Reconnaissance de la parole et modélisation statistique : expérience du CNET (*)

Speech recognition and statistical approach : CNET's experience

C. GAGNOULET
CNET LAA/TSS/RCP
BP 40, F-22301 Lannion

Christian GAGNOULET : né le 2 septembre 1952. Diplômé de l'École Nationale Supérieure des Télécommunications, il entre au CNET en 1976. Il s'intéresse à la réalisation de différents systèmes de reconnaissance de parole, et à la mise en application de ces systèmes. Il est aujourd'hui responsable des études de reconnaissance au département Recherches en Communications par la Parole.

D. JOUVET
CNET LAA/ TSS/RCP
BP 40, F-22301 LANNION

Denis JOUVET : né le 23 juillet 1956. Ancien élève de l'École Polytechnique et ingénieur des Télécommunications, il entre au CNET en 1981. A partir de 1985, il conçoit et développe le système PHIL86 et participe à sa mise en application dans plusieurs projets. Il soutient en 1988 une Thèse de Doctorat de l'ENST sur la reconnaissance statistique indépendante du locuteur.

RÉSUMÉ

Cet article décrit les travaux menés au CNET ces dernières années, dans le domaine de la reconnaissance de la parole. Après avoir rappelé le contexte de cette recherche, on décrit le logiciel PHIL86 destiné à reconnaître des vocabulaires de petite taille, indépendamment du locuteur, et les développements matériels qui lui ont été associés. Deux expérimentations de la reconnaissance dans le domaine des Télécommunications sont ensuite présentées, en insistant principalement sur les enseignements qui en ont été tirés et les résultats des évaluations menées sur le terrain.

MOTS CLÉS

Applications de la reconnaissance de parole, modèles de Markov cachés, reconnaissance indépendante du locuteur, serveurs vocaux interactifs.

SUMMARY

This paper presents the work done at the CNET in speech recognition during the last few years. The authors present the recent generation of speaker-independent systems, based on statistical modeling using the Markov models (PHIL86 *software). Several applications of these systems in the Telecommunications area are described, as well as the lessons drawn from them.*

KEYWORDS

Speaker-independent speech recognition applications, hidden Markov modeling, speech-activated audiotex.

(*) *Nota :* Cet article reprend plusieurs fragments d'un article publié dans l'Écho des Recherches n° 135 du 1er trimestre 1989.

1. Introduction

Les démonstrations en laboratoire de systèmes de reconnaissance sont chaque jour plus impressionnantes (très gros vocabulaires reconnus, langue quasi naturelle...). Le nombre de produits industriels ne cesse de croître (environ 200 produits différents disponibles aujourd'hui), et les sociétés proposant des systèmes de reconnaissance sont nombreuses (notamment aux USA). Les prix sont en baisse, les performances annoncées en hausse. Les études de marché [1, 2] prédisent avec obstination depuis 1980 une explosion du marché dans un futur proche. Malgré cela, il n'existe toujours pas d'application de grande diffusion. L'impact dans le grand public demeure quasiment nul, la reconnaissance y étant souvent totalement inconnue, ou perçue comme un gadget amusant.

Les études de reconnaissance menées au CNET depuis 1981 ont pour vocation de répondre aux besoins spécifiques des Télécommunications. C'est pourquoi certains choix ont été faits depuis plusieurs années, limitant les efforts de recherche à quelques domaines jugés prioritaires : la **reconnaissance indépendante du locuteur,** en présence de **canaux de transmission téléphoniques,** ainsi que la **prise en compte des facteurs humains** cruciaux pour la mise en place d'applications réelles.

A partir de 1985, une nouvelle génération de systèmes de reconnaissance a été conçue au CNET, reposant sur une modélisation statistique du vocabulaire de chaque application. Ces systèmes utilisent des variantes d'un même logiciel, appelé PHIL86, qui ont été implantées sur des dispositifs matériels adaptés (cartes RDP), avant d'être évaluées en situation dans plusieurs applications expérimentales et transférées depuis dans le milieu industriel français.

Après quelques rappels sur les principes de cette modélisation statistique, nous décrirons les spécificités du logiciel PHIL86. Puis, pour deux applications différentes, nous détaillerons la mise en œuvre, les problèmes rencontrés, et l'évaluation qui en a résulté.

2. Description de PHIL86

2.1. MODÉLISATION STATISTIQUE

L'approche par modèles de Markov cachés [3, 4] est utilisée au CNET depuis 1985, et a donné naissance au système PHIL86 [5]. Ce système s'est rapidement avéré nettement plus performant que le système antérieur SERAPHINE qui reposait sur la technique de comparaison dynamique entre formes acoustiques [6]. Ce système PHIL86 permet l'introduction dans les modèles de connaissances phonétiques explicites [7, 8]. Dans ce système, les fonctions de densité de probabilité sont associées aux transitions, et sont supposées continues et gaussiennes, avec une approximation diagonale pour les matrices de covariance. L'analyse acoustique calcule toutes les 16 ms (fenêtres de Hanning de 32 ms avec un recouvrement de 50 %) 6 coefficients cepstraux obtenus à partir de l'échelle Mel (MFCC), complétés par un paramètre d'énergie et sa variation temporelle (entre trame suivante et trame précédente).

Les **modèles de Markov** employés sont définis par les états (q_i) de la chaîne de Markov sous-jacente, les probabilités (a_{ij}) des transitions et les paramètres (vecteurs moyennes m_{ij} et matrices de covariance diagonales Σ_{ij}) des fonctions de densité de probabilité gaussiennes (B_{ij}) associées aux transitions. En notant $X[\tau]$ la τ-ième trame (vecteur de p coefficients) du mot inconnu $X[1 \ldots T]$ de T trames, $B_{ij}(X[\tau])$ représente la probabilité d'observation de la trame $X[\tau]$, durant la transition de l'état q_i vers l'état q_j.

Au cours de la reconnaissance, on s'intéresse à la probabilité maximale d'observation de l'ensemble des trames du mot (ou de la phrase) inconnu, la chaîne de Markov étant donnée. En notant $\Phi[\tau, q_i]$ la **probabilité maximale d'observation** des τ premières trames, le long des chemins atteignant l'état q_i au temps τ, on peut utiliser l'algorithme de Viterbi (algorithme de programmation dynamique) pour calculer $\{\Phi[\tau, q_i], \forall i\}$, à partir de $\{\Phi[\tau - 1, q_i], \forall i\}$. Pour chaque trame, et pour tous les états de la chaîne, nous utilisons la formule de réestimation suivante qui établit qu'un chemin de longueur τ résulte de la prolongation d'un chemin de longueur $\tau - 1$ par une transition entre états et l'observation de la trame $X[\tau]$ au cours de cette transition.

$$\Phi[\tau, q_i] = \operatorname{Max}_{q_j} \Phi[\tau - 1, q_j] \cdot a_{ji} \cdot B_{ji}(X[\tau]) .$$

Ainsi, en notant q_F le dernier état de la chaîne, $\Phi[T, q_F]$ est la probabilité cherchée.

Bien que l'algorithme de Baum-Welch calcule la probabilité d'observation exacte (et non pas la probabilité maximale le long d'**un** chemin), l'algorithme de Viterbi est beaucoup plus maniable, notamment avec une arithmétique en virgule fixe. C'est pour cette raison que dans PHIL86, nous utilisons cet algorithme, aussi bien durant la reconnaissance que l'apprentissage.

2.2 MODÉLISATION DES APPLICATIONS

Dans PHIL86, les modèles de chacun des mots sont regroupés en un réseau unique, compilé pour chaque application, qui décrit toutes les séquences de mots autorisées (dans le cas d'une reconnaissance de mots enchaînés). Ce réseau inclut des modèles de silence en début et en fin de phrase, ce qui autorise l'emploi d'un algorithme de séparation bruit-parole relativement simple.

Pour une application donnée, on doit associer à ce réseau les densités de probabilité sur l'espace des trames acoustiques. Pour illustrer cette modélisation, et afin de ne pas alourdir la figure, nous avons choisi l'exemple d'une application « fictive » destinée à reconnaître les nombres à 2 chiffres, compris entre 00 et 69.

Pour construire ce modèle unique, on introduit et on traite successivement :

- La *syntaxe* des phrases possibles (*fig.* 1*a*).

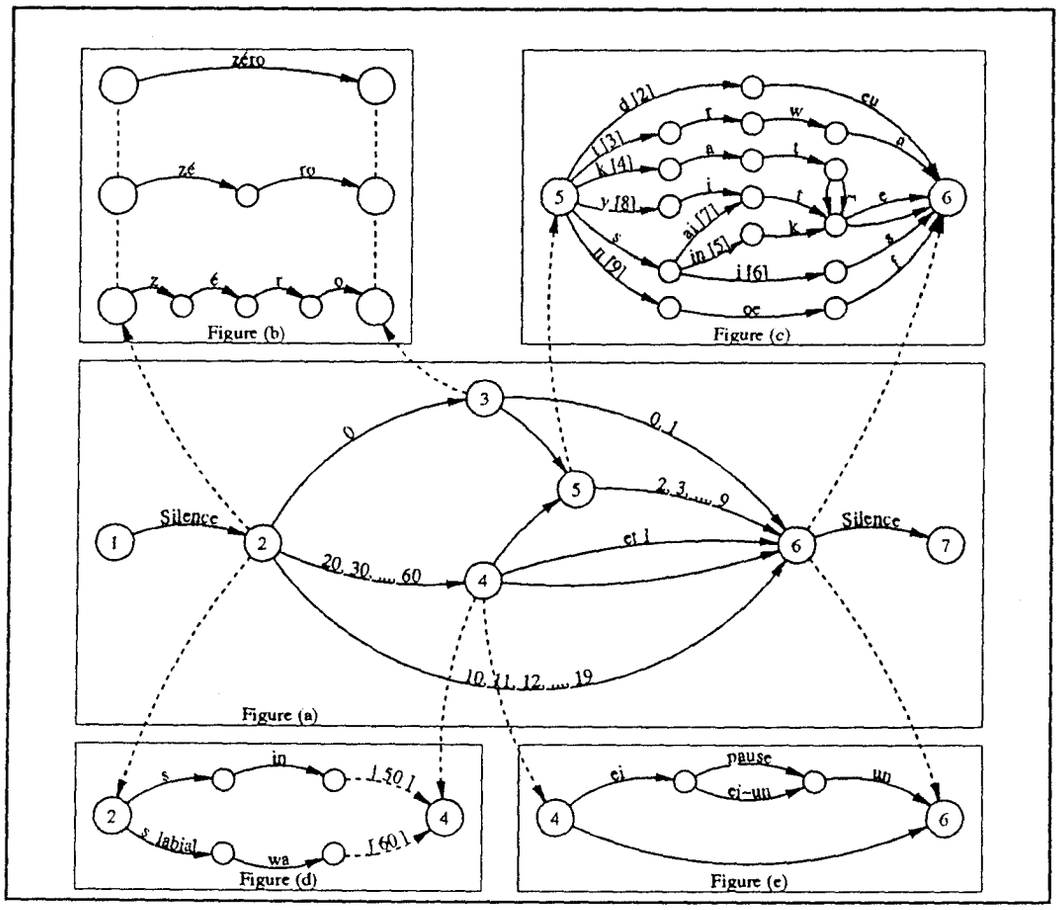

Figure 1. — Modélisation d'une application : « Nombres compris entre 00 et 69 ».

• La décomposition éventuelle de chaque mot (*fig.* 1*b*) en *unités plus petites* (syllabes, phonèmes ou diphones).

• Les *règles phonologiques* qui permettent de traiter les coarticulations entre mots (*fig.* 1*e*) et de définir éventuellement des unités allophoniques (*fig.* 1*d*).

• La description des *modèles de Markov* pour chaque unité de base (chaîne de Markov et association des densités de probabilité).

Afin de réduire les calculs lors de la phase de reconnaissance, le réseau est optimisé en regroupant les éléments communs (*fig.* 1*a* et 1*c*).
Pour des petits vocabulaires (quelques dizaines de mots) et pour une prononciation en mots isolés, les modèles utilisant les mots comme éléments de base sont suffisants. En revanche, pour de plus gros vocabulaires ou bien pour le traitement des mots enchaînés, l'emploi d'unités de taille inférieure au mot devient nécessaire (pour des raisons de taille du modèle et de qualité de reconnaissance). En effet, on obtient ainsi des modèles plus compacts et l'introduction des *règles phonologiques* permet d'améliorer sensiblement la qualité de reconnaissance.

2.3. IMPLANTATION SUR CARTES

Trois cartes ont été conçues en 1986 et 1987 pour recevoir et exploiter ce logiciel PHIL86, sous des formes plus ou moins simplifiées [9], comme le résume le tableau 1.

La carte RDP50 est une carte au format PC, organisée autour d'un processeur Texas Instruments TMS32020 ou TMS320c25, et dotée d'une capacité mémoire de 2 fois 64 Kmots de 16 eb. La parole est numérisée au moyen d'un COFIDEC adapté aux conditions d'exploitation téléphoniques. Pour une fréquence d'horloge de 40 MHz, on peut reconnaître en temps réel un vocabulaire d'environ

TABLEAU 1
Cartes de reconnaissance adaptées à PHIL86

Cartes	Logiciels	Interface
RDP50	PHIL86	bus PC
RDP20	PHIL86	bus PC
RDP8	PHIL86 simplifié (mots isolés)	RS232

70 mots. On notera qu'en utilisation téléphonique, et pour des raisons de qualité de reconnaissance, il convient de se limiter à une dizaine de mots différents à chaque instant. Cette carte est aujourd'hui commercialisée par plusieurs sociétés françaises (XCOM, SEFER, MEDIAVISUEL...) à un prix voisin de 15 000 F, logiciel compris.

La carte RDP20 est une version réduite de la précédente. C'est également une carte au format PC, mais elle utilise un processeur moins puissant : le TMS32010 de Texas Instruments. Avec une capacité mémoire réduite à 2 fois 8 Kmots, on peut traiter des vocabulaires d'une trentaine de mots, avec des performances identiques à celles obtenues sur la RDP50. Cette carte, dotée d'une interface téléphonique intégrée, est commercialisée par la société ACSYS à un prix de 15 000 F environ.

En bas de gamme enfin, et pour des applications grand-public tolérantes où le coût est le critère dominant (jeux), la carte RDP8 a été réalisée, à titre d'exercice de style essentiellement. En effet, cette carte, exploitable par liaison RS232, utilise un monochip Motorola 6805. Seuls des modèles par mots très simples sont pris en compte, et l'analyse acoustique est ici réduite à un calcul... d'histogrammes de passages par zéro. On aboutit ainsi, pour un coût approximatif de 100 F, à un système (presque) indépendant du locuteur, capable d'identifier (presque) correctement jusqu'à 8 mots isolés (!).

2.4. ÉVALUATIONS EN LABORATOIRE

Plusieurs tests de laboratoire ont été effectués sur PHIL86, avec des bases de données enregistrées à travers le réseau téléphonique interurbain (avec des locuteurs d'accents régionaux différents) :

Chiffres : Chiffres isolés (0 ... 9), 450 locuteurs,
Trégor : 36 mots isolés (mots de commande), 510 locuteurs,
Nombres : Nombres à deux chiffres (00 ... 99), 720 locuteurs.

Une base de données complémentaire a été enregistrée à travers le réseau téléphonique **local** pour étudier l'influence de la taille du vocabulaire sur les performances :

Mots : les 500 mots les plus courants du Français, base décomposée en séries aléatoires de 100 mots, 10 locuteurs, et 3 répétitions par locuteur. Pour cette base, les résultats sont fournis pour des sous-ensembles de 100, 300 et 500 mots (Mots_100, Mots_300 et Mots_500).

Les trois premières bases ont été découpées en deux parties sensiblement égales, une pour l'apprentissage, l'autre pour les tests, contenant évidemment des locuteurs différents pour des tests en mode « indépendant du locuteur » (« Xloc »). La dernière base a servi à des tests « plurilocuteurs » (liste fermée de locuteurs, « Ploc »). On présente dans le tableau 2 deux ensembles de tests :

Pour les tests PHIL86/RDP, on utilise des modèles par pseudo-diphones, et l'analyse acoustique standard de la carte RDP50 (6 MFCC, l'énergie et sa variation).

TABLEAU 2
Évaluations de PHIL86 en laboratoire

Base	Type	Taille du corpus de test	PHIL86/ RDP	PHIL86/ VAX
Chiffres	XLoc	2 100 mots	2,5 % [± 0,7 %]	1,3 % [± 0,5 %]
Trégor	XLoc	8 400 mots	2,5 % [± 0,3 %]	0,9 % [± 0,2 %]
Nombres à 2 chiffres	XLoc	6 700 mots	8,5 % [± 0,7 %]	5,5 % [± 0,5 %]
Mots_100	Ploc	1 000 mots	5,2 % [± 1,4 %]	
Mots_300	Ploc	3 000 mots	13,3 % [± 1,2 %]	
Mots_500	Ploc	5 000 mots	21,1 % [± 1,1 %]	

Pour les tests PHIL86/VAX, on emploie des modèles par mots avec 30 états par mot, et une analyse acoustique étendue, calculant toutes les 16 ms : 8 MFCC, l'énergie, et les dérivées temporelles de ces 9 paramètres obtenues par régression linéaire sur 5 trames adjacentes (80 ms).

3. Applications aux Télécommunications

Les premières applications de la reconnaissance vocale dans les Télécommunications ne sont apparues en France qu'en 1988. Elles reposent toutes sur les techniques présentées ci-dessus : logiciel PHIL86 et cartes associées (ou dérivées de celles-ci). Ces applications, initialisées par le CNET, ont servi en particulier à sensibiliser les industriels aux possibilités offertes par la reconnaissance. Le savoir faire étant maintenant entre les mains des industriels, de nouvelles applications de complexité équivalente devraient se développer en 1990, notamment dans le domaine des serveurs vocaux interactifs.

Nous décrirons ici les deux premières de ces applications, lancées par le CNET entre 1985 et 1987 : la cabine téléphonique **PUBLIVOX** commandée à la voix (reconnaissance locale) et le serveur interactif **MAIRIEVOX** (reconnaissance à travers le réseau téléphonique).

3.1. PUBLIVOX : CABINE PUBLIQUE COMMANDÉE A LA VOIX

En supprimant le clavier et le combiné téléphonique dans une cabine téléphonique publique, on réduit d'autant les risques de vandalisme, tout en améliorant le confort grâce à la conversation en mode mains-libres. Le projet PUBLIVOX [10], mené avec la participation industrielle de la société CROUZET, reposait sur cette hypothèse, et avait également pour objectif d'étudier les limites, en situation réelle, des systèmes de reconnaissance indépendants du locuteur et de mieux mesurer l'importance des facteurs humains dans un dialogue homme-machine. Il faut noter qu'en même temps, on plaçait la reconnaissance dans une situation techniquement très risquée (environnement acoustique difficile, utilisation d'un vocabulaire difficile et très peu compétitif face aux claviers : les nombres).

Une maquette de cabine à commande vocale où la numérotation était obtenue en prononçant des chiffres isolés ayant été favorablement accueillie lors d'une exposition au musée postal à Paris, en 1984, il fut décidé de réaliser 10 prototypes industriels, où la numérotation serait faite par groupes de deux chiffres. Ces prototypes devaient être évalués sur le terrain, auprès du grand public, dans plusieurs villes françaises.

Extérieurement, les cabines PUBLIVOX restent très proches des cabines publiques conventionnelles. Seules les vitres de l'habitacle ont été renforcées pour améliorer l'isolation acoustique (une amélioration de 6 dB a ainsi été obtenue). Le combiné est remplacé par un microphone et un haut-parleur dissimulés derrière des grilles de protection. La barrette d'affichage a été conservée pour guider l'usager. La parole synthétique codée à bas débit est réservée aux cas où l'usager hésite ou commet une erreur au cours du dialogue avec le publiphone. Le paiement par cartes à mémoire est bien entendu conservé. Les nouveautés essentielles concernent l'emploi de la reconnaissance de la parole durant la phase de numérotation, et d'un téléphone mains-libres durant la phase de communication.

La reconnaissance est effectuée **en local (large bande).** La numérotation se fait en prononçant des nombres de deux chiffres, selon les habitudes des utilisateurs français. Ainsi le système de reconnaissance autorise la prononciation des nombres de 00 à 99 (en mode mots enchaînés), auxquels s'ajoutent quelques mots isolés pour certaines fonctions particulières (appels directs des numéros d'urgence : POMPIERS, SAMU...) ou pour le contrôle du dialogue (ENVOI, CORRECTION...).

Un exemple de dialogue entre un usager et la cabine PUBLIVOX est représenté sur la figure 2. En fonction des actions mécaniques de l'usager et des mots prononcés (commandes vocales), on indique l'état du dialogue et l'affichage correspondant sur la barrette.

Action mécanique	Affichage	Commande vocale	État du dialogue
Introduction carte, verrouillage	Décrochez ou insérez votre carte		Inactif Validation de la carte
	Prononcez votre numéro par groupes de 2 chiffres 96 -- 96 07 -- 96 -- 96 05 -- 96 05 11 -- 96 05 11 11	**96** **05** **CORRECTION** **05** **11** **11** **ENVOI**	Numérotation
Appui sur bouton Récupération de la carte			Communication Raccrochage Inactif

Figure 2. — Exemple de dialogue entre un usager et PUBLIVOX.

Les dix prototypes ont été installés fin 1988 dans 6 villes françaises : Paris, Rennes, Montpellier, Valence, Lannion et Perros-Guirec. Le comportement des usagers a été suivi et analysé durant 6 mois dans 3 de ces villes.

Évaluation de la reconnaissance : Le taux d'erreur de reconnaissance observé en situation avec des utilisateurs novices est voisin de 25 % pour les nombres à deux chiffres. Ce chiffre est à comparer à ceux obtenus lors des évaluations en laboratoire (3,3 % d'erreur pour des nombres à 3 chiffres enregistrés en large bande), et lors d'une évaluation faite à Lannion dans un PUBLIVOX avec des locuteurs expérimentés (7 % d'erreur). La plupart des erreurs de reconnaissance proviennent :

— d'une part, du non-respect des consignes d'élocution par les usagers (mode d'emploi non lu ou mal interprété),

— d'autre part, du manque de robustesse du système face aux perturbations acoustiques (à Paris notamment).

Évaluation du service : Les moyennes observées révèlent qu'au premier essai, 19 % des utilisateurs abandonnent en cours de route, ou composent un mauvais numéro. Certains des utilisateurs ont été conviés à faire une seconde tentative. Lors de ce second essai, ce pourcentage tombe à 9 %. Ceci montre bien l'effet d'adaptation **du** locuteur. Le téléphone mains-libres, et le guidage par messages vocaux en cas de problèmes sont très appréciés. Globalement, le trafic observé sur les PUBLIVOX s'élève à 65 % de celui obtenu avec des cabines conventionnelles situées à proximité. Très peu de réactions de rejet ont été observées. Même si le coût des prototypes ne permet pas une généralisation de ces cabines dans l'état actuel, cette expérience reste très positive par les améliorations techniques qu'elle a suscitées et la meilleure connaissance des conditions réelles d'exploitation de la reconnaissance qu'elle a permise.

3.2. MAIRIEVOX : SERVEUR VOCAL INTERACTIF

Les progrès de reconnaissance ont permis d'envisager dès 1987 des applications centralisées **dans le réseau téléphonique,** sous forme de serveurs interactifs à commande vocale. La qualité du réseau analogique, et la diversité des microphones et des terminaux entraînent cependant des limitations sévères sur la taille des vocabulaires reconnus à chaque instant (une dizaine de mots isolés seulement).

Les problèmes ergonomiques posés par des dialogues entièrement vocaux pour ces serveurs destinés au grand public (définition des prompts, choix des menus, procédures de récupération des erreurs de reconnaissance...) ont été soigneusement étudiés.

Un serveur sur PC, démontrant l'intérêt de la reconnaissance dans le domaine de l'information au grand public, a ainsi été réalisé et installé par le CNET à la mairie de Lannion dès Avril 1988. Ce serveur MAIRIEVOX [11] est destiné à fournir aux habitants de cette ville des renseignements vocaux sur les loisirs régionaux ou sur les services d'urgence.

Le dialogue entre le système et l'usager est de type arborescent, à menus explicites : l'usager doit dans chaque cas prononcer une des commandes qui lui sont proposées.

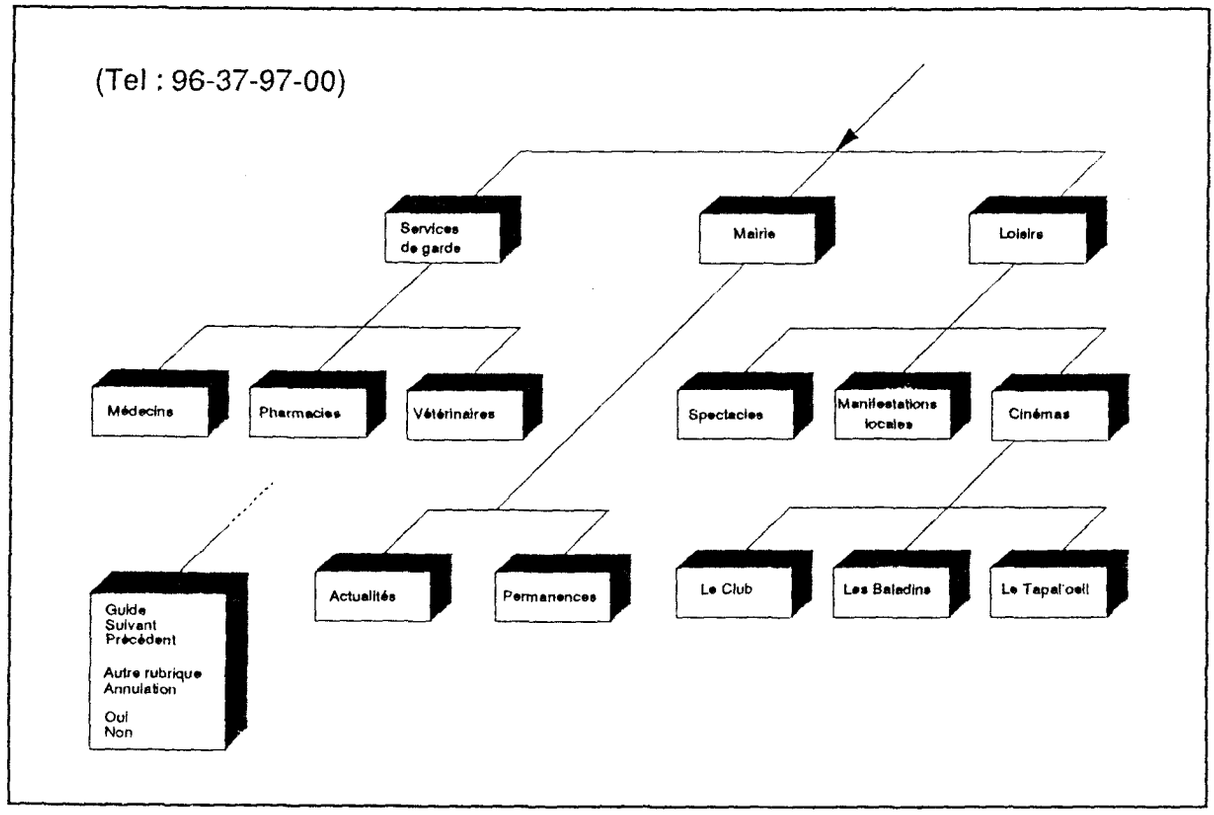

Figure 3. — Description du service MAIRIEVOX.

Ce type de dialogue limite de fait la complexité de l'arbre décrivant le service, puisque pour des raisons de rapidité d'accès à l'information, l'arbre ne doit pas être trop profond, et pour des raisons de mémoire auditive, il ne doit pas être trop large. Une profondeur et une largeur de 3 semblent un compromis tout à fait acceptable. La structure de cet arbre est précisée sur la figure 3.
La figure 4 illustre un exemple de dialogue entre un usager et MAIRIEVOX. On notera que le système autorise l'usager à interrompre les messages à tout moment, ce qui, compte tenu de la mauvaise séparation entre les voies d'émission et de réception, a nécessité d'introduire un annuleur d'écho pour ne pas dégrader la qualité de reconnaissance.

Évaluation de la reconnaissance. Le vocabulaire de reconnaissance comporte 21 mots isolés, mais seulement 6 d'entre eux sont valides à un instant donné (3 choix possibles + 3 mots de gestion). Ici encore, comme pour PUBLIVOX, on observe un écart important entre les taux d'erreur en laboratoire (2,5 %) et ceux observés en situation (de l'ordre de 20 % d'erreur).

Une procédure d'écoute des mots prononcés a été récemment mise en place. Un premier relevé des erreurs réellement imputables au système de reconnaissance (hormis celles dues au locuteur) laisse supposer que plus de 50 % de ces erreurs résultent de mots tronqués par suite d'une erreur du processus de séparation bruit-parole (automate d'états fini travaillant sur l'énergie du signal de parole). Ceci se produit principalement durant les phases de double parole (interruption d'un message en cours), conditions évidemment très différentes de celles que l'on utilise durant l'apprentissage, ou durant les tests de laboratoire. Ceci montre une fois de plus le peu de signification des tests effectués en laboratoire, mais aussi la nécessité de s'adapter dynamiquement aux conditions réelles de prise de son en phase d'exploitation de ces serveurs.

Évaluation du service. Un des points fondamentaux relevés lors des expérimentations est la nécessité de valider la reconnaissance durant tout le dialogue (y compris durant l'émission des messages vocaux). Grâce à cette possibilité, le serveur s'avère plus rapide et plus naturel qu'un serveur identique utilisant un dialogue par clavier à touches multifréquences ou qu'un serveur VIDEOTEX.

Depuis plus d'un an, ce serveur monovoie (une seule ligne téléphonique) est utilisé régulièrement par les habitants de Lannion. Plus de 6 000 communications ont été relevées entre avril 88 et février 89. Plus de la moitié de ces appels (62 %) concernent les renseignements relatifs aux loisirs, ce qui tend à prouver l'intérêt du service. Aujourd'hui encore, sans aucune publicité extérieure au CNET, MAIRIEVOX reçoit environ 150 appels par semaine, essentiellement en soirée (horaires des cinémas).

Hormis l'intérêt purement local du service, MAIRIEVOX a permis de démontrer la faisabilité de telles applications

S(erveur) : *Bonjour, ici MAIRIEVOX. A tout moment, pour obtenir des renseignements sur MAIRIEVOX, utilisez le mot GUIDE. Et maintenant, à vous :*

Dites "SERVICES DE GARDE", "LOISIRS" ou "MAIRIE"

U(tilisateur) : **LOISIRS**

S : *Dites "SPECTACLES", "MANIFESTATIONS LOCALES" ou "CINEMAS"*

U : **CINEMAS**

S : *Dites "LE CLUB", "LE TAPAL'OEIL" ...*

U : **LE CLUB**

S : *Programmes du cinéma LE CLUB' :*

"Rain Man", le 25, à ...

U : **SUIVANT**

S : *"Chinatown" ...*

U : **PRECEDENT**

S : *"Rain Man ...*

U : **AUTRE RUBRIQUE**

S : *Dites "LE CLUB", ...*

U : **GUIDE**

S : *Guide : ...*

...

Figure 4. — Exemple de dialogue entre un usager et MAIRIEVOX.

sur le réseau téléphonique. Parmi les domaines directement concernés, il faut citer : les télésondages, les jeux interactifs, les systèmes de messagerie, l'accueil automatique au niveau des standards d'entreprise...

D'ores et déjà, plusieurs systèmes industriels comparables, mais de plus forte capacité (60 voies téléphoniques ou plus) ont été mis en place et sont en cours d'expérimentation (certains avec plusieurs milliers d'appels par jour). D'autres sont attendus pour les prochains mois.

4. Conclusion et perspectives

Des progrès sensibles ont été accomplis ces dernières années, et des applications réelles ont montré que les techniques de reconnaissance ont à présent atteint, sous certaines conditions, un niveau de qualité acceptable par le grand-public. Cependant, elles ont aussi mis en lumière l'importance de l'ergonomie du dialogue pour la réussite d'une application vocale, et l'importance de problèmes techniques dont ne se préoccupaient guère les chercheurs jusque-là.

Ainsi, le **rejet des mots étrangers** au vocabulaire à reconnaître reste un problème pour lequel aucune solution satisfaisante n'a encore été trouvée (en mode indépendant du locuteur). Une meilleure tolérance vis-à-vis des **« défauts d'élocution »** des locuteurs (hésitations, répétitions, mots parasites...) paraît également un préalable nécessaire au développement de services nouveaux bien acceptés du grand public. C'est sur ces points prioritairement que porteront nos efforts de recherche au cours des prochaines années.

REMERCIEMENTS

Les auteurs tiennent à associer à ce papier tous les membres de l'équipe de reconnaissance du CNET ayant contribué à divers titres à ces travaux.

Manuscrit reçu le 14 décembre 1989.

BIBLIOGRAPHIE

[1] STONERIDGE TECHNICAL SERVICES, *Understanding Voice I/O markets, opportunities in the 80s,* 1984.

[2] PROBE RESEARCH Inc., *Speech recognition, the major market thrusts, 1988-1995,* New York, décembre 1988.

[3] F. JELINEK, *Continuous speech recognition by statistical methods,* Proc. IEEE, vol. 64, avril 1976.

[4] L. R. RABINER, B. H. JUANG, *An introduction to Hidden Markov Models,* IEEE ASSP Magazine, 1986.

[5] D. JOUVET, J. MONNE, D. DUBOIS, *A new network-based speaker independent connected word recognition system,* IEEE ICASSP-86, Tokyo, 1986.

[6] C. GAGNOULET, M. COUVRAT, *SERAPHINE, a Connected Word Speech Recognition System,* Proc IEEE ICASSP-82, Paris, 1982.

[7] D. JOUVET, *Reconnaissance de mots connectés, indépendamment du locuteur, par des méthodes statistiques,* Thèse Doctorat ENST, Paris, juin 1988.

[8] K. BARTKOVA, D. JOUVET, *Speaker-Independent Speech-Recognition Using Allophones,* Proc ICPhS 1987, Tallin, USSR, août 1987.

[9] J. P. TUBACH C. GAGNOULET, J. L. GAUVAIN, *Advances in speech recognition products from France,* Speech Technology Conference, New York, avril 1989.

[10] C. GAGNOULET, F. ZURCHER, J. TIRBOIS, T. SERRADURA, *PUBLIVOX : a voice controlled card pay-phone,* European Conf. on Speech Technology, Edinburgh, septembre 1987.

[11] C. GAGNOULET, J. DAMAY, *MAIRIEVOX : a speech activated voice information system,* Eurospeech, Paris, septembre 1989.