L'INTELLIGENCE ARTIFICIELLE :

SYNTHÈSE DE LA PAROLE : L'EFFORT FRANÇAIS

(suite, voir n° 1416)

LA diffusion des systèmes à réponse vocale a été jusqu'ici très lente, mais on dénote aux Etats-Unis particulièrement, une nette tendance à l'accélération : il y a, outre-Atlantique, plus de 600 installations, le marché étant dominé par IBM qui en détient 80 %. Parmi les constructeurs d'ordinateurs, seuls Burroughs et Honeywell proposent des matériels de ce type. On rencontre, en outre, de nombreux constructeurs spécialisés (Datatrol, Periphonics, Phonoplex...). Ces derniers proposent des matériels à vocabulaires limités (souvent 30 à 50 mots) ; en outre, certains de ces matériels sont conçus pour fournir un message vocal à caractère digital. Ainsi, au lieu d'annoncer la date sous la forme classique « 15 septembre », ils énonceront : « 1509 », l'auditeur, averti, reconnaissant dans les deux premiers chiffres, le numéro du jour, et, dans les deux derniers chiffres, le numéro du mois.

De nombreuses unités de réponse vocale sont destinées à être intégrées dans des systèmes auxquels auront accès de nombreux utilisateurs ; il y a lieu, dans certains cas, de pouvoir stocker, dans la mémoire de ces unités, des phrases entières, et non d'envisager d'assembler des bribes de phrase, ou des mots isolés, pour faciliter la compréhension des messages énoncés. Cette possibilité n'est pas offerte dans tous les systèmes commercialisés.

Une autre caractéristique des unités de réponse vocale est leur « pouvoir communicatif ». Il est clair qu'une telle unité ne se conçoit que dans des situations où il faut transmettre un volume important d'informations : c'est le cas dans les transactions bancaires. Ainsi, la Banque Régionale d'Escompte et de Dépôts (la BRED) a installé deux autocommutateurs IBM 2750, autorisant 650 postes téléphoniques à touches à se connecter à son ordinateur IBM 370/145, une unité IBM 7770 fournissant les réponses sous forme vocale. L'unité 7770 est essentiellement une unité de sortie capable de délivrer des messages composés de mots pré-enregistrés sous forme analogique, sur un support magnétique. Le système a été calculé pour supporter 1 500 messages à l'heure sans attente. Le vocabulaire de la BRED comprend : les chiffres de 0 à 9, des numéros de comptes, des codes, les 26 lettres de l'alphabet et 75 mots prédéterminés, servant à préciser le sens des réponses. Ce vocabulaire a été enregistré par un speaker de Radio-Monte-Carlo au laboratoire IBM de la Gaude, et envoyé à l'usine de Raleigh, aux U.S.A., pour le montage du tambour servant de mémoire à l'unité.

LE TÉLÉPHONE A TOUCHES : UN VÉRITABLE TERMINAL D'ORDINATEUR

Dans de nombreux systèmes faisant appel à des unités de réponse vocale, le périphérique d'entrée-sortie proposé à l'utilisateur est le clavier à touches. Le prix des téléphones à touches est peu différent des classiques postes téléphoniques à cadran. Un poste téléphonique est en outre une machine capable d'envoyer au central des numéros d'appel, donc des chiffres. Actuellement, la numérotation entre le poste d'abonné et le central s'effectue suivant l'un des trois principes suivants :

- la signalisation par train d'impulsions. C'est le type de signalisation le plus répandu en France ;
- La signalisation à courant continu : l'impédance du poste est fonction du chiffre à transmettre.
- La signalisation à fréquences vocales : chaque chiffre est converti en 2 fréquences parmi 7, émises simultanément.

Pour des applications de téléinformatique, la signalisation à fréquences vocales paraît la plus intéressante, car elle est la seule à pouvoir traverser telle quelle le réseau jusqu'à l'abonné demandé. Le même poste à clavier pourrait donc, en ce cas, non seulement établir la communication, mais ensuite transmettre des « données ».

Terminal à réponse vocale : c'est un simple poste téléphonique à touches.
(Cliché Post Office Research Station, Londres.)

11 L

TABLEAU 5
Qui participe, en France, au développement de la synthèse de la parole ?

- CEA - Saclay
- CII - Vélizy/Grenoble
- CIT-Alcatel
- CNET - Lannion
- ENSER - Grenoble
- ENST - Paris
- IBM - La Gaude
- Laboratoires de Marcoussis
- LIMSI - Orsay
- SLE-CITEREL - Lannion
- Thomson-CSF

TABLEAU 6. — Les unités à réponse vocale IBM

	7770	7772
Nombre de lignes de sortie	4, 8, 12, ... 48.	2, 4, 6, 8.
Enregistrement du vocabulaire	ANALOGIQUE	NUMÉRIQUE
Taille du vocabulaire	32, 48, 64, ... 128 mots de durée fixe (0,5 seconde).	Illimitée : le nombre et la durée des mots, ou phrases sont choisis par le client.
Stockage du vocabulaire	Tambour magnétique interne à l'unité à réponse vocale.	Mémoires à accès sélectif de l'ordinateur.
Transfert du vocabulaire en mémoire	A partir d'une bande magnétique enregistrée par un locateur.	L'enregistrement analogique est numérisé par un analyseur de vocodeur à canaux.

TABLEAU 8. — Enquête de marché pour les synthétiseurs de marché effectuée en 1971 par la C.I.T. (25 sociétés interrogées)

Système proposé : — installation desservant 50 à 100 terminaux vocaux ;
— entrée des informations par clavier téléphonique ;
— prix : 100 millions de francs ;
— vocabulaire 200 mots.

Résultat

Intérêt très certain	Intérêt très moyen	Intérêt très médiocre	Pas d'intérêt manifesté
40 %	20 %	28 %	12 %

TABLEAU 7

Type de synthétiseur / Qualités	Enregistrements		Synthétiseurs à canaux				
	Terminal vocal à mémoire holographique	IBM 7700	CIT DECLAM	CNET U.R.V.	S.P.S.	IBM 7772	IBM synthèse à 200 bits/s
Intelligibilité intrinsèques de la parole	Excellente		←		Bonne		
Qualité intrinsèques de la parole	Excellente		←		Moyenne		
Qualité d'information	Signal analogique		←		2400 bits/seconde		
Nombre et nature des paramètres de commande	"		← 12 canaux + pitch →			← 15 canaux + pitch	
Complexité du matériel	Moyenne	Simple	←		Moyenne		
Etat d'achèvement du matériel	Etude	Commercialisé	←	Opérationnel	→	Commercialisé	Opérationnel
Méthodes — Phrases			oui	oui			
Méthodes — Mots	oui	oui	oui	oui		oui	
Méthodes — Diphonèmes					oui		oui
Méthodes — Phonèmes	←		matériel pas approprié				
Etat d'achèvement du système		Commercialisé	Opérationnel	Développement	Etude	Commercialisé	Opérationnel
REMARQUES	Ces méthodes ne permettent pas l'accès à la prosodie ni à aucun paramètre lié à la sémantique de la parole. Il s'agit plus de recherches sur les mémoires que sur la synthèse de la parole.		Pas de prosodie	Prosodie	Prosodie	Pas de prosodie	Prosodie

Possibilité d'action sur les paramètres spectraux pour corriger le vocabulaire.
Possibilité facile d'action sur le pitch et le rythme, tant en correction du vocabulaire qu'en cours de synthèse pour l'introduction de la prosodie.

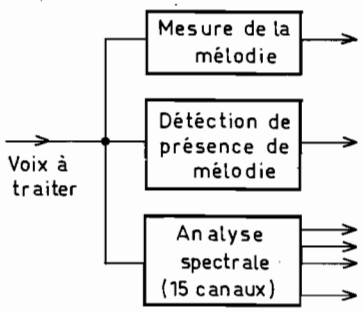

Fig. 5

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 l'existence de cette vibration. Lorsqu'elle est présente (cas des voyelles), il faut en mesurer la fréquence, deuxième paramètre fondamental. Cette information, appelée mélodie, donne la hauteur, au sens musical, de la voyelle. Le timbre du son est donné par les cavités résonnantes du système phonatoire humain, qui renforcent ou atténuent les divers harmoniques du son fondamental produit par les cordes vibrantes. Dans le cas des sons non laryngés (consonnes sourdes), tout se passe comme si une source de bruit était modulée par ces cavités résonnantes. Par suite, une dernière série de mesures définit la forme du spectre d'énergie du signal en fonction de la fréquence.

La figure 5 montre le schéma d'un analyseur de codeur de voix (ou VOCODEUR, compression de « voice coder »), où l'on reconnaît un système d'analyse spectrale associé à un détecteur de mélodie.

Les possibilités d'un tel « terminal » sont suffisantes pour une application en informatique : le CNET a réalisé un réseau expérimental permettant, à partir d'un poste téléphonique, d'interroger un calculateur pour effectuer des opérations du type « calcul de bureau ». Les résultats des calculs demandés sont renvoyés au client sous forme verbale grâce à une unité de réponse vocale. C'est le système SCT, ou « Service de Calcul par Téléphone ».

EN FRANCE, PLUSIEURS LABORATOIRES S'INTÉRESSENT A LA SYNTHÈSE DE LA PAROLE

L'Institut de Recherche d'Informatique et d'Automatique (IRIA) a organisé, le 2 février 1973, une journée d'études consacrée aux perspectives de recherche et de développement en France de la synthèse de la parole, à laquelle ont participé de nombreux responsables français de travaux en ce domaine. On y a montré que les unités faisant la synthèse de la parole se classent dans l'une des cinq classes suivantes :

● Synthétiseur analogique : on enregistre directement le signal vocal, qui a donc la même qualité que le son à la sortie d'un magnétophone. Des exemples de cette méthode sont l'horloge parlante, le système IBM 7770, ainsi que le terminal vocal à mémoire holographique réalisé à Toulouse : dans ce dernier cas, les syllabes à mémoriser sont enregistrées photographiquement

		Synthétiseurs à formants			Simulateurs de conduit vocal		Codage prédictif		
ICOPHONE	Vocoder numérique S. L. E.	ENSERG	CGE	CIPHON Thomson/ CSF	ENSERG	CNET	ENSERG	CNET	ENST
→ / - - - - → / →	→ / →	← Bonne → / ← Bonne →			Espérées bonnes à très bonnes		Espérées bonnes à très bonnes		
1700 bits/s	2400 bits/s	1000 bits/s	1200 bits/s	7000 bits/s	Objectif final 50 bits/s		à étudier		
47 canaux binaires	12 canaux + pitch	11 paramètres		7 paramètres	26 ou 17 aires de section du conduit ou environ 10 paramètres		12 paramètres de commande + paramètres d'entrée		
Simple	Moyenne	Moyenne →			?	?	?		
pérationnel	Développement	Développement			Etude		Etude		
		oui	oui	Utilisé	Sans intérêt		Toutes les méthodes sont envisageables		
		oui	oui	en					
oui		oui	mais	télécom-					
→	→	oui	mais	munications	oui	oui			
Opérationnel		Etude			Recherche fondamentale		Etude		
as de pitch, inc pas de osodie. es diphonèmes nt en mémoires ne sont pas aités en temps el.	Susceptible des mêmes applications que les vocoders analogiques.	La méthode actuellement optimale d'emploi de ce type de synthétiseurs est la synthèse par mots. La synthèse par règles avec ces matériels peut être étudiée à fin de recherche fondamentale sur les transitions de la parole. Mais la complexité finale des règles de transition en termes de formants ne semble pas ouvrir la voie à une utilisation réelle du synthétiseur à formants par association de phonèmes ou diphonèmes.			C'est la source d'une recherche fondamentale sur la physiologie du conduit vocal, le fonctionnement des cordes vocales. On se rapproche de l'étude des commandes réelles de l'organe phonatoire par l'homme.				

TABLEAU 4. — LES UNITÉS A RÉPONSE VOCALE (SAUF IBM)

Fabricant	Modèle	Vocabulaire			Mémoire de stockage	Nombre de lignes téléphoniques accessibles	Temps d'accès (milliseconde)	Extension possible ?
		Nombre de mots	Nombre de messages	Durée des mots (seconde)				
Advanced Communications Inc.	VS6464/1	64	64			1	< 1	
	VS6464/1M	64	64			1		
American Systems	Voicemaster 3000	3 000			Disque magnétique	32		Jusqu'à 10 000 mots
Cognitronics	630	10	10		Tambour magnétique	1	625	
	631, 632, 672/4/6/8	31	31	0,5-0,6		1	625	
	636	31		0,5-0,6		1	625	
	MARS	31 à 189		0,5-0,6		4 à 48	260	
	STAR	32				1	625	
	Horloge parlante	70				1	10 000 (par message)	
Co-System, Inc.	TAU16-W	16			Ruban magnétique		600	
	TAU16-Wx	64					600	
	DTD-16W	16					600	
	DTD-16Wx	64					600	
	DTD-8M		8				600	
	TAU-8M		8				4 600	
Datatrol Inc.	CS-1400	31 à 256		0,5 à 1,5	Tamb. magn.	2 à 64		
Métrolab Inc.	Digitalk 256		26 à 256	0,5-0,6	Tamb. magn.	1	600 à 1 200	
	Digitalk 3100		32 à 64	0,5-0,6	Tamb. magn.	1	600 à 1 200	Par bloc de 27 mots ou phrases
	Digitalk 4000	36 à 64		0,5-0,6	Tamb. magn.	1	500	
Periphonics Corp.	PAC2000	2 000		Diverse	Disque magn.	256	16,5	Jusqu'à 10 000 mots
Phonoplex Corp.	7050	50		0,5	Semi-conduct.	4 à 256	125	Extension non limitée
Wavetek Data Communications	A500	32 à 256		0,5-0,6	Tamb. magn.	8 à 64		

sous forme d'une modulation d'amplitude, et on réalise à partir de chacune des photographies ainsi obtenues, des microholographes ; la lecture des hologrammes se fait grâce à un faisceau laser.

● Synthétiseur 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 la bande passante audible, et d'amplitude variable. Ce signal, pour un locuteur donné, dépend du phonème (*) prononcé, mais également de la fréquence fondamentale, dite « pitch », qui caractérise une voix grave ou aiguë. L'absence de pitch correspond à une voix chuchotée.

Ce type de synthétiseur est le plus répandu. Il est à la base du système DECLAM de la CIT-Alcatel, de l'Unité de Réponse Vocale (URV) du CNET, du système de synthèse par syllabes (SPS), également étudié au CNET, de l'IBM 7772, du système IBM à 200 bits par seconde, du Vocoder de la SLE et du système ICOPHONE. Ce dernier système se distingue des précédents par le fait qu'il n'y a pas commande en amplitude des filtres, mais commande par tout-ou-rien, ce qui fait passer le nombre de canaux de 12 ou 15, à 45.

● Les synthétiseurs à formants : on simule les qualités de résonance du conduit vocal par des circuits électriques. Les pics de résonance correspondent à ce qu'on appelle les formants. Le naturel des voix synthétisées ainsi est incontestablement meilleur qu'avec les synthétiseurs à canaux. Des synthétiseurs de ce type ont été réalisés à l'ENSERG, aux laboratoires de Marcoussis, ainsi que par Thomson-CSF.

● Les simulateurs de conduit vocal : ils simulent le comportement acoustique d'un conduit de section variable correspondant à un conduit vocal simplifié, à l'aide de circuits électriques. A l'ENSERG et au CNET, des études sur ce type de synthétiseurs sont menées.

● Les synthétiseurs à codage prédictif : ils simulent de façon globale la fonction de transfert du conduit vocal. Trois études sont en cours, à l'ENSERG, au CNET et à l'ENST.

(à suivre)

Marc FERRETTI.

* Voir H.-P. précédent « L'ordinateur parle ».