TECHN 29

Publication technique de la SmalS-MvM

04/2005

Reconnaissance vocale
Les systèmes de dictée continue

Le langage est la peinture de nos idées…
(le Comte de Rivarol)[1]

1. Introduction

Hervé Haut

est ingénieur civil et docteur en sciences physiques. Après plusieurs années de recherche en physique théorique à l'UCL, il a occupé diverses fonctions informatiques dans le secteur privé. Il a rejoint SmalS-MvM en 1998 comme consultant à la section des Recherches où il effectue principalement des missions de consultance pour des projets de gestion documentaire et de workflow.

Contact : 02 509 58 16
Herve.haut@smals-mvm.be

Le traitement automatique des langues est un vaste domaine de recherche où se côtoient des spécialistes de nombreuses disciplines : des linguistes, des informaticiens, des logiciens, des psychologues, des traducteurs… C'est aussi un domaine économiquement porteur dont les applications sont nombreuses dans des secteurs aussi divers que la bureautique, l'aide aux handicapés, l'enseignement, la domotique, la traduction, l'aide à la navigation, la documentation…

Dans la suite de cet article, nous nous limiterons aux technologies de traitement de la parole et, plus particulièrement, à la reconnaissance vocale dans les systèmes de dictée continue.

Si l'on écoute le son émis par un modem qui se connecte à un serveur, ce sifflement aigu accompagné de parasites nous est complètement inintelligible. A l'inverse, sans traitement approprié, nos paroles sont tout aussi incompréhensibles pour la machine. Or le langage est de loin notre moyen de communication le plus rapide et le plus expressif ; il n'est donc pas étonnant que depuis les débuts de l'informatique, des recherches aient été effectuées dans le but de communiquer avec l'ordinateur par ce moyen tellement naturel pour nous.

Après bien des années, ces recherches ont finalement abouti et aujourd'hui la reconnaissance de la voix humaine par l'ordinateur se banalise et quitte le banc de laboratoire pour se retrouver dans les rayons de nos supermarchés et devenir ainsi accessible au grand public dans divers types d'applications utiles.

Après un rapide historique de l'évolution de la recherche en technologie vocale, nous consacrerons quelques paragraphes aux domaines d'application des technologies de traitement de la parole avant de nous consacrer exclusivement aux systèmes de dictée continue où l'utilisateur peut dicter son texte de manière fluide et naturelle avec un vocabulaire suffisamment évolué.

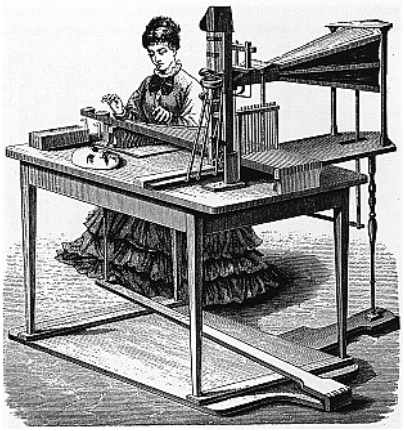
Nous verrons comment fonctionnent (dans les grandes lignes) les systèmes de reconnaissance vocale avec les problèmes que ces technologies doivent résoudre pour atteindre un niveau suffisant de fiabilité.

[1] Discours sur l'universalité de la langue française (1784).

Nous décrirons l'état actuel de la technologie et ses perspectives d'avenir avant de présenter les résultats que nous avons obtenus avec le logiciel (versions française et néerlandaise) "Dragon Naturally Speaking Preferred" de Scansoft Inc., leader actuel du marché. Enfin nous terminerons par quelques recommandations quant à l'utilisation d'un logiciel de dictée continue.

2. Historique

On a coutume de fixer l'origine des recherches en reconnaissance vocale aux années 1950. C'est à cette époque en effet qu'IBM commence à investir dans ce domaine avec comme objectif de développer une nouvelle forme d'interaction entre l'homme et la machine.



Orgue vocal (1846)

Il est cependant amusant de mentionner que, un siècle plus tôt, on s'intéressait déjà au problème connexe qu'est la synthèse vocale, c'est-à-dire aux possibilités de faire parler des machines. C'est ainsi qu'en 1846, un certain Joseph Faber construisait à Londres un "orgue vocal" capable de reproduire des phrases ordinaires et même de chanter le "God Save the Queen" ! Plus tard, en 1890, Edison mettait sur le marché une poupée parlante à 10$ (une somme équivalente au salaire de deux semaines de travail de l'époque) capable de réciter quelques vers d'une comptine ; c'était le début de l'histoire du phonographe.

A la fin des années cinquante, IBM développe le premier ordinateur entraîné à écouter des modèles spécifiques de sons et à dégager des corrélations statistiques entre ces sons et les mots qui y correspondent. En 1964, IBM fait la première démonstration de reconnaissance vocale : le logiciel "Shoe Box" permet de reconnaître une série de chiffres dictés. Cette démonstration incite le ministère américain de la Défense à financer un programme de recherche pour développer cette nouvelle technologie. C'est également ainsi que naît l'approche statistique dans le domaine de la reconnaissance vocale et que les techniques d'apprentissage voient le jour, techniques basées sur des algorithmes statistiques habituellement utilisés dans les théories de l'information. Ces techniques statistiques sont encore aujourd'hui considérées comme la meilleure approche et sont celles qui ont abouti à des produits concrets (par opposition aux techniques basées sur les réseaux neuronaux dont nous ne parlerons pas).

En 1984, IBM présente le premier système de reconnaissance vocale au monde disposant d'un lexique de 5000 mots et bénéficiant d'un taux de reconnaissance de 95%. Ce logiciel nécessite 3 processeurs vectoriels et un grand système 4341 avec une interface utilisateur fonctionnant sur un ordinateur Apollo. Le logiciel permet à un utilisateur expérimenté de dicter ses textes en mode discret, c'est-à-dire en marquant une pause entre chaque mot. La même année, Philips commence le développement de "SPICOS", un logiciel de reconnaissance avec un vocabulaire de 1000 mots.

Dans les années suivantes, les développements vont s'accélérer. La puissance croissante des processeurs (et leur diminution de coût) va en effet permettre d'améliorer constamment les performances des algorithmes utilisés et également de traiter ces algorithmes par des logiciels et non plus par du hardware dédicacé. Plus tard encore, l'émergence de la carte son Soundblaster de Creative Labs comme standard de fait va encore favoriser le développement et la diffusion de ces logiciels sur les postes de travail PC compatibles.

A partir des années nonante, de nouveaux acteurs se lancent dans ce marché et de nouveaux produits font leur apparition. Dragon Systems annonce la sortie de son premier logiciel de dictée en 1990 ; Apple lance en 1993 son produit "Plain Talk" ; en 1994, IBM commercialise "IBM Personal Dictation System" pour PC OS/2.

3. Domaines d'application

Bien que notre sujet principal soit les systèmes de dictée continue, il nous a paru utile de présenter une classification et un bref résumé des différents domaines d'application où les technologies de traitement de la parole jouent un rôle important. On distinguera essentiellement la synthèse vocale et la reconnaissance vocale. On notera cependant que dans la plupart des applications courantes, ces deux technologies sont souvent associées.

3.1. La synthèse vocale

La synthèse vocale peut être définie comme la communication de la machine à l'homme. Pour qu'un texte puisse être transformé en paroles par une machine, il importe de découper le texte en morceaux correspondant de manière univoque à une unité de son. On conçoit facilement que si cette découpe se faisait par exemple au niveau des mots, il serait nécessaire de stocker en mémoire la prononciation de tous les mots d'une langue, ce qui n'est guère concevable. C'est la raison pour laquelle cette découpe se fait généralement au niveau des phonèmes[2] , le texte étant ainsi "traduit" de façon phonétique. Des modules de production du son (synthétiseurs) peuvent alors, sur la base de cette analyse, "lire" le texte. Cette technologie est aujourd'hui bien maîtrisée au niveau de la prononciation des mots. Au niveau des phrases, il reste encore pas mal de développements à réaliser pour obtenir une prosodie correcte, c'est-à-dire pour interpréter les phrases avec le ton, le timbre, le phrasé, le rythme et l'emphase qui caractérisent le langage humain ; les améliorations dans ce domaine nécessitent l'usage de dictionnaires, d'analyses grammaticale et sémantique analogues à celles utilisées en reconnaissance vocale. Nous nous y attarderons plus loin.

Parmi les applications de la synthèse vocale, nous citerons :

- l'aide aux personnes handicapées : une personne privée de la parole peut par exemple communiquer par téléphone avec un tiers en tapant son message sur PC qui peut le lire pour son correspondant, ou encore un malvoyant peut avoir un ordinateur qui lui lit des textes ;
- l'interaction par téléphone avec une base de données de produits pour en obtenir une description ou avec une centrale d'aide en ligne ;
- les bornes interactives à vocation touristique par exemple ;
- la possibilité de consulter son courrier électronique à distance via une communication téléphonique ;
- la possibilité d'inclure des messages vocaux dans des applications de bureautique ou dans des pages Internet.

3.2. La reconnaissance vocale

La reconnaissance vocale recouvre tous les aspects liés à l'interprétation, par la machine, du langage humain. Dans ces applications de reconnaissance vocale, nous distinguerons les systèmes de commandes vocales et les systèmes de dictée.

[2] Un phonème correspond à la plus petite unité de la chaîne linguistique (sonore) qui peut être considérée comme unité distinctive sans être elle-même nécessairement significative ; c'est un son élémentaire non ambigu.

3.2.1. Les systèmes de commandes vocales

Les applications de ce type permettent notamment à l'utilisateur de contrôler verbalement des équipements. Par complexité croissante, on peut classifier ces systèmes en trois groupes :

1. Les systèmes à reconnaissance discrète
   Ce sont les applications où soit un nombre limité de mots soit de courtes phrases peuvent être utilisés pour commander le système. On y retrouve par exemple les applications téléphoniques où l'on peut choisir vocalement un point de menu (navigation interactive), le contrôle vocal des commandes de menus dans les logiciels ("fermer fichier, sortir"), certains logiciels de saisie automatique de données où les valeurs sont à choisir dans une liste limitée connue.

2. Les systèmes à reconnaissance "à la volée"
   Ces systèmes permettent à l'utilisateur de s'exprimer par des phrases mais sont entraînés à repérer certains mots de la phrase, mots qui se trouvent dans son dictionnaire interne et sur lesquels ils basent leur action. La consultation d'un horaire de chemin de fer est un exemple de ce type de système : sur base de la phrase "je voudrais me rendre de Bruxelles à Paris lundi prochain", le système repérera "Bruxelles", "Paris" et "lundi" pour proposer l'horaire correspondant.

3. Les systèmes à reconnaissance continue
   On trouvera ici les applications les plus avancées de commande vocale où l'on peut s'adresser au système en langage naturel. On trouvera des exemples dans les systèmes évolués de dictée où l'on pourra inclure des commandes vocales évoluées comme "souligner et mettre en gras le troisième mot de ce paragraphe".

3.2.2. Les systèmes de dictée

Les systèmes de dictée constituent le problème le plus difficile à résoudre dans le domaine de la reconnaissance vocale. Comme pour les systèmes de commande vocale, les applications de dictée peuvent être divisées en plusieurs catégories en fonction de leur complexité :

1. Les systèmes de reconnaissance discrète
   On classe ici les systèmes où l'utilisateur doit parler avec de courtes poses entre chaque mot ; ce sont les premiers systèmes de dictée qui ont été développés dans les années quatre-vingts. Ces systèmes n'ont plus beaucoup de succès aujourd'hui étant donné les progrès accomplis dans la dictée continue.

2. Les systèmes de reconnaissance continue
   C'est la "quête du Graal" des chercheurs en reconnaissance vocale : permettre à un utilisateur de dicter son texte à l'ordinateur, de façon continue, avec un vocabulaire riche, à une vitesse de locution normale et avec une reconnaissance proche de 100%. C'est à ce type de système que nous nous intéressons dans la suite.

4. Les systèmes de dictée continue : théorie et technologie

Que l'on veuille utiliser la reconnaissance vocale pour composer un numéro de téléphone, naviguer parmi les fenêtres sur notre PC, entrer des données dans un logiciel ou dicter une lettre dans un traitement de texte, le problème de base reste le même : identifier le sens d'un flux de paroles prononcées souvent dans un bruit de fond plus ou moins important.

Cette tâche est rendue difficile non seulement par les déformations induites par l'usage d'un micro mais aussi par une série de facteurs inhérents au langage humain :

- les homonymies[3] : une même séquence de sons peut correspondre à plusieurs mots (par exemple le son "s-en" [s ɑ̃] dans 100, cent, sans, san [Francisco], cents, sang, [je] sens, [il] sent) ;
- les accents locaux ;
- les habitudes du langage (comme certaines élisions qui rendent difficile la séparation des mots : "j'vais l'chercher" pour "je vais le chercher") ;
- les différences de vitesse entre les locuteurs ;
- les imperfections d'un micro…

Pour notre oreille humaine, ces facteurs ne représentent généralement pas de difficultés. Notre cerveau jongle avec ces déformations de la parole en prenant en compte, quasi inconsciemment, des éléments non verbaux et contextuels qui nous permettent de lever les ambiguïtés.

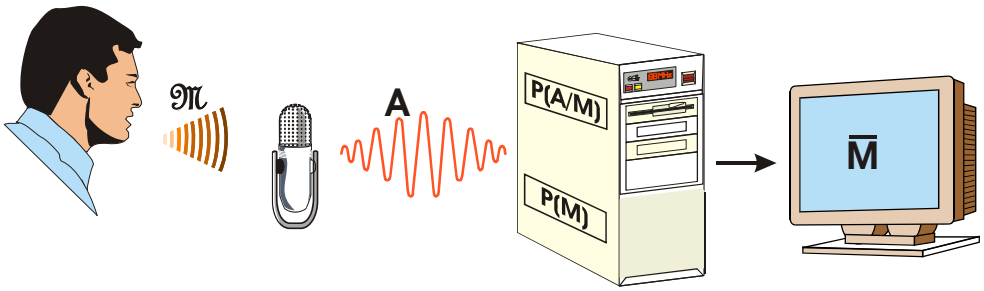
Ce n'est qu'en prenant en compte ces éléments extérieurs au son proprement dit que les logiciels de reconnaissance vocale pourront atteindre des degrés élevés de fiabilité.

Aujourd'hui, les logiciels de reconnaissance vocale qui donnent les meilleurs résultats sont tous basés sur une approche probabiliste.

Le but de la reconnaissance vocale est de reconstituer une séquence de mots $\mathfrak{M}$ à partir d'un signal acoustique enregistré **A**.

Dans l'approche statistique, on va considérer toutes les suites de mots **M** qui pourraient correspondre au signal **A.** Dans cet ensemble de suites possibles, on choisira alors celle ($\overline{M}$) qui est la plus probable, c'est-à-dire celle qui maximise la probabilité[4] **P(M/A)** que **M** soit l'interprétation correcte de **A**, ce que l'on notera :

$$\overline{M} = \arg\max_{M} P(M/A) .$$



[3] Homonyme: se dit des mots de prononciation identique et de sens différents, qu'ils soient de même orthographe ou non.

[4] **P(A/B)** représente la probabilité de l'événement **A** si l'événement **B** a eu lieu. L'axiome de Bayes permet de calculer la probabilité du concours de deux événements **A** et **B** par les égalités suivantes :

P(A et B) = P(A/B) P(B) = P(B/A) P(A)

où **P(A)** représente la probabilité que l'événement A ait lieu.

L'axiome de Bayes permet de réécrire l'expression précédente :

$$\overline{M} = \arg\max_{M} \frac{P(A/M)\,P(M)}{P(A)}$$

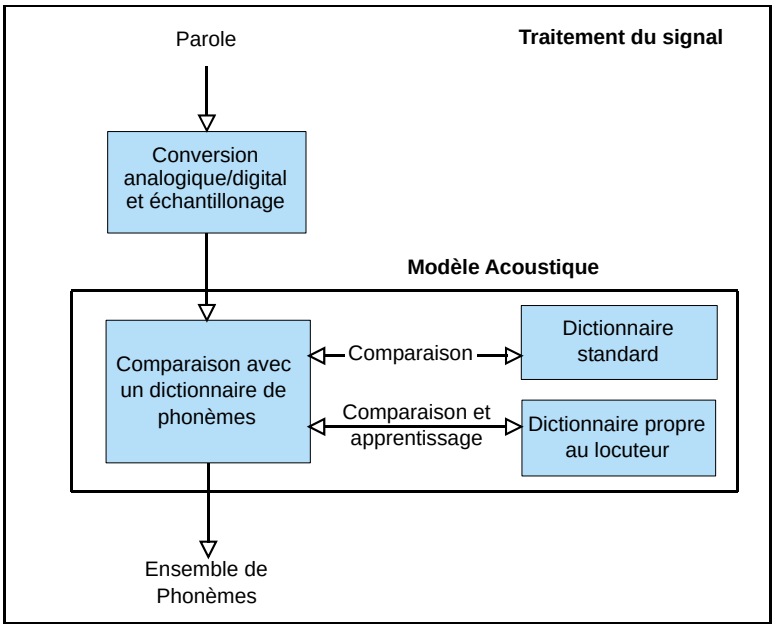
et comme **P(A)** est une constante dans la recherche du meilleur M, on a finalement :

$$\overline{M} = \arg\max_{M} P(A/M)\,P(M)\,.$$

Cette dernière équation est la clé de l'approche probabiliste de la reconnaissance vocale. En effet, le premier terme **P(A/M)** représente la probabilité d'observer le signal acoustique **A** si la séquence de mots **M** a été prononcée : c'est un problème purement acoustique ; le second terme **P(M)** représente la probabilité que c'est la suite de mots **M** qui a été effectivement énoncée : c'est un problème de nature linguistique. L'équation ci-dessus nous enseigne donc que l'on peut scinder le problème de reconnaissance vocale en deux parties indépendantes : on modélisera séparément les aspects acoustiques et les problèmes linguistiques. Dans la littérature, on parle généralement d'orthogonalité entre les modèles acoustiques et les modèles linguistiques.

4.1. Le modèle acoustique

Le modèle acoustique est basé sur la notion de phonèmes. Les phonèmes peuvent être considérés comme les unités sonores de base dans le langage verbal. Le premier stade de la reconnaissance vocale est de reconnaître un ensemble de phonèmes dans un flux de paroles.

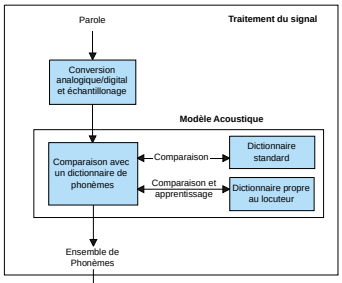


Le signal de la parole (capté à l'aide d'un micro) est d'abord digitalisé : il est échantillonné par une transformation de Fourier qui calcule les niveaux d'énergie du signal par bandes de 25 millisecondes[5], bandes qui se recouvrent entre elles de 10 millisecondes (valeurs

[5] La fréquence d'échantillonnage doit être au moins égale au double de la fréquence maximale du signal à numériser ; la voix couvre environ la bande de 60 Hz à 10 kHz.

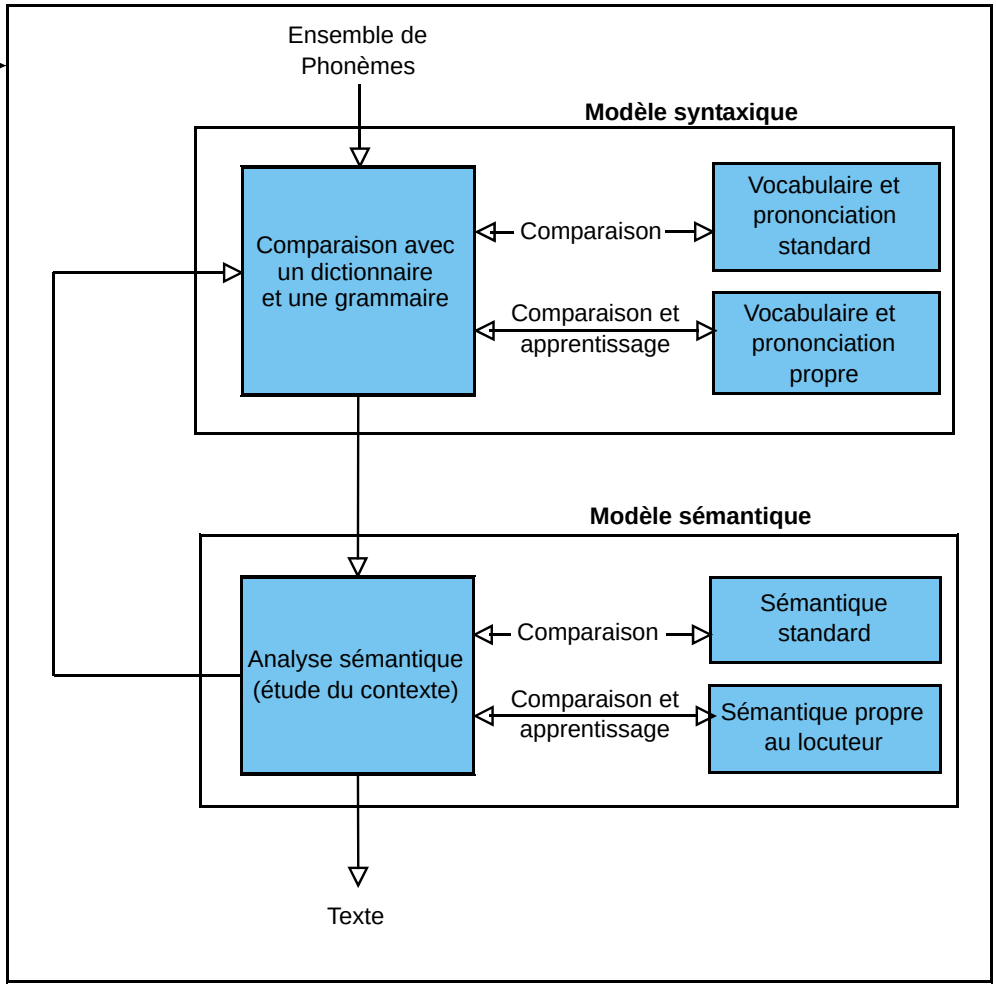
typiques). Le résultat[6] est comparé avec des prototypes stockés en mémoire de l'ordinateur à la fois dans un dictionnaire standard et dans un dictionnaire propre au locuteur. Ce dernier dictionnaire est construit au départ par des séances de dictée de textes standards que le locuteur doit effectuer avant d'utiliser efficacement le logiciel. Ce dictionnaire propre est régulièrement enrichi par auto-apprentissage lors des utilisations du logiciel. Il est intéressant de noter que l'empreinte vocale ainsi constituée est relativement stable pour un locuteur donné et peu influencée par certains facteurs extérieurs comme le stress, le rhume…

4.2. Le modèle linguistique



On scinde généralement en deux la partie liée au langage : une partie syntaxique et une partie sémantique.

Lorsque le modèle acoustique a identifié au mieux les phonèmes "entendus", il faut encore rechercher le message $\overline{M}$ le plus probable qui y correspond, c'est-à-dire la probabilité P(**M**) définie plus haut. C'est le rôle des modèles syntaxique et sémantique.



[6] Dans la pratique, pour identifier un phonème, la machine procède par analyse statistique. Les modèles utilisés pour cette identification sont les "modèles de Markov cachés".

A partir de l'ensemble des phonèmes issus du modèle acoustique, le modèle syntaxique va assembler les phonèmes en mots. Ce travail se fait également sur base d'un dictionnaire et d'une grammaire standards ainsi que d'un dictionnaire et d'une grammaire propres au locuteur ; ces derniers tiennent compte des "habitudes" du locuteur et s'enrichissent continûment.

Ensuite, le modèle sémantique cherche à optimiser l'identification du message par une analyse du contexte des mots et en se basant à la fois sur une sémantique courante propre à la langue et sur une sémantique (un style) propre au locuteur. Cette sémantique propre s'enrichira au fur et à mesure de l'utilisation du logiciel ; la plupart des logiciels permettent également de l'enrichir par l'analyse de textes qui reflètent les habitudes stylistiques du locuteur.

Ces deux modules travaillent de pair et on conçoit facilement qu'il existe un feedback entre eux. Initialement, les dictionnaires liés à ces deux modules étaient basés sur des *modèles de langage à syntaxe fixe*, c'est-à-dire calqués sur une grammaire définie par un ensemble rigide de règles.

Ensuite, les logiciels de reconnaissance vocale ont évolué vers l'utilisation de *modèles probabilistes locaux* : la reconnaissance ne s'effectue plus au niveau d'un mot mais au niveau d'une suite de mots, appelée n-gram où n représente la longueur en mots d'une séquence[7]. Les statistiques de ces modèles sont obtenues à partir de textes types et peuvent s'enrichir progressivement. Ici aussi ce sont les modèles de Markov cachés qui sont utilisés pour décrire les aspects probabilistes.

Actuellement, les logiciels les plus évolués tendent à combiner les avantages des modèles statistiques et des modèles à syntaxe fixe dans ce que l'on nomme les "*grammaires probabilistes*", l'idée étant de dériver à partir de grammaires fixes des probabilités pouvant être combinées avec celles d'un modèle probabiliste[8]. Dans ces dernières approches, il devient difficile de distinguer le modèle syntaxique du modèle sémantique et l'on parle plutôt d'un seul modèle linguistique.

4.3. L'état actuel et l'avenir de la technologie

Les systèmes de reconnaissance vocale ont fait d'importants progrès ces dernières années. Ces progrès ont pu être réalisés grâce aux recherches sur les modèles acoustiques et linguistiques mais aussi grâce à la croissance continue de la puissance des processeurs. Cet accroissement de puissance a non seulement permis de tirer le maximum de profit des algorithmes utilisés mais également de pouvoir exécuter les logiciels de reconnaissance vocale sur les configurations actuelles des PC, ce qui a ouvert un très large marché et, par conséquent, incité au développement de la recherche.

Aujourd'hui, les systèmes fonctionnels sont basés sur une approche statistique. Les leaders du marché proposent généralement des logiciels de reconnaissance du langage continu avec des possibilités de commandes pour les fonctions courantes des applications de traitement de texte. Ils annoncent des tailles de vocabulaire allant de 30 000 à 60 000 mots par langue, en permettant des dictées à la vitesse de 120 à 160 mots par minute (ce qui correspond à un débit de paroles relativement élevé) et avec un succès de reconnaissance supérieur à 95%. Nous verrons plus loin, lors des tests, ce qu'il en est dans la réalité.

---

[7] Si l'on considère l'exemple simple du mot "passé", on lui attribuera deux étiquettes grammaticales possibles auxquelles on attachera une probabilité lexicale qui rend compte de leur usage en français : (NOM, prob.= 0.7) et (PARTICIPE PASSÉ, prob.=0.3). Si l'on analyse ce mot non pas seul mais dans son contexte (par exemple dans un di-gram), le fait qu'il soit précédé d'une forme fléchie de l'auxiliaire être renforcera la probabilité de l'étiquette PARTICIPE PASSÉ.

[8] Par exemple, on attri buera un poids plus important à la séquence VERBE + ARTICLE + NOM qu'à la séquence NOM + ARTICLE + NOM toutes deux valables dans une grammaire fixe.

Parallèlement au vocabulaire courant, certains fournisseurs proposent des dictionnaires propres à certaines catégories professionnelles : médecins, juristes, financiers.

La plupart des logiciels actuels sont dépendants du locuteur, ce qui nécessite un apprentissage initial qui peut varier sensiblement d'un constructeur à l'autre. Pratiquement tous les bons logiciels permettent l'amélioration de cet apprentissage au fur et à mesure de leur utilisation.

La recherche dans le domaine de la reconnaissance vocale reste très active et il y a encore beaucoup de défis à relever afin d'améliorer les performances futures :

- Améliorer les modèles acoustiques qui aujourd'hui sont encore fort influencés par les conditions externes telles la position du micro, le fond sonore ambiant…
- Améliorer les modèles linguistiques : nous avons déjà parlé plus haut des recherches en matière de "grammaire probabiliste" ; des recherches sont également entreprises pour utiliser simultanément différentes approches comme les techniques statistiques et les réseaux neuronaux.
- Rendre les modèles indépendants du locuteur : des recherches dans ce sens sont effectuées afin de construire des modèles valables pour une large classe de population (par exemple : toutes les personnes dont l'anglais est la langue maternelle) ; cela permettrait de supprimer les phases d'apprentissage. En pratique, on tend vers un compromis qui permettrait au locuteur de se passer de l'apprentissage initial tout en améliorant la reconnaissance au cours du temps.

5. Tests et recommandations

En 1999, nous avions réalisé une étude[9] du marché d'où il ressortait que le logiciel de la société Dragon Systems se détachait nettement de ses concurrents de l'époque pour des dictées en français et nous l'avions recommandé. Dragon Systems ne proposait pas de reconnaissance vocale en néerlandais et c'est le logiciel FreeSpeech de Philips qui avait donné les meilleurs résultats. Toutefois, au vu des résultats et surtout du manque de convivialité, nous n'avions pas recommandé ce logiciel pour des dictées en néerlandais.

Depuis, la société Dragon a d'abord été absorbée par Lernout & Hauspie puis sauvée de la faillite de ce dernier en étant rachetée en 2001 par Scansoft Inc., leader actuel du marché de la reconnaissance vocale. C'est la raison pour laquelle nous avons choisi de reprendre les mêmes tests qu'en 1999 avec la version 7 de "Dragon Naturally Speaking Preferred" que Scansoft propose aujourd'hui[10] tant en français qu'en néerlandais.

5.1. Caractéristiques techniques de "Dragon Naturally Speaking Preferred v7"

Les logiciels de reconnaissance vocale sont des logiciels gourmands en ressources techniques. Les exigences techniques minimales de "Dragon Naturally Speaking Preferred" sont les suivantes :

- Microsoft Windows XP, Me, 2000, NT 4.0 SP6 avec Internet explorer 5 ou plus ;
- Intel Pentium III 500 Mhz (impératif !) ;

[9] H. Haut, Reconnaissance vocale – les dictées continues, rapport d'étude, février 1999.

[10] Au moment d'écrire cette note, Scansoft a mis sur le marché la version 8 de son logiciel. On peut raisonnablement penser que les résultats ne peuvent que s'améliorer avec cette nouvelle version.

- 128 Mb RAM ;
- de préférence 1 Gb de libre sur le disque dur ;
- carte son "Creative Labs Sound Blaster 16" ou carte son équivalente prenant en charge l'enregistrement 16 bits ;
- un casque microphone à réduction de bruits (fourni avec le logiciel).

Les performances mises en avant par le constructeur sont :

- un premier apprentissage en quelques minutes ;
- une grande convivialité de l'interface ;
- une vitesse de lecture pouvant aller jusqu'à 160 mots par minute ;
- une précision proche de 99% (nos tests donnent plutôt 95%) ;
- la possibilité de dictée dans MS Word et dans pratiquement toutes les applications Windows avec prise en charge d'un nombre important de commandes de formatages ;
- un vocabulaire extensible avec une adaptation au style du locuteur ;
- des possibilités étendues de contrôle vocal du poste de travail ;
- l'utilisation possible d'un enregistreur portable (dictaphone, Pocket PC ou Palm) avec transfert de la dictée sur le PC.

5.2. Tests en langue française

5.2.1. Scénario de test

Dans le but d'évaluer l'évolution de "Dragon Naturally Speaking Preferred FR v7" depuis nos tests de 1999, nous avons repris le même schéma de test.

Nous avons rigoureusement suivi le scénario suivant :

1. Installation du logiciel et du microphone conformément aux options par défaut préconisées par le constructeur.
2. Création d'un nouvel utilisateur et adaptation afin de permettre à "Dragon Naturally Speaking" de reconnaître la voix de cet utilisateur. Ce premier entraînement requiert environ 10 minutes de dictée.
3. Familiarisation avec le logiciel pendant 15 minutes.
4. Dictée dans MS Word d'un texte d'une page : nous avons choisi le résumé (464 mots, 2721 caractères) d'un rapport annuel de l'ONSS. Après cette dictée, le texte sera corrigé (sans apprentissage) et l'on notera :
   - le temps total mis pour dicter le texte ;
   - le nombre total de fautes et le temps total mis pour corriger le texte.

   On en déduira :
   - une vitesse de lecture réelle correspondant à un texte sans erreur : (nombre total de mots) / (temps de lecture + temps de correction) ;
   - une vitesse de frappe équivalente : (nombre total de caractères) / (temps de lecture + temps de correction) ;
   - la précision : (nombre de mots - nombre de fautes) / (nombre de mots).
5. Entraînement du système pendant une heure (dictée et corrections comprises). Nous avons choisi comme texte la brochure "Technologies Florissantes" publiée par SmalS-MvM. La correction sera faite avec apprentissage afin d'améliorer la reconnaissance du logiciel ; une optimisation acoustique sera effectuée après cette dictée.

6. Dictée dans MS Word d'une autre page (394 mots, 2297 caractères) de la même brochure. Pour cette dictée, on calculera les mêmes paramètres que ci-dessus après une correction sans apprentissage.

7. Dans ces logiciels, on peut raffiner le modèle linguistique en lui faisant analyser les textes que l'on a l'habitude de produire de façon à ce que le modèle s'adapte non seulement à notre vocabulaire mais également à notre style d'écriture. Pour simuler cet apprentissage, nous avons optimisé le vocabulaire et le modèle linguistique du logiciel avec le texte intégral du roman "Germinal" d'Emile Zola. Nous avons ensuite dicté un passage de ce roman (713 mots, 3683 caractères) et noté les mêmes caractéristiques que ci-dessus. La correction sera faite sans apprentissage.

5.2.2. Résultats

Le tableau ci-dessous reprend l'ensemble des résultats obtenus.

	Etude 2005	Etude 1999
Dictée du rapport ONSS		
temps de lecture (min.)	5.3	4.9
nombre de fautes	46	40
Précision	90%	91%
temps de correction (min.)	7.3	22
temps total (min.)	12.6	26.9
vitesse de lecture réelle (mots/min.)	**37**	**17**
vitesse de frappe équivalente (car./min.)	**216**	**102**
Dictée d'une page de la brochure "nouvelles technologies"		
temps de lecture (min.)	4.6	4.2
nombre de fautes	27	46
Précision	93%	88%
temps de correction (min.)	6.2	9
temps total (min.)	10.8	13.2
vitesse de lecture réelle (mots/min.)	**37**	**30**
vitesse de frappe équivalente (car./min.)	**214**	**174**
Dictée dans Word d'une page de "Germinal"		
temps de lecture (min.)	8.2	7.3
nombre de fautes	27	49
Précision	96%	93%
temps de correction (min.)	9.7	14.6
temps total (min.)	17.9	21.9
vitesse de lecture réelle (mots/min.)	**40**	**33**
vitesse de frappe équivalente (car./min.)	**206**	**170**

Le nombre d'erreurs n'est pas un critère très significatif, car certaines erreurs peuvent être corrigées beaucoup plus rapidement que d'autres. Ce que nous avons défini comme étant la "vitesse de frappe équivalente" en nombre de caractères par minute nous paraît être une meilleure mesure en ce sens qu'elle tient compte à la fois du temps de lecture et du temps de correction.

Par rapport à notre étude de 1999, on constate une amélioration significative (de l'ordre de 20%) de cette vitesse de frappe équivalente. On constate aussi que cette vitesse est acquise dès le premier test alors que l'entraînement initial lors de la création du locuteur n'a nécessité qu'une dizaine de minutes.

La dictée du texte de Zola donne un résultat équivalent aux autres tests : cela démontre un bon apprentissage, ce texte comprenant des mots difficiles et parfois surannés avec un style élaboré.

5.3. Tests en langue néerlandaise

5.3.1. Scénario de test

L'évaluation[11] de "Dragon Naturally Speaking Preferred NL v7" s'est faite suivant le même schéma de test que pour le français en prenant la version néerlandaise des textes utilisés et en remplaçant "Germinal" d'Emile Zola par "De Kerels van Vlaanderen" de Hendrik Conscience.

5.3.2. Résultats

Le tableau ci-dessous reprend l'ensemble des résultats obtenus.

	Etude 2005	Etude 1999
Dictée du rapport ONSS		
temps de lecture (min.)	6.7	7.0
nombre de fautes	32	80
Précision	93%	83%
temps de correction (min.)	8.6	25.2
temps total (min.)	15.3	32.2
vitesse de lecture réelle (mots/min.)	**31**	**15**
vitesse de frappe équivalente (car./min.)	**187**	**88**
Dictée d'une page de la brochure "nouvelles technologies"		
temps de lecture (min.)	6.2	4.5
nombre de fautes	37	55
précision	89%	84%
temps de correction (min.)	12.3	14.7
temps total (min.)	18.5	19.2
vitesse de lecture réelle (mots/min.)	**18**	**18**
vitesse de frappe équivalente (car./min.)	**116**	**112**
Dictée dans Word d'une page de "De Kerels van Vlaanderen"		
temps de lecture (min.)	10.3	9.1
nombre de fautes	32	72
précision	95%	90%
temps de correction (min.)	13.5	26.2
temps total (min.)	23.8	35.3
vitesse de lecture réelle (mots/min.)	**29**	**21**
vitesse de frappe équivalente (car./min.)	**155**	**121**

[11] Je tiens à remercier Joëlle Heris qui a gentiment accepté, comme en 1999, de participer à ces tests.

Les résultats sont moins bons qu'en français. Néanmoins on peut constater une nette amélioration par rapport aux résultats de 1999 hormis pour la dictée extraite de la brochure "nouvelles technologies" ; peut-être est-ce dû au vocabulaire plus technique de cette dernière et qu'il faudrait d'abord "enseigner" ce vocabulaire au logiciel.

Il est intéressant de constater que le meilleur résultat est obtenu d'emblée (après une dizaine de minutes d'entraînement) avec le texte dont le vocabulaire est le plus courant.

Par ailleurs, le résultat de la dictée "De Kerels van Vlaanderen" confirme la bonne capacité d'apprentissage du logiciel étant donné que le vocabulaire et le style de ce texte sont, comme chez Zola, plutôt difficiles et démodés.

Nous pensons qu'en continuant à entraîner "Dragon Naturally Speaking Preferred NL v7", les résultats s'amélioreront et se rapprocheront de ceux obtenus en français.

5.3.3. Remarques générales

Dans l'ensemble, le logiciel offre une bonne convivialité et est facile à installer et à utiliser. Même si, au premier abord, la façon de corriger les textes en mode apprentissage peut paraître lourde, cette correction se fait assez rapidement lorsque l'on a acquis une certaine habitude et réglé quelques paramètres techniques.

En 1999, nous avions trouvé que le modèle grammatical présentait des lacunes assez importantes avec pour effet un grand nombre d'erreurs de type accord de genre et pluriels tant pour les adjectifs que pour les verbes. On peut constater aujourd'hui une nette amélioration de ce modèle grammatical.

Les résultats obtenus sont appréciables : une vitesse de frappe de 200 caractères à la minute est considérée comme performante pour un(e) secrétaire et est probablement largement supérieure à ce que la plupart d'entre nous peut atteindre.

A côté de la reconnaissance vocale en mode dictée, le logiciel "Dragon Naturally Speaking Preferred" permet de pratiquement tout contrôler au niveau du poste de travail : démarrer un programme, ouvrir et fermer des menus, sélectionner des icônes ou des options, se déplacer dans les fenêtres, contrôler la souris…. Nous n'avons pas testé de manière intensive ces fonctionnalités que nous jugeons secondaires par rapport à notre sujet. Cependant, les quelques tests que nous avons faits se sont révélés conformes aux dires du constructeur.

Enfin, une nouvelle version 8 est actuellement disponible et ne devrait qu'améliorer encore les résultats que nous avons obtenus.

5.4. Recommandations

Sur base des tests réalisés, nous pouvons conseiller le logiciel "Dragon Naturally Speaking Preferred" de Scansoft, Inc.

Malgré son prix relativement élevé (environ 200$) et ses exigences en ressources techniques, ce logiciel atteint ses objectifs en permettant à la majorité d'entre nous de dicter un texte à une vitesse largement supérieure à celle que nous pouvons atteindre en le dactylographiant.

"Dragon Naturally Speaking Preferred" permet de dicter dans un enregistreur numérique portable (un PC de poche, un Palm Tungsten ou un dictaphone à choisir dans une liste de modèles agréés par Scansoft) ; de retour au bureau, le logiciel permet de transcrire

automatiquement les dictées après synchronisation avec le PC et de les corriger ou de les faire corriger par une tierce personne. Nous pensons qu'il serait très utile de tester cette fonctionnalité qui pourrait intéresser certains de nos managers.

Nous voulons cependant attirer l'attention du lecteur sur un certain nombre de points qui nous paraissent importants.

L'utilisation d'un logiciel de reconnaissance vocale requiert un investissement personnel lors des premières utilisations. C'est à ce prix que des résultats satisfaisants peuvent être atteints.

Cela demande également une discipline personnelle :

- Prendre la peine de bien positionner et de tester le micro avant chaque séance de dictée.
- S'efforcer (surtout dans les premières dictées) de corriger ses textes avec la fonction d'apprentissage plutôt que de les corriger simplement dans un traitement de texte (ce qui est plus rapide), afin d'améliorer les modèles acoustiques et linguistiques.
- Adopter une dictée continue, régulière dans le ton et avec une prononciation correcte, c'est-à-dire dicter plutôt que parler spontanément : ne pas manger la fin des mots, éviter les hésitations comme "euh", "mmm"..., essayer d'avoir un flux de paroles constant et ne pas revenir sur ce que l'on a dit. Par contre, il est conseillé de parler avec un débit normal plutôt que de ralentir l'élocution.
- Apprendre un minimum de vocabulaire de commandes et le respecter : "à la ligne", "nouveau paragraphe", "en maj", "toutes maj", "ceci en maj", "corrigez ça",...

  Par contre, nous conseillons de ne pas abuser de commandes sophistiquées de mise en page (comme "mettre en gras l'avant-dernier mot"…): il est souvent plus rapide de faire manuellement cette mise en page lorsque la dictée est terminée.
- Dicter dans un environnement calme, les logiciels étant assez sensibles au bruit ambiant.
- Faire régulièrement une copie de sauvegarde de ses données personnelles (acoustiques et linguistiques).

6. Conclusion

Après un bref historique de l'évolution des recherches en reconnaissance vocale, nous avons décrit les différents domaines où ce type d'application présente un réel intérêt.

Nous avons vu que, depuis le début de ces recherches, la "Quête du Graal" des chercheurs a été le développement de logiciels capables de reconnaître un texte dicté de façon continue, avec un vocabulaire suffisamment riche et une vitesse de locution normale.

Parmi les diverses pistes de recherche explorées, c'est l'approche statistique basée sur les chaînes de Markov cachées qui s'est avérée la plus prometteuse.

Pendant longtemps cependant, la puissance de calcul nécessaire a été un obstacle majeur à la production de logiciels suffisamment performants à un coût abordable. L'évolution permanente des processeurs pendant ces dix dernières années a permis de surmonter ce

problème et nous trouvons, aujourd'hui, sur le marché, des logiciels de reconnaissance vocale ayant des performances tout à fait acceptables.

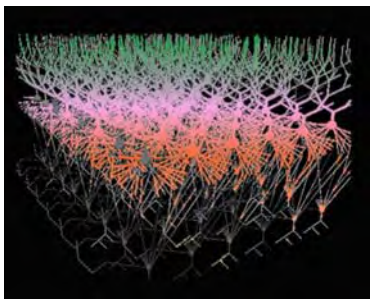
Le logiciel "Dragon Naturally Speaking Preferred v7" que nous avons testé s'avère suffisa<mment performant pour le recommander tant en français qu'en néerlandais : les résultats obtenus dépassent les capacités dactylographiques de la plupart d'entre nous !

Cependant nous avons vu que l'utilisation d'un tel logiciel requiert encore, de la part du locuteur, un ensemble de contraintes et une discipline assez stricte. Nous en sommes encore au stade de la parole dictée (ou lue) qui est malgré tout plus proche de la langue écrite que ne l'est la parole spontanée.

Cette parole spontanée est réellement celle utilisée par l'être humain pour communiquer. Elle se caractérise par ses hésitations, ses reprises et un taux de constructions asyntaxiques assez élevé.

La reconnaissance de cette parole spontanée est peut-être la nouvelle Quête des chercheurs. Mais ce n'est plus aujourd'hui la puissance de calcul qui est un frein : ce n'est pas en raffinant ou en augmentant la complexité des modèles statistiques que des progrès majeurs seront accomplis. La plupart des chercheurs s'accordent à dire qu'il nous faut d'abord mieux comprendre comment fonctionne le cerveau humain, trouver ce qu'il lui donne cette extraordinaire faculté de s'adapter immédiatement à la parole d'autrui en en saisissant immédiatement toutes les nuances.

Il est rare qu'en sciences une seule voie de recherche conduise à une découverte majeure ; le dualisme semble être une constante de notre univers. Heisenberg constatait déjà : "The most fruitful developments have always emerged where two different kinds of thinking met". En reconnaissance vocale, l'approche neuronale fournira peut-être cette voie complémentaire, mais ceci est une autre histoire !



Colonnes de neurones dans le cortex visuel

LES PUBLICATIONS TECHNIQUES DE LA SMALS-MVM

La recherche publie régulièrement des Technos consacrés à des sujets d'actualité. La liste de ces technos est reprise ci-dessous.

28 Isabelle Boydens, *La préservation à long terme de l'information numérique*, Septembre 2004.
27 Bob Lannoy, *Open Source Software - un nouveau modèle de logiciel*, Mai 2004.
26 Isabelle Boydens, *Un retour riche d'expériences*, Novembre 2003.
25 Nick Marly, *L'enjeu des derniers kilomètres*, Juillet 2003.
24 Hervé Haut, *Le e-learning : Un nouvel espace d'apprentissage*, Avril 2003.
23 Michel Laloy, *Le marché des Télécommunications après 4 ans de libéralisation*, Novembre 2002.
22 Marco Saerens, *L'intelligence artificielle : Quelques éléments de base - Deuxième partie*, Mai 2002.
21 Isabelle Boydens, *La recherche d'information sur Internet*, Décembre 2001.
20 Marco Saerens, *L'intelligence artificielle : Quelques éléments de base - Première partie*, Juin 2001.
19 Alex De Koning, *Les Agents Intelligents*, Février 2001.
18 Françoise Vanden Bossche, *Les logiciels de recherche documentaire*, Octobre 2000.
17 Marc De Decker, *Method Engineering*, Juin 2000.
16 Michel Laloy, *Introduction à la normalisation*, Mars 2000.
15 Hervé Haut, *Gestion électronique des documents*, Novembre 1999.
14 Isabelle Boydens, *La gestion des flux d'information administratifs*, Septembre 1999.
13 Guy Geerts, *Vidéoconférence*, Avril 1999.
12 Alex De Koning, *Datamining*, Décembre 1998.
11 Denis Francotte, *Le concept IP*, Septembre 1998.
10 Olivier Tribel, *JAVA*, Juin 1998.
9 Philippe Vanderheyden, *Edi/Edifact*, Mai 1998.
8 Gilles Kempgens, *L'Audit Informatique*, Mars 1998.
7 Isabelle Boydens, *Evaluer et améliorer la qualité des bases de données*, Janvier 1998.
6 Marc De Decker, *Software Process Improvement*, Novembre 1997.
5 Dominique Thomas, *ISDN/RNIS*, Septembre 1997.
4 Leo V. Broekhoven, *Sosenet*, Juillet 1997.
3 Luisa Anzalone, *Technologie Client/Serveur*, Juin 1997.
2 Alex De Koning, *Datawarehousing*, Mai 1997.
1 Isabelle Boydens, *Les systèmes de méta-information*, Avril 1997.

Vous pouvez commander ces technos à l'adresse suivante :
SmalS-MvM
Secrétariat "Clients et Services"
Mme Joëlle Ankaer
Joelle.ankaer@smals-mvm.be
02/509.58.62
Rue du Prince Royal 102 à 1050 BRUXELLES

ou les consulter sur les sites http://documentation.smals-mvm.be et http://www.smals-mvm.be