Aujourd'hui, l'ordinateur donne de la voix!

Il savait tout faire : calculer, dessiner, composer, bref, simuler toutes les activités de l'intelligence humaine. Il lui manquait juste la parole. Mais parce qu'on a su synthétiser la voix humaine, l'ordinateur sait aujourd'hui simuler le langage des hommes.

D'abord, ils apprirent à calculer, à décider puis à écrire. Quelque temps plus tard, ils furent capables de lire, puis de dessiner. Aujourd'hui, ils parlent. Et bien. Sans un mot plus haut que l'autre, d'une voix calme et impersonnelle, une voix d'ailleurs, du monde des machines.

Les ordinateurs, puisque c'est d'eux qu'il s'agit, se sont en effet mis à parler. Pas tout seuls bien sûr, mais parce que l'homme l'a voulu et a créé pour cela un périphérique de plus, véritable organe vocal, du nom barbare de synthétiseur de parole. Ces instruments forment une entité en eux-mêmes, ils peuvent fonctionner seuls, mais sans la puisance de mémoire d'ordinateurs derrière eux les synthétiseurs de parole seraient restés des gadgets de laboratoire, sans aucune possibilité de pouvoir pénétrer un jour dans la vie courante. Aujourd'hui, alliés aux ordinateurs les synthétiseurs prennent la parole dans les banques, les centres de renseignement, les centres de calcul, etc.

On distingue deux types de périphériques à réponse vocale, les uns dits analogiques sont dédaignés des chercheurs car trop proches des magnétophones et au vocabulaire trop limité. Les autres dits digitaux semblent promis à un bel avenir.

Les synthétiseurs analogiques portent d'ailleurs très mal leur nom, car ils ne synthétisent rien du tout. Les mots et les phrases sont enregistrés par un speaker — on dit un locuteur — sur un tambour magnétique. L'ordinateur se contente de retrouver l'adresse d'un mot lorsqu'il a besoin de lire, de composer les phrases en indiquant l'ordre dans lequel les mots doivent être émis en intercalant des temps de silence.

Deux systèmes de réponses vocales fonctionnent sur ce principe : l'unité 7770 d'IBM et l'Audio-Reponse-System mis au point par Burroughs. La qualité de la voix émise est excellente, aussi bonne que celle donnée par un magnétophone. Mais le grand inconvénient de ce genre d'unité à réponse vocale est la limitation du vocabulaire. On ne peut enregistrer que 128 mots sur l'IBM-7770 et 189 sur le Burroughs Audio-System.

Pour des raisons de volume et de coût, impossible de dépasser ce nombre, ce qui limite évidemment le nombre des applications. Impossible par exemple d'utiliser ce genre d'appareils dans un magasin de vente par correspondance où le nombre d'articles se chiffre par milliers.

Deux chercheurs de l'Université de Toulouse, A. Bruel et J.C. Cazaux, ont donc songé, pour éviter cet inconvénient, à employer des mémoires holographiques. Elles permettent d'enregistrer sous de très petites dimensions des mots ou des syllabes. Ces dernières sont enregistrées photographiquement sous forme d'une modulation d'amplitude comme sur les pistes sonores des films cinématographiques.

A partir de chaque photographie d'une syllabe, on réalise une plaque de microhologramme de la taille d'une tête d'épingle. Pour obtenir de nouveau la voix, il suffira de

lire les plaques constituant un mot, à l'aide d'un faisceau laser. Cette étude est loin d'être terminée et un problème majeur reste à résoudre : celui du codage son-image avant l'enregistrement des hologrammes. Le choix de la piste sonore de film cinématographique ne satisfait pas les chercheurs de l'Université de Toulouse.

Les synthétiseurs analogiques actuels ne semblent pas constituer la solution d'avenir, l'étroitesse de leur vocabulaire en est sans doute la cause. Mais d'ores et déjà deux installations ont été réalisées à Paris. Qui n'a pas formé sur son cadran, parce qu'il était en retard, INF 84.00, et qui n'a pas entendu une voix impersonnelle déclarer : qu'« au quatrième top il serait exactement 8 heures 43 minutes et 15 secondes ». L'horloge, pilotée par un ordinateur fut probablement la première machine française à parler.

Une banque, la B.R.E.D. (Banque Régionale d'Escompte et de Dépôts) possède depuis trois ans une unité à réponse vocale analogique, l'IBM-7770, qui permet à chacune de ses agences, à chacun de ses guichets, d'entrer en liaison directe avec l'ordinateur central pour obtenir par exemple la situation du compte d'un client.

L'employé envoie les éléments de sa question par le clavier même de son poste téléphonique à touches et il reçoit la réponse vocale de l'ordinateur, un IBM-370/145, par son écouteur. Malgré ces applications spectaculaires et rentables les recherches sur les synthétiseurs analogiques en sont restées là.

Et tous les espoirs de ceux qui travaillent sur les unités vocales se tournent vers les synthétiseurs numériques qui créent vraiment la parole à partir de chiffres binaires contenus dans la mémoire des calculateurs.

Mais avant de pouvoir réaliser cette unité à réponse vocale, il a fallu analyser, et de près, la parole, afin de pouvoir la coder.

Lorsque nous parlons, l'air venant de nos poumons passe par le larynx. Là se trouvent nos cordes vocales. Elles excitent l'air de notre pharynx, de notre bouche, de notre nez ; ce sont nos cavités résonnantes qui par un jeu de muscles peuvent changer continuellement de forme. En se modifiant, elles favorisent la création de fréquences qui sont des multiples (on dit aussi des harmoniques), de la fréquence de vibration des cordes vocales, mais multiples qui évoluent constamment en intensité, en nombre et en rang.

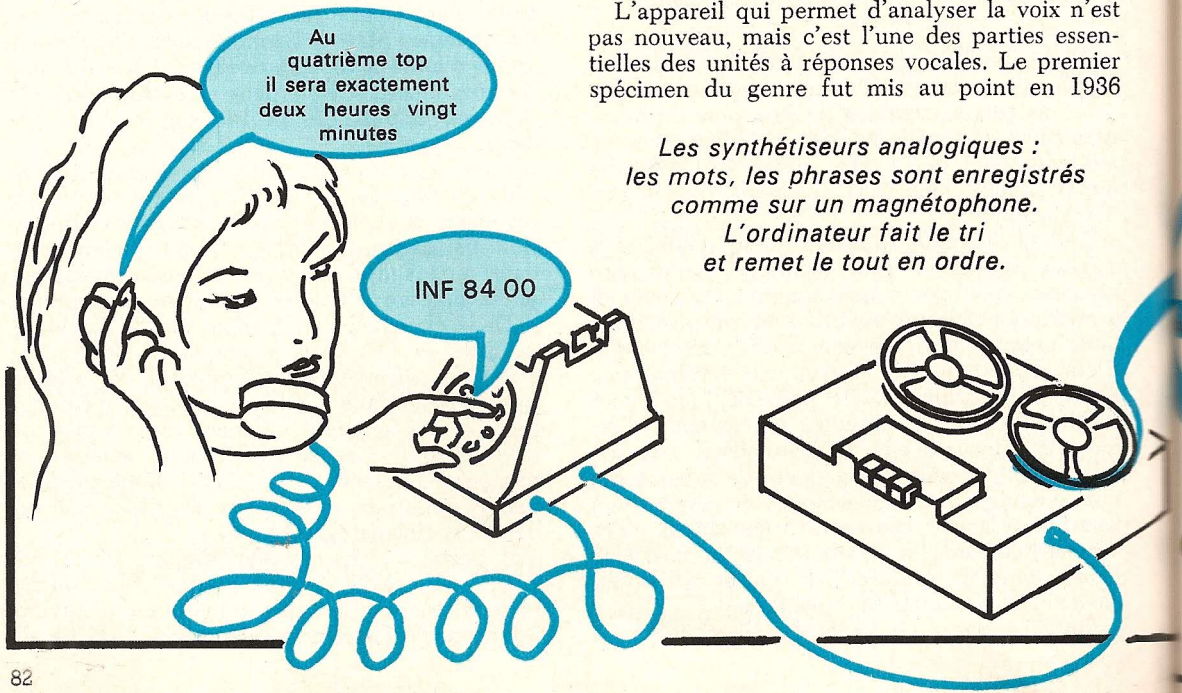
C'est ainsi que notre bouche émet les voyelles et les consonnes sonores. Pour telle position des cavités résonnantes, tels harmoniques de la fréquence de vibration des cordes vocales sont favorisées et nous entendons la voyelle « a », pour d'autres nous prononçons la voyelle « e ». Pour certaines voyelles dites « sourdes » et certaines consonnes, les cordes vocales ne vibrent pas du tout et les sons sont uniquement façonnés par les cavités résonnantes.

La voix humaine est donc une suite de silences, de parties mélodiques dues à la vibration des cordes vocales et de bruits sourds. Scientifiquement, quantitativement la voix ou signal sonore est donc caractérisée par deux facteurs : d'une part la fréquence de vibration des cordes vocales qui varie de 300 à 3 400 hertz et que l'on appelle aussi mélodie de la voix, et d'autre part, l'intensité des vibrations.

Pour analyser un son de la voix humaine, il faut donc pouvoir répondre aux quatre questions suivantes :

- Les cordes vocales ont-elles vibré ?
- Si oui, à quelle fréquence ?
- Quels sont les harmoniques en rang et en nombre ?
- Quelle est l'intensité de la vibration ?

L'appareil qui permet d'analyser la voix n'est pas nouveau, mais c'est l'une des parties essentielles des unités à réponses vocales. Le premier spécimen du genre fut mis au point en 1936



Les synthétiseurs analogiques : les mots, les phrases sont enregistrés comme sur un magnétophone. L'ordinateur fait le tri et remet le tout en ordre.

aux U.S.A. par H. Dudley un ingénieur des Bell Laboratories. Son nom : le vocoder, contraction de « voice coder » et devenu un France vocodeur. Aujourd'hui, la plupart des vocodeurs (dits canaux) sont identiques dans leur principe à celui mis au point par Dudley.

Ainsi le C.N.E.T. (Centre National d'Etudes des Télécommunications) à Lannion, haut lieu de la recherche sur la synthèse et la reconnaissance de la parole, a conçu en 1968 un vocodeur à canaux. Il se compose :

- d'un détecteur de pitch ou détecteur de mélodie, nom savant donné au capteur placé sur la gorge d'un locuteur et qui permet de déterminer la fréquence de vibration des cordes vocales ;
- de 12 filtres très sélectifs de bande passante allant de 200 à 4 000 hertz. A la sortie de ces filtres on mesure l'énergie donc l'intensité de la vibration.

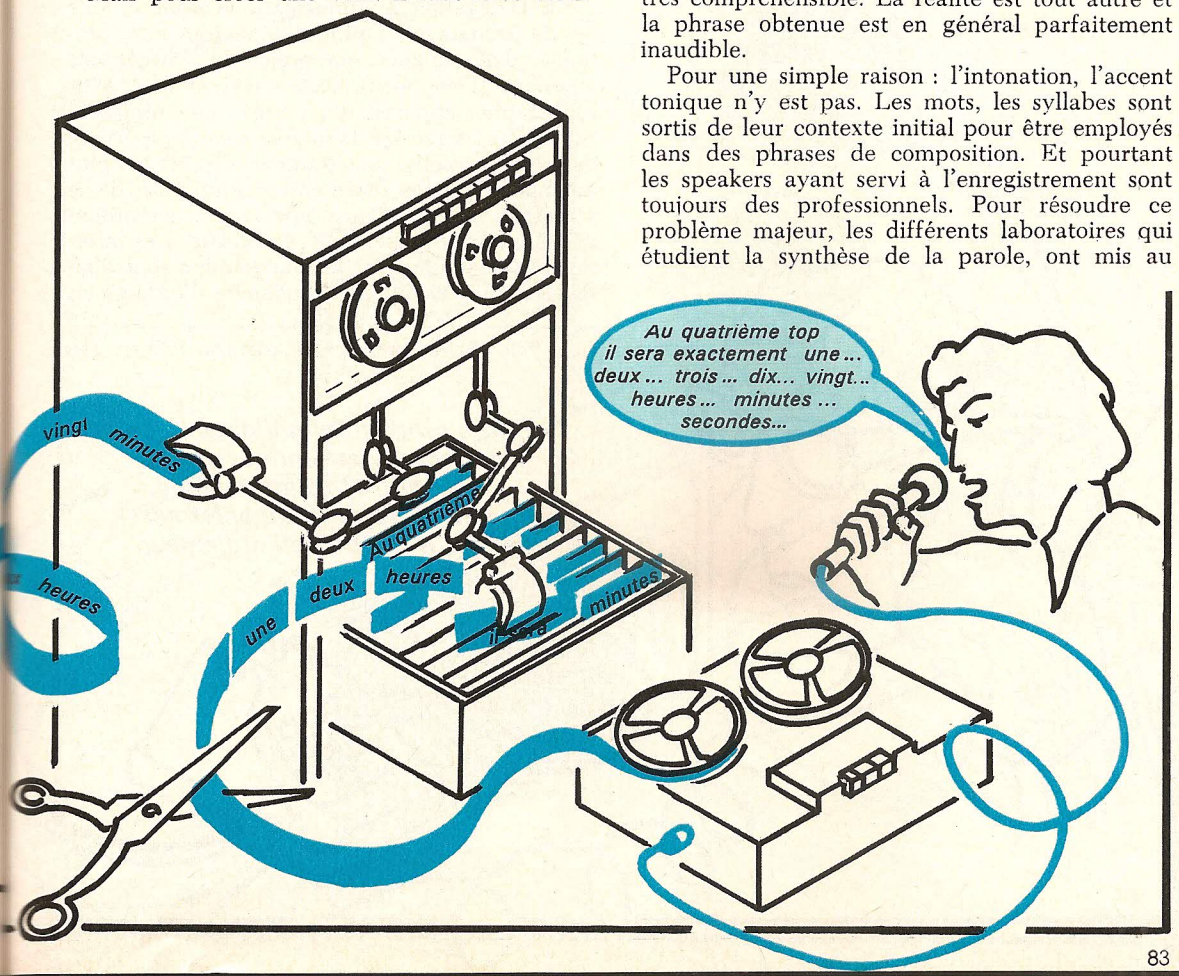
Ces deux informations, fréquence de vibration et énergie, sont transformées en langage binaire et transmises vers l'ordinateur, qui conservera dans sa mémoire les images numériques des différents sons ou des différents mots. Pour une seconde de langage parlé, il est nécessaire de mesurer entre cinquante et cent fois par seconde la valeur de ces deux grandeurs.

Mais pour créer une voix, il faut faire sortir ces images numériques et les recomposer, donc effectuer le processus inverse de celui de l'analyse. Un synthétiseur de parole à canaux comprend donc un jeu de filtres dont le gain sera commandé par la valeur des énergies enregistrées en ordinateur. Parallèlement, on envoie un signal périodique ou simplement bruyant selon la nature du son que l'on veut émettre. La somme de ces signaux est alors dirigée vers un haut-parleur. C'est ainsi que l'on fait parler les machines. Il y a plusieurs manières de synthétiser la parole. On peut analyser, enregistrer dans la mémoire de l'ordinateur des mots entiers, ou simplement des sortes de syllabes que l'on appelle des diphonèmes.

Dans le premier cas, on synthétise les phrases à l'aide de mots, dans le second on crée d'abord les mots avec lesquels on construira des phrases. Prenons un exemple : « papa va au bureau ». Lors d'une synthèse par mot, l'ordinateur retrouve les quatre mots et les ordonne afin de construire la phrase.

Lors d'une synthèse par syllabe, la machine cherche les syllabes, pa, va, au, bu, reau, puis elle recomposera les mots avant de former la phrase. En théorie donc, il suffit que l'ordinateur place bout à bout les mots ou les diphonèmes pour obtenir à la sortie une expression très compréhensible. La réalité est tout autre et la phrase obtenue est en général parfaitement inaudible.

Pour une simple raison : l'intonation, l'accent tonique n'y est pas. Les mots, les syllabes sont sortis de leur contexte initial pour être employés dans des phrases de composition. Et pourtant les speakers ayant servi à l'enregistrement sont toujours des professionnels. Pour résoudre ce problème majeur, les différents laboratoires qui étudient la synthèse de la parole, ont mis au



point de très importants programmes d'ordinateurs qui modifient instantanément les intonations des mots en fonction de leur place au sein d'une phrase et les accents toniques en fonction de la situation du diphonème dans le mot.

Actuellement en France, la plupart des applications emploient des synthétiseurs à base de vocodeurs à canaux.

Le Département d'études acoustiques du C.N.E.T. à Lannion a mis au point un système de réponse vocale à mot permettant à chaque abonné au téléphone de connaître le coût de sa dernière communication, le montant de son compteur téléphonique, ou le numéro d'appel des communications aboutissant à des numéros désaffectés. Cette application est opérationnelle sur le réseau expérimental « Platon ».

A Issy-les-Moulineaux, le Service Informatique du C.N.E.T. a réalisé un service bureau par téléphone, grâce auquel chaque abonné pourrait avoir la puissance d'un ordinateur au bout de son téléphone. Le cœur de ce système est un synthétiseur à canaux qui fournit à l'abonné les résultats des travaux demandés aux calculateurs (1). Il fonctionne actuellement sur le réseau interne du C.N.E.T. Il suffirait du feu vert du ministère des P.T.T. pour que ces deux expériences s'étendent au niveau national.

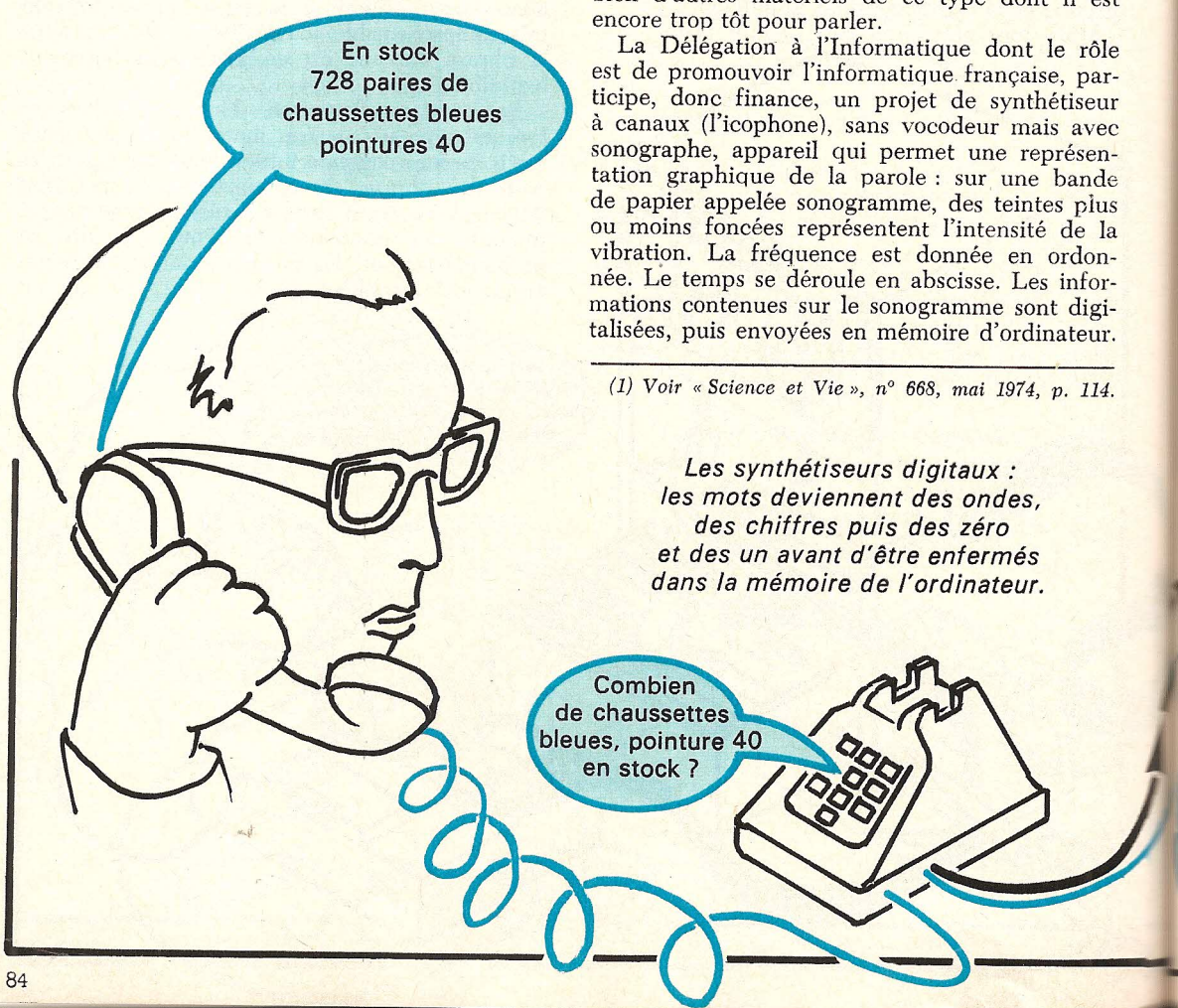
La C.I.T. (Compagnie Industrielle des Télécommunications) a mis au point, pour le compte du service technique de la navigation aérienne, le système DECLAM (Dispositif Automatique d'Emission en Clair de l'Assistance Météorologique en vol) qui transmet aux avions par l'intermédiaire d'un synthétiseur à canaux, des informations météorologiques.

La Société IBM possède à La Gaude un centre de recherche qui est responsable, pour le monde entier (excepté les Etats-Unis) des études sur la synthèse de la parole. C'est lui qui a développé voici cinq ans l'IBM-7772, l'un des seuls synthétiseurs de paroles commercialisés en France, et l'un des meilleurs aussi. Et pourtant, il n'existe aucune installation. Sans doute le marché ne s'est-il pas jusqu'à présent révélé suffisamment rentable, car depuis le lancement de l'IBM-7772 ; la société se montre très discrète sur ses activités dans ce domaine.

On sait par exemple que le laboratoire de La Gaude met au point un synthétiseur sans vocodeur, donc sans analyse de la voix et sans enregistrement humain. C'est l'ordinateur lui-même qui crée la représentation codée de la voix à partir d'une représentation phonétique des mots. Mais peut-être IBM a-t-elle dans ses cartons bien d'autres matériels de ce type dont il est encore trop tôt pour parler.

La Délégation à l'Informatique dont le rôle est de promouvoir l'informatique française, participe, donc finance, un projet de synthétiseur à canaux (l'icophone), sans vocodeur mais avec sonographe, appareil qui permet une représentation graphique de la parole : sur une bande de papier appelée sonogramme, des teintes plus ou moins foncées représentent l'intensité de la vibration. La fréquence est donnée en ordonnée. Le temps se déroule en abscisse. Les informations contenues sur le sonogramme sont digitalisées, puis envoyées en mémoire d'ordinateur.

(1) *Voir « Science et Vie », n° 668, mai 1974, p. 114.*



Les synthétiseurs digitaux : les mots deviennent des ondes, des chiffres puis des zéro et des un avant d'être enfermés dans la mémoire de l'ordinateur.

Dessins Jean-Marie Carron

L'icophone les restitue. C'est le Laboratoire de Mécanique-Physique de l'Université de Paris VI qui étudie et développe ce type d'unité à réponse vocale.

Les vocodeurs à canaux basés sur le repérage de l'énergie et de la fréquence d'une vibration ne sont pas les seuls sur le marché. Certains laboratoires préfèrent caractériser la voix par la fréquence et l'amplitude des maximums d'intensité des vibrations que l'on appelle les formants (voir schéma). Le synthétiseur à formants est donc conçu à partir de circuits résonnants dont la fonction globale est équivalente à celle du conduit vocal.

Le Laboratoire de la communication parlée de l'E.N.S.E.R.G. à Grenoble a longuement étudié ce type de matériel avant de céder ses études au Laboratoire de Marcoussis appartenant à la C.G.E. (Compagnie Générale d'Electricité) qui probablement le commercialisera. Aux Etats-Unis les Bell Laboratories ont réalisé plusieurs applications à l'aide d'unités de réponse à formants, applications très semblables à celles développées par le C.N.E.T.-Lannion avec les synthétiseurs à canaux.

Les tendances à la mode actuellement poussent les différents laboratoires de recherches à orienter leurs études sur un quatrième type d'unité à réponse vocale : les simulateurs de conduit vocal. En effet, le son est produit par la vibration de cordes dans des cavités résonnantes en mouvement. Pour rendre la voix artificielle, proche de la voix naturelle, pourquoi ne pas tenter de simuler notre appareil vocal ? C'est ce qu'essaient de faire, en France le C.N.E.T.-Lannion et l'E.N.S.E.R.G. à Grenoble.

Mais comment simuler le fonctionnement du conduit vocal ? D'abord en le simplifiant, en le mettant en équations et en lui substituant des circuits électroniques répondant aux mêmes équations. Les différentes études dans ce domaine qui donneront peut-être une vraie voix à l'ordinateur en le dotant véritablement d'un appareil vocal, en sont encore au stade de la recherche fondamentale.

Cependant, la synthèse de la parole seule, ouvre la porte à de nombreuses applications : la réservation de places d'avions, d'hôtels, de théâtres, la vente par corespondance, les renseignements en tous genres.

Alors pourquoi les unités à réponse vocale sont-elles encore dans les laboratoires ? Peut-être tout simplement parce que le progrès technologique va trop vite...

Françoise Harrois-Monin ■