Informatique et Santé
Collection dirigée par P. Degoulet et M. Fieschi
Paris, Springer-Verlag France

Nouvelles Technologies et Traitement de l'Information en Médecine

Rédacteurs :
H. Ducrot, E. Martin et J. -R. Scherrer

Volume 4 - 1991

Reconnaissance automatique de la parole

H. Cerf-Danon, M. El-Bèze, B. Merialdo

Centre Scientifique IBM - France, 3 et 5 Place Vendôme - 75021 Paris Cedex 01

Abstract

Computer Speech Recognition comes of age. Several Speech Recognition products are now available, and many are being prepared all over the world. This paper presents both a state of the art of this dicipline as well as the current status of the research conducted by the Automatic Dictation Group of the IBM France Scientific Center.

Informatique et Santé, 1991 (4) : 84-100

1 Introduction

La reconnaissance automatique de la parole pose de nombreux problèmes d'un point de vue théorique. Leur complexité fait que seuls des sous-problèmes ont pu être à ce jour résolus. Ces solutions partielles correspondent à des contraintes plus ou moins fortes, et les systèmes existants supposent une coopération plus ou moins grande des utilisateurs.
Pour classer les systèmes de reconnaissance automatique, on a généralement recours aux critères suivants :

- le mode d'élocution *(1)*
- la taille du vocabulaire
- la dépendance plus ou moins grande vis-à-vis du locuteur
- l'environnement protégé ou non

En outre, il n'est pas inintéressant de les différencier selon deux points qui ont aussi leur importance :

- La compréhension est-elle requise ou non ?
- Le discours est-il naturel, ou la syntaxe des phrases doit-elle être contrainte ?

Dans une première partie, nous présentons un état de l'art du domaine suivant ces critères. La deuxième partie présente en détail les problèmes que pose la Reconnaissance de la Parole en général, ainsi que les problèmes plus spécifiques de la langue française.
Une troisième partie, plus technique, concerne les réalisations du Centre Scientifique d'IBM France dans le domaine et, plus généralement les techniques markoviennes. Enfin, la quatrième partie donne un aperçu des perspectives à court et moyen terme.

2 Etat de l'art en Reconnaissance de la Parole

Dans le domaine de la reconnaissance automatique de la parole, on distingue trois grands types d'applications:

- les systèmes de commandes vocales
- les machines à dicter
- les systèmes de compréhension

(1) des syllabes ou mots isolés aux mots connectés, jusqu'à une parole dite, "continue" c'est-à-dire sans pauses artificielles.

2.1 Les commandes vocales

On trouve aujourd'hui, un grand nombre de produits fiables sur le marché qui permettent de contrôler l'environnement (informatique ou non) au moyen d'une entrée vocale. La presse spécialisée, et dans une large mesure la grande presse, rendent compte de l'ensemble des produits aujourd'hui disponibles. Les applications multiples et variées vont du jouet gadget à l'outil de travail sophistiqué. Il ne nous est pas possible de dresser une liste exhaustive de ces produits. Nous nous contenterons de citer quelques exemples d'applications, puis nous décrirons leurs traits communs.

- Voiture (projet R 25)
- Jeux vidéo
- SNCF (noms de gares)
- Aide aux handicapés *(2)*
- Reconnaissance de chiffres *(3)*

Pour des raisons plus ergonomiques que purement techniques, la taille du vocabulaire, dont disposent au moment du décodage les systèmes de commande vocale, est de façon générale, limitée à quelques centaines de mots. Avec un vocabulaire dépassant le millier de commandes, les utilisateurs se heurteraient alors à deux problèmes pratiques :

- S'il est donné à tout individu moyen de mémoriser le contenu d'une liste de cent mots, il ne lui est pas du tout évident d'en mémoriser mille avec la même précision. Si un conducteur doit hésiter entre Klaxon et avertisseur, feux de position, feux de croisement, feux de routes et lanternes, codes ou phares ... il peut se faire tard avant que ses lumières ne s'allument ou s'éteignent,
- Deuxièmement, dans le cas des systèmes mono-locuteurs, les utilisateurs trouveraient fastidieux un apprentissage qui consiste à prononcer plusieurs fois chacun des mots de la liste.

Bien qu'une grande souplesse soit donnée à l'utilisateur quant au choix du vocabulaire, il est recommandé de choisit des mots contrastés pour réduire les risques d'ambiguïté. Ces systèmes sont d'autant plus performants (reconnaissance fiable à 99 %) que les mots sont bien différenciables par la longueur, ou leur transcription phonétique. Certains de ces systèmes sont multi-locuteurs, et ne nécessitent donc pas d'apprentissage préalable. Dans ce cas, le taux de succès avoisine les 95 %.

La méthode privilégiée était pendant longtemps essentiellement la comparaison de références par programmation dynamique [1 (pp.516, 533, 540)]. Depuis le début des années 80, on utilise de plus en plus, des modèles markoviens [2, 3], auxquels certains chercheurs préfèrent aujourd'hui les réseaux neuronaux *(4)* .

(2) Les applications sont de deux types. Les unes apportent une aide aux personnes ayant un organe défaillant, alors que les autres sont des outils de rééducation, comme par exemple Speech Viewer, système destiné à la rééducation de la parole des enfants sourds. Parmi le grand nombre de jeux qu'il propose, citons celui du mobile se déplaçant dans un labyrinthe vers quatre directions [4] associées à quatre sons différents.

(3) Parmi les systèmes multi-locuteurs existants, on peut citer comme exemple intéressant les travaux menés aux USA chez ATT, et en France la cabine vocale Publivox [5] développée par le CNFT: il suffit à l'utilisateur de prononcer les chiffres du numéro qu'il veut appeler pour que la communication téléphonique s'établisse.

(4)Ces derniers modèles sont utilisés en parole, entre autres par il. Bourlard, T. Kohonen et A. Waibel. Leurs principes ont été passés en revue de façon claire et synthétique à ICASSP 88,

Des systèmes d'un autre type s'intéressent à la recherche de quelques mots clés ("word-spotting") en ignorant délibérément le reste des mots qui apparaissent dans le flot d'un discours non contraint. S'ils fonctionnent de façon satisfaisante, ils détectent la présence dans l'énoncé des mots du dictionnaire. Trois cas de dysfonctionnement de surface sont. possibles : rester passif face à un mot connu, en détecter un autre à la place, prendre à tort un mot inconnu pour un mot connu. Plus profondément, et même en cas de fonctionnement parfait, il est difficile d'exploiter les détections. En effet, un mot extrait d'un contexte négatif petit entraîner l'effet contraire à celui souhaité.

2.2 Les systèmes de compréhension

Bien que de nature profondément différente, les systèmes de compréhension se caractérisent aussi par un vocabulaire limité à quelques centaines de mots, et donc un domaine sémantique fermé. Plus gênant, la syntaxe est aussi contrainte. Pour faciliter la gestion du dialogue par le système, les phrases acceptables pour le système doivent se conformer à des schémas grammaticaux simplifiés. La fermeture du domaine sémantique peut être tolérable si l'application est bien ciblée.

En attestent les applications généralement choisies comme l'interrogation d'une base de données, les standards téléphoniques automatisés' qui donnent des renseignements météo, ou même permettent de réserver des places.

Ces systèmes sont connectés à des modules d'interprétation du message reconnu, dont le but est de réagir soit par l'émission d'une réponse vocale soit par une action mécanique sur l'environnement, après prise de décision. De ce fait, la performance de ces systèmes doit être jugée sur la base du nombre de phrases reconnues. Le critère pourrait sembler sévère si l'on devait compter pour fausse toute phrase dont le moindre mot a été mal reconnu. Comme en pratique les paraphrases sont acceptées, la difficulté s'en trouve atténuée, mais l'évaluation ici n'est pas chose aisée et inclut, de fait, une part de subjectivité.
Beaucoup de grands systèmes ont été conçus autour du projet ARPA lancé en 1977 aux USA. Aujourd'hui, plusieurs laboratoires travaillent sur le projet DARPA (CMU, MIT, BBN, SRI, Bell labs, ...). Par exemple, le système SPHINX développé à CMU, suppose l'emploi d'un petit vocabulaire (les mille mots du corpus « Ressource Management Data Base »). Il a par contre le mérite de permettre aux locuteurs une parole continue et ne nécessite pas d'apprentissage préalable. Le taux de reconnaissance sur les mots est excellent [8], puisqu'il est légèrement supérieur à 96 %.
En France, a été aussi explorée l'idée que les imperfections du décodage Acoustico-phonétique n'empêchaient pas d'accéder au niveau supérieur de compréhension, mais qu'au contraire les faiblesses des niveaux "inférieurs" pouvaient être épaulées par le recours à un des niveaux plus "élevés". C'est dans cette optique qu'ont été conçus les systèmes KEAL au CNET [1 (pp.500, 562, 674)], Esope au LIMSI [1 (pp.500, 627)], Myrtille au CRIN [9], Arial au CERFIA [10].

---

par R. Lippmann [6]. On se reportera aussi à l'excellent article d'E . Levin (Bell Labs), qui compare et combine modèles markoviens et neuronaux pour reconnaître des chiffres [7].
Installé depuis 1988 à la mairie de Lannion, le système multi-locuteur MAIRIE-VOX [2, 5] du CNET permet aux utilisateurs d'un téléphone ordinaire l'obtention d'informations diverses sur les loisirs, les services de garde ou l'actualité municipale. Les personnes qui appellent sont amenées par un ensemble de menus à n'utiliser à un instant donné qu'un vocabulaire de six mot,, extraits, d'une liste de 21 mots.

2.3 Les systèmes de dictée automatique

Les machines à dicter ont pour but de retranscrire un texte dicté par un locuteur devant un microphone aussi bien qu'une secrétaire pourrait le faire, c'est-à-dire, en respectant au mieux les règles d'usage et d'accord orthographique propres à la langue utilisée, La compréhension des phrases n'est nullement requise. Aussi, la plupart des systèmes de dictée automatique ne savent pas discriminer les différents sens d'un mot donné. S'ils le pouvaient, leurs performances seraient certainement meilleures.
On peut remarquer que ce domaine occupe un lieu charnière, à la frontière de l'oral et de l'écrit. De fait, les registres de langue traités ne sont pas ceux du langage parlé, mais plutôt ceux de l'écrit. En fonction de l'application envisagée, seront dictés des rapports, des articles de journaux, des lettres administratives ... Il en résulte que la complexité est moindre que s'il fallait retranscrire des dialogues à l'état brut. Dans le vif d'une conversation, les phrases agrammaticales se mêlent aux phrases incomplètes, tandis que fourmillent hésitations, reprises, retours en arrières, ou autres répétitions.
Cependant, l'exercice même de la dictée sous-entend l'utilisation de plusieurs dizaines voire plusieurs centaines de milliers de formes fléchies. Les mots sont pris en contexte, et régis par une syntaxe aussi libre que la grammaire de la langue naturelle le permet. De plus, s'il nous faut comparer la séance de dictée à une scène connue, il s'agit de reproduire plutôt le scénario du médecin dictant une lettre à sa secrétaire, que celle de l'institutrice vérifiant les connaissances de ses élèves. Cette précision est importante dans la mesure où l'utilisateur n'a pas forcément comme l'enseignant, devant les yeux un texte déjà écrit, Il improvise, et donc doit pouvoir se tromper, revenir en arrière afin de corriger un mot ou remanier une tournure.
Actuellement, l'état de l'art en reconnaissance de parole ne permet pas l'affranchissement conjoint de l'ensemble des quatre contraintes majeures décrites en tête de chapitre. Les systèmes qui gèrent des vocabulaires de grande taille (plusieurs milliers de mots), ne sont pas indépendants du locuteur. Ils nécessitent un apprentissage préalable, et ne peuvent encore aujourd'hui supporter un mode d'élocution non-contraint. Nous reviendrons très largement sur ces points dans les pages qui suivent.
Historiquement, l'équipe de recherche IBM dirigée par F. Jelinek, est la première à avoir montré qu'un système grand vocabulaire (*Tangora* 5 000 mots en 1995, 20 000 en 87) pouvait tenir dans une petite boîte "portable". Par la suite, l'ensemble des grands systèmes développés ici et là se sont inspirés peu ou prou du système *Tangora*. Avec un taux de réussite supérieur à 95% pour un vocabulaire de 20 000 mots, *Tangora* tend à devenir un système multi-lingue existant pour l'anglais [11], l'italien [12,13], aujourd'hui le français [14], et l'allemand [15], bientôt l'espagnol.
Le système Dragon [16] est l'un des rares produits présents aujourd'hui sur le marché. Il fonctionne en mots isolés avec un vocabulaire de base de 16 000 entrées extensible à 30 000. Un de ses points forts est sa capacité d'adaptation (par validation au clavier) en cours de décodage. Bien qu'un peu fastidieuse (il faut attendre le décodage d'un mot avant de prononcer le mot suivant), l'adaptation doit être faite convenablement sinon, revers

de la médaille, le système diverge. Dans de telles conditions, il est difficile d'interpréter le taux de reconnaissance (estimé aux alentours de 90 %) si sa variance n'est pas connue. Dragon a été conçu pour la langue anglaise telle qu'on la parle outre-atlantique, mais une transposition du système dans les langues européennes est en voie d'être confiée [17] à la société belge Lernout et Hauspie.
Aux USA, le second grand système vendu est la machine de Kurzweil (1 000 à 10 000 mois). Le Voice Terminal de Kurzweil (KVT) ne s'est pas vraiment détaché de la commande vocale. Ses concepteurs affirment qu'il offre la possibilité de dicter en mots isolés un texte, propre à un domaine spécialisé (radiologie), tout en permettant de contrôler des machines, véhicules et autres robots. Cependant, la presse [18] ne lui confère pas le statut de système de dictée, les critères de définition n'étant pas atteints.
En France, la machine à dicter développée au LIMSI (5 000 à 10 000 mots) autour du circuit µPCD a abouti au produit DATAVOX (5 000 mots) commercialisé par la société VECSYS. Le taux de reconnaissance publié [5] s'élève à 95% pour un locuteur masculin. Le système Hamlet [19] est une maquette développée parallèlement. Ce prototype petit traiter un vocabulaire de 7 000 mots en supposant une élocution en mots isolés.
Par ailleurs, de nombreux laboratoires de recherche ont mis ou mettent encore au point les prototypes de produits futurs. Nous décrirons par la suite le système *Parsyfal* développé depuis 1985 au Centre Scientifique IBM-France de Paris. Disons qu'une de ses caractéristiques fortes est de pouvoir traiter un dictionnaire de très grande taille (quelques centaines de milliers de formes en entrée). La liste des systèmes qui peuvent concurrencer *Parsyfal* quant à la taille du dictionnaire est réduite : le système développé l'INRS (Bell Northern) par M. Lennig [20] fonctionne en anglais avec une capacité de 86 000 mots. Ensuite, on trouve les systèmes dédiés aux langues asiatiques comme celui réalisé pour le mandarin [21], et pouvant traiter 60 000 mots.
Les thèmes majeurs de recherche portent sur:

- un mode d'élocution naturel
- l'ouverture maximale du domaine de l'énoncé
- les méthodes connexionistes concurrentes des modèles markoviens
- les approches de type intelligence artificielle *(6)*
- l'adaptation rapide au locuteur, voire l'indépendance vis à vis du locuteur.

Ainsi, des recherches sur le mode de parole dit continu sont menées au laboratoire IBM de Yorktown. Des expériences [22] portant sur un vocabulaire limité à 5 000 mots, ont donné un taux d'erreur sur les mots de 11%.

(6) Le système expert APHODEX du CRIN pour le décodage acoustico-phonétique. Système de lecture de spectrogrammes au LIMSI. Les travaux de Il. Méloni à l'université de LUMINY (Marseille).

3 Complexité du problème

3.1 Problèmes indépendants de la langue

Le signal de parole n'est pas un signal ordinaire : il s'inscrit dans le cadre de la communication parlée, un phénomène des plus complexes. Afin de souligner les difficultés du problème, nous ferons ressortir essentiellement quelques caractéristiques notoires de ce signal :

- un débit intense
- une extrême redondance
- une grande variabilité
- un lieu d'interférences

D'un point de vue mathématique, il est ardu de modéliser le signal de parole, car ses propriétés statistiques évoluent au cours du temps.

3.1.1 Redondance du signal de parole

Quiconque a vit une représentation graphique de l'onde sonore a certainement été frappé par le caractère répétitif du signal de parole. Un grossissement à la loupe d'une brève émission de parole donne à voir une succession de figures sonores semblant se répéter à l'excès. Lin peu de recul laisse apparaître des zones moins stables qu'il convient, de qualifier de transitoires. Ce qui semblerait de prime abord superflu, s'avère en réalité fort utile. Les répétitions confèrent à ce signal une robustesse. La redondance le rend résistant au bruit. Dans une certaine mesure, elle fonctionne comme un code correcteur d'erreur, puisqu'un interlocuteur humain sait décrypter un message même s'il est entaché de bruits dus à de possibles interférences.

3.1.2 Une grande variabilité

Une même personne ne prononce jamais un mot deux fois de façon identique. La vitesse d'élocution en détermine la durée. Toute affection de l'appareil phonatoire peut altérer la qualité de la production. Un rhume teinte les voyelles de nasalité; une simple fatigue et l'intensité de l'onde sonore fléchit, l'articulation perd de sa clarté. La diction évolue dans le temps: l'enfance, l'adolescence, l'âge mûr, puis la vieillesse, autant d'âges qui marquent la voix de leurs sceaux.
La variabilité inter-locuteur est encore plus flagrante. La hauteur de la voix, l'intonation l'accent diffèrent selon le sexe, l'origine sociale, régionale ou nationale.
Enfin toute parole s'inscrit dans un processus de communication où entrent en jeu de nombreux éléments comme le lieu, l'émotion, l'intention, la relation qui s'établit entre les interlocuteurs. Chacun de ces facteurs détermine la situation de communication, et influe à sa manière sur la forme et le contenu du message.

3.1.3 Les interférences

- l'acoustique du lieu (milieu protégé ou environnement bruyant)
- la qualité du microphone et sa position par rapport à la bouche
- les bruits de bouche

3.1.4 Les effets de co-articulation

La production "parfaite" de chaque son suppose théoriquement un positionnement précis des organes phonatoires. Or, lorsque le débit de parole s'accélère, le déplacement de ces organes est limité par une certaine inertie mécanique. Les sons émis dans une même chaîne acoustique subissent l'influence de ceux qui les suivent ou les, précèdent.
Ces effets de co-articulation sont des interférences. lis entraînent l'altération des formes sonores en fonction des contextes droits ou gauches, selon des règles étudiées par les acousticiens d'un point de vue articulatoire ou perceptif.

3.2 Particularités de la langue française

Le français est mis en regard de cinq autres langues selon sept axes dans la table qui suit.
Bien que la valuation des critères soit binaire là où il faudrait mesurer des degrés, on peut l'utiliser pour dégager quelques caractères particuliers des langues qui y figurent. Pour ce qui est du français, son originalité réside dans la combinaison de quatre caractéristiques : innexions, dérivations, liaisons et apostrophes. La formation de mots composés ou "composition" qui opère de façon mineure en français au regard de l'allemand risque de prendre, avec la réforme de l'orthographe en cours, une toute autre envergure. Sans que ceci soit un " apriori " on risque de voir le "statuquo" actuel bouleversé par l'usage de mots devenus "ultrafréquents" comme certaines tournures "passepartouts" employées dans les "cinéromans".
La *dérivation* est l'opération morphologique la plus complète qui permet de rattacher un mot i sa racine après dépouillement de ses éventuelles préfixes ou suffixes. L'ensemble des dérivés rassemble ce qu'il est convenu d'appeler communément les mots de la même famille comme par exemple « ***Jardin*** », "***jardinier***" ou « ***jardinage*** ».
Les *inflexions* d'un mot correspondent à toutes les formes obtenues par conjugaison de ce mot (on parie alors de lemme). Par exemple, "***pomme***" et "***pommes***" sont deux flexions du lemme "***pomme***", mais «***poire*** » et "***poirier***" ne proviennent pas d'un même lemme bien qu'étant de la même famille.

Comparaison du français avec d'autres langues						
	français	allemand	italien	espagnol	anglais	arabe
inflexion	+	+	+	+	-	+
composition	+-	+	-	-	-	-
apostrophe	+	-	+	-	-	-
liaison	+	-	-	-	-	+
dérivation	+	+	+	+	+	+
ordre libre	-	+	-	+	-	-
agglutination	-	-	+-	+	-	+

3.2.1 Une langue fléchie

La beauté d'une langue réside en grande partie dans sa richesse. Le français vérifie doublement cette assertion. En effet, dans la langue de Molière, la variété du vocabulaire permet la nuance dans le concept, les touches subtiles dans la description. Ecrivains, poètes, hommes d'état, femmes de lettres et chanteurs de rue lui ont donné ces qualités qui lui sont universellement reconnues.
Mais, cette richesse provient aussi de sa complexité tant sur le plan grammatical que morphologique. Les tables de conjugaison par leur taille et leur nombre en fournissent une preuve remarquable.

3.2.2 Problème de couverture

La *couverture* d'un texte par un dictionnaire est le pourcentage de mots de ce texte contenus dans le dictionnaire. Ce pourcentage se calcule, de façon classique [23], de deux manières, selon que l'on considère le dictionnaire comme statique ou dynamique.

Selon la méthode choisie, un mot absent du dictionnaire est compté comme inconnu, respectivement

- à chacune de ses occurrences
- seulement lors de sa première occurrence

L'agglutination ("clitic attach") se traduit par la concaténation syntagmatique de deux mots remplissant une fonction grammaticale complémentaire (verbe)-(pronom personnel objet) ou (nom)-(adjectif possessif). Par exemple, dans la langue internationale qu'est devenue le jargon informatique, ~~g~~ive me" s'agglutine en "gime". En outre, chacun des deux mots peut subir une modification graphique et phonétique.
Si l'on pense aux concaténations de mots (doublées de contractions) qui foisonnent cri argot américain, on peut s'étonner de ne pas voir la case agglutination marquée du signe + polir la langue anglaise, La réponse tient dans le fait que la dictée automatique privilégie les registres de la langue écrite.

Couverture d'un dictionnaire en fonction de sa taille		
taille	couverture statique	couverture dynamique
215 000	99,7 %	99,8 %
135 000	99,6 %	99,7 %
85 000	97,9 %	98,6 %
22 000	96,2 %	97,3 %

Les couvertures ci-dessus ont été calculées sur un test de 35 000 mots n'ayant pas servi à la constitution du dictionnaire. Quatre dictionnaires sont envisagés ici. Le plus grand d'entre eux est constitué de l'ensemble des 215 000 différentes formes rencontrées dans un corpus de 38 millions de mots. Après rejet des formes qui n'apparaissent qu'une fois, il n'en reste que 135 000. L'intersection des 215 000 et de 200 000 formes contenues par un dictionnaire plus général donne 80 000 formes auxquelles 5 000 noms propres ont été rajoutés pour obtenir le dictionnaire intermédiaire de 85 000 formes. Enfin, on a choisi parmi les 215 000 formes, les 22 000 formes les plus fréquentes pour construire le dictionnaire le plus petit. On trouve au total selon les dictionnaires 109, 141, 720 ou1 326 mots inconnus (couverture statique) et 96, 114, 492 ou 951 mots inconnus différents (couverture dynamique).
Le choix du dictionnaire en reconnaissance est critique, car les systèmes actuellement ne savent que reconnaître les mots présents dans leur dictionnaire. Pour une langue fléchie comme le français, où le taux moyen de flexions par lemme avoisine le chiffre 7, le problème devient crucial. Aussi performant que soit le système, il ne pourra faire moins de 27 fautes sur 1 000 mots (8), s'il ne sait traiter que 20 000 formes différentes.

3.2.3 Homographes et homonymes

La correspondance entre graphèmes et phonèmes n'est pas univoque. Certaines graphies de la langue française peuvent donner lieu à deux prononciations différentes. Par exemple, le graphème "***est***" se prononcera [est] (***"à l'est comme à l'ouest"***) s'il s'agit du substantif, mais il se prononcera [e] dans le cas de l'auxiliaire (***"tout est affaire de décor"***). Il en va de même des exemples classiques "***couvent***" et "***président***". Il arrive parfois que la classe grammaticale ne suffise pas à lever l'ambiguïté phonétique. Il faut recourir au sens pour savoir comment prononcer la graphie "***fils***" lorsqu'il est employé en tant que substantif masculin pluriel. Des exemples tels ***" les fils du président cousent avec des fils de soie*** " restent l'exception, une curiosité qui amuse.
A l'inverse, il est très courant de pouvoir écrire un même son de diverses façons. Les différentes graphies correspondant à une même Phonétique sont dites homonymes. Ainsi, "***sot***" a pour homonymes les mots "***sots***",

"***saut***", "***sauts***", "***seau***", "***seaux***", "***sceau***", "***sceaux***", "***Sceaux***" ... En général, une entrée du dictionnaire donne lieu par le jeu des inflexions grammaticales à plusieurs formes fléchies. Certaines d'entre elles se prononcent de la même façon bien que s'écrivant différemment. Par exemple, un substantif singulier entretient le plus souvent avec son pluriel une relation d'homonymie. Cette non-univocité est encore plus manifeste dans le cas des formes conjuguées d'un verbe ("***penser***", "***pensé***", "***pensée***", "***pensés***", "***pensées***" ...)
L'ampleur (9) de ce phénomène nous oblige à considérer comme entrée du dictionnaire le couple (graphie , phonétique). L'analyse de la relation graphèmes-phonèmes montre qu'en français, il y a peu d'homographes mais, un nombre élevé d'homophones.

(8) 1 000 mots représentent un peu moins de trois pages du présent article.
(9) En atteste la difficulté de la dictée, exercice redouté par tant d'écoliers, véritable épreuve polir nombre d'adultes une fois leur scolarité terminée.

3.2.4 Ambiguïtés

La langue est un lieu majeur d'ambiguïtés. Les mots qui la composent entretiennent entre eux des liens complexes: les sons, les graphies, les classes grammaticales, le sens ou le "référent" (c'est-à-dire le réel auxquels ceux-ci renvoient) ne se discernent pas de façon univoque et immédiate. Une parole, un écrit prêtent facilement à équivoque.
Dans la pratique, les concepts saussuriens de "signifiant" et «signifié » s'articulent sous des formes plurielles chacune des faces de la pièce faisant miroiter dans leur multiplicité les autres facettes qui la façonnent. Et quand bien même le contexte requis permet de lever l'ambiguïté, une signification seconde peut encore subsister motivée par des lectures plus ou moins conscientes. Un tel foisonnement d'hypothèses rend le traitement du langage naturel d'autant plus complexe que les frontières de mots ne sont pas marquées de façon explicite dans la chaîne phonétique en entrée. Que l'on parle en syllabes isolées ou en continu, on ne trouve pas de marque explicite de fin ou de début de mot. Ces carences combinées avec les ambiguïtés phonétiques mettent tout système de reconnaissance en difficulté.

4 La reconnaissance de la parole au Centre Scientifique IBM-France

4.1 La dictée automatique au CS IBM-France

Depuis 1985, le groupe de reconnaissance de la parole du Centre Scientifique IBM-France (CS) s'est fixé comme objectif de montrer la faisabilité d'une machine à dicter pour le français. Rapprochant une constatation (un système de dictée automatique ne peut reconnaître que les mots" présents dans son dictionnaire) d'une intention (il est souhaitable de pouvoir traiter à terme un vocabulaire complet de la langue) le groupe du CS a conçu le système *Parsyfal*, système original, adapté au français et permettant l'utilisation d'un très grand vocabulaire (200 000 formes).
En retour, il a bien fallu resserrer les autres contraintes. Par suite, le système est mono-locuteur et le mode d'élocution retenu pour cette application est le mode syllabes isolées. Il est donc momentanément demandé aux utilisateurs de ménager lors de la diction une pause entre les syllabes. Mais, l'objectif ultime est bien l'affranchissement de cette contrainte.
Parallèlement, depuis le milieu de l'année 1989, la décision a été prise de transposer le système *Tangora* à la langue française. Bénéficiant de l'expérience acquise avec le développement du système *Parsyfal*, cette adaptation a été rapidement menée à bien. Le système fonctionne aujourd'hui avec la même tenue que les systèmes anglais ou italiens,

4.2 Les principes de la reconnaissance de la parole probabiliste

4.2.1 Généralités

Les systèmes de reconnaissance de la parole actuellement développés chez IDM, sont fondés sur une approche probabiliste introduite par F. Jelinek [24]. Soit $W_n = w_1 w_2 \ldots w_n$, une suite de mots prononcée par le locuteur. Un processeur de signal extrait de cette prononciation une suite d'informations acoustiques: $A = a_1 a_2 \ldots a_m$. Il s'agit alors de trouver la suite de mots W dont la probabilité étant donnée la suite acoustique A, soit maximum. C'est A dire trouver W telle que:

(10) Par "mot" on entend ici forme fléchie.

$$P(W'/A) = \max_{W} P(W/A)$$

En utilisant la règle de Bayes, on obtient:

$$P(W/A) = \frac{P(A/W) \cdot P(W)}{P(A)}$$

où

- P(A) est la probabilité d'occurrence de la chaîne acoustique A, donc indépendante de W, le problème revient donc à maximiser P(A/W).P(W).
- P(A/W) est la probabilité d'observer la suite acoustique A si on prononce la suite de mots W. (Problème de modélisation acoustique).
- P(W) est la probabilité que la suite de mots W apparaisse dans le langage considéré.

On voit donc apparaître clairement les trois composantes d'un système de reconnaissance probabiliste:

- Un processeur de signal qui extrait une suite d'éléments acoustiques du signal.
- Un ou plusieurs modèles acoustiques probabilistes qui permettent de calculer pour toute suite de mots la probabilité d'une suite acoustique donnée (terme P(A/W)).
- Un modèle de langage qui calcule la probabilité dans la langue d'un suite de mots.

Les modèles théoriques utilisés sont, aussi bien pour les modèles acoustiques que pour les modèles de langage des modèles de Markov (automates probabilistes d'états finis)[25].

4.2.2 Modèles de Markov

Une source de Markov *(11)* est un automate probabiliste d'états finis qui se définit, dans le cas d'une approche discrète, par la donnée de :

- E : un ensemble fini d'états *(12)* , dont un état initial e1 et un état final eF
- L : un alphabet fini de labels *(13)*
- et un sous ensemble T de E x L x E

Les éléments r e T, r = (ei, lk , ej) sont appelés des transitions. Une transition donnée r désigne le passage de l'état Ici à l'état ej accompagné de l'émission (on dit parfois aussi la production) du label Ik. Les transitions pour lesquelles ei = ej, sont appelées des boucles.

Paramètres de la source.

q(,r) =- q(ei . lk , ej) est la probabilité d'émission du symbole Ik lors du passage de l'état ei à l'état ej . Une distribution est attachée à chaque état de telle sorte que

$$\forall e_i \qquad \sum_{e_j,\, l_k} q(e_i, l_k, e_j) = 1$$

(11) Pour l'ensemble de ce chapitre, on pourra se reporter aux ouvrages ou articles référencés [1, 23, 26, 27, 28, 29, 30].
(12) les termes états ou sommets sont employés indifféremment
(13) les termes symboles, labels ou étiquettes sont équivalents

La probabilité q (e_i, l_k,e_j) peut se décomposer en un produit de deux facteurs: la probabilité $a_{i,j}$ de transiter de e_i vers e_j et la probabilité $b_{i,j,k}$ d'émettre le Symbole l_k sachant que l'on a pris cette transition particulière -r.

$$\begin{aligned} a_{i,j} &= p(e_j \mid e_i) \\ b_{i,j,k} &= p(l_k \mid e_i, e_j) \\ q(\tau) &= q(e_i, l_k, e_j) = a_{i,j} \times b_{i,j,k} \quad \forall\, l_k \neq \phi \end{aligned}$$

Ces probabilités s'appellent les paramètres de la source de Markov.

Définition.
Un chemin à travers une (ou plusieurs) machine(s) de Markov est la donnée d'une suite finie d'éléments de T telle que : Dep (Ti) = ei et Arr (ri) = Dep (,ri+1) pour i = 1 où Dep et Arr sont des applications de T dans E vérifiant
Dep (e, , 1, e,) = ci et Arr (e, , 1, ej) = ej . Leur fonction est de déterminer le départ et l'arrivée d'une transition (on emploie aussi bien les termes origine et but).

Chemin de probabilité maximale.
Lorsqu'est recherché le chemin allant d'état initial à état final et ayant assuré la production d'une suite de labels avec une probabilité maximale, il est classique d'utiliser l'algorithme de Viterbi. Les applications de l'alignement de Viterbi dans le cadre du traitement de la parole sont nombreuses et variées :

- segmentation et étiquetage phonétiques automatiques
- étiquetage grammatical et morphologique
- simplification de l'apprentissage ou du décodage

Estimation des paramètres.
L'estimation des paramètres d'une source de Markov selon un maximum de vraisemblance, consiste à chercher les paramètres maximisant la probabilité que cette source ait produit une collection d'observations relatives à l'unité modélisée par la source.
Or, le calcul d'une telle probabilité suppose que les paramètres soient connus. On résout ce problème par une procédure itérative, dite algorithme de Baum-Welch ou 'Forward-Backward", proposée par Baum [31] qui en a prouvé la convergence vers un optimum local. Son apport a été décisif pour rendre la modélisation stochastique applicable à la résolution de problèmes concrets.

Sources de Markov phonétiques.
L'intérêt d'une approche probabiliste est de pouvoir considérer la parole comme "produite par une hiérarchie de sources de Markov" [1 (p.364)] au niveau des phonèmes, des syllabes, des mots ou des phrases. Pour la commodité de l'exposé, nous donnerons ici un exemple de modélisation phonétique.
Par définition, le phonème est la plus petite unité constitutive d'un mot, auquel elle petit faire perdre Son sens par substitution. Il est bon de rajouter à cette définition du phonème, un point de vue structuraliste, qui restitue le phonème comme élément d'un système dans lequel il entretient des relations d'oppositions avec les autres éléments. On dit alors que le phonème est une unité composite, combinaison de traits distinctifs. Un des systèmes utilisés au CS [32] est construit autour d'un système phonétique courant pour le français de 33 éléments. Ce système se compose de 14 voyelles, 16 consonnes et 3 semi-voyelles, auxquelles il faut rajouter 7 machines Pseudo-phonétiques modélisant le silence, les occlusions, les chutes énergétiques ou l'impulsion glottale.

Machines " phonétiques ".
De façon classique (cf. les systèmes *Tangora* ou *Sphinx*), chaque phonème est représenté par une machine de Markov comptant sept états. Les phonèmes p, t et k nécessitent la concaténation de deux machines:

- la première relative à l'occlusion (oP, oT ou oK)
- la seconde relative à l'explosion (bP, bT ou bK).

La structure de la machine phonétique peut se représenter comme suit

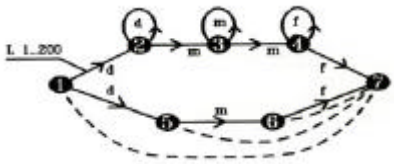

Figure 1. Une machine de Markov sept états

Une flèche pleine correspond à K transitions entraînant chacune l'émission d'un des K labels (K -200 dans notre cas). Un label est ici un événement acoustique centiseconde.
Une flèche pointillée représente une transition vide.

Puisque les états ne traduisent pas une réalité physique directement observable, de telles chaînes de Markov sont dites cachées ou autrement dénommées sources de Markov.
Malgré tout, on peut donner une lecture de la structure même des machines phonétiques en terme de comportement physique. Dans le schéma ci-dessus, les six transitions basses (partie inférieure de la machine) correspondent à une prononciation rapide du phonème, les sept transitions du haut à une prononciation plus lente.
L'axe horizontal quant à lui, représente l'axe temporel et peut être découpé en trois segments relatifs à la prononciation du phonème:

- Une phase transitoire : l'attaque
- Une phase stable : le cœur du phonème
- Une phase transitoire : la chute

4.3 Les systèmes Tangora

4.3.1 L'architecture.

Le système est implémenté Sur un ordinateur personnel IBM PS/2 qui contient une carte d'entrée sortie ainsi qu'une carte spécialisée pour les algorithmes spécifiques utilisés [33].

4.3.2 Le traitement du signal.

Le signal est amplifié et digitalisé (20K échantillons à la seconde, 12 bits par échantillon) par la carte de traitement de signal, puis les données sont soumises à une transformée de Fourrier qui s'applique toutes les 10 tris. sur un fenêtre de 25.6 tris. de signal. Un vecteur de 20 éléments est alors extrait du spectre ainsi obtenu, chaque élément représentant l'énergie dans une bande de fréquence entre 200 Hz et 8 Khz. Ces vecteurs sont enfin comparés A 200 prototypes dépendants du locuteur et on remplace chacun d'eux par l'indice du prototype le plus proche (distance euclidienne). Ce processus permet de réduire le débit des données de 30 000 à 100 octets par seconde.

4.3.3 Les processus de comparaison acoustiques.

Deux passes de comparaison acoustique sont effectuées: une passe relativement grossière qui sélectionne une première liste de mots candidats (la taille maximum de cette liste peut être choisie par l'utilisateur, elle est en moyenne de l'ordre de 500 mots). Une deuxième passe plus précise est alors effectuée uniquement sur les éléments de cette liste. Ces deux processus utilisent des modèles de Markov de différents éléments acoustiques de la langue (exemple: phonèmes) (De l'ordre de 80 modèles pour la première passe, 200 modèles pour la seconde).

4.3.4 Le modèle de langage.

C'est un modèle probabiliste dit trigramme, qui affecte à une suite de trois mots une probabilité obtenue par une combinaison des probabilités des mots et des suites de deux ou trois mots. Les probabilités sont estimées à partir d'un grand corpus de mots (plusieurs millions de mots).

4.3.5 L'algorithme de décodage.

Il fonctionne de gauche à droite en combinant les divers scores obtenus à tous les niveaux du processus pour déterminer la suite de mots qui a le plus probablement été prononcée.
A chaque instant, on étend la meilleure hypothèse présente non encore étendue .

4.3.6 Les limites de Tangora.

Comme tout système mono-locuteur, les systèmes *Tangora* nécessitent un apprentissage d'une centaine de phrases courtes avant utilisation. Etant donné la taille du dictionnaire géré, ils sont utilisables dans les domaines d'application qui ont servi à la collecte de leur modèle de langage (correspondance de bureau, journaux, rapport de radiologie etc. ..), mais fonctionnent en temps quasi réel.
Le système français est, comme les autres, un système en mots isolés, où les liaisons sont impossibles. Par ailleurs, comme nous l'avons déjà souligné, le français possède beaucoup d'homophones non homographes (mots dont la prononciation est la même niais dont l'orthographe diffère: parti, partis, partie, parties) qui entraînent des erreurs du système.

Les ambiguïtés sont moindres par exemple en anglais où le "s" final du pluriel se prononce.
De plus, dans une perspective moyen terme, il est évidemment souhaitable de pouvoir traiter Lin vocabulaire complet de la langue: 20 000 mots couvrent en moyenne 95% des textes, il en faut 200 000 pour couvrir 99 %.
Ces considéra t ions ont conduit le groupe du CS, à la construction d'un système original au français et accédant un très grand vocabulaire : *Parsyfal*.

4.4.1 L'approche syllabique

Dès le début de sa recherche sur la Dictée Automatique, le groupe du CS a choisi un élément acoustique original: la syllabe. De nombreux éléments acoustiques sont possibles (phonèmes simples, phonèmes contextuels, triphones, mots etc. ...) [34]. Mais la syllabe présente un certain nombre d'avantages [30].
Tout d'abord, étant plus longue que le phonème, elle fournit une première contrainte sur les suites de phonèmes possibles. En tant que sous ensemble du mot, elle permet de couvrir un très grand dictionnaire avec un nombre relativement restreint d'éléments (5 200 syllabes phonétiques suffisent à décrire acoustiquement un dictionnaire de 200 000 mots français). Enfin, elle permet simplement d'avoir accès aux liaisons et apostrophes puisqu'en rajoutant 1 200 syllabes supplémentaires, on couvre ce même dictionnaire, y compris liaison et apostrophe [35].

4.4.2 Le modèle triclasse

La spécificité du système *Parsyfal* est la taille du vocabulaire qu'il peut gérer. Plutôt qu'un système réservé à un certain domaine (correspondance administrative pour *Tangora* américain, articles économiques ou rapports de radiologie pour *Tangora* italien, dépêches de l'AFP pour *Tangora* français), nous avons voulu construire un système plus général dont le vocabulaire soit beaucoup plus étendu, et qui soit relativement aisé à mettre à jour.
Au lieu de considérer les mots rencontrés le plus fréquemment dans un certain corpus, le dictionnaire est construit à partir d'un dictionnaire de base en rajoutant toutes les formes déclinées fréquentes. Il est clair que la taille des données à stocker pour un tel système, si l'on voulait utiliser des statistiques trigrammes est prohibitive. De plus, pour un modèle trigramme, la mise à jour pour ajouter un mot non rencontré est très difficile.
Nous avons donc introduit un modèle de langage dit triclasse [36] qui utilise les probabilités de suites de classes grammaticales. Plus précisément, nous définissons un ensemble de 103 classes pseudo grammaticales. Il est alors possible, pour chaque mot du dictionnaire, de stocker ses fréquences relatives pour toutes les classes qu'il peut avoir.
Exemple: "partis" peut être un substantif masculin pluriel ou un participe passé. "couvent" peut être un substantif masculin singulier ou un verbe conjugué etc. ..
De plus il est tout à fait possible, pour 103 classes de calculer et stocker les comptes des suites de deux et trois classes. Un tel modèle permet d'éviter certaines des erreurs d'accord simple à contexte proche que ne peut pas toujours contourner un modèle trigramme. Par ailleurs il est relativement aisé de rajouter un mot nouveau, puisqu'il suffit de le rajouter dans le lexique, les comptes triclasses et biclasses étant inchangés.
Un modèle 'morphologique", basé sur la notion de lemme a aussi été défini plus récemment et implémenté dans *Parsyfal* afin de permettre la gestion de très grands vocabulaires tout cri gardant une composante sémantique [30].

4.4.3 Architecture de Parsyfal

Il est à noter que, dans cette première étape, le système utilise un mode d'élocution en syllabes isolées (contrairement aux mots isolés de *Tangora*) mais permet par ailleurs l'utilisation de liaisons.
Seuls le traitement de signal et l'approche probabiliste sont communs avec *Tangora*.
Toutes les étapes ultérieures du traitement sont différentes et l'algorithme de décodage 'Multi Level Decoding" est adapté à l'emploi de la syllabe aussi bien comme imité acoustique que comme unité d'accès lexical [23].

4.4.4 Implémentation et résultat

Parsyfal est implémenté sur un ordinateur personnel IBM connecté à un IBM 370. Le traitement du signal est effectué sur la même carte spécialisée que dans *Tangora*. La suite du traitement a lieu sur le système VM qui renvoie la phrase décodée au PC pour affichage. Le système peut décoder en temps réel avec un taux d'erreur de 9,5% et un dictionnaire de 200 000 mots.

5 Conclusion

5.1 Les applications envisageables

L'application de prédilection est la dictée de lettres, de rapports, d'articles ou même de livres. Elle concerne doublement le secteur de la presse et de l'édition, certainement toute entreprise relevant du secteur tertiaire pour améliorer le traitement des documents circulant dans le flux de la communication interne ou externe, mais de façon plus générale, le secrétariat de toute entreprise d'envergure quelle que soit son activité.
Certains secteurs médicaux semblent bien se prêter à une utilisation effective de la dictée automatique. En particulier, la transcription de rapports médicaux, par exemple en radiologie, est le sujet d'un certain nombre d'expérimentations. Elle présente les deux caractéristiques favorables d'un environnement calme, qui conditionne la qualité de la reconnaissance, et d'un grand volume de production, qui rend important tout gain de productivité. De plus, le langage utilisé pour ces rapports semble bien convenir à l'état de l'art actuel de la technologie de la dictée automatique: la taille du vocabulaire utilisé et la variété des tournures syntaxiques sont assez grandes pour dépasser le champ des systèmes de commande vocale, mais toutefois assez limitées pour en faire une application raisonnable.
La machine à dicter peut s'avérer fort utile pour des personnes handicapées au niveau des membre, supérieurs ou de l'appareil auditif. Au risque de plagier la parabole de l'aveugle et du paralytique, retenons qu'aujourd'hui une machine à dicter connectée à un téléphone permet à un sourd de communiquer à distance avec un tétraplégique, fut-il non voyant.
Le sous-titrage en direct d'émission télévisée est une tâche complexe, lourde en implications techniques et économiques. Si l'on peut identifier dans cette tâche un caractère répétitif, la machine à dicter devient un outil idéal pour la réaliser. on peut d'ores et déjà imaginer qu'un locuteur ayant entraîné le système à sa voix résume devant la machine placée dans un studio voisin (sans être entendu de quiconque) les propos du présentateur ou de la personne interviewée. Grâce à cc locuteur intermédiaire (qui pourrait être aussi un traducteur), des téléspectateurs français pourraient voir s'afficher en temps réel, incrustées en bas de l'écran, les paroles prononcées pal, un chanteur espagnol, un écrivain anglais ou même un présentateur français voulant que son message soit perçu par les malentendants.

5.2 Les perspectives

Les progrès technologiques réalisés ces dernières années ont permis la réalisation de prototypes de dictée automatique basés sur des configurations matérielles de faible coût, une station de travail et des cartes spécialisées. Ces prototypes présentent encore des restrictions, l'entraînement d'un nouveau locuteur au système, l'élocution en mots isolés, la définition du vocabulaire spécifique à l'application... Il est toutefois possible que ces prototypes soient utilisables pour certaines applications précises, où ces contraintes sont compensées par les commodités induites par l'utilisation d'une entrée vocale. Les prochaines années verront sans doute se développer ces système, spécialisés qui serviront de précurseurs à d'autres systèmes Plus évolués où les restrictions d'utilisation disparaîtront pou à peu.

Bibliographie

[1] CALLIOPE CNET-ENST eds. *La parole et son traitement automatique*. Masson, septembre 89.
[2] Gagnoulct C, Jouvet D. Développements récents en reconnaissance de la parole. *L'écho des recherches*. CNET-ENST, N° 135, 1989, pp. 27-36
[3] Levinson S, Rabiner L, Sondhi M. *Speaker-Independant Isolated Digit Recognition Using Hidden Markov Models*. ICASSP 1993, Boston, pp 1049-1052.
[4] El-Bèze M. *Caractérisation des sons stables et assistance automatique pour améliorer leur production.* Thèse pour obtenir le diplôme d'ingénieur, IIE CNAM, Paris juin 1984.
[5] Tubach JP, Gagnoulet C, Gauvain JL. *Advances in speech recognition products from France* . Conférence Speech Tech 1989.
[6] Lippmann RP. *Neural Nets for Computing*. ICASSP 1988, New-York, pp. 1-6.
[7] Levin E. *Word Recognition using Hidden Control Neural Architecture.* ICASSP 1990, Albuquerque, pp. 433-436.
[8] Lee KF, Hon HW, Reddy R. *An Overview of the SPHINX Speech Recognition System* . IEEE Trans on ASSP, 38 N° 1, janvier 90, pp. 35-45.
[9] Pierrel JM *Utilisation des contraintes linguistiques en compréhension de parole continue le système Myrtille II.* TSI, Vol 1, N° 5, 1982, pp. 403-421.

[10] Pérennou G. The ARIAL II Speech Recognition System In: Haton JP ed, *Automatic Speech Analysis and Recognition* . 1982, pp. 269-275.
[11] Averbuch A. et al. *Experiments with the TANGORA 20,000 word speech recognizer* . ICASSP 1987, Dallas, pp. 701-704.
[12] Alto P, Brandetti M, Ferretti M, Maltese G, Scarci S. *Experimenting Natural-Language Dictation with 20,000- Word Speech Recognizer* , Proceedings of IEEE, CompEuro 1989, Section 2 pp. 78-91, Hambourg.
[13] D'Orta P, Ferretti M, Martelli A, Melecrinis S, Scarci S, Volpi G. *A Speech Recognition System for the Italian Language* . ICASSP 1987, Dallas, pp 941-943.
[14] Cerf-Danon H, de La Noue P, Diringer L, El-Bèze M, Marcadet JC. *A 20,000 words, automatic speech recognizer*. Adaptation to French of the US TANGORA system , Nato 1990.
[15] Wothke K, Bandara U, Kempf J, Keppel E, Mohr K, Walch G. *SPRING Speech Recognition System for German* . Eurospeech, Paris, septembre 1989, Vol 2, pp. 9-12.
[16] Baker J. DRAGON DICTATE (TM) -30K *Natural Language Speech Recognition statistical methods* . *Eurospecch* 89 Vol 2, Septembre 1989, pp. 161-163.
[17] Fanchettc F. Speech recognition .- Talk about progress . *Language Technology*, N° 19, p. 4-5.
[18] Le Breton JL. *Le premier traitement de texte vocal* . L'ordinateur individuel N° 98, décembre 97 p. 108-110.
[19] Mariani .J. Hamlet *un prototype de machine à écrire à entrée vocale*, Journal d'acoustique 2 mars 1989, Paris, pp. 79-83.
[21] Lee LS, Tseng CY, Gu HY, Liu FH, Chang CH, Hsich SH, Chen CH *A Real- Time Mandarin Dictation Machine for Chinese Language with Unlimited Texts and Very Large Vocabulary* . ICASSP 1990, Albuquerque, pp 65-68.
[22] Bahl L, Bakis R, Bellagarda J, Brown P, Burshtein D, Das S, de Souza P, Gopalakrishnan P, Jelinck F, Kanevsky D, Mercer R, Nadas A, Nahamoo D, Picheny M. *Large Vocabulary Natural Language Continuous Speech Recognition*. ICASSP 1989, Glasgow, Vol SI pp 465-467.
[23] Wrialdo B. Multi Level Decoding for Very-Large-Size-Dictionary speech recognition . *IBM J. Res. Develop*., Vol 32, N° 2, Mars 1988, pp. 227-237 .
[24] Jelinck F. *Continuous speech recognition by statistical methods.* Proceedings of the IEEE, Vol 64, Avril 1976, pp 532-556.
[25] Levinson S E. *Structural methods in automatic speech recognition*. Proceedings of the IEEE, Vol 73, N° 11, Novembre 1985, pp 1625-50.
[26] Jelinck F, Mercer RL, Bahl LR. *Continuous Speech Recognition: statistical methods* In: Krishnaiah PR, Kanal LN, eds. Handbook of statistics, Volume 2, Classification, pattern recognition and reduction of dimensionality. North-Holland, 1982.
[27] Levinson S, Rabiner L, Sondhi M. An Introduction to the Application of the theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition *ATT Bell System Technical Journal*, Vol 62, N° 4, Avril 1983, pp. 1035-1074.
[28] Poritz A *Hidden Markov Models: a Guided Tour* . ICASSP 1988, New York , Vol S1, pp. 7-13.
[29] Roseaux *Exercices et problèmes résolus de recherche opérationnelle, T2, phénomènes aléatoires en recherche opérationnelle* . Masson 1987, pp. 1-76.
[30] El-Bèze M, *Choix d'unités appropriées et introduction de connaissances dans des modèles probabilistes pour la reconnaissance automatique de la parole* . Thèse de doctorat d'état, Université Paris 7, Novembre 1990.
[31] Baum L. *An Inequality and Associated Maximization Technique in Statistical Estimationfior Probabilistic Functions of Markov Process* . Academic Press, Inequalities, Vol III pp 1-8, 1972.
[32] Cerf H, Dcrouault AM., El-Bèze M, Mcrialdo B, Soudoplatoff S. *Speech Recognition experiment with 10,000 word vocabulary* . NATO Advanced Institute on Pattern Recognition, 18-20 juin 1986, Bruxelles, pp. 204-209.
[33] Schichman G. Personal instrument (PI) - A PC based signal processing system. *IBM .J. Res. Develop.* Vol 29, N° 2, Mars 1985, pp. 158-169,
[34] Dcrouault AM. *Context-dependant phonetic Markov Models for large vocabulary speech recognition*.. ICASSP 97, Dallas.
[35] Cerf H, Derouault AM, El-Beze M, Wrialdo M, Soudoplatoff S. *Reconnaissance de la parole par des modèles markoviens application aux grands vocabulaires.* 15ème JEP Aix en Provence, 27-29 Mai 1986.
[36] Dcrouault AM, Wrialdo B. *Language modeling at the syntactic level* . 7th International Conference on Pattern Recognition, Août 1984, Montreal.