L'INTELLIGENCE ARTIFICIELLE

comment faire parler un ordinateur?

(Suite : voir n° 1424)

LA parole est le moyen de communication le plus naturel : elle est encore rarement utilisée comme support de sortie d'informations provenant d'ordinateurs. Or, ce moyen s'impose déjà pour fournir à l'abonné au réseau téléphonique des indications issues d'un central (« le numéro que vous avez demandé n'est pas attribué actuellement »), d'un centre de renseignements (tel le service d'informations parlées, en composant INF 1), de l'horloge parlante ou d'autres sources (banques, systèmes de réservation de places...). L'abonné aux prises avec des difficultés de trafic ou de numérotation peut, depuis longtemps, entendre des conseils enregistrés.

DE LA BANDE MAGNETIQUE A L'HOLOGRAMME

La méthode la plus immédiate de synthèse de la parole est celle où l'on enregistre directement le signal vocal sur un support adéquat, une bande magnétique par exemple. La qualité de la voix restituée ne sera altérée, que par le manque de mélodie, dans le cas où la phrase est constituée par l'assemblage d'éléments de phrases ou de mots, initialement disjoints.

L'unité à réponse vocale IBM 7770, connectable à un ordinateur IBM360 ou 370, fonctionne suivant ce principe. Les informations d'entrée ont pour origine un poste téléphonique à clavier ou un terminal similaire.

Les signaux venant du terminal sont transmis par le réseau téléphonique, démodulés à l'arrivée, et présentés à l'unité à réponse vocale.

Ces signaux sont ensuite acheminés vers l'unité centrale de l'ordinateur pour y être traités.

La réponse est assemblée par la sélection, dans un dictionnaire, de mots ou de parties de phrases, et l'unité à réponse vocale transmet au terminal la réponse sous forme de voix humaine.

Dans l'IBM7770, le vocabulaire est enregistré sous forme analogique sur un tambour magnétique qui, par ses dimensions, et sa vitesse de rotation, limite le vocabulaire à un maximum de 128 mots d'une durée de 0,5 seconde. Pour préparer la réponse, l'ordinateur définit les adresses des mots à prononcer, et l'ordre dans lequel les mots doivent être émis : au moyen de ces adresses, l'unité 7770 lit l'information analogique et transmet la réponse, sous forme de voix humaine, au demandeur.

Le transfert des mots sur le tambour s'effectue à partir d'une bande magnétique enregistrée par un locuteur. Toutefois, on réduit à 0,5 seconde, les mots qui, à l'origine, peuvent durer jusqu'à 0,7 seconde : la réduction s'opère par des méthodes analogiques; elle est basée sur la redondance dans la partie voisée des mots.

Un mot plus long occupe plus d'une piste du tambour.

A Toulouse, A. Bruel et J.-C. Cazaux proposent d'utiliser les propriétés de l'enregistrement holographique comme mémoire adressable, permettant de générer un ensemble de mots constituant un vocabulaire spécialisé. L'holographie permet, en effet, d'enregistrer, sous de très petites dimensions, les représentations graphiques de syllabes, par exemple, d'associer côte à côte plusieurs de ces enregistrements, permettant un adressage aléatoire à chaque syllabe.

Les syllabes à mémoriser sont enregistrées, tout d'abord, suivant des méthodes similaires à celles des pistes sonores de films cinématographiques. De chaque photographie ainsi obtenue, on réalise des microhologrammes d'un millimètre de diamètre : c'est l'écriture de la « mémoire vocale ».

La lecture de cette mémoire se fait grâce à l'adressage d'un faisceau laser, le faisceau venant éclairer l'hologramme de la syllabe sélectionnée par une adresse.

L'un des problèmes majeurs est celui du codage son-image, avant l'enregistrement des hologrammes. Le choix de la piste sonore ne semble pas en effet, très satisfaisant (la longueur d'image est trop importante), et il a été envisagé de confier à l'ordinateur cette transmission par l'intermédiaire d'un analyseur de parole.

LES SYNTHETISEURS A CANAUX

Dans les synthétiseurs à canaux, le signal vocal est considéré comme la somme de signaux émis par une batterie de générateurs émettant un signal de fréquence donnée, dans une bande passante audible, et d'amplitude variable. Le signal, pour un locu

L.R. Rabiner, des Bell Laboratories, analyse la qualité d'une voix synthétiq en la « visualisant » sur écran cathodique.

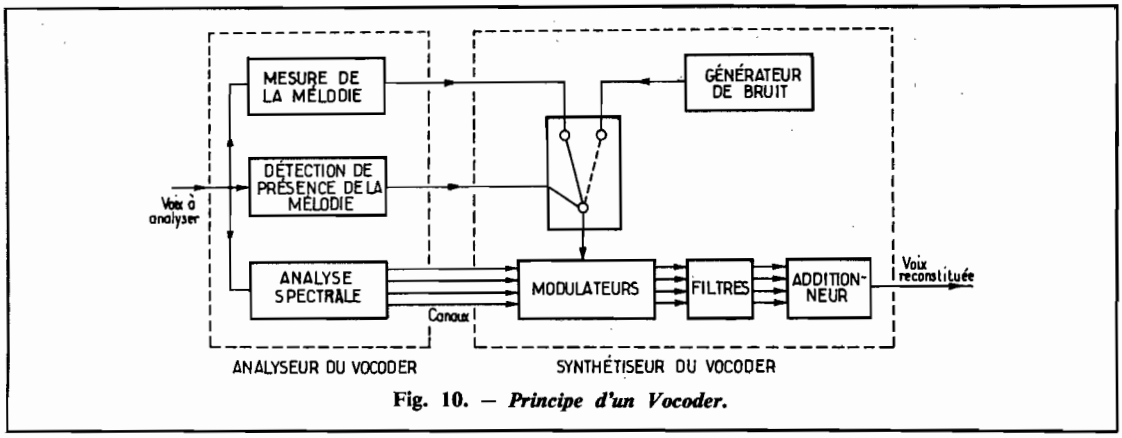

Fig. 10. — ***Principe d'un Vocoder.***

teur donné, dépend du phonème prononcé, mais également de la fréquence fondamentale, dénommée « pitch », caractérisant le caractère grave ou aigu de la voix. L'absence de pitch correspond à une voix chuchotée.

Les sons vocaux peuvent être divisés en deux grandes catégories suivant qu'il y a, ou non, vibration des cordes vocales du larynx. Un premier paramètre fondamental concerne donc l'existence de cette vibration. Lorsqu'elle est présente (cas des voyelles), il faut en mesurer la fréquence, second paramètre fondamental. Cette information, appelée mélodie, donne la hauteur, au sens musical, de la voyelle. Le timbre du son est donné par les cavités résonantes du système phonatoire humain, qui renforcent ou atténuent les divers harmoniques du son fondamental produit par les cordes vocales. Dans le cas des sons non laryngés (consonnes sourdes), tout se passe comme si une source de bruit était modulée par ces mêmes cavités résonantes. Par suite, une dernière série de mesures donne le troisième groupe de paramètres, qui définit la forme du spectre d'énergie du signal, en fonction de la fréquence.

Ces trois groupes de paramètres sont déterminés par l'analyseur d'un « vocoder » (compression de « voice-coder », soit codeur de voix). L'appareil (Fig. 10) permet d'extraire de la voix, les paramètres fondamentaux, caractéristiques d'un son, pendant un intervalle de temps significatif (par exemple, 20 millisecondes). Les trois groupes de paramètres sont conservés en mémoire par l'ordinateur.

Dans le synthétiseur du vocoder, un jeu de filtres passe-bande, couvrant tout le spectre à reproduire, est alimenté en parallèle, soit par des impulsions créées à la fréquence de la mélodie, soit par un générateur de bruit (cas des sons non laryngés). A chaque impulsion de commande, les filtres résonnent, chacun à sa fréquence propre, pendant quelques périodes. en modulant le signal dans chaque filtre par le niveau d'énergie mesuré à l'analyse, dans la bande de fréquences correspondante, on lui fait restituer une partie du spectre vocal d'origine. L'addition de tous ces éléments reconstitue un son proche du signal étudié.

L'artifice décrit réduit considérablement la quantité d'informations à traiter en machine, car il suffit souvent de moins de 2 500 éléments binaires, par seconde de parole, soit moins d'un vingtième du taux de départ.

Ce type de synthétiseur, dit à canaux est actuellement le plus répandu.

U.R.V., ICOPHONE, DECLAM... LA SYNTHESE PAR CANAUX

Une « unité de réponse vocale » (U.R.V.) a été étudiée par le Centre national d'études des télécommunications (*), pour être mise en service dans le réseau expérimental de commutation téléphonique « Platon ». La réalisation pratique de l'U.R.V. aurait dû permettre de fournir aux abonnés, le coût de leur dernière communication, et le contenu de leur compteur. Ce service devait être le premier d'une liste à compléter ultérieurement (heure, service du réveil, changement de numéro).

Dans l'U.R.V., la parole est représentée par son spectre à court terme; la connaissance de ce spectre toutes les 25 millisecondes permet de reconstituer un signal de parole de qualité acceptable, parfaitement intelligible. Chaque spectre est lui-même défini par l'énergie dans 12 bandes de fréquences.

Le vocabulaire de l'U.R.V. est constitué de quelques portions de phrase (« Votre compteur indique... », « le numéro que vous avez appelé... ») et d'une trentaine de chiffres et portions de nombres (« zéro », « un »,... « vingt », « trente ») : on peut ainsi prononcer tout nombre quelconque inférieur à un million. Lors de la synthèse, des corrections automatiques sont effectuées sur le rythme de la parole et la hauteur de la voix pour les mots, en fonction de leur emplacement dans le nombre ; les membres de phrase sont, quant à eux, restitués sans correction.

L'U.R.V. du C.N.E.T. est constitué de deux synthétiseurs à canaux, d'une mémoire de masse contenant 512 000 éléments binaires et d'un calculateur CII 10010.

Le C.N.E.T. étudie également un système dit de « synthèse par syllabes », à l'aide d'un matériel similaire au précédent. La parole est, ici, obtenue en mettant bout à bout des segments de parole; ceux-ci sont, en général, des diphonèmes (**). Des éléments plus importants (triphonèmes) peuvent s'imposer dans le cas de groupes consonnantiques complexes. Une première version, partielle, du vocabulaire a été obtenue par enregistrement et segmentation automatique; elle permettra d'aborder l'étude de la prosodie, mais devra être corrigée, diphonème par diphonème, avant de fournir une parole de bonne qualité.

Dans le modèle 7772 d'IBM, le vocabulaire est stocké sous forme numérique (contrairement au 7770), dans une mémoire à accès sélectif de l'ordinateur. Pour préparer une réponse vocale, l'ordinateur y lit la représentation numérique des mots et phrases à prononcer, et les transmet, dans l'ordre, à l'unité à réponse vocale. Celle-ci, au moyen d'un synthétiseur de vocoder à canaux, transforme les représentations numériques en mots et phrases parlés, qui sont transmis par le réseau téléphonique à l'utilisateur.

Deux ingénieurs du Centre d'études et recherches d'IBM, à la Gaude, A. Németh et R. Buron, ont réalisé une expérience de synthèse automatique de la voix à 200 bits par seconde de parole, au lieu de 2 500 bits par seconde dans les autres unités. Dans ces travaux, on a cherché à atteindre deux buts : supprimer, d'une part, la nécessité de l'enregistrement humain pour la préparation du vocabulaire des unités à réponse vocale; c'est l'ordinateur, lui-même qui génère la représensation codée de la voix, à partir d'une représentation phonétique des mots. En outre, il fallait réduire au minimum le taux d'information du code de la voix, en se limitant à un seul type de voix.

Les principes de base sur lesquels repose l'unité de Németh et Buron tiennent compte du fait que sur le millier de diphonèmes (ou phonatomes), seuls 200 d'entre eux, environ, sont utilisés en français. Dans la parole réelle, la transition entre deux sons élémentaires (les phonèmes) est continue ; dans la parole synthétique IBM, un échantillonnage adéquat permet de reproduire toute transition avec au plus trois états intermédiaires pouvant être communs à plusieurs transitions. Il en résulte qu'en français, 90 sons de transition suffisent, ce qui avec quelque 36 phonèmes, conduit à 126 éléments sonores différents pour générer le langage. En outre, dans la parole IBM, l'évolution de la mélodie n'est nullement continue (alors que c'est le cas dans la parole naturelle), mais approximée par 15 segments de droite (segments mélodiques), de pentes et de longueurs différentes.

Ainsi, le codage d'une voix peut faire appel à :

- 126 sons élémentaires, chacun d'eux étant représenté par 45 éléments binaires, décrivant la distribution de l'énergie contenue dans le son;
- 15 segments mélodiques représentés, chacun, par 4 éléments binaires.

Au total, toute l'information sur la parole peut être stockée dans une mémoire de 7 000 bits, située dans l'unité à réponse vocale. L'information est reçue, au niveau de cette unité, à la cadence de 200 éléments binaires par seconde; mais elle y est transformée en une information à environ 3 000 bits par seconde : en effet, l'information incidente

contient par exemple une adresse de spectre d'énergie (soit 7 bits) et cette adresse donne accès à 45 bits utilisés par le synthétiseur ; ou encore une adresse, en 4 bits, d'un segment mélodique, générant, au niveau du synthétiseur, un élément mélodique équivalent à 60 bits. Finalement, l'information à 200 bits par seconde servant aux communications entre l'ordinateur et le synthétiseur, est transformée en une information à 3 000 bits par seconde dans le synthétiseur, et la qualité de la voix reconstituée est équivalente à celle des autres unités à réponse vocale.

Un autre exemple pratique de système à canaux est fourni par le système « DECLAM » de la Compagnie industrielle des téléphones. Il s'agit d'un dispositif destiné à émettre par radio des informations météorologiques à l'usage des avions commerciaux. Ces informations arrivent de divers points du territoire, et sont acheminées vers un ordinateur central par liaisons télégraphiques. A partir des données qu'elles contiennent (température, pression, vent, etc.) l'ordinateur compose un message « vocal » du type « Orly, vent de dix mètres par seconde ». Les informations vocalisées sont constamment rafraîchies par l'arrivée de nouveaux messages sur les lignes télégraphiques. Le système actuel permet d'alimenter 10 voies de sortie vocale avec des programmes de vocalisation différents correspondant à des sélections particulières d'informations météorologiques, et à des émissions en langue française et en langue anglaise.

Au S.L.E.-Citerel, à Lannion, un synthétiseur à 12 canaux est également mis au point. En raison de sa structure entièrement numérique permettant de faire travailler les circuits en temps partagé, ce système est capable de synthétiser simultanément jusqu'à huit voix différentes.

Le système Icophone V réalisé à l'Université de mécanique physique, à Saint-Cyr-l'Ecole, se distingue des précédents du fait qu'il n'y a pas commande en amplitude des filtres, mais commande par tout ou rien. Ceci doit être compensé par l'augmentation du nombre de canaux, qui passe de 12 (ou 15) à 45.

LA SERIE DES ICOPHONES

Le modèle I de l'Icophone, construit en 1965 utilisait le principe du « Sonographe » : dans cet appareil, un signal est enregistré en boucle pendant 2,4 secondes, sur une piste magnétique, puis répété un certain nombre de fois pour l'analyse. Un procédé hétérodyne permet de n'utiliser qu'un seul filtre pour couvrir tout le spectre. Dans le modèle II, la fréquence sinusoïdale issue d'un générateur est utilisée dans un circuit de mélange de sons provenant de plusieurs canaux, lorsqu'une zone opaque, sur un sonagramme, défile près de l'une des cellules photoélectriques (Fig. 11). Un sonagramme, obtenu par un sonagraphe, est une représentation visuelle des sons. Par exemple, un son sinusoïdal de 1 000 Hz est représenté par un trait horizontal, fin en bande étroite, épais en bande large ; un son de fréquence décroissante donne un trait descendant, une impulsion est représentée par un trait vertical, un bruit blanc se traduit par un grisé plus ou moins uniforme, etc. L'Icophone II était donc un relecteur de sonagrammes, à 44 voies.

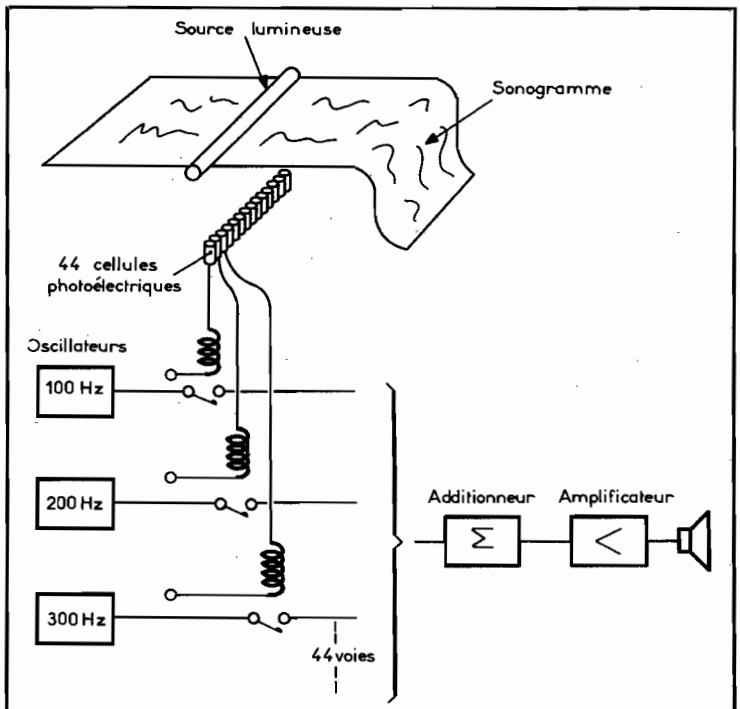

Fig. 11. — *L'Icophone II à commande optique.*

Le développement de l'Icophone a donné naissance à un troisième modèle, utilisant un ordinateur IBM 1130. Les cellules photoélectriques sont remplacées par des portes électroniques dont la fermeture est commandée par des signaux binaires. L'appareil sort sous forme vocale l'information contenue dans un texte écrit en français, et introduit dans l'ordinateur à partir de la machine à écrire ou d'un lecteur de cartes. Le texte orthographié est d'abord traduit, automatiquement, en une suite de symboles phonétiques, qui permet d'appeler les phonatomes correspondants, stockés dans une mémoire à disque. Les phonatomes sont ensuite juxtaposés et édités sur l'Icophone (***).

L'adjonction au message synthétisé d'éléments esthétiques tels que l'intonation et le rythme ont fait l'objet d'une nouvelle version de l'appareillage, l'Icophone IV.

Le prototype de l'Icophone V est en cours de réalisation avec l'aide de la Délégation à l'informatique : ce sera une unité autonome à réponse immédiate et vocabulaire illimité, qui pourra être connecté en lieu et place de n'importe quel télescripteur.

Marc FERRETTI

(*) On lira, à ce sujet, l'article de M. Cartier, J. Génin, P. Lorand, publié par le C.N.E.T., dans sa revue « L'Echo des recherches » en juillet 1971.

(**) Pour la définition des « phonèmes », à voir l'article « L'Ordinateur parle », paru, voici trois mois, en septembre 1973, dans le Haut-Parleur.

(***) J.-S. Lienard et D. Teil ont donné, en octobre 1970, dans la revue « Automatisme » une description détaillée de cet appareillage.