIGURE 1



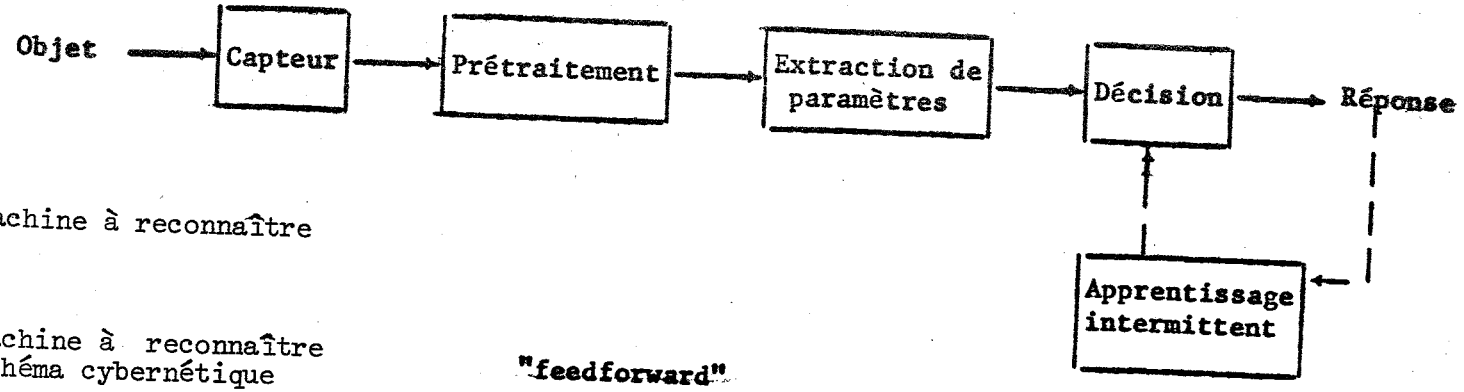
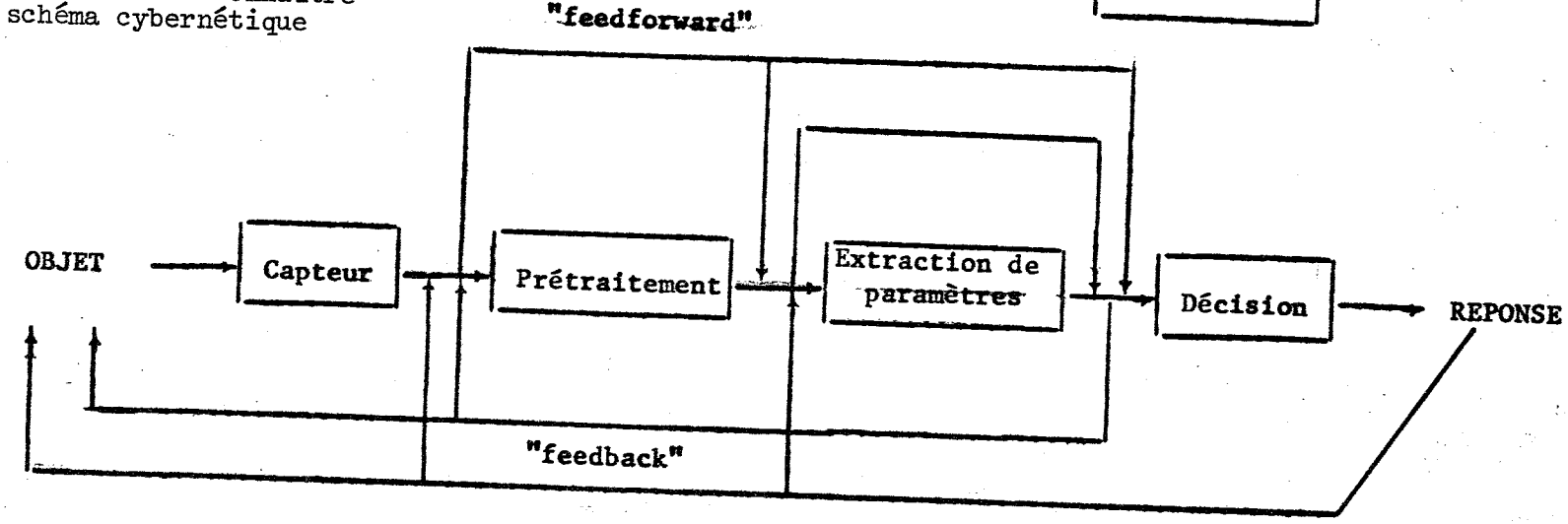


FIGURE 2 - Machine à reconnaître

FIGURE 3 - Machine à reconnaître
schéma cybernétique



conforme au contexte acoustique ou au contexte linguistique si l'on suppose que toute suite de sons prononcée devant la machine appartient au langage à reconnaître. Il est donc utile de prévoir une rétroaction ("feed back") de l'étage de décision sur les étages précédents (*). On conçoit aussi l'utilité d'une action anticipée ("feed forward"). Bien que cette dernière puisse s'interpréter comme la mise en parallèle de plusieurs machines à reconnaître, il m'a semblé intéressant de lui donner un rôle symétrique de celui de la rétroaction. La figure 3 montre les nouvelles liaisons fonctionnelles. L'objet lui-même peut être transformé pour être reconnu. La reconnaissance est alors une véritable "Ubereinstimmung" entre l'objet et la machine. Ce schéma semble aller à l'encontre des principes d'économies qui guident habituellement la réalisation des systèmes de reconnaissance. Mais les progrès de la technologie des calculateurs analogiques ou numériques sont rapides... D'autre part, il se trouve que les expériences qui, depuis dix ans marquent un certain progrès en reconnaissance des formes, peuvent s'interpréter à l'aide d'un tel schéma (16) (66).

- FIGURES 2 et 3 -

On peut imaginer des schémas plus complexes où le schéma précédent serait le composant fondamental.

On peut aussi trouver des schémas entièrement différents. Rien ne permet de les exclure. L'efficacité étant en ce domaine le seul critère de jugement.

On peut revenir au schéma initial et penser que, si médiocre soit-il, il est suffisant. Cela dépend de l'usage que l'on veut faire de la reconnaissance.

Après avoir évoqué les problèmes posés par chacun des composants d'une machine à reconnaître (capteur, étages de traitement et de prétraitement, étage de décision), je parlerai de l'apprentissage, puis de l'utilisation de la reconnaissance vocale. Celle-ci a des répercussions théoriques et pratiques, car c'est elle qui motive, pour une partie d'entre nous, les recherches sur la reconnaissance.

a) LES CAPTEURS

Le compte de la journée d'étude organisée par l'AFCET en juin 1969 contient une description détaillée des capteurs utilisés en France. On pourra s'y reporter. A deux exceptions près aucun de ces appareils n'a été conçu pour la reconnaissance, mais pour la synthèse ou la transmission à faible débit de la parole. Ce sont, peut-être, de bons appareils d'analyse mais ils ne semblent pas être bien employés à l'heure actuelle : ils fournissent un code trop fin

---

(*) On peut objecter à ce raisonnement rapide en faveur d'une approche cybernétique, un raisonnement analogue en faveur d'une approche probabiliste : la probabilité d'apparition d'un phénomène ϕ étant, entre autres, fonction du contexte C_1C_2 (limité) de ce phénomène. C'est une sommation de $P_r\ (C_1\phi C_2/C_1C_2)$. L'argument essentiel en faveur de la première me semble sa relative simplicité de mise en oeuvre face à l'énormité des calculs que demanderait la seconde. On pourra se reporter à propos de la seconde aux études limpides de CHOW.

à un système qui en général ne sait qu'en faire. Il y a une désapdaptation nette entre eux et les étages ultérieurs de la machine. Cela traduit peut-être une évolution historique, mais ne devrait-on pas là aussi mettre en pratique les principes d'économie qui, jusqu'ici, semblent caractériser toute expérience de reconnaissance automatique - notamment utiliser les méthodes les plus simples (18) (54) (59) comme celle du comptage des passages par zéro (48).

La distinction est assez arbitraire entre un capteur et un extracteur de paramètres. On peut définir le capteur comme un codeur et un compresseur d'entrée : il traduit l'objet présenté en une série d'indications assimilables pour la suite du traitement, à un rythme adapté au phénomène variable qu'est la parole et à la capacité d'absorption du système. Il est souhaitable qu'il n'élimine pas toute redondance acoustique.

Pour les phénomènes quasi-stationnaires 20 ms semble être une bonne période d'échantillonnage alors que 4 ms convient aux sons transitoires. Faut-il échantillonner à un rythme fixe intermédiaire (REDDY considère des segments minima de 10 ms) ou à un rytme variable (difficile à mettre en oeuvre), ou encore à un rythme lent, des dispositifs particuliers se chargeant de repérer et éventuellement de traiter les phénomènes transitoires (c'est la solution adoptée par M. DREYFUS GRAF avec un détecteur de plosives) (11a).

Qu'ils soient composés de circuits numériques ou analogiques (ou même optiques), les capteurs sont des organes spécialisés de traitement parallèle qui semblent jusqu'ici dotés d'une puissance de calcul supérieure aux machines universelles. Ils sont utiles tant qu'il n'existe pas de calculateur capable de gérer à un coût assez faible en temps réel le débit d'un codeur A - D de type MIC ou semblable. Malheureusement ils manquent considérablement de souplesse. Lorsque l'on veut innover, il faut souvent recourir à la simulation lente et coûteuse. Combien de temps encore les machines à reconnaître seront-elles hybrides ?

b) EXTRACTION DE PARAMETRES

1°) Tentative de définition

Dans la plupart des cas, on ne peut classer directement la bouffée de code brut provenant du capteur. On commence en général par essayer d'éliminer certains défauts de celui-ci par des techniques de lissage. Puis, si cela est possible, on cadre grossièrement l'information obtenue. C'est le prétraitement : phase importante où l'on cherche à éliminer le "bruit" introduit par l'appareil d'analyse. Pour O. SELFRIDGE c'est l'essentiel de la reconnaissance des formes malheureusement il y a peu à en dire car en ce domaine les règles sont très empiriques.

L'extraction de paramètres répond à deux principes d'utilité : simplifier le problème de classement en réduisant la quantité d'information traitée, dégager au cours de cette simplification des éléments caractéristiques des classes envisagées. Les transformations de l'objet qui définissent l'extraction des paramètres sont donc à rechercher parmi le groupe de celles qui respectent l'invariance des classes (voir les travaux sur les groupes de LIE, et l'approche de U. GRENANDER). La difficulté tient à la nature même de ces classes qui sont des classes d'objets et non de symboles. On a trop souvent tendance à l'oublier et l'on établit souvent des parallèles hâtifs entre les symboles que l'on manipule habituellement (phonèmes, digrammes, etc) et les objets qui les supportent. La théorie des ensembles vagues de ZADECH (71) assainit un peu cette optique mais elle n'a pas jusqu'ici, encore prouvé son utilité pratique.

2°) Considérations pratiques

J'ai déjà parlé de la dépendance qui existe entre segments acoustiques successifs. La nature et la longueur de ces segments étant d'ailleurs définies en fonction de cette dépendance. Cette dépendance peut apparaître comme un phénomène purement acoustique, ou mécanique ou même linguistique. Elle se manifeste alors, entre paramètres de même niveau. Ce peut être la redondance de la parole considérée comme un signal brut (dégagé de toute signification) : la parole que l'on transmet sur une voie téléphonique.

Mais des paramètres de plusieurs niveaux peuvent dépendre les uns des autres de manière qu'il y ait aussi redondance : ainsi la description à l'aide de 15 phonèmes d'un mot parmi 20. On peut envisager des processus de décision hiérarchisés où les décisions d'un niveau permettent de corriger celles erronées du niveau précédent (*). Bien sûr, cette correction n'a d'intérêt que si elle est utilisée par la suite.

Pour assigner une valeur correcte aux paramètres recherchés, il faut pouvoir échanger des informations à des niveaux différents du traitement de reconnaissance. Cela se traduira pour un programme par un enchevêtrement plus ou moins délicat de procédures récursives, ou pour un dispositif analogique par l'utilisation du "feed back" ou du "feed forward".(29)(44)(66)

Dans la plupart des dispositifs, on se contente souvent, par mesure d'économie, d'une seule famille de paramètres. C'est-à-dire des paramètres dont la valeur est accessible par des algorithmes voisins ou identiques. Les expériences semblent indiquer que les valeurs trouvées sont en général correctes à 80 ou 85 %. On voit comment les performances d'un tel système peuvent se dégrader encore si l'on place en série plusieurs étages de ce genre. Il est donc intéressant d'avoir des processus parallèles compétitifs mettant en jeu des paramètres de nature différente, susceptibles de se corriger mutuellement.

3°) Le choix des paramètres

Pour réduire la quantité d'information à traiter par la machine à reconnaître, on recherche des paramètres caractéristiques, c'est-à-dire des grandeurs qui tiennent compte de la structure de la parole.

On peut employer des méthodes systématiques de recherche plus ou moins automatisées (Taxonomie numérique, analyse factorielle), ou interroger les sciences de la parole. Le dernier procédé étant le plus économique c'est celui qu'on utilise le plus fréquemment. Une recherche efficace doit certainement utiliser les deux méthodes, la progression se faisant alternativement sur des ensembles de plus en plus réduits. On a peu d'exemples de recherche systématique (46). Cela traduit-il la gratuité des recherches sur la reconnaissance vocale ?

(*) Ainsi des indications syntaxiques et sémantiques permettraient de corriger une séquence erronée de segments phonétiques.

Actuellement, la mise en oeuvre d'une telle procédure ne peut être que limitée à des langages de structure bien connue (2) (42) (61) (65b).

Les théories de la perception peuvent fournir des mises sur la voie, aussi bien sur les recherches des paramètres que sur la structure d'un automate de reconnaissance (à vrai dire, le schéma initial proposé d'une telle machine repose sur un schéma perceptif erroné). Mais il faut noter, comme l'a fait M. WAJSKOP dans sa conférence, qu'aucune de ces théories ne fournit d'explication complète, c'est-à-dire entièrement modélisable. Elles sont de nature hémistique (70).

Les théories acoustiques ou linguistiques sont une autre source. Elles contiennent une description "fine" de la parole susceptible d'être utilisée à un certain niveau : le discours est une concaténation d'éléments sonores assez brefs aux propriétés individuelles connues (phonations = phonèmes ou ou digrammes). Parmi ces éléments ceux que l'on dit stationnaires ont été bien étudiés (voyelles soutenues et certaines consonnes) isolément ou dans un contexte réduit (VCC ou CVCVC) (30) (58). La notion de formant y est apparue comme particulièrement féconde : on a pû relier espace de formants et espace perceptif, valeur des formants et allure du canal vocal lors de la phonation (12) (14) (19a) (21) (22) (23) (24) (33) (40) (49).

Les phénomènes transitoires ou les formes évolutives ont été jusqu'ici un peu délaissés, notamment l'étude des plosives (8) (9b).

Les déformations dûes au contexte n'ont pas à ma connaissance fait l'objet d'études systématiques d'une certaine importance. La segmentation de la parole continue en dépend. Il est d'ailleurs probable qu'elles soient un obstacle à la reconnaissance vocale comme à la reconnaissance de l'écriture manuscripte. Une manière de contourner le problème est d'élargir la durée des éléments considérés, ou d'étudier les transitions elles-mêmes comme l'équipe de E. LEIPP. A mon avis cela n'est peut être pas suffisant, aux digrammes, il serait intéressant d'ajouter des trigrammes ou d'autres formes moins simples qui coexisteraient avec les phonèmes. Il faut bien sûr s'assurer que cela n'augmente pas le nombre des classes à considérer dans des proportions géométriques. C'est souvent le cas lorsqu'on se limite à une langue particulière (42) (43a) (43b).

La théorie motrice (36) (37) (39) fournit une approche différente où l'on reconnaît la parole en reconstituant la forme du canal vocal et l'excitation initiale. Elle est malheureusement difficile à mettre en oeuvre et surtout utilisée à des fins d'analyses (31) (36) (37) (39).

On entend habituellement par paramètre une grandeur issue d'une des théories. Ces grandeurs sont heuristiques, en ce sens qu'il existe rarement un algorithme permettant d'atteindre leur valeur, sans d'ailleurs, que l'on mette en doute leur objectivité. On peut lors d'un processus de reconnaissance rechercher une approximation aussi bonne que possible de la valeur réelle ou se contenter de la valeur obtenue de ce qui est détecté. On dispose alors d'un paramètre-machine ou d'un pseudo-paramètre. Une telle grandeur peut être aussi utile qu'une grandeur réelle (66) (67).

4°) La généralisation à un grand nombre de locuteurs

Lorsque l'on cherche à étudier l'emploi d'un système de reconnaissance simultanément à un grand nombre de locuteurs on assiste à une dégradation rapide de ses performances. Cela est-il dû à l'impropriété des paramètres "individuels" utilisés ou à la structure même des systèmes actuels, c'est-à-dire une mauvaise utilisation de ces paramètres ?

Faut-il se livrer à une analyse plus fine de la parole ou raffiner les algorithmes de traitement ? Dans la situation actuelle on peut penser que nous n'avons pas encore su tirer parti des indications que nous sommes capables d'extraire du signal vocal (*) : d'une part les paramètres habituels sont mal exploités, d'autre part aucune analyse approfondie de la parole n'a été faite pour la reconnaissance, tout au moins en France, où je ne connais pas d'enquête comparable à celle de K.N. STEVENS (56) (57) (58).

c) IDENTIFICATION, DECISION

Les critiques que l'on peut adresser aux méthodes de séparation ou de décision, dites de reconnaissance des formes, ne sont pas particulières aux études sur la parole. Elles portent sur deux points : la compatibilité entre le type de séparation et la nature des classes, le temps nécessaire au calcul des coefficients de séparation. Pratiquement le seul cas où la convergence de l'apprentissage est démontrée - celui de la séparation linéaire - est de peu d'utilité pratique. Même dans ce cas il n'existe pas de théorème sur le temps nécessaire à l'apprentissage. On sait seulement qu'il croît très vite avec le nombre de classes considérées.

Cela est aussi valable pour d'autres méthodes globales, non linéaires, où la convergence de l'apprentissage n'est pas établie.

Lorsque le nombre de classes à étudier croît, on a tendance à hiérarchiser le processus de décision : on regroupe les classes en familles que l'on séparera plus finement ensuite. A chaque étage de séparation l'espace où l'on opère peut être continu (paramètre) ou discret (caractéristiques).

Jusqu'ici on a considéré des vocabulaires peu étendus, de quelques dizaines de mots. Si on cherche à les élargir, il faudra améliorer le rendement des méthodes de séparation notamment accélérer la recherche parmi les solutions possibles, pour obtenir la réponse en un temps acceptable. C'est un problème classique d'intelligence artificielle et l'on peut espérer tirer un assez grand profit des méthodes propres à ce domaine. Le programme de VINCENS en montre la fécondité. On peut aussi se reporter à la thèse de Guzman (16) qui, malheureusement pour nous, a trait à des formes visuelles et idéales.

Quelle relation y a-t-il entre ces méthodes et le schéma proposé dans l'introduction ?

J'avoue ne pas bien savoir s'il y a plus qu'une simple différence de formalisme. Le schéma initial décrit une machine capable de classer un petit nombre d'objets réels (et non de symboles). On peut certainement le traduire par une méthode simple de cheminement dans un graphe. Les problèmes accessibles aux méthodes d'intelligence artificielle sont d'un ordre de complexité supérieur.

(*) Ainsi les changements de rythme ou d'intonation

d) L'APPRENTISSAGE

Ce qu'on entend habituellement par apprentissage est relatif aux méthodes de reconnaissance des formes : c'est le calcul ou l'ajustement automatique, supervisé ou non, des coefficients de séparation des classes observées.

La notion de temps réel y a deux significations : l'une relative au temps de calcul des coefficients une fois les données suffisantes introduites, l'autre à l'introduction même de ces données.

L'apprentissage nécessite des calculs très importants. Pour cela, il semble peu envisageable de l'exécuter exclusivement en temps réel. On peut pour des expériences pratiques le préparer en temps différé à l'aide d'échantillons enregistrés au préalable puis le terminer en temps quasi-réel à l'aide d'échantillons introduits de vive voix, c'est-à-dire procéder à un réajustement de coefficients à chaque expérience nouvelle. Nous procédons ainsi au C.R.L.. La méthode n'est sans doute pas applicable au grand public mais elle nous est commode.

On peut généraliser la notion d'apprentissage à un processus où la structure même de la machine à reconnaître est modifiée. Cela peut se faire automatiquement ou au cours d'un processus de dialogue élaboré avec l'expérimentateur. C'est selon un processus rapide ou très lent où la machine s'éveille lentement à des possibilités nouvelles de la même manière qu'un enfant au cours de son développement intellectuel. Elle reconnaîtra d'abord les phonatomes, puis des mots, ensuite des phrases qu'elle sera capable d'interpréter et peut-être de traduire en une action concrète.

Nouvelle définition :

L'apprentissage a été défini jusqu'ici comme l'adaptation de la machine à l'homme. Si on lui donne le sens général d'amélioration des performances de la machine, il ne faut pas négliger la faculté d'adaptation de l'expérimentateur. Il me semble que certains systèmes lui doivent beaucoup plus leurs succès qu'à la méthode de reconnaissance utilisée. Cela me porte d'ailleurs à contester fortement tout résultat d'expérience où l'expérimentateur est un familier de la machine (*).

Il faudrait à mon avis deux nombres pour exprimer les performances d'un système de reconnaissance vocale : un taux de reconnaissance, et le temps mis par un locuteur non averti à atteindre ce taux. Cela nécessite la rédaction d'un manuel d'apprentissage. En testant la reconnaissance, on apprécierait l'efficacité de ce manuel.

C'est un premier pas vers une optique saine de la reconnaissance vocale : la placer dans un contexte réel d'utilisation.

L'apprentissage de l'utilisateur peut-il aller jusqu'à une normalisation de la voix ? Qu'est-ce que la normalisation de la voix ? Est-ce autre chose qu'une bonne "articulation" ?

(*) Je propose que l'on définisse l'apprentissage comme l'adaptation réciproque de l'homme et de la machine.

e) UTILISATION DE LA RECONNAISSANCE VOCALE

Utilisation pratique

La liste ne manque pas des applications où la reconnaissance vocale serait intéressante si elle était suffisamment bon marché et si elle était fiable. Ce sont en effet ses deux défauts majeurs. Dans l'état actuel des recherches les systèmes sont extrêmement limités quant à l'extension du vocabulaire traité, quant au nombre de locuteurs concernés simultanément et quant aux performances. Je ne connais aucun système qui s'adapte à la dérive de la voix. Cette dérive, qu'elle se produise lentement, ou rapidement à la suite d'un changement d'état émotionnel du locuteur est un facteur important.

Un système faible mais doté d'un vocabulaire limité (les deux termes sont peut-être contradictoires car la réduction du vocabulaire diminue la redondance dans les mots) serait utile à la commande vocale d'une machine, dans le cas où l'opérateur a les quatre membres occupés, c'est le cas du pilotage spatial ou du tri postal.

Un système capable d'analyser un langage simple (dans l'état actuel des recherches, il est utopique d'envisager la reconnaissance de la parole) peut être un organe de dialogue avec un calculateur pour un centre de renseignements téléphoniques, une banque de données, un système d'enseignement assisté. L'entrée vocale de données numériques ou de programmes me semble peu intéressante, à moins que le langage de programmation soit particulièrement concis (on peut envisager un type mixte d'entrée). Un tel système est capable de compenser, comme je l'ai déjà dit, par des indications syntactiques ou sémantiques les insuffisances d'un système de reconnaissance de phonatomes. Paradoxalement, il est peut-être plus facile à réaliser qu'un système du premier type (voir figure 4). Toutefois la généralisation à un grand nombre d'usagers est un problème crucial.

- FIGURE 4 -

Préambule à l'étude des langages de communication vocale avec calculateur

L'utilisation de la langue naturelle est exclue, parce qu'aux difficultés syntactiques et sémantiques habituelles viennent s'ajouter les difficultés propres de la reconnaissance. Il est d'ailleurs certain comme je l'ai souligné plusieurs fois que ces difficultés ne sont pas indépendantes. Il faut donc employer des langages réduits, qui, par leur syntaxe et par leurs propriétés acoustiques soient assimilables facilement par un calculateur. On peut vraisemblablement se contenter de taux de reconnaissance assez bas, pour que le langage soit suffisamment redondant.

On peut imaginer la présence de mots drapeaux qui, facilement repérables dans la phrase, coïncident avec les articulations syntactiques ou sémantiques.

Deux qualités sont à rechercher : la concision, le naturel. La première est commune à tous les langages de programmation. Elle ne doit pas aller jusqu'à l'élimination de toute redondance. La seconde est souvent oubliée, elle me semble essentielle : les règles de production du langage devraient être assez simples pour qu'on puisse les utiliser "de tête" et que l'on puisse énoncer les expressions d'une manière naturelle - quelques observations ont montré qu'il était moins naturel de hacher le débit de la parole en mots, que de bien articuler .

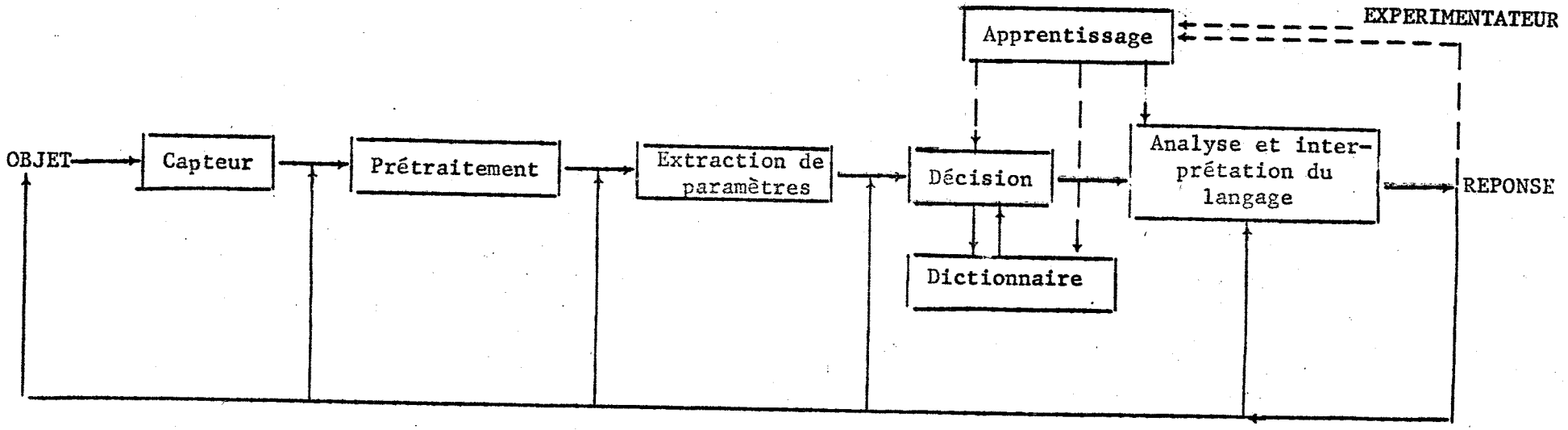


FIGURE 4 : Machine Effective

Il existe un problème d'introduction de mots nouveaux.

Il convient maintenant de rechercher quelle définition formelle donner à l'efficacité d'un langage.

Le contexte réel

Je répéterai ici, ce que j'ai dit plus haut. La reconnaissance des formes n'a d'intérêt que dans un contexte réel, c'est du moins sa justification économique. Comme d'autre part le contexte peut vraisemblablement s'accommoder de performances médiocres, on devrait dès maintenant lui accorder beaucoup d'attention. On a, en effet, la possibilité d'améliorer les performances d'un système d'utilisation, par l'instauration d'un dialogue homme-machine, si simple soit-il. Ainsi à LANNION le système le plus primitif mais aussi le plus rapide à mettre en oeuvre atteint pour un vocabulaire de 15 mots et pour un locuteur 80 à 85 % de reconnaissance correcte. Il permet, malgré ce taux médiocre, d'utiliser dans des conditions acceptables un petit compilateur d'expression arithmétique, grâce à l'utilisation d'un mot d'effacement facilement reconnaissable .

Cela permettrait peut-être un nouveau saut dans les progrès de la reconnaissance vocale (figure 1).

A titre de comparaison du "naturel" des langages de communication verbale, je cite, sans prendre parti, deux exemples qui pourraient être tirés de la machine DESCAL de VICENS et de la machine dite "de bureau" utilisée au CRL (figure 6). Chaque "/" indique un arrêt dans l'énoncé du texte.

A titre d'exemple de système simple, je cite (figure 5) ce qu'est ou ce que sera dans un proche avenir le système de CRL fonctionnant sur le calculateur maison RAMSES 1 L. J'ai expliqué plus haut comment il est tenu compte d'une manière très primitive de la possibilité de dialogue entre l'homme et la machine de manière à améliorer sensiblement la commodité du système.

- FIGURES 5 et 6 -

III. MODALITES D'UNE COLLABORATION ENTRE LES EQUIPES

a) NORMALISATION DU MATERIAU DE TEST

A partir d'un certain degré de développement tous les dispositifs automatiques de reconnaissance devraient être capables de reconnaître les mêmes éléments sonores : par exemple les mots ou certaines parties de mots, de courtes phrases. Nous devrions étudier l'enregistrement d'un certain nombre de bandes que nous pourrions constituer en fond commun. En premier temps, il faudrait établir une liste des éléments sonores, puis fixer les conditions d'enregistrement.

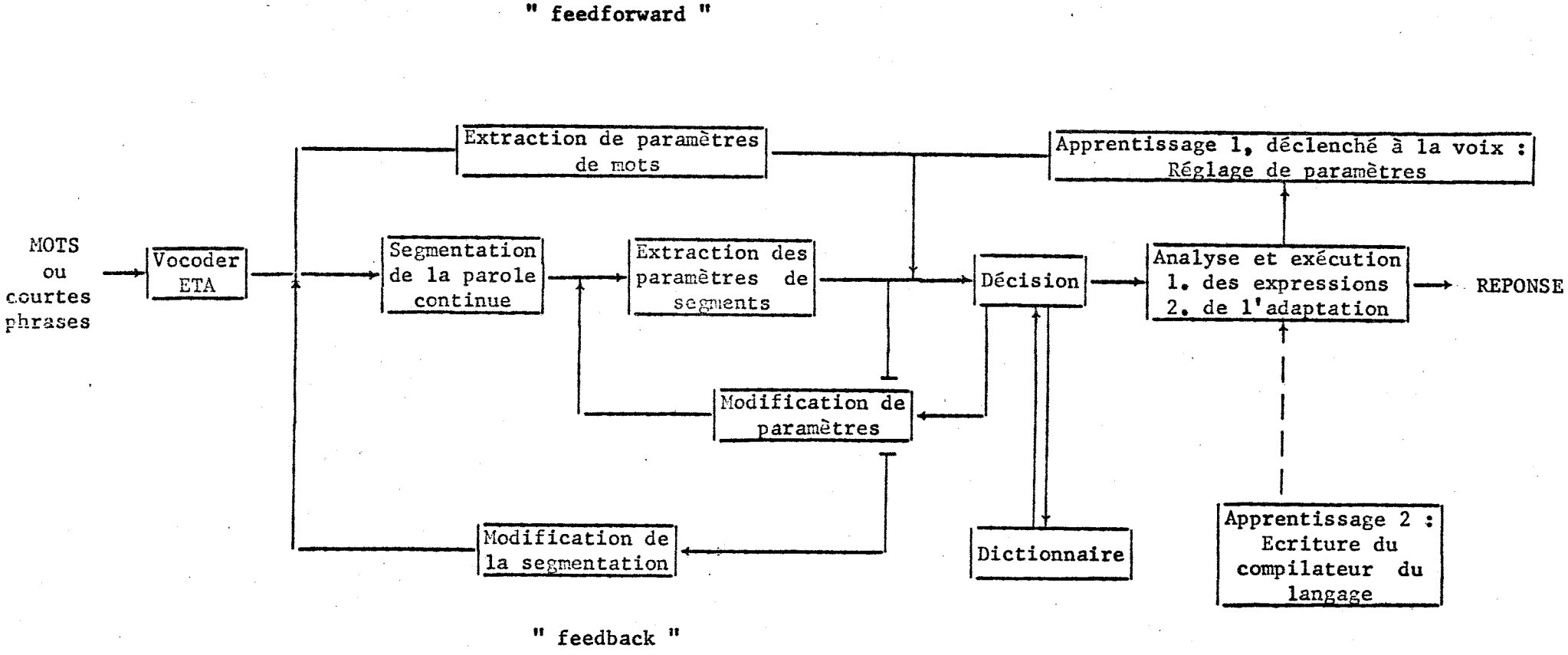


FIGURE 5 : La "Machine de Bureau" du C. N. E. T. (Centre de Recherches de Lannion)

```
PUT  2  INTO  STAR  /          A/1/EGAL/2/EFFECTUEZ/
PUT  5  INTO  ALPHA  /         A/3/EGAL/5/EFFECTUEZ/
MULTIPLY  STAR  BY  STAR  /    A/1/EGAL/A/1/FOIS/A/1/PLUS/A/3/FOIS/A/3/EFFECTUEZ/
MULTIPLY  ALPHA  BY  ALPHA  /
ADD  STAR  TO  ALPHA  /

                               A1 = 2
                               A3 = 5
                               A1 = A1 * A1 + A3 * A3
```

FIGURE 6

VOCABULAIRE DE VISSOTSKY

Ноль	nol	zéro
Один	adin	un
Два	dva	deux
Три	tri	trois
Четыре	tchetiri	quatre
Пять	piat'	cinq
Шесть	chest'	six
Семь	siem	sept
Восемь	vosiem	huit
Девять	dievit'	neuf
Сложить	slajit'	additionner
Умножить	umnojit'	multiplier
Вычесть	vitchest'	soustraire
Разделить	razdilit'	diviser
Синус	sinus	sinus
Логарифм	logarifm	logarithme
Пуск	pusk	départ
Арксинус	arksinus	arcsinus
Печать	pietchat	imprimer
Стоп	stop	stop
Косинус	kosinus	cosinus
Запись	zapis	écrire
Ячейка	yatcheika	mot mémoire
Экспонента	eksponenta	exposant
Целых	tsielic'h	entier
Десятых	diesiatic'h	dixième
Точка	totchka	point
Лента	lenta	bande
Конец	kaniets	fin
Корень	korien	racine
Барабан	baraban	tambour
Слушай	sluchaĭ	allo
Плюс	plius	plus
Минус	minus	moins
Арккосинус	arkkosinus	arccosinus
Тангенс	tangens	tangente
Котангенс	kotangens	cotangente
Ввести	vviesti	introduire
Вывести	viviesti	extraire
Сумма	summa	somme
Градус	gradus	degré

b) RASSEMBLEMENT DE DONNEES CHIFFREES SUR LE FRANCAIS PARLE

Dans l'état actuel pour cinq équipes travaillant à la reconnaissance, il y a quatre types de capteurs différents (cela fera cinq pour six quand M. LIENARD commencera ses études). Il apparaît encore utopique d'utiliser sans adaptation les résultats ou les données d'une équipe à l'autre. Pourtant il semble dommage de répéter des explorations, intéressantes mais fastidieuses.

c) "CREDIBILITE" DES RESULTATS OBTENUS

Souvent le manque de détails dans la présentation des résultats fait que l'on se demande souvent si l'on est en présence d'un succès ou d'un échec. On devrait accorder un soin particulier à la description des conditions d'expériences :

- nombre de mots employés, nombre d'exemplaires de ces mots, mode de présentation (répété ou aléatoire), conditions d'enregistrement (ou de présentation si l'expérience se fait en temps réel) ;

- nombre de locuteurs, hommes, femmes, degré de familiarité du locuteur avec le système ;

- conditions d'apprentissage.

Il est important de préciser le degré d'automatisme de la machine à reconnaître et le temps d'exécution pour un élément.

Il faudrait éviter l'utilisation des mêmes mots pour l'apprentissage et la reconnaissance (cela est facile lorsque l'on travaille en temps réel), la présence exclusive de locuteurs familiers avec le système (homme ayant tendance à s'adapter plus vite à la machine), de faire des statistiques sur des mots prononcés une dizaine de fois chacun.

Chaque système "intégré" de reconnaissance devrait fournir un manuel d'utilisation.

Le texte de la conférence s'arrête là. J'ai l'impression en me relisant de n'avoir parlé de reconnaissance vocale qu'à mots couverts. Cela tient sans doute à l'ambiguité des échecs des études sur la reconnaissance. Le mirage du PERCEPTRON a longtemps égaré les chercheurs. Il m'a paru important de signaler que l'on ne pouvait plus se contenter d'une conception simpliste de la reconnaissance des formes. Il m'a semblé moins important de faire un choix sur les différentes conceptions de la parole. Pour le moment toutes les voies nous sont ouvertes. L'efficacité étant le seul critère.

La reconnaissance en temps réel d'un vocabulaire limité, ou d'un langage réduit est une chose acquise, il ne tient qu'à nous d'étudier si elle doit rester un "gadget" de laboratoire, en attendant la généralisation des systèmes, ou si elle peut déjà trouver des applications.

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

Cette bibliographie n'est pas exhaustive. Elle est très limitée dans le temps (je n'ai pas voulu remonter jusqu'à l'article de DAVID, sur la reconnaissance automatique des chiffres parlés, paru en 195[illegible] dans le JASA). Le congrès de Stockholm en 1962 m'a semblé un point de départ suffisamment éloigné.

Il n'y a aucune référence sur la reconnaissance des formes au sens où on l'entend habituellement. On peut se reporter à l'article de G. NAGY paru en Mai 1968 dans les "IEEE Proceedings" - "State of the Art in Pattern Recognition".

OUVRAGES DE BASE :

FLANAGAN J.L.	(1965)	Speech Analysis, Synthesis and Perception - Springer Verlag
MALMBERG B.	(1963)	Structural linguistics and Human communication - Springer Verlag
JACOBSON R., HALLE M., FANT G.	(1961)	Preliminaries to speech analysis. The distinctive features and their correlates (réédité au MIT Press, récemment).

REVUES DE BASES :

JASA , les références bibliographiques sont sous la rubrique 9 - 10
IEEE , Transactions on Electroacoustics

La Revue de GALF (notamment le numéro 3 - 4 de 1968 consacré au colloque sur les structures acoustiques de la Parole, Grenoble Avril 1967).

Les articles cités sont tirés de :

(A) - 5e Congrès International d'Acoustique Liège, 7 - 4 septembre 1965

(B) - 6e Congrès International d'Acoustique Tokyo, 21 - 28 Août 1968

(C) - Speech Communication Seminar Stockholm, 29 août - 1er septembre 1962

(D) - 1967 Conference on Speech Communication and Processing, 6 - 8 novembre 1967, Cambridge (Mass.)

(E) - JASA

(F) - IEEE - Trans. on Audio and Electroacoustics.

Classement alphabétique par auteurs

1. - A.S. ABRAMSON — Voice onset in stop consonants - Acoustic Analysis & synthesis (A).

2. - R. ALTER — Utilization of Contextual Constraints in Automatic Speech Recognition (D).

3. - ANON, C.L. CLAPPER — New Pattern Recognizer can identify both spoken and written inputs. Electron, 40, 91-102 (oct. 1967)

4a.- D.G. BOBROW, D.H. KLATT — "A limited speech Recognition System" F.J.C.C. Proceedings (1968).

4b.- P.W. BECKER — Reduction of data flow using Pattern recognition Colloque International sur la Téléinformatique (Paris 1969).

5.- C. SACERDOTE, G. SACERDOTE — Statistical properties of individual voices (B).

6.- M.W. CANNON — A Method of Analysis and Recognition for Voiced Vowels (D).

7.- S. CHIBA — Spoken word recognition by multiple linear separation (B).

8.- L.A. CHISTOVICH — Direction of transition as a perceptual parameter of time - varying stimuli (B).

9a.- J.E. DAMANN — Dispersion Analysis as applied to Speech Spectra (A).

9b.- P.C. DELATTRE, A.M. LIBERMAN, F.S. COOPER — Acoustic Loci and Transitional Cues for Consonants JASA, (1955).

10.- P.B. DENES, T.G. von KELLER — Articulatory segmentation for automatic recognition of speech (B).

11a.- J.A. DREYFUS GRAF — Phonétographe et Cybernétique (A) Spectre phonétographiques, Revue d'Acoustique n° 3 et 4.

11b.- J.W. FALTER, K.W. OTTEN — Cybernetics in Speech Communication - Survey of Russian Litterature (F) (mars 1967).

12.- G. FANT et al. — Formant Amplitude Measurements (C).

13.- G. FANT — Current topics of speech research (B)

14.- H. FUJISAKI, T. KAWASHIMA — The Roles of Pitch and Higher Formants in the Perception of Vowels (D).

15a.- L. GILLI, A.R. MEO — Sequential System for recognizing Spoken digits in Real Time. Acustica (1969)

16.- A. GUZMAN — Scene Analysis by computer-Ph. D. Thesis MIT (1969).

17.- F.J. HILL et al. — Speech Recognition as a Function of Channel Capacity in a Discrete Set of Channels (1968) (E).

18.- K. HIRAMATSU, T. MATSUNO — On moments as parameters for speech recognition (A).

18a.- K. HIRAMATSU, R.K. WACKERBARTH, C.L. COATES — Classification by the Distinctive Features (D).

19a.- N. HIRATO, O. KAKUSHO, K. KATO, T. KOBAYASHI — Harmonic components and vowel perception (B).

19b.- W.J. HURD — Use of sample moments to detect speech in noise (D).

20.- Y. KADOKAWA, J. SUZUKI — A simple calculation of the vocal tract configuration from three formant frequencies (B).

21.- Y. KANAMORI, S. HIKI, H. KAWAI, J. OIZUMI — On the formant information in connected speech (B).

22.- D.D. KALIC — On the relation between fondamental frequency of the vowel and its first three formants (A).

23.- H. KASUYA, H. SUZUKI, K. KIDO — On auditory space model of vowel perception (B).

24 - K. KIDO, H. KASOYA, H. SUZUKI — Discrimination of Japanese vowels in connected speech (B).

25.- A. KUREMATSU, S. INOUE — Speech recognition with time-normalized frequency pattern (B).

26.- A. LANDERCY, M. WAJSKOP — La segmentation temporelle : forme de prélèvement et perturbation spectrale (B).

27.- M. LECOURS, J.J. SPARKES — Adaptative Spectral Analysis for Speech Sound Recognition (D).

28.- J. LILJENCRANTS — A few experiments on voiced voiceless identification and time segmentation of speech (C).

29.- J. MEINHARDT — Beitrag zur Frage der redondanzansmetzung bei der AutomatischenSpracherkennung
Technische Hochschule wissenschaft Z. 13, n° 2, pp. 135 - 138 (1967).

30.- K.M. MENON et al. — Acoustic Properties of certain VCC utterances, vol. 46 ≠ 2 (part 2). 1969 (E).

31.- P. MERMELSTEIN — Determination of smoothed cross sectional area functions of the vocal tract from formant frequencies (A).

32.- E.N. MYASNIKOVA — Objective recognition of speech Sound, pp. 131 Joint Publications Res. Ser., Wash. DC, (1968) JPRS - 43926.

33.- M. NAKATSUI, J. SUZUKI — Fast formant frequency tracking technique (B).

34.- Texas Univ. (Austin) Electronics Research Center, C. OHLENDORF, C. L. COATES — Recognition of spoken digits utilizing sequential patterns. Technical Rept. Novembre 1968 (83 b) Rept. n° TR-57 AFOSR-68 - 24-68

35.- J. OISUNI, S. HIKI, H. SATO, T. IGARASHI — Dynamic model of vowel perception (B).

36.- A.V. OPPENHEIM, R.W. SCHAFER — Homomorphic Analysis of Speech (D).

37.- A.V. OPPENHEIM — Speech analysis -synthesis system based on homomorphic Filtering (E).

38.- G.E. PETERSON — Research on speech communication Automatic speech regognition
PP 21. Univ. Michigan Ann Arbor (janv. 1966).
AD - 478 - 122

39.- E.N. PINSON — Computing vocal tract shapes to yield specific tract transfer functions (A) .

40.- L.C. POLS et al. — Perceptual and Physical Space of vowel Sounds Vol. 46 ≠ (part 2) - 1969, (E)

41.- G.S. RAMISHWILI — Ob automatisaticheskom Uznavanii Golosov Akad. Nauk. Izv. Tekniksk. Kibern. N° 5, 87-92 (1966), Automatic Voice Recognition Engineering Cybernetics, n° 5, 84-49 (1966).

42.- P.R. REDDY, R.B. NEELY — Contextual analysis of phonemes of english Stanford Univer. Calif. Dept of Computer Science Janv. 69, AI. Memo - 79.

43a.- D.R. REDDY — Segmentation of speech sounds. JASA, 40, 307, 312 (1966)
Phoneme grouping for speech recognition JASA, 41, 1295 (1966).

43b.- D.R. REDDY, P. VICENS — A procedure for segmentation of connected speech J. Eng. Soc. 16 : 4 (1968)

44.- D.R. REDDY, A.E. ROBINSON — Phoneme-to-Grapheme Translation of English (D).

45.- P.W. ROSS — A limited Vocabulary Adaptative Speech Recognition System. J. Audio Eng. Soc. 15, n° 4, 414-419 (1967).

46.- P.W. ROSS, E.S. ROGERS — The application of Pattern Recognition Techniques to the Evaluation of Speech Recognition Systems (B).

47a.- T. SAKAI, S. DOSHITA — The automatic speech Recognition System for conversational sound (E).

47b.- T. SAKAI, S. DOSHITA, K. NAGATA, T. SEKIHOTO — Phonetic Typewriter (C).

48.- R.W.A. SCARR — Zero Crossings as a Means of Obtaining Spectral Information in Speech Analysis (D).

49.- J.N. SHEARME — Analysis of the performance of an automatic formant measuring system.

50.- S. SINGH — Distinctive feature : a context-free element (B).

51.- M.M. SONDHI — New Methods of Pitch Extraction (D).

52.- M.R. SCHROEDER — Recent Studies in speech Research at Bell Telephone Laboratoires (A).

53.- A.M. NOLL — Recent Studies in speech Research at Bell Telephone Laboratoires (A).

54.- M.R. SCHROEDER — Period Histogram and Product Spectrum : New Methods for Fondamental Frequency Measurement vol. 43, ≠ 4, 1968 (E)

55.- M.R. SCHROEDER — Similarity measure for Automatic Speech and Speaker Recognition Vol. 43,≠ 2, 1968. (E)

56.- K.N. STEVENS — Acoustics Correlates of Certain Consonantal Features (D).

57.- K.N. STEVENS et al. — Study of Acoustic Properties of speech Sounds BBN Techn. Rept 1968.

58.- H. SUZUKI, H. KASUYA, K. KIDO — The Acoustic Parameters for Vowel Recognition without distinction of speakers

59.- T. SUGIMOTO, S. HASHIMOTO — The voice fondamental pitch and formant tracking computer program by short term autocorrelation functions (C).

60.- V.N. SOROKIN — Speech Recognition using Pattern Analysis Eng. Cybernetics, n° 5, 90-94 (1966).

61.- S. SUZUKI, S. ITAHASHI, K. KIDO — The Effectiveness of Utilization of Word Lexicon in Recognition of Japanese Spoken Language (D).

62.- T. TAKEDA, Y. SATO, K. NAGATA — Measurements of vocal tract transfer Response (A).

63.- Y. TAKEFUTA — Automatic detection of intonational signals in American English.

64.- C.F. TEACHER, H.G. KELLET, L.R. FOCHT — Experimental; limited vocabulary, speech Recogniger IEEE. Trans. on Electro A. 15, n° 3, 127-130 (1967).

65a.- H.G. TILLMANN, G. HEIKE, H. SCHNELLE, G. UNGEHEUER — Dawid-I Ein Beitrag zur automatischer "spracherkennung" (A).

65b.- J.P. TUBACH — Reconnaissance des chiffres parlées -Colloque International sur la Téléinformatique, PARIS 1969.

66.- P. VICENS — Aspects of Speech Recognition by computer, Stanford U. Memo AI-85 (Ph. D. Thesis), avril 1969.

67.- G.R. VISSOTSKY et al. — АЛГОРІТМ ОПОЗНАВАНИЯ 40 СЛОВ НА ЦВМ БЕСМ-3М ВЦ АН СССР (Работы по технической кибернетике)
(algorithme de reconnaissance d'un vocabulaire de 40 mots sur BESM - 3M - Centre de Calcul de l'Académie des Sciences d'URSS - MOSCOU).

68.- S. WASHIZA
W.A. OKAMOTO
T. SHIGA
S. INOMATA — A spoken digit recognition scheme and its computer simulation (B).

69.- H. YILMAZ — A program of research directed toward the efficient and accurate machine recognition of human speech perception, final report ; Nov. 1967 N 68 - 16749.

70.- ZWICKER
HESS
TERHARDT — Recognition of spoken Numbers with functional Model and Electronic Computer. Kybernetik -3 n° 6

71.- ZADEH — Fuzzy sets and Fuzzy algorithms (différents articles parus depuis 1967 dans la revue INFORMATION AND CONTROL).

D I S C U S S I O N

1°) Travaux à l'étranger :

CASTAN fait remarquer que le professeur SAKAÏ (Kyoto) travaille en temps réel. La NEC possède un appareil pour la reconnaissance des dix chiffres et un système fonctionnant à l'aide de magnétophone.

2°) A propos de la notion de paramètre

PERRENNOU pense qu'il faut avoir une définition simple pour être acceptée par tous mais qu'il est difficile d'en dire plus.

3°) Utilisation de la reconnaissance

BURON donne des précisions sur le contrat de RCA, application très réelle mais particulière, avec l'administration des postes américaines. Le but est la reconnaissance dans une ambiance bruyante de 3 mots à la seconde (chiffres et quelques lettres). C'est un contrat renouvelable tous les ans, qui dure depuis 2 ou 3 ans.

RISSET : certains, c'est le cas des chercheurs des BELL Telephone Laboratories, pensent que le problème de la reconnaissance automatique de la parole est insoluble sans recours aux indices syntaxiques et sémantiques, et l'on ne sait faire tenir compte de ces indices par un programme de reconnaissance que d'une façon très rudimentaire. Les BELL Telephone Laboratories ont pratiquement abandonné les études de reconnaissance automatique de parole après une longue exploration, en dépit du parti que le Bell System pourrait tirer commercialement d'un système de reconnaissance. Sans ces indices, on ne peut, comme il a été indiqué dans l'exposé, espérer un taux parfait dans des conditions suffisamment larges, aussi est-il effectivement intéressant d'étudier une reconnaissance imparfaite dans un contexte d'utilisation pour voir si cette reconnaissance imparfaite peut être utilisée. NEELY a incorporé un système de réponse vocale au système de REEDY et VICENS à STANFORD ; cela vise à placer la reconnaissance automatique dans un contexte de communication vocale homme-machine.

M. DREYFUS-GRAF a résumé dans ce qui suit l'ensemble de ses interventions.

En 1952, nous avions proposé le néologisme de "phonétographe" pour désigner l'ensemble des machines transformant des langages écrits phonétiquement.

Au cours des développements consécutifs, nous avons admis l'importance primordiale des compresseurs d'amplitude sélectifs : ceux-ci fournissent des spectres logarithmiques auto-régulés par leurs compositions spectrales et ils réduisent préalablement les quantités d'information à traiter.

En conséquence, depuis 1961, et selon nos publications, nous désignons par "phonétographiques" tous les systèmes de reconnaissance automatique de la parole qui sont basés sur de tels compresseurs.

Mais, au stade actuel, il conviendrait de distinguer deux d'appareils qui diffèrent par leurs buts et par leurs complexités.

a) L'actuateur phonétique, ou phonacteur, transformant en actions mécaniques (2) un nombre limité de mots entiers, tels que des chiffres ou des ordres, et qui pourrait, par exemple, "phonactionner" l'entrée d'un calculateur ou d'un tri postal. (En anglais, on pourrait dire : phonetic actuator, phonactor, to phonactuate).

b) Le phonétographe proprement dit qui doit transformer en textes lisibles le nombre pratiquement illimité des mots d'un discours ou d'une dictée. A son stade le plus perfectionné, il doit aboutir au phonémographe, c'est-à-dire qu'il reconnaîtrait, à l'intérieur de chaque mot, chacun des quelques 32 phonèmes (voyelles ou consonnes) qui suffisent à reconstituer une langue complète alphabétiquement.

Le champ d'application pratique d'un dispositif commandé par la parole est d'autant plus vaste qu'il peut utiliser un plus grand nombre de réseaux existant de télécommunications.

Pour préparer le développement d'un Phonacteur I, nous nous sommes donc placés d'emblée dans les conditions réelles d'une transmission par ligne téléphonique locale, limitant la bande passante de 200 à 3600 Hertz environ, et le microphone étant disposé dans un milieu moyennement bruyant (de l'ordre de 60 décibels).

Nous avons commencé les essais avec un microphone ordinaire à charbon. Mais, par la suite, nous avons préféré un microphone du type magnétique (déjà utilisé pour l'écouteur normal) qui est plus stable et linéaire. Il est incorporé dans un nouveau poste téléphonique expérimental (à semi-conducteurs) type S 63 qui est compatible avec les anciens.

Le pourcentage de réponses justes a été de 96,6 % avec 30 séries à 18 mots, prononcés par une personne, c'est-à-dire avec 540 mots prononcés dans des ordres variés, à travers la ligne téléphonique et dans un bruit ambiant de l'ordre de 60 décibels.

L'optimalisation vise à obtenir le maximum d'information, donc de précision, avec le minimum d'action, c'est-à-dire d'énergie et de temps. Elle dépendra d'abord du but recherché, c'est-à-dire de l'étendue du vocabulaire, mais aussi des niveaux de bruit, ainsi que des apprentissages réciproques admissibles entre l'usager et la machine.

Les paramètres à optimaliser se répartissent parmi les compresseurs sélectifs, les classes de phonèmes et les programmes logiques.

Les calculateurs en fournissent maintenant les statistiques de base.

D'une manière générale, les études décrites ci-dessus rentrent dans le cadre d'extracteurs d'information vocale, incluant l'intonation, l'accent, le chant, le chuchotement, l'identification de personnes etc.

Dans le cas du Phonacteur I, dont le vocabulaire est limité à quelques dizaines ou centaines de mots, on cherchera à éviter le réglage de la machine selon chaque voix individuelle. Par contre, si le vocabulaire est très étendu, voir illimité, comme dans le cas du phonétographe proprement dit, on admettra l'adaptation rapide de la machine à l'usager pour minimaliser le taux d'erreur.

Cependant, les erreurs sont d'autant moins tolérables que les vocabulaires sont plus limités, et il faut les éliminer par tests répétitifs utilisant les redondances.

Nous avons relevé les prédictions suivantes dans la Revue 01 - Informatique, déc. 1968, basées sur des sondages d'opinion : "les entrées verbales deviendront possibles, dans la pratique, en 1979. A ce moment, la carte perforée et la bande perforée auront vécu en tant que supports de communication".

Sans être aussi dilatoire et radical tout à la fois, on peut espérer que, dans un proche avenir, les réseaux de télécommunications permettront de métamorphoser la parole humaine en actions précises.