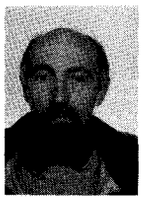
Intégration d'un système de reconnaissance analytique de la parole dans une console sonar : vers un dialogue naturel (**)

Integration of analytical speech recognition system in a sonar console : towards a natural dialogue (**)

E. GALLAIS
THOMSON-SINTRA/DASM BP 5
F-06801 Cagnes-sur-Mer Cedex

Après des études conventionnelles en informatique de Système à l'Université de PARIS VI, E. GALLAIS est diplômée de Conception de Systèmes (1974) et se spécialise en Formalisation de Systèmes d'E/S et Gestion de Bases de données. En 1979 elle obtient le diplôme d'ethno-linguistique à l'Institut National des Langues et Civilisations Orientales, s'étant spécialisée dans les relations entre phonologie et cosmogonie dans les langages proto-maori. De 1971 à 1978, est responsable des systèmes de bases de données de l'un des deux centres de calcul d'AIR FRANCE. Elle contribue également à l'introduction à AIR FRANCE des méthodes de programmations structurées. Elle entre à THOMSON Activités Sous-Marines (CAGNES/MER) en 1980 pour y exercer la responsabilité des logiciels des Systèmes de Reconnaissance Analytique de Parole.



P. ALINAT
THOMSON-SINTRA/DASM BP 5
F-06801 Cagnes-sur-Mer Cedex

Pierre ALINAT est diplômé de l'École Nationale Supérieure des Télécommunications (1966). Son intérêt pour le traitement de la parole débuta dès cette époque grâce à un projet sur le décodage de la « voix hélium ». Travaillant à Thomson ASM depuis 1968, il y mène en parallèle des études sur d'une part la reconnaissance analytique de la parole, d'autre part, en ASM, l'initialisation, la poursuite, le regroupement de pistes sonars et la classification des contacts obtenus.

G. SOUVAY
CRIN/INRIA-Lorraine BP 239
F-54506 Vandœuvre-Lès-Nancy Cedex

Gilles SOUVAY est titulaire d'un DEA en informatique. Il prépare une thèse depuis 1987 au CRIN-INRIA Lorraine dans l'équipe *Reconnaissance de Forme et Intelligence Artificielle* dans le cadre du projet DIALOGUE. Il travaille plus particulièrement à la réalisation de systèmes de dialogue pour des applications de type commande de processus.

(**) Cette étude a été financée par la Direction de Recherches Études et Techniques du Ministère de la Défense.

Intégration d'un système de reconnaissance analytique

J. M. PIERREL
CRIN/INRIA-Lorraine BP 239
F-54506 Vandœuvre-Lès-Nancy Cedex

Jean-Marie PIERREL, 38 ans, est professeur d'informatique à l'université de Nancy I. Depuis plus de 15 ans, il travaille en reconnaissance et compréhension de la parole continue au sein du CRIN (Centre de Recherche en Informatique de Nancy), URA 262 du CNRS. Après avoir successivement défini et mis en œuvre les systèmes MYRTILLE I et II, travail qui fut couronné par le prix scientifique IBM-France en informatique, il orienta ses recherches plus particulièrement vers le dialogue oral homme-machine puis aujourd'hui vers le dialogue multi-mode. Il est aujourd'hui directeur adjoint du CRIN et responsable, au sein de l'équipe *Reconnaissance des Formes et Intelligence Artificielle*, d'un projet de recherche « Parole et dialogue » commun au CRIN et à l'INRIA.

RÉSUMÉ

Nous présentons dans cet article l'intégration d'un système de reconnaissance analytique de la parole dans une console sonar. Cette application correspond à un besoin des opérateurs qui ont continuellement les yeux occupés à scruter l'écran sonar. Le système DIAPASON présente deux particularités :

— le décodage acoustico-phonétique utilisé dans le système n'est pas fondé sur une reconnaissance phonétique classique ; son objectif est d'obtenir pour chaque segment de parole une description précise en termes de traits acoustico-phonétiques. Ce niveau de décodage est associé à une procédure spécifique d'accès lexical et de reconnaissance de phrases ;

— le système DIAPASON est de plus un véritable système de dialogue homme-machine et ne se limite pas à une simple reconnaissance de phrases comme c'est le plus souvent le cas. L'historique du dialogue est considéré comme une source de connaissances à part entière durant la phase de reconnaissance et cela permet d'augmenter de façon notable les performances globales du système.

Cet article détaille l'architecture de DIAPASON et décrit ses diverses composantes. Il présente aussi les résultats expérimentaux obtenus en mode multilocuteurs et compare les systèmes de dialogue avec parole et sans parole pour une console sonar réelle.

MOTS CLÉS

Reconnaissance de la parole, méthodes analytiques, dialogue homme-machine, reconnaissance multilocuteur.

SUMMARY

We present in this paper the integration of a analytical speech recognition system to the control of a sonar console by a human operator. This application is really useful since it does correspond to a pratical need of the operator who has his eyes busy looking at the sonar screen. The DIAPASON system presents two original features :

— the acoustic-phonetic decoding part of the system is not based on a classical phoneme labeling process but it yields for each segment of speech a set of acoustic phonetic labels that describe this segment very precisely. This phonetic labelling is associated with a special procedure for lexical access and sentences recognition ;

— the DIAPASON system is moreover a genuine man-machine dialogue system and not only a system capable of understanding a single sentence as it is often the case. The history of the dialogue is used as a special knowledge source during the understanding process. Pragmatic knowledge is thus intimately associated with the analysis of sentence ; this point highly encreas as the overall performance of the system.

The paper presents the architecture of DIAPASON and its various components. It is also discusses experimental results obtained in a multispeaker mode and compares the voice dialog system with the without voice system for a real sonar console.

KEY WORDS

Speech recognition, analytical methods, man-machine dialogue, multi-speaker recognition.

1. Introduction

Depuis de nombreuses années, le traitement de la parole fait l'objet d'un effort de recherche important. Pourtant, dans le cadre de la communication homme-machine, l'utilisation de la parole reste très limitée et peut apparaître aujourd'hui comme une demi-réussite ou un demi-échec. Cela provient sans doute du fait que durant trop longtemps (jusqu'au début des années 1980) l'effort de recherche a porté quasi exclusivement sur la reconnaissance et la compréhension de phrases. Or, pour permettre une utilisation effective de la parole dans la communication homme-machine (accès à une base de données, de renseignements ou de connaissances, commande de processus industriels, dialogue avec une console...), il faut certes un système de reconnaissance de la parole robuste, fiable et fonctionnant en temps réel, mais cela ne suffit pas, il faut aussi exploiter le résultat de cette compréhension dans le cadre de l'application. Cela nécessite [1] :

— de valider les résultats de la reconnaissance et/ou de la compréhension. Divers types d'erreurs ou d'ambiguïtés peuvent, en effet, apparaître au terme de la phase de reconnaissance ou de compréhension : mauvaise compréhension globale ou partielle, ambiguïtés lexicales, syntaxi-

ques ou sémantiques, score de reconnaissance trop faible, etc. ;

— d'effectuer une compréhension contextuelle, ce qui est particulièrement important si l'on souhaite traiter les références, ellipses ou anaphores ;

— d'engendrer les réponses aux demandes de l'utilisateur et d'intégrer la parole aux autres media de communication : graphisme, clavier, système de désignation... ;

— de prévoir les enchaînements qui permettent d'atteindre le but spécifié par la demande initiale de l'usager et, par voie de conséquence, de déterminer le but à atteindre et d'être capable de l'intégrer au mieux dans l'application mise en œuvre.

Comme il ne serait pas réaliste de vouloir gérer et comprendre des dialogues quelconques, nous nous limiterons aux dialogues coopératifs, orientés par la tâche, ou dialogues finalisés. Dans de telles situations, le but à atteindre et la tâche à gérer ne limitent pas uniquement la diversité des actes de parole ou de discours [2], ils influent également sur le niveau de langage utilisé par l'usager et conduisent à la définition de sous-langages qui sont, avec les langages artificiels (éventuellement à consonance naturelle), les seuls types de langages que l'on puisse espérer comprendre automatiquement à ce jour. La structure et le lexique de tels sous-langages sont assez restreints et peuvent être facilement définis à partir des spécifications de l'application mise en œuvre. Cela simplifie donc le problème général de compréhension et de gestion de dialogues — certaines difficultés non encore surmontées sont évacuées — et correspond au domaine des applications actuellement envisageables. On trouvera dans [3] une présentation critique et détaillée, que nous ne reprendrons pas ici, des diverses solutions possibles pour introduire une composante de dialogue oral dans un système de communication homme-machine.

Pour montrer l'intérêt de l'introduction de la parole dans un système de dialogue avec une console, nous avons choisi une application réelle, le dialogue d'une console sonar et monté une expérience incluant à la fois un système de reconnaissance très robuste et un module de dialogue souple. Après la présentation générale du contexte d'application, nous détaillons dans la suite le système de reconnaissance utilisé et le système de dialogue (DIAPASON) mis en œuvre en précisant les résultats obtenus tant en termes de scores de reconnaissance qu'en termes de validation ergonomique.

2. Généralités et architecture du système

2.1. DESCRIPTION DES INTERACTIONS PRISES EN COMPTE

2.1.1. Contexte de l'étude

En vue d'étudier l'intégration d'une entrée vocale aux autres moyens d'interface homme-machine d'une console et de pouvoir se rendre compte des avantages et inconvénients de ce mode d'entrée, on s'est volontairement placé dans le cas réaliste de la console d'un matériel existant, en l'occurrence un sonar. L'introduction d'une entrée vocale sur une console nécessite des modifications très importantes du système d'interaction avec l'opérateur et il était donc hors de question de modifier directement le matériel. Compte tenu de cela, on a préféré simuler, sur un PC connecté à un système de reconnaissance, les images et commandes de la console sonar réelle. En disposant le sonar réel et le PC côte à côte dans le même local, il a été possible de faire des comparaisons entre l'interaction avec commande vocale et celle sans, pour les mêmes sous-tâches élémentaires.

Les commandes du sonar réel (sans entrée vocale) sont essentiellement du type menu. L'opérateur dispose pour cela d'un clavier logiciel et de 2 commutateurs. Il dispose également d'une boule pour déplacer un curseur. Sur le sonar choisi, il y a 2 images : l'image « initialisation » et l'image « veille ».

— L'image initialisation permet la gestion du contenu des mémoires de visualisation et la gestion des paramètres de base du sonar.

— L'image veille est composée d'une image (azimut, temps) sur laquelle sont représentées les pistes, c'est-à-dire les détections successives dans le temps d'un même but, avec plus ou moins de détails et d'une image annexe grâce à laquelle l'opérateur peut avoir des renseignements complémentaires sur les pistes. Sur cette image de veille, l'opérateur peut régler les échelles, décider des informations à visualiser, pointer des traitements particuliers dans des directions déterminées, régler les paramètres de certains traitements et envoyer des renseignements vers l'extérieur.

2.1.2. Présentation du sonar simulé sur PC

Le sonar simulé sur PC (avec entrée vocale) dispose des mêmes images et commandes que le sonar réel. Le correspondant vocal des menus est constitué par des phrases de « mots enchaînés ». L'opérateur peut également donner les mêmes ordres au clavier mais, par souci de simplification, il ne lui est pas donné la possibilité de passer aussi ses commandes au moyen de menus à cliquer sur l'écran puisqu'il existe l'équivalent sur le sonar réel. Par ailleurs, la souris permet de déplacer le curseur. Au point de vue commande vocale, un vocabulaire de 125 mots est suffisant pour permettre toutes les commandes du sonar réel. La syntaxe est du type « mots enchaînés » utilisant 3 sous-vocabulaires-syntaxes différents :

— quel que soit le contexte,
— en contexte initialisation seulement,
— en contexte veille seulement.

Bon nombre de mots appartiennent à deux ou trois des sous-vocabulaires et à chaque instant, deux des sous-vocabulaires sont actifs. Les phrases de commande sont bâties selon les modèles :

mot de commande - mot paramètre - mot paramètre...

avec la possibilité de quatre versions différentes selon l'état du dialogue et du contexte, par exemple :

— modèle de base : « remplacer mémoire pointée par Demon 1 »
(dans cet exemple « *Demon 1* » et « *mémoire pointée* » sont des objets manipulés par l'opérateur du sonar)
— modèle elliptique : « remplacer par Demon 1 »
« mémoire pointée par Demon 1 »
« par Demon 1 »
— modèle de correction : « négatif par Demon 1 »
« négatif mémoire pointée »
— modèle de réponse à question : « mémoire pointée »
« par Demon 1 ».

Contrairement au cas du sonar réel et en grande partie du fait de l'utilisation de l'entrée vocale, sur le sonar simulé, l'opérateur dispose d'un dialogue constitué d'une suite d'échanges homme-machine via les divers modes de communication offerts. Le dialogue élémentaire, c'est-à-dire le dialogue nécessaire à la compréhension d'une commande, se déroule en 3 phases :

— formulation par l'opérateur de l'ordre initial,
— affichage graphique de ce qui a été compris par le système. Cet affichage se traduit le plus souvent possible par une modification de l'image plutôt que par un commentaire alphanumérique en bas de l'écran,
— mises au point diverses par l'opérateur (s'il y a lieu) : lever d'ambiguïtés, correction d'erreur, confirmation pour les ordres ayant des conséquences graves.

L'ordre peut alors être envoyé au sonar proprement dit dans la mesure où les 2 parties sont d'accord.

Voici un exemple de dialogue élémentaire où l'on considère des pistes baptisées par une lettre suivie d'un chiffre :

Opérateur :	« effacer piste A1 »	(en fait, il se trompe car A1 n'existe pas : il voulait dire A2)
Système :	« A1 n'existe pas »	(l'opérateur s'en aperçoit essentiellement du fait que la piste A2 ne s'efface pas)
	« que dois-je faire ? »	
Opérateur :	« négatif A2 »	étant donné qu'il n'y a pas besoin de confirmation pour cet ordre (qui est sans gravité) le système exécute en effaçant A2. L'opérateur constate directement sur la visualisation que la piste A2 qu'il était en train de regarder s'efface.

Les modèles de correction et de réponse sont donc utilisés dans les dialogues élémentaires.

Une mémoire du dialogue à moyen terme (ici de profondeur 1, c'est-à-dire que l'on tient compte de l'avant-dernier ordre donné par l'opérateur) permet d'enchaîner des ordres syntaxiquement incomplets faisant référence à des ordres précédemment exécutés. C'est dans ces cas que les modèles elliptiques sont utilisés. Ci-après un exemple de tel dialogue à moyen terme (ici « *lofar 1* » est le nom d'un des objets manipulés par l'opérateur du sonar. « *Graphique annexe* » est le nom d'une des zones graphiques de l'image veille) :

« graphique annexe lofar 1 gamme A »
puis « gamme B » (c'est-à-dire graphique annexe-lofar 1 gamme B)
puis « lofar 2 gamme D » etc.

L'opérateur peut changer de mode d'expression (voix ou clavier) entre les échanges d'un dialogue élémentaire. Ceci est en particulier nécessaire en cas de bruit d'ambiance devenant nettement trop fort. La désignation d'un objet à l'aide de la souris doit précéder l'énoncé (ou la frappe) de l'ordre qui portera sur cet objet. En revanche, la mixité des modes d'expression est interdite à l'intérieur d'une même commande.

2.2. ARCHITECTURE DU SYSTÈME

La figure 1 illustre l'architecture générale du système. Il se décompose en 4 parties :

— Les senseurs d'entrée : clavier-souris, reconnaissance de parole. Ce dernier système fonctionne avec un vocabulaire-syntaxe fourni par le système de compréhension selon l'état du dialogue, du contexte sonar ou de la visualisation. Lorsqu'une phrase a été prononcée, le système de reconnaissance envoie les n ($n \leq 3$) phrases reconnues de meilleurs scores dans la mesure où ces scores sont suffisants.

— Le dispositif de sortie : le seul disponible est l'écran de visualisation. La synthèse vocale ne pouvait être utilisée ici car les opérateurs ont leur ouïe occupée à d'autres tâches.

Sur l'écran, 2 types d'informations sont représentées, parfois de façon mêlée :
+ les sorties du système sonar,
+ ce qui est nécessaire au dialogue, la plupart du temps « en place » c'est-à-dire là où l'opérateur est censé regarder.

— Le système de compréhension multimodal reçoit les ordres donnés par l'opérateur via les senseurs et cherche à en saisir la signification grâce à la connaissance du vocabulaire, de la syntaxe et de la sémantique, de l'état du sonar et de la visualisation, de l'état du dialogue et de son passé récent. Lorsqu'un ordre semble correctement compris, le système de compréhension l'envoie au sonar.

— Le système sonar : il s'agit du sonar proprement dit, c'est-à-dire, le système avec lequel l'opérateur interagit.

3. Le système de reconnaissance analytique

3.1. GÉNÉRALITÉS

3.1.1. Reconnaissance analytique versus globale

Le système de reconnaissance utilisé est de type analytique ce qui signifie qu'il met en jeu des connaissances relatives à l'analyse de la parole (traits, phonèmes, syllabes, mots) notamment et spécialement pour les plus bas niveaux.

Les méthodes analytiques s'opposent aux méthodes dites globales qui, après analyse de la parole (DFT ou modèle AR ou Cepstre...) cherchent à reconnaître directement des mots en les comparant à de simples copies déterministes ou à des chaînes de Markov, sans se soucier des niveaux traits (c'est-à-dire, par exemple, voisement, friction, position des formants...), phonèmes et syllabes.

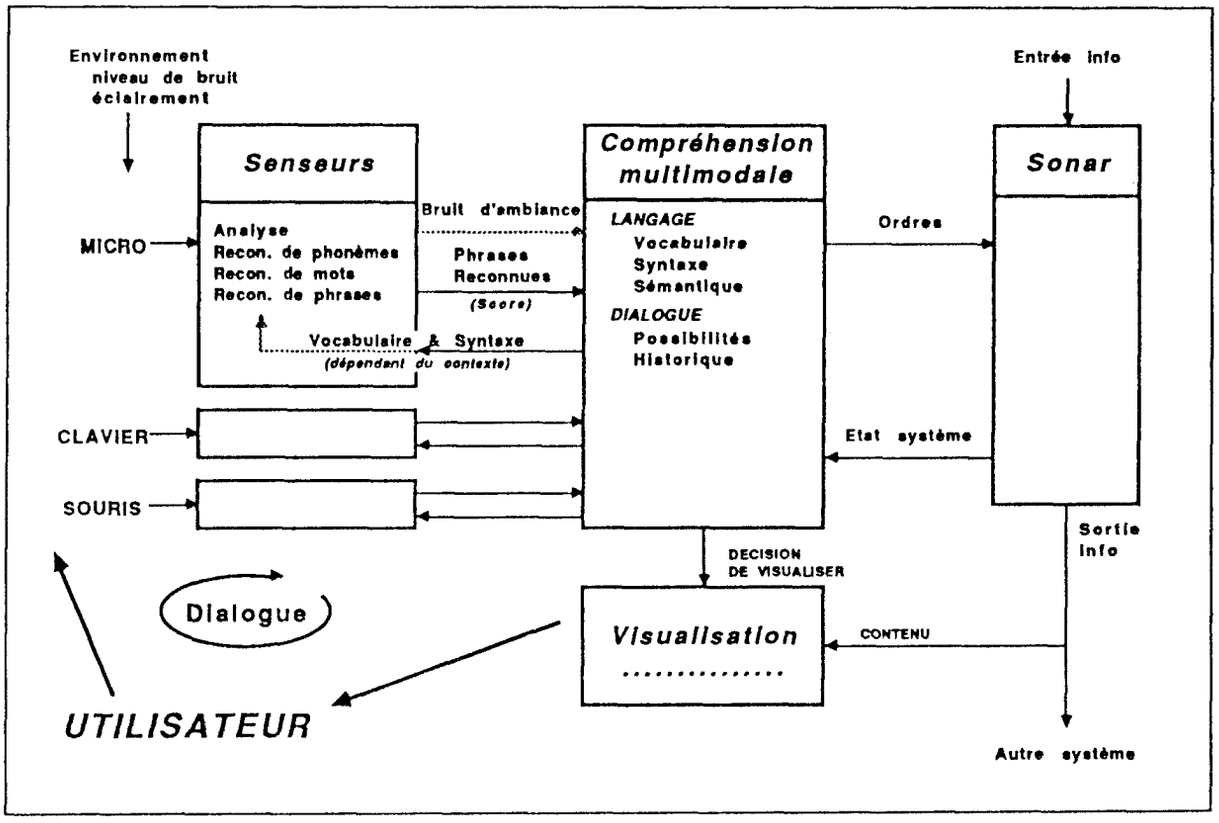


Figure 1. — Architecture du système DIAPASON.

Bâtis sur des principes simples, les systèmes globaux sont efficaces dans la mesure où ils sont employés dans des conditions stables. De plus, ils présentent l'avantage d'être valables pour des ensembles de langues voisines. Du fait de leur simplicité, les systèmes globaux sont de loin les plus étudiés et les plus répandus, notamment pour ce qui est des systèmes en vente actuellement. Toutefois, n'utilisant que peu de connaissances relatives à la parole, ces systèmes globaux font plutôt de la reconnaissance de sons que de parole et il en découle nombre de défauts et limitations. Tout d'abord, ces systèmes sont sensibles aux variations de prononciation intra ou interlocuteur et aux variations de bruit d'ambiance, d'où la nécessité de l'apprentissage et un compromis entre performances, nombre de mots du vocabulaire et nombre de locuteurs utilisant le système. Par ailleurs, ces systèmes sont mal adaptés à la prise en compte des coarticulations entre phonèmes, des liaisons entre mots et des mots outils (articles, propositions, pronoms...) inévitables dès qu'on laisse un minimum de libertés aux locuteurs.

Enfin, la part du volume de calculs proportionnelle au nombre de mots possibles est trop importante, ce qui est gênant dès qu'on veut utiliser un vocabulaire étendu (avec facteur du branchement, c'est-à-dire un nombre de possibilités à chaque instant, important).

Les systèmes analytiques cherchent à éviter ces défauts en prenant mieux en compte les connaissances qu'on peut avoir sur la nature profonde de la parole. Ils ont leur origine dans un effort de recherche très ancien : la phonétique qui a permis en particulier d'établir les classifications de phonèmes. Plus près de nous, dès les années 40, les moyens d'analyse et de synthèse de la parole déduits du Vocoder ont permis de commencer l'étude quantitative des traits aux Haskins Laboratories en particulier [4]. Des études de systèmes de reconnaissance plus ou moins analytique ont lieu depuis longtemps (par exemple [5], le phonétographe [6], le système Keal [7] ou le présent système [8]) mais du fait de la complexité des connaissances à mettre en œuvre, la progression a été (et est) bien plus lente que dans le cas des systèmes globaux. L'effort en reconnaissance analytique continue actuellement malgré la difficulté de l'approche, en particulier le projet SUMMIT au MIT [9] utilisant un modèle de cochlée comme analyseur, le travail [10] au CMU, l'utilisation de réseaux d'automates flous par Gubrynowicz [11], le système de GEC combinant l'utilisation de traits phonétiques et des modèles de Markov cachés [12], les systèmes experts de lecture de spectrogrammes [13, 14] couplé à une reconnaissance de phonèmes par réseaux neuronaux dans le cas de [15], le travail de Méloni au GIA de Luminy [16] sans oublier les travaux plus fondamentaux en décodage acoustico-phonétique du GRECO-PRC « Communication homme-machine » [17]. Cet effort vaut la peine car il apparaît que les systèmes vraiment analytiques (c'est-à-dire prenant en compte le niveau trait) tel celui utilisé ici,

présentent de réels avantages. Tout d'abord, ils permettent d'éviter la phase d'apprentissage pour les locuteurs, c'est-à-dire, qu'ils sont tolérants aux variations de prononciation de chaque locuteur et, de plus, multilocuteur. De plus, ils sont adaptés à une prise en compte simple et efficace (au point de vue calcul) des coarticulations, des liaisons et des mots outils en parole continue. Enfin, bien que cela ne soit point encore employé, ils devraient pouvoir permettre une lente adaptation pour un locuteur utilisant le sustème suffisamment longtemps et permettre une prise en compte des diverses caractéristiques des bruits d'ambiance. Ces avantages se paient par quelques inconvénients : le coût du calcul dû à la mise en œuvre des niveaux traits et phonémes et la restriction à une seule langue. En particulier, pour passer à une autre langue, il faut adapter ou modifier toutes les connaissances mises en œuvre.

3.1.2. Difficultés de la reconnaissance analytique

On sait depuis longtemps que la parole normale n'est généralement pas constituée d'une suite de phonèmes adjacents et indépendants [4]. En effet, bien qu'il soit raisonnable de penser qu'une telle suite existe plus ou moins quelque part dans le cerveau des interlocuteurs, les sons transmis sont le résultat d'un codage effectué sur cette suite de phonèmes. Ce codage est complexe et peut varier selon l'accentuation des phonèmes et leur importance pour la compréhension de la phrase compte tenu du contexte, c'est-à-dire, tout ce qui est censé être connu des auditeurs, sans oublier le bruit d'ambiance. Ce codage se traduit par des déformations de certains traits en fonction des phonèmes voisins (coarticulation) même pour de la parole articulée avec soin, par des relâchements de prononciations des phonèmes les moins accentués qui peuvent se réduire à une simple modification d'un phonème voisin ou à la limite qui peuvent ne pas être prononcés du tout. Il résulte de tout cela que la parole est certes redondante par rapport aux informations transmises mais bien moins que ne le serait la suite des phonèmes parfaits. Tout ceci est la conséquence de ce que le locuteur cherche à se fatiguer le moins possible et à transmettre son message le plus vite possible tout en restant compréhensible pour les auditeurs.

On sait également qu'un auditeur humain ne perçoit pas de façon séquentielle la suite des phonèmes plus ou moins codés qu'il reçoit [18]. En fait, pour pouvoir traiter la parole suffisamment rapidement les auditeurs humains semblent utiliser une stratégie de test cherchant à établir quelle est l'hypothèse qui correspond le mieux à la parole reçue parmi les hypothèses faites à partir du contexte, notamment de ce qui a été compris jusque-là. Ces tests ne nécessitent normalement pas de percevoir dans le détail tout ce qui a été prononcé. A ce niveau, il faut noter l'importance du rythme syllabique et des variations du fondamental.

Pour mieux appréhender ces phénomènes, il est possible de faire un parallèle avec l'écriture manuscrite dont en général toutes les lettres sont loin d'être bien formées et à la limite peuvent même être remplacées par un trait informe. Là encore la perception, même dans le cas de l'écriture imprimée, n'est pas séquentielle [19]. Les blancs séparant les lettres et les mots et la largeur régulière des lettres jouent un grand rôle.

Pour être efficace, un système analytique doit tenir compte des remarques précédentes. Il faut en particulier, choisir la liste des traits utiles à partir de la nombreuse littérature sur le sujet, mais en sachant se limiter aux plus pertinents. Il faut également savoir détecter ces traits et, pour certains, estimer correctement les paramètres correspondants. Enfin, il faut savoir différer les décisions le plus longtemps possible car, dans un système de type hypothèse-test, la vraie décision se prend à un niveau élevé. En particulier, il est illusoire de vouloir prendre des décisions définitives trop tôt.

3.2. LE SYSTÈME DE RECONNAISSANCE MIS EN ŒUVRE

3.2.1. Description de la reconnaissance phonétique

Le système de reconnaissance analytique utilisé ici est composé d'une chaîne de traitement ascendante — du signal acoustique vers les phonèmes — analyseur, estimation des paramètres acoustiques, localisation des phones (on appelle ainsi par simplification les phonèmes reconnus), estimation des paramètres des phones, et d'une chaîne de traitement descendante — de la phrase vers les phonèmes — vocabulaire-syntaxe, analyse syntaxique, comparaison chaîne de phones (reconnus), chaînes de phonèmes (vocabulaire). De plus la stratégie générale est dans l'ensemble gauche-droite avec de légers retours en arrière possibles à l'intérieur des syllabes.

L'analyseur est constitué par un banc de 48 filtres passe-bande linéaires suivis chacun d'une détection intégration. Les fonctions de transfert de ces filtres ont été choisies d'après ce qui est connu de la partie mécanique de la cochlée humaine tant par mesure directe des déformations de la membrane basilaire [20], [21] que par mesure de psycho-acoustique [22]. Il s'agit d'un modèle grossier de cochlée artificielle ne prenant en compte ni les non-linéarités diverses, ni les traitements effectués par les premières couches du système nerveux [8]. Malgré cela, ce modèle de cochlée donne déjà des résultats d'analyse qui permettent de rendre plus simples ou plus robustes certains critères. Il a été ainsi possible de mettre au point un système très simple, bien qu'efficace, de classification des voyelles (les zones formatiques) [8] et de préciser la détermination de la place d'articulation des plosives (labiale, dentale, palatale) [23]. De ce fait, le système d'analyse est un maillon de la chaîne de reconnaissance aussi important que les autres et à ce titre il conditionne pour une bonne part les résultats.

Un certain nombre de paramètres dits acoustiques sont estimés toutes les 8 ms. Il s'agit de paramètres qui sont utilisés par la suite pour vérifier l'existence des traits. Bien que certains ne seront utilisés que sur quelques intervalles de temps, c'est par souci de simplification qu'ils sont estimés systématiquement toutes les 8 ms. Il s'agit de l'énergie, du voisement et de la friction (présence ou absence), des explosions, du premier et du second formant, du degré de nasalisation, de la forme du burst, de la forme des phases soutenues des plosives (sourde, sonore,

nasale) et de la période du fondamental. Les formants sont définis à partir des pics du « spectre » fourni par l'analyseur en utilisant tout un ensemble de règles visant d'une part à affranchir le premier formant F1 des déformations dues au peigne de raies du fondamental, notamment pour les voix de femmes, et d'autre part à choisir correctement le deuxième formant F2 parmi toutes les possibilités. Il faut noter qu'il ne s'agit pas forcément ici des 2 premiers pôles de la fonction de transfert du conduit vocal mais plutôt d'informations déduites de façon déterministe de la forme du spectre. En particulier, il peut arriver que les 2ème et 3ème pôles interviennent ensemble dans la détermination de F2 lorsqu'il est élevé. Lorsque cela semble nécessaire, des hésitations sont notées pour chacun des 2 formants. A ce niveau, le mot formant n'est utilisé que par commodité car il ne signifie pas ici qu'une décision phonétique a été prise au sujet de la présence d'une voyelle ou d'une semi-voyelle.

La détection de présence de parole est basée essentiellement sur l'occurrence de signaux voisés avec un S/N acoustique suffisant. On suppose donc ici avoir affaire à de la parole normale et non de la parole chuchotée pour laquelle il faudrait introduire d'autres critères.

La localisation de phones mise en place ici est très différente de la segmentation traditionnelle [9], [10], [15]. En effet, il ne s'agit pas de découper le signal d'entrée en segments adjacents mais plutôt de détecter la présence de telle ou telle catégorie de phones. Les différentes catégories (voyelles, semi-voyelles, fricatives, plosives et plosives nasales, phonème R) sont recherchées en parallèle en utilisant des traits propres à chaque catégorie. Chacun de ces traits est la traduction même de la définition de la catégorie de phones en question et se retrouve généralement plus ou moins dans le nom de cette catégorie. Par exemple, une consonne fricative est repérée par un maximum d'énergie de friction et une plosive par la présence d'une explosion. Normalement, ces traits sont recherchés en relatif par rapport au voisinage (temporel ou fréquentiel) de façon à disposer d'une certaine robustesse vis-à-vis des variations de prononciation, du bruit et des distorsions de transmission. La recherche des phones est effectuée en deux étapes : d'abord les phones les mieux prononcés (entre la moitié et les deux tiers) puis les autres. La recherche étant effectuée indépendamment pour chaque catégorie, il faut ensuite introduire la notion de syllabe et régler les conflits. Parfois, des hésitations peuvent être conservées. Il peut également arriver qu'un phone ainsi détecté corresponde en fait à deux phonèmes successifs : par exemple un phone voyelle peut être le résultat de la fusion de deux phonèmes voyelle à la jonction entre 2 mots donnant plus ou moins un phone diphtongue.

Pour chaque phone détecté et selon sa catégorie un certain nombre de paramètres sont estimés. Ces paramètres sont relatifs aux différents traits impliqués dans la description de chaque catégorie. Certains concernent le phone en tant que faisant partie de sa catégorie (par exemple pour les voyelles : amplitude, durée, force de voisement...), les autres concernent la classification fine du phone, c'est-à-dire à l'intérieur de sa catégorie (par exemple pour les voyelles description de la position des formants et degré de nasalisation).

3.2.2. Résultats de la détection de phones

On obtient ainsi une chaîne de phones, chacun étant décrit par son nom et par ses paramètres estimés, étant entendu que des hésitations sont possibles et que certains phones comportent des indications permettant de penser en fait qu'ils correspondent à 2 phonèmes adjacents. Étant donné, comme cela a été dit précédemment, que la parole émise est un codage complexe et variable d'une chaîne phonémique idéale, il est malaisé de chiffrer les performances d'un système de reconnaissance à ce niveau car elles doivent être évaluées non par comparaison à la chaîne idéale mais par comparaison à la chaîne réelle telle qu'un spécialiste la perçoit. De plus, l'évaluation porte sur un grand nombre d'informations parfois dépendantes. De façon succincte en multilocuteur (12 locuteurs, 8 locutrices), les erreurs sont 9 % de phones en trop, 4 % de phones omis et 4 % d'erreurs ou rejets de la catégorie. En général, les paramètres sont extraits correctement sauf la nasalisation des voyelles et la place d'articulation des plosives.

3.2.3. Principe de la reconnaissance de mots et de phrases

Pour chaque mot hypothèse fourni par l'analyseur syntaxique, une comparaison chaîne de phones reconnus-chaîne de phonèmes du mot vocabulaire est effectuée. La meilleure correspondance entre les 2 chaînes est obtenue par programmation dynamique en utilisant les notes fournies par les comparaisons élémentaires phone reconnu-phonème vocabulaire. Ces comparaisons élémentaires font intervenir :

- des règles de fonction,
- des règles de coarticulation,
- les forces des phones et accentuation des phonèmes.

En tenant compte, bien entendu de :

- l'identité du phonème vocabulaire,
- l'identité du phone reconnu.

Les règles de coarticulation sont ici au nombre d'une vingtaine environ, ce qui est suffisant pour couvrir une large proportion des coarticulations qui se produisent. Ces règles ne sont jamais obligatoires mais en définitive elles s'avèrent très utiles pour la reconnaissance. Elles sont prises en compte en se basant sur la chaîne de phonèmes. Il s'agit par exemple des règles telles que :

— (/S/ ou /Z/) suivi d'une voyelle (ou vice-versa)
⇒ F2 de la voyelle peut être déplacée vers le haut.

— Voyelle orale suivie d'une plosive nasale ou voyelle nasale (ou vice-versa)
⇒ nasalisation de la voyelle orale peut être augmentée.

De même une dizaine de règles de jonction entre mots sont employées. Par exemple :

— [mot i terminé par une voyelle] suivie [même voyelle débutant le mot $i + 1$]
⇒ les 2 voyelles identiques peuvent être fusionnées en une seule plus longue.

Les règles de classification ne sont rien d'autre que l'utilisation de la définition (sous forme de traits) des différents phonèmes.

La syntaxe de base est une syntaxe de mots enchaînés. Des extensions permettent, lorsqu'elles sont activées, des élisions de certains mots, des ordres de correction et des réponses aux questions du système. La recherche de la phase prononcée est effectuée en faisceau, c'est-à-dire qu'on conserve à chaque étape les N meilleures hypothèses de phrases. Dans le cas des phrases elliptiques, il peut arriver qu'on obtienne deux hypothèses correspondant à la même suite de mots mais à des analyses grammaticales différentes. L'ambiguïté ne pourra alors être éventuellement levée qu'à l'étage compréhension.

3.2.4. Résultats de la reconnaissance de mots et de phrases

Avec un vocabulaire de 100 mots et des facteurs de branchement de l'ordre de 15 à 30, les résultats suivants ont été obtenus en multilocuteurs (7 locuteurs, 3 locutrices) :

— phrases reconnues correctement 85 %
— mots reconnus correctement 95 %
— nombre de phrases nécessaires pour qu'une commande soit correctement exécutée : 116 %

Il n'y a pas de différence marquée de taux d'erreur entre les locuteurs et ces résultats ont été obtenus en ambiance de laboratoire, pour des vitesses de prononciation normale laissées au libre choix de chaque locuteur.

4. Compréhension et gestion de dialogue

4.1. LE DIALOGUE HOMME-MACHINE DANS DIAPASON

Il est constitué d'une suite d'échanges entre l'utilisateur et le système. Il se déroule en trois phases. L'opérateur donne un ordre, vient ensuite sa mise au point et enfin son exécution. DIAPASON est alors prêt à recevoir un nouvel ordre.

4.1.1. Notion d'ordre dans DIAPASON

Un ordre correspond à une tâche bien définie que l'opérateur veut réaliser. Quel qu'il soit, on distingue tout d'abord deux éléments ; la **commande** et la **liste des paramètres.** Le premier détermine le type d'action que l'on veut effectuer (affectation/libération d'une mémoire, affichage d'un bruiteur). Le second précise les objets sur lesquels la commande agit. Exemple : afficher (commande) bruiteur alpha 1 (paramètres). Leur nombre varie de zéro à deux.

4.1.2. La phase de mise au point

Afin de ne pas alourdir le dialogue, elle devra être la plus courte possible. Dès que DIAPASON a analysé une demande, il envoie en réponse un message. Dans le cas général, il s'agit de l'interprétation de l'ordre. Si le locuteur est satisfait on passe alors à la phase d'exécution. Dans le cas contraire, mauvaise reconnaissance ou erreur de l'opérateur, le système offre alors la possibilité de corriger la demande. La **correction** porte sur un ou plusieurs des éléments composant l'ordre.

Exemple : O : Afficher bruiteur alpha 2
D : Afficher bruiteur alpha 1
O : Négatif alpha 2
D : Afficher bruiteur alpha 2

Il se peut que le système détecte une incohérence dans les propos du locuteur. Il envoie alors un message indiquant exactement la nature du problème.

Exemple : O : Afficher bruiteur alpha 2
D : Il n'y a pas de bruiteur alpha 2

Le système peut demander un complément d'information.

Exemple : O : Afficher...
D : Afficher quel bruiteur ?
O : Alpha 2
D : Afficher bruiteur alpha 2

L'opérateur a la possibilité d'annuler un ordre.

Exemple : O : Effacer bruiteurs
D : Raz bruiteurs
O : Négatif
D : Que dois-je faire ?

4.1.3. Exécution d'un ordre

Lorsque la commande et les paramètres répondent à l'attente du locuteur, il faut confirmer l'ordre pour l'exécuter. La **confirmation** valide le couple (commande, paramètres). Elle peut être **explicite :** on impose au locuteur la confirmation systématique de l'ordre, on ne pourra rien faire d'autre tant qu'il ne sera pas validé ou infirmé ; **implicite :** si le système n'observe pas de contestation, il exécute l'ordre au bout d'un temps t ou lorsque la phrase suivante ne contient pas de correction ; **inexistante :** l'ordre s'exécute immédiatement sans intervention de l'opérateur.

Selon l'importance de la commande, on va lui associer un type de confirmation. Cela se fait en fonction de la gravité de l'ordre. Une libération de mémoire est une action qui fait perdre irrémédiablement des données. On lui associe donc une confirmation explicite. Un changement de paramètre (la cap par exemple) n'a pas de conséquence grave sur son fonctionnement. On lui associe une confirmation inexistante.

Exemple : libération d'une mémoire
Avec confirmation explicite
O : Libérer mémoire VPR
D : Confirmez : libérer mémoire VPR
O : Ok
D : Ok
exécution de libérer mémoire VPR
Avec confirmation implicite
O : Libérer mémoire VPR
D : Libérer mémoire VPR
O : Cap bâtiment
exécution de libérer mémoire VPR
D : Cap bâtiment
Avec confirmation inexistante
O : Libérer mémoire VPR
D : Libérer mémoire VPR
exécution de libérer mémoire VPR

Un ordre est donc composé de trois éléments :

ordre = commande + paramètres + confirmation

4.1.4. Enchaînement des ordres dans DIAPASON

Un point fort des systèmes de dialogue homme-machine finalisés est de pouvoir omettre certains éléments d'une phrase en fonction du contexte. DIAPASON permet ainsi l'ellipse sur la commande ou un des paramètres. L'ellipse sur la commande est utilisée pour effectuer une même opération sur un objet différent.

Exemple : O : Libérer mémoire VPR
D : Libérer mémoire VPR
O : Mémoire CLASS
D : Libérer mémoire CLASS

L'ellipse de paramètre est possible lorsque l'on effectue plusieurs actions sur un même objet.

Exemple : O : Afficher bruiteur alpha 1
D : Afficher bruiteur alpha 1
O : Lever de doute gauche
D : Lever de doute bruiteur alpha 1 gauche

4.2. LE FONCTIONNEMENT DE DIAPASON

DIAPASON est donc un système de dialogue oral homme-machine chargé d'exécuter des ordres. On distingue trois niveaux dans l'architecture du système : la reconnaissance de phrase **RECONN,** la gestion de la tâche **TACHE** et la gestion du dialogue **DIALOGUE.** Ce type d'architecture composé de modules non liés directement les uns aux autres entraîne une grande indépendance de DIAPASON vis-à-vis de l'application.

4.2.1. Le module de reconnaissance RECONN

Il y a deux modes de communication avec le système : l'écrit ou la parole. Il existe pour chacun un sous-module de reconnaissance :

— Le module de reconnaissance écrite **SYNTAXE,** reprend les principes développés pour MYRTILLE I [24], [25]. Il s'agit d'un système guidé par la syntaxe qui à chaque étape, émet des hypothèses sur les mots à reconnaître et les valide par un sous-module de reconnaissance de mots. SYNTAXE travaille à partir de grammaires fournies par DIALOGUE et lui transmet la représentation sémantique de la phrase analysée.

— Le module de reconnaissance orale **VOCAL,** est celui décrit dans le paragraphe 3. Tout comme SYNTAXE il reçoit des grammaires de travail et transmet la structure sémantique de la phrase reconnue.

La structure sémantique permet la transmission des données entre les niveaux RECONN et DIALOGUE. Compte tenu du caractère artificiel du langage utilisé dans DIAPASON, nous nous sommes fondés sur une modélisation par grammaires dites sémantiques. La construction de la structure se fait donc en recueillant les éléments pertinents dans les arbres syntaxiques fournis par SYNTAXE ou VOCAL.

4.2.2. Le module de gestion de la tâche : TACHE

Il dirige le fonctionnement de DIAPASON et contrôle le reste du système en formulant à DIALOGUE des requêtes. Son but est d'obtenir les éléments nécessaires à l'exécution d'un ordre : commande, paramètre et confirmation. Il s'efforce de gérer au mieux les incohérences qu'il pourrait détecter.

4.2.2.1. *Le fonctionnement de TACHE*

TACHE est un automate d'états finis dont le but est de se mettre dans un état de satisfaction (*fig.* 2).

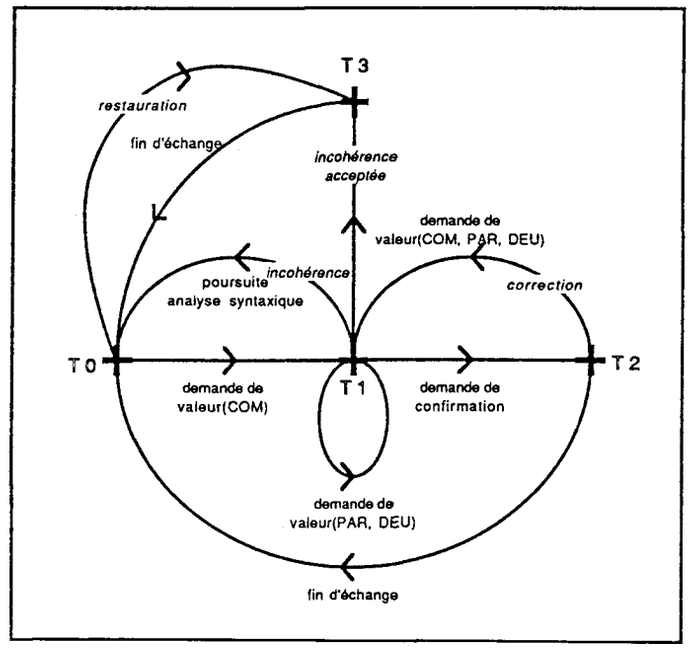


Figure 2. — Module de gestion de la tâche.

Les transitions sont constituées d'appels au module DIALOGUE, qui lui fournira successivement la commande, les paramètres (demande de valeur T0 → T1, T1 → T1) et demandera s'il peut se mettre dans l'état de satisfaction en fonction du type de confirmation associé à la commande (demande de confirmation T1 → T2).

Le fonctionnement ne se fait pas de manière linéaire. Il est influencé par les réponses fournies par DIALOGUE et revient dans un état antérieur dans le cas d'une correction (T2 → T1). L'incohérence peut provoquer un retour à l'état initial (T1 → T0) ou faire avancer dans un état de satisfaction (T1 → T3) paragraphe 4.2.2.2.

Enfin il retourne dans son état initial lorsqu'un ordre est accepté, signal de fin d'échange envoyé à DIALOGUE. Il l'exécute s'il n'est pas incohérent (T2 → T0) ou envoie un message d'erreur (T3 → T0).

4.2.2.2. *Le traitement de l'incohérence*

Il existe des conditions de validité pour les doublets (commande, paramètres). Lorsque l'une de ces règles est violée on dit qu'il y a **incohérence.** Elle a deux origines possibles : le niveau reconnaissance a proposé une

hypothèse ne correspondant pas aux propos du locuteur ou l'opérateur ne tient pas lui-même des propos cohérents. Une manière approximative de distinguer ces deux cas est de se fier aux scores de reconnaissance. Si l'on est au-dessus d'un certain seuil on considère que le locuteur s'est trompé et l'on envoie un message d'incohérence (T1 → T3). Si l'on est en-dessous on la refuse et on poursuit l'analyse en ramenant TACHE dans son état initial (T1 → T0). Dans ce cas de figure, on explore successivement toutes les hypothèses fournies par RECONN. On arrête la recherche lorsque l'on rencontre un ordre cohérent ou lorsqu'il n'y a plus d'hypothèse. Dans ce dernier cas TACHE reçoit un signal de restauration qui lui fait récupérer la première hypothèse rejetée (T0 → T3).

4.2.3. Le module de gestion du dialogue : DIALOGUE

Le module de gestion du dialogue joue un rôle entre RECONN et TACHE. Il doit répondre aux requêtes formulées par TACHE. Pour cela il fait appel soit au module RECONN, soit à sa mémoire suivant une stratégie préétablie.

4.2.3.1. *Le fonctionnement de DIALOGUE*

DIALOGUE est un automate d'états finis. Ses transitions sont plus complexes que dans le cas de TACHE car elles sont liées à des appels aux deux modules TACHE et RECONN et à des événements internes.

Les échanges DIALOGUE → RECONN sont de trois natures. La **demande d'hypothèse** est formulée lorsque l'on détecte de la parole ou un accès au clavier. L'**acceptation d'hypothèse** indique au module de reconnaissance concerné qu'il a terminé son travail. Il est réinitialisé à partir de données fournies par DIALOGUE. Le **rejet d'hypothèse** demande à l'analyseur de fournir l'hypothèse suivante.

Les échanges DIALOGUE → TACHE correspondent tout d'abord aux réponses aux requêtes formulées par TACHE : l'**envoi de valeur** pour la demande de valeur, la **réponse de confirmation** pour la demande de confirmation. Il existe deux autres requêtes permettant la synchronisation : le **signal d'activation** qui met TACHE dans son état initial et le **signal de restauration** pour la gestion de l'incohérence.

4.2.3.2. *La mémoire du dialogue*

La mémoire du dialogue a deux composantes. La première correspond à la mise au point de l'ordre courant. Le but de DIAPASON est d'arriver dans l'état de satisfaction de TACHE en recueillant commande, paramètres et satisfaction. Cette collecte peut se faire avec plusieurs interventions de l'opérateur (demande de complément d'information du système, corrections de la part de l'utilisateur). La **mémoire à court terme** conserve toutes les hypothèses proposées par le niveau reconnaissance au cours de cette mise au point. A chaque phrase on associe un **statut** qui est fonction de la réponse fournie par TACHE (refus de la commande ou du paramètre pour cause d'incohérence) ou fonction des corrections du locuteur (négation de commande, refus du paramètre).

La deuxième correspond au dialogue global. On stocke dans la **mémoire à long terme** tous les ordres exécutés par le système. On y conserve aussi tous ceux qui étaient incohérents mais finalement proposés à l'opérateur.

Stratégie de recherche de valeur-résolution des ellipses

Pour répondre à la demande de valeur formulée par TACHE, DIALOGUE dispose de quatre sources : la mémoire à long terme, la mémoire à court terme, les valeurs par défaut et le niveau reconnaissance. Il consulte tout d'abord la mémoire à court terme, puis la mémoire à long terme. Il regarde ensuite si une valeur par défaut existe. Enfin il fait appel en dernier ressort au module de reconnaissance. Une des caractéristiques de DIAPASON et de ne pas proposer consécutivement deux fois le même ordre en cas de correction. Cela est possible grâce au statut évoqué dans le paragraphe précédent, DIALOGUE ne propose plus de commande ou de paramètre qui ont été niés ou refusés une première fois pour cause d'incohérence ou de correction.

Exemple : O : Libérer mémoire vpr0
D : Confirmez : libérer mémoire vpr ?
O : Négatif vpr0
Le statut de vpr = NIEPAR (paramètre nié). Supposons alors que deux hypothèses de phrase se présentent dans l'ordre suivant :
(1) Négatif VPR
(2) Négatif VPR0
DIAPASON élimine la première à cause du statut de VPR, d'où la réponse :
D : Confirmer : Libérer mémoire VPR0 ?

Lorsque DIALOGUE reçoit une hypothèse de niveau reconnaissance il l'ajoute aussitôt dans la mémoire à court terme. L'appel à un des analyseurs revient à faire une recherche dans la mémoire à court terme le cycle suivant (DIALOGUE repasse dans son état initial). Ce mécanisme résoud les ellipses.

Incohérence et mémoire

Lorsque nous avons traité l'incohérence, nous n'avions pas encore indiqué que les valeurs pouvaient avoir deux origines, mémoire à court terme ou à long terme. Le message envoyé à l'opérateur était un constat d'échec. DIALOGUE proposait des valeurs erronées et on n'y pouvait rien. En cas d'incohérence un traitement supplémentaire est effectué qui permet de revenir sur un choix. Le principe est le suivant : si une valeur provient de la mémoire à long terme et provoque une incohérence c'est qu'il ne fallait pas la proposer. On pose une question au lieu d'envoyer un message d'erreur.

Exemple :
sans traitement
O : Afficher bruiteur pointé
D : Afficher bruiteur pointé
exécution de afficher bruiteur pointé
O : Afficher
D : Le bruiteur pointé est déjà affiché
avec traitement
O : Afficher bruiteur pointé
D : Afficher bruiteur pointé
exécution de afficher bruiteur pointé
O : Afficher
D : Afficher quel bruiteur ?

4.2.3.4. *La résolution des ambiguïtés*

Le langage de commande ne contient qu'une seule ambiguïté. Elle concerne la commande afficher/effacer lorsque l'on fait un enchaînement d'ordres.

Exemple : O : Afficher baptême *b* 2.
D : Afficher baptême bruiteur *b* 2.
O : *c* 3.

DIAPASON a le choix entre les deux interprétations suivantes :

D : Afficher bruiteur *c* 3. (1)
D : Afficher baptême bruiteur *c* 3. (2)

Dans cet exemple, DIAPASON ne sait pas a priori si l'opérateur désire afficher le bruiteur ou son baptême. L'ambiguïté se retrouve au niveau humain, le problème du choix est le même. Il semble néanmoins qu'il y aie une tendance à garder le plus d'éléments possibles. La première interprétation serait donc la (2) et la seconde ne serait envisagée qu'en cas de non-satisfaction de la (1). Dans le cas où il faudrait afficher le bruiteur la phrase serait plutôt « afficher *c* 3 ». La solution retenue pour DIAPASON consiste à garder le maximum d'éléments et donc de proposer l'interprétation (2). La solution (1) ne sera jamais envisagée même en cas d'incohérence de la (2).

4.2.4. Le module d'aide à l'opérateur

Le module AIDE assure deux fonctions. Il participe à la formation d'un nouvel opérateur et assiste l'utilisateur déjà formé. Il permet d'accéder à la syntaxe de toutes les phrases qui peuvent être prononcées. Un menu donne la possibilité de choisir entre une aide contextuelle et une des commandes concernant l'image courante. Trois catégories de phrases sont disponibles : la syntaxe complète de l'ordre, la syntaxe de correction et enfin les enchaînements possibles après cet ordre. Elle est valable pour les deux modules de reconnaissance SYNTAXE et VOCAL.

L'appel se fait à l'aide de la souris de même que tous les choix qui suivent. DIAPASON représente la forme sélectionnée avec un graphe (*fig.* 3).

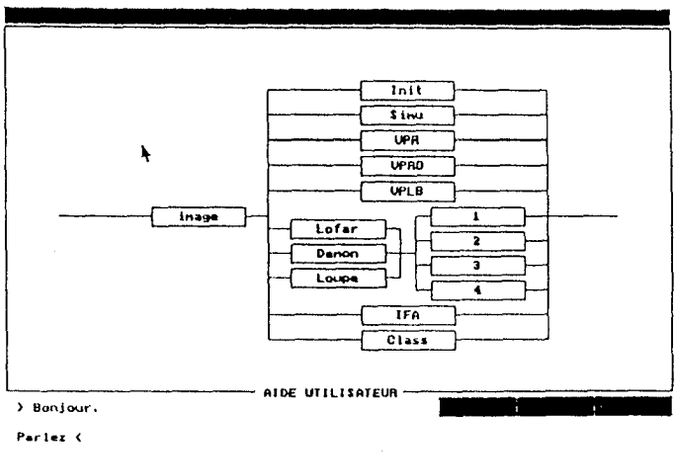


Figure 3. — Exemple d'aide.

5. Présentation des résultats

Afin d'évaluer les performances du système nous avons procédé à une comparaison des dialogues de commande du sonar réel et de la maquette implantée sur PC.

5.1. CONDITIONS DE L'EXPÉRIMENTATION

La manipulation s'est déroulée sur la plate-forme d'intégration des sonars. Le bruit d'ambiance était relativement important, six bases d'électronique dans la pièce et des conversations fréquentes de cinquante centimètres à deux mètres du micro (omnidirectionnel pour l'expérimentation). Les opérateurs étaient entraînés soit sur le sonar, soit sur le PC, le temps ne permettait pas de les entraîner sur les deux systèmes. Le sonar fonctionnait en simulation, les tests n'ont donc pu être effectués en environnement opérationnel sur une tâche opérationnelle, mais sur des sous-tâches simulées.

5.2. MESURES ERGONOMIQUES

Vu l'impossibilité de disposer du sonar réel pendant un temps important et vu le fait que les tests ne pouvaient être faits sur une tâche réelle, nous avons concentré nos efforts de comparaison sur quelques sous-tâches exécutées par quelques opérateurs en gardant bien à l'esprit que seules les comparaisons sur les tâches réelles en ambiance réelle permettent des conclusions quasi définitives.

Trois sous-tâches ont été effectuées par six opérateurs sur le sonar, puis sur le PC. Pour chaque sous-tâche on a mesuré la durée totale et le nombre d'opérations effectuées. En plus de ces essais, une vingtaine de locuteurs non entraînés ont pu essayer plus ou moins longuement le système de commande vocale montrant ainsi les qualités du système : multilocuteur sans apprentissage, résistance aux variations de prononciation (accents régionaux, rapidité d'élocution, silence ou liaisons entre mots, hésitation...), résistance au bruit d'ambiance et aux conversations de proximité.

5.3. RÉSULTATS ACQUIS

La comparaison entre la console du sonar et sa simulation sur PC nous a permis d'acquérir des résultats relatifs à l'emploi et aux avantages de la commande vocale pour les consoles.

5.3.1. Mode d'emploi de l'entrée vocale

Les **temps de réponse** cumulés des systèmes de reconnaissance de la parole et de compréhension de l'ordre ne doivent pas dépasser la seconde. Au-delà l'opérateur perd du temps à attendre.

Les **demandes de confirmation** d'ordre doivent le plus souvent possible être faites « en place » c'est-à-dire se matérialiser par une modification de l'image proche de l'endroit où l'opérateur est censé observer à cet instant.

Nous avions réservé une zone de l'écran pour les dialogues où les phrases reconnues et leur interprétation s'affichait. A l'usage il s'est avéré qu'elle faisait perdre une partie des avantages principaux de la parole qui est que le regard reste fixé sur l'objet manipulé.

L'**utilisation d'un commutateur** (« pédale micro ») permettant à l'opérateur d'indiquer quand il donne un ordre vocal s'avère un moyen simple, pratique à utiliser et très efficace pour se protéger de la réception d'ordres parasites dus aux conversations se tenant près du microphone.

Pour les ordres vocaux l'opérateur n'est plus guidé dans son choix par des possibilités affichées sur l'écran. Dans ce cas les **moyens d'aide** deviennent très utiles, surtout pour les opérateurs débutants.

Dans les systèmes à commande par mots enchaînés, il faut limiter la taille du vocabulaire et la richesse de la syntaxe de façon que l'opérateur puisse l'apprendre. La taille maximale semble être de cent cinquante à deux cents mots, ce qui est suffisant pour les applications de type console.

5.3.2. Aspects ergonomiques de l'avantage à employer une entrée vocale

Le regard de l'opérateur peut rester fixé sur l'objet manipulé, autrement dit il n'est pas nécessaire, pour un ordre vocal, de regarder le clavier ou des menus affichés à l'écran. Malgré l'usage de la zone dialogue on a remarqué que le regard était quatre fois moins distrait ; on devrait arriver à quinze fois en effectuant systématiquement la confirmation en place. L'usage de la commande vocale permet de concentrer en une seule phrase de nombreuses manipulations élémentaires sur le sonar.

Les échanges sont plus rapides. Dans le cas de système avec un temps de réponse de une seconde (actuellement deux à quatre secondes), on obtient pour l'ensemble des opérateurs et des sous-tâches une amélioration de la vitesse de travail de 1,66 en faveur de la commande vocale (entre 1,1 et 2 selon le cas). Cet avantage est d'autant plus net que les ordres correspondent à des phrases longues.

Les menus occupent de la place sur l'écran. Cette occupation peut être permanente ou temporaire. Dans tous les cas c'est de la surface perdue pour l'affichage des informations proprement dites. La commande vocale permet de récupérer cette surface.

6. Conclusion

L'étude décrite ici a montré l'intérêt :

— de la reconnaissance analytique de la parole qui permet une bonne résistance d'une part aux variations de prononciation pour une même fonction (plus ou moins rapide, silences ou liaisons entre mots, hésitations), d'autre part aux différences de prononciation entre les locuteurs (sans apprentissage). Ce type de reconnaissance est de plus bien adapté à la prise en compte des coarticulations, des fusions entre mots et des mots outils, c'est-à-dire en un mot à la parole continue ;

— du dialogue, particulièrement nécessaire dans le cas de l'emploi d'une entrée vocale du fait du taux d'erreur plus élevé que dans les autres modes et du fait du « naturel » de ce mode qui permet à l'opération des ordres incomplets. On peut affirmer que même un système de reconnaissance de parole parfait (sans erreur) ne servirait pas à grand chose si son usage n'était complété par un dialogue ;

— de l'intégration de l'entrée vocale aux autres modes d'entrée. L'entrée vocale doit être vue comme complémentaire du clavier et de la souris et l'opérateur doit pouvoir à tout moment librement choisir entre la voix et les autres modes. La possibilité d'ordres mixtes voix + souris serait très utile.

En tenant compte des conclusions précédentes, il a été possible de montrer l'intérêt de l'ajout d'une entrée vocale à une console. Dans ce cas, le vocabulaire nécessaire n'est pas très étendu (< 200 mots) et on peut se contenter de mots enchaînés avec quelques libertés pour les différentes phases du dialogue. Les avantages liés à l'emploi de l'entrée vocale sont alors importants.

— Il est plus naturel pour l'opérateur de donner ses ordres sous forme d'une phrase de n mots enchaînés que de cliquer successivement n mots sur l'écran surtout si le locuteur dispose de libertés (élisions, articles) et d'un dialogue évolué.

— Le regard de l'opérateur peut rester fixé sur l'objet auquel il s'intéresse et au sujet duquel il donne des ordres, plutôt que de regarder des menus en d'autres régions de l'écran.

— Les 2 points précédents conduisent à une plus grande rapidité pour donner les ordres (d'un facteur 1 à 2 selon les cas).

— L'écran n'est pas systématiquement surchargé par des menus au détriment de l'image principale.

Pour des applications plus ambitieuses, telles que la communication avec des bases de données ou des systèmes experts complexes, l'entrée vocale sera encore plus utile. Mais une parole plus naturelle que les simples mots enchaînés deviendra souhaitable. De ce fait, la taille du vocabulaire, les libertés et les facteurs du branchement augmentent et il faudra alors améliorer encore le système de reconnaissance et le dialogue.

Manuscrit reçu le 9 octobre 1989.

BIBLIOGRAPHIE

[1] J. M. Pierrel, N. Carbonnel, *Vers des dialogues oraux naturels Homme-machine*, dans « Ergonomie et intelligence artificielle », à paraître chez Masson, 1990.

[2] M. Argile, A. Furnaham, J. A. Graham, *Social Situations*, Cambridge University Press, 1981.

[3] J. M. Pierrel, N. Carbonnel, J. P. Haton, K. Smaili, « Vers une meilleure intégration de la parole dans des systèmes de communication homme-machine », Revue *Traitement du signal*, à paraître, 1990.

[4] A. M. Liberman *et al. Perception of the speech code,* Psychological Review, vol. 74, n° 6, pp. 431 à 461, 1967.

[5] P. Denes, *The Design and Operation of the Mechanical Speech Recognition at University College London,* Journal Beit IRE, pp. 219 à 229, April 1959.

[6] J. Dreyfus-Graf, *Phonétographe : Présent et Futur,* Bull. Tech. PTT Bern n° 5, 1961.

[7] J. Y. Gresser, G. Mercier, *Automatic segmentation of speech into syllabic and phonème units.* Symposium on auditory analysis and perception of speech, Academic Press, pp. 359 à 382, 1975.

[8] P. Alinat, *Étude des phonèmes de la langue française ou moyen d'une cochlée artificielle. Application à la reconnaissance de la parole,* Revue Technique Thomson CSF vol. 7, n° 1, mars 1975.

[9] V. Zue *et al., Acoustic Segmentation and Phonetic Classification in the SUMMIT System,* IEEE-ICASSP, pp. 389-392, 1989.

[10] R. Cole, L. Hon, *Segmentation and Broad Classification of Continuous Speech,* IEEE-ICASSP, pp. 453-456, 1988.

[11] R. Gubrynowicz *et al., Reconnaissance de mots isolés par la méthode descriptive de traits phonétiques,* JEP, GALF-SFA, pp. 235-238, 1986.

[12] K. Frimpong-Ausah *et al., A Stochastic/Feature Based Recogniser and its Training Algorithm,* IEEE-ICASSP, pp. 401-404, 1989.

[13] V. Zue *et al., An Expert Spectrogram Reader : a Knowledge-based Approach to Speech Recognition,* IEEE-ICASSP, pp. 1193-1196, 1986.

[14] N. Carbonnel *et al., APHODEX, Design and Implementation of Acoustic-Phonetic Decoding Expert System,* IEEE-ICASSP, pp. 1201-1204, 1986.

[15] K. Hatazaki *et al., Phonems Segmentation Using Spectrogram Reading Knowledge,* IEEE-ICASSP, pp. 393-396, 1989.

[16] H. Méloni, R. Bulot, *Décodage acoustico-phonétique en Prolog,* JEP, GALF-SFA, pp. 231-234, 1986.

[17] GRECO-PRC, *Décodage acoustico-phonétique,* Actes du séminaire décodage acoustico-phonétique du GRECO-PRC « Communication Homme-Machine, Nancy, 98 p., 1988.

[18] C. Sorin, *Perception de la parole continue,* Psychoacoustique of Perception auditive. Vol. II de la série l'Audition. A paraître chez Lavoisier.

[19] A. Levy-Schoen *et al., Le regard de la lecture,* La Recherche n° 211, pp. 744 à 753, juin 1989.

[20] G. Von Bekesy, *Experiments in Hearing,* New York McGraw-Hill, 1960.

[21] B. M. Johnstone *et al., Mechanics of the guinea pig cochlea,* JASA, vol. 47, pp. 504 à 509, 1970.

[22] D. D. Greenwood, *Critical Bandwidth and the frequency coordinates of the basilar membrane.* JASA, vol. 33, pp. 1344 à 1356, octobre 1961.

[23] P. Alinat, *Étude du trait permettant de distinguer entre les 3 classes de consonnes explosives PB, TD, KG, JEP, GALF,* Lannion, pp. 297-303, 31 mai-2 juin 1978.

[24] J. M. Pierrel, *Un système de compréhension automatique du discours continu utilisant des contraintes morphologiques syntaxiques et sémantiques,* RAIRO informatique, vol. 12-2, pp. 83-105, 1978.

[25] J. M. Pierrel, *Dialogue oral homme-machine,* Hermès, Paris, 240 p. 1987.