

# MODELISATION DU LANGAGE POUR LES SYSTEMES DE RECONNAISSANCE DE LA PAROLE : APPLICATION A MAUD

**Imed ZITOUNI**

LORIA/INRIA-Lorraine  
B.P.239 54506 Nancy, France  
Mél : zitouni@loria.fr

Le traitement automatique de la parole suscite actuellement un grand intérêt ; il est considéré comme une branche importante de l'interaction homme-machine. Nous éprouvons le besoin de communiquer avec nos ordinateurs, de la façon la plus naturelle et la plus directe qui soit : le langage parlé ; l'interaction et l'échange d'informations s'en trouvent grandement facilités. Nous cherchons à rendre ces machines accessibles par la voix au téléphone ou au microphone, pour délivrer et recueillir de l'information sans clavier ni écran de visualisation. Ces techniques d'entrées/sorties vocales sont également d'un intérêt évident pour les applications d'aide aux handicapés, la reconnaissance de cris de détresse, la prise de commandes avec mains libres dans la conduite de l'automobile, etc.

Comprendre un énoncé, c'est d'abord l'entendre, puis le reconnaître et enfin l'interpréter. L'étape de reconnaissance est la phase que les systèmes de Reconnaissance Automatique de la Parole (RAP) cherchent à maîtriser. Celle-ci vise à transformer le signal acoustique en une séquence d'unités linguistiques. Pour modéliser l'acoustique, la plupart des systèmes de RAP actuels utilisent les modèles de Markov cachés (HMM). Nous notons également que d'autres techniques commencent à être de plus en plus utilisées et combinées avec les HMM : les réseaux de neurones et les réseaux bayesiens. Concernant la modélisation du langage, c'est surtout les modèles n-grammes qui sont très employés de nos jours. Des progrès significatifs ont été réalisés ces vingt dernières années dans le domaine de la reconnaissance automatique de la parole. Le marché des logiciels offre aujourd'hui des produits qui prétendent effectuer une reconnaissance de la parole continue pour un vocabulaire important. En réalité, les performances de ces systèmes sont encore largement inférieures à celles de l'être humain, particulièrement au niveau de la modélisation du langage.

Le travail que nous présentons s'inscrit dans le cadre de la modélisation du langage pour les systèmes de reconnaissance de la parole continue destinés aux grands vocabulaires. Nous proposons de nouveaux modèles statistiques : le modèle multiclasse, le modèle hiérarchique et un modèle fondé sur des séquences de mots de longueur variable. Ces séquences représentent des structures langagières qui s'apparentent à des syntagmes linguistiques. Elles sont détectées automatiquement, à partir d'importants corpus de textes, en utilisant des mesures issues de la théorie de l'information. Nous proposons également une approche hybride combinant les modèles de langage probabilistes, utilisés dans la plupart des systèmes de reconnaissance actuels, avec des connaissances linguistiques explicites supplémentaires. L'évaluation de l'ensemble de ces modèles est effectuée en terme de perplexité et en terme de prédiction à l'aide du jeu de Shannon. Pour tester leurs performances au niveau de la reconnaissance, nous avons développé un système de reconnaissance vocale nommé MAUD : Machine AUTomatique à Dicter ; il se fonde sur les modèles de Markov cachés de second ordre et utilise un vocabulaire de 20000 mots. Le corpus utilisé pour l'estimation des modèles statistiques regroupe 50 millions de mots extraits du journal « Le Monde ».

Le modèle multiclasse, inspiré de l'approche multigramme, se fonde sur des classes syntaxiques. Il suppose qu'une source, gouvernant l'activité d'une langue, émet des séquences de mots dont les classes syntaxiques correspondantes ne sont pas indépendantes et que leurs dépendances sont de longueur variable. Chaque séquence de classes, nommée multiclasse, est supposée correspondre à une entité sous-jacente issue d'un ensemble fini  $M_C^* = \{m_\alpha^*\}$ . Ces entités sont des séquences de classes syntaxiques de longueur variable, indépendantes entre elles. Dans le cas idéal,  $M_C^*$  sera l'ensemble des syntagmes représentant la structure du langage naturel. Le modèle extrait donc à partir d'un corpus d'apprentissage l'ensemble fini  $M_C^*$  des entités et attribue ensuite une valeur de vraisemblance pour chacune d'entre elles. Ainsi, pour estimer la vraisemblance d'une suite de mots, il suffit au modèle multiclasse de l'étiqueter avec des classes syntaxiques et de calculer le produit des probabilités des entités qui la composent. Un intérêt de ce modèle réside dans le fait qu'il permet une structuration non supervisée de successions de mots (appartenant à un vocabulaire donné  $V$ ) en unités plus longues et de longueur variable, à partir de simples considérations statistiques. Ce modèle est capable d'utiliser de grands vocabulaires sans avoir besoin de machines puissantes ni d'énormes corpus.

En utilisant le modèle ci-dessus, les multiclassés d'une phrase sont supposés indépendants, ce qui est en contradiction avec la structure du langage naturel. Dans la langue, une phrase est caractérisée par la cohésion interne de ses mots. Les mots se regroupent et constituent des sous-ensembles (syntagmes), qui eux-mêmes s'assemblent pour former d'autres sous-ensembles, et ainsi de suite jusqu'à la construction de la structure de la phrase. Pour remédier au problème d'indépendance entre les séquences (multiclasse) et pour se rapprocher le plus possible de la structure du langage naturel, nous proposons un nouveau modèle de langage : le modèle hiérarchique. Le modèle hiérarchique suppose que les classes syntaxiques des mots dans une phrase soient dépendantes ; le regroupement des classes et leurs rattachements à des entités forment le premier niveau de la hiérarchie. A leur tour les entités de ce premier niveau vont être rattachées à d'autres et construisent ainsi un deuxième niveau, ainsi de suite jusqu'à un niveau donné. Chaque entité, d'un niveau de hiérarchie  $j$  donné, est supposée appartenir à un ensemble fini  $M_{C_j}^*$ . Les résultats d'évaluation en terme de perplexité confirment l'intérêt de prendre en compte la dépendance entre les multiclassés d'une phrase. En effet, le modèle hiérarchique a augmenté d'environ 21% les performances du modèle multiclasse ; ce dernier qui pourtant dépasse de 43% les performances d'un modèle biclasse conventionnel, est loin derrière le modèle triclasse de

6%. En revanche, par rapport à ce modèle triclasse, le modèle hiérarchique est meilleur de 15%. L'utilisation des modèles multiclasse et hiérarchique, lors des premières étapes de reconnaissance, n'est pas évidente. Ces modèles ont besoin de la totalité de la phrase (ou presque) pour agir. Ainsi, nous ne les avons utilisés qu'au niveau du filtrage des hypothèses. L'exploitation du modèle hiérarchique dans MAUD a permis d'améliorer les performances de 4%. En revanche, les performances apportées par le modèle multiclasse ne sont pas significatives.

L'amélioration des performances d'un système de RAP reste limitée si nous n'agissons qu'au niveau du filtrage des hypothèses. Nous proposons ainsi un nouveau modèle de langage pouvant intervenir dès les premières étapes de la prédiction et de la sélection. Ce modèle se fonde sur des séquences de mots qui sont naturellement de longueur variable et qui permettent d'avoir une élocution naturelle. Ces séquences sont extraites automatiquement à partir d'un corpus d'apprentissage, en utilisant des mesures issues de la théorie de l'information. Tout d'abord, nous avons considéré l'ensemble de ces séquences comme des unités de la langue. Ensuite, nous les avons ajoutées au vocabulaire de base, construisant ainsi un nouveau vocabulaire qui sert à l'apprentissage des modèles de langage. Par conséquent, lors de la prédiction et même de la sélection, les modèles de langage utilisant ce vocabulaire se fondent sur un historique d'unités où chacune d'entre elles peut être, soit un mot, soit une séquence. Ceci donne aux modèles la possibilité d'utiliser un historique plus important et de mieux prendre en compte le rôle prédictif de ces séquences. Les modèles de langage correspondants sont également capables de prédire la totalité d'une séquence, et de ne plus se limiter à la prédiction d'un seul mot. Nous avons adapté les modèles de langage les plus utilisés (n-grammes, n-classes, cache et triggers) pour qu'ils prennent en compte ces séquences lors de la prédiction et la sélection dans un système de dictée vocale. L'utilisation de ces séquences a amélioré la perplexité d'environ 23% et le taux de reconnaissance de MAUD d'environ 14%.

L'approche hybride que nous proposons, combine les modèles de langage probabilistes à des connaissances langagières explicites. Ces connaissances sont modélisées par une grammaire d'unification, appartenant à la famille des grammaires à traits, construisant ainsi le modèle formel. Le formalisme des réseaux de transition augmentés correspond à la notion d'automate à pile ; il est, de ce fait, assez simple à mettre en œuvre sur le plan informatique. Ces réseaux ont été choisis pour implanter notre grammaire. Préalablement, nous exposons la classification et la méthode d'étiquetage pour lesquelles nous avons opté. En effet, les connaissances linguistiques utilisées reposent sur des classes syntaxiques de la langue française. L'utilisation de cette approche pour le filtrage des N meilleures hypothèses produites par notre système de reconnaissance vocale MAUD a permis d'augmenter les performance d'environ 3%.

La version de base de MAUD se fonde sur les modèles de Markov cachés non contextuels de second ordre et sur un modèle de langage n-grammes. Cette version, mais avec un modèle diphone, a été classée seconde dans la première campagne d'évaluation de l'Aupelf-UREF. Dans la nouvelle version de MAUD, un modèle se servant des séquences, est utilisé lors des deux premières étapes pour fournir les N meilleures hypothèses, correspondant à la phrase prononcée. Ces hypothèses sont ensuite filtrées par l'approche hybride, pour être enfin réévaluées avec le modèle hiérarchique. Le résultat du système est la meilleure hypothèse restante. Dans cette version de MAUD, les performances se sont améliorées considérablement par rapport à celles de la version de base : 20% environ.

## Référence bibliographique

[Zitouni, 2000] Zitouni I. (2000). Modélisation du langage pour les systèmes de reconnaissance de la parole destinés aux grands vocabulaires : application à MAUD. Thèse de l'université Henri Poincaré, Nancy.