UNIVERSITE PIERRE ET MARIE CURIE
Laboratoire d'Informatique de Paris 6

La reconnaissance vocale et son mentor : l'évaluation

Habilitation à diriger des recherches

par

Marie-José CARATY

Soutenue publiquement le 1er février 1999 devant le jury composé de :

Joseph	MARIANI	Président
Madeleine	BONNET	Rapporteur
Hervé	BOURLARD	Rapporteur
Renato	DE MORI	Rapporteur
Pierre	ESCUDIER	Examinateur
Jean-Yves	JAFFRAY	Examinateur
Xavier	RODET	Directeur

Sommaire

Présentation du mémoire

Les travaux présentés dans ce mémoire couvrent dix années de recherche post-doctorale au GR22, devenu le LAFORIA (Laboratoire Formes et Intelligence Artificielle), puis le Pôle IA (Intelligence Artificielle) du LIP6, le laboratoire d'Informatique de Paris 6.

Le titre du mémoire présente l'évaluation comme un guide sage et expérimenté pour la recherche en reconnaissance vocale. Sans avoir cherché à en faire un thème récurrent, l'évaluation revient très naturellement au cours des trois chapitres de ce mémoire.

Le premier chapitre est évidemment consacré à l'**évaluation**. Une synthèse des différents modes d'évaluation est illustrée dans des domaines d'application aussi variés que l'ingénierie de la langue, l'architecture parallèle et distribuée et la réalité virtuelle. L'émergence d'un **paradigme d'évaluation en reconnaissance vocale** est ensuite expliquée. Son efficacité dans le développement des systèmes de reconnaissance vocale se distingue par le partage de ressources et d'outils d'évaluation. Ses diverses composantes sont décrites : les spécifications de la tâche de reconnaissance, les bases de données acoustiques, les ressources linguistiques et l'outil d'évaluation de performance. Un bref historique des premiers plans d'évaluation en reconnaissance vocale illustre l'influence du paradigme d'évaluation sur le développement technologique. Les succès, mais également les limites et dangers de l'évaluation, sont discutés.

Le deuxième chapitre regroupe les travaux concernant la **reconnaissance markovienne**. Les systèmes fondés sur les modèles de Markov cachés se sont peu à peu imposés comme les systèmes de l'état de l'art en reconnaissance vocale grand vocabulaire, indépendamment du locuteur. Pour un tel problème, la maîtrise de la technologie markovienne n'est pas simple. C'est à travers le développement de notre propre système de dictée vocale, des options choisies, et de notre expérience d'une campagne d'évaluation de dictée vocale, que sont décrits les travaux. D'autres études sont présentées dans le cadre de systèmes hybrides fondés sur la modélisation markovienne. Le premier système se caractérise par son contrôle temporel dans les modèles de Markov cachés. Le second système utilise la règle de décision des K-plus proches voisins comme alternative à l'estimation gaussienne des probabilités d'émission.

C'est après ce que l'on peut considérer comme le « must » de la reconnaissance vocale qu'au troisième chapitre les travaux des plus anciens aux plus récents sont présentés ; ils concernent **l'espace de représentation et de décision** en reconnaissance de la parole. Les premiers travaux concernent la paramétrisation de la parole en formants et la conception d'une mesure de dissimilarité interspectrale à critères perceptifs, adaptée à cet espace de représentation, ainsi que les évaluations dont cet espace métrique a fait l'objet. Une seconde série de travaux, concernant le numérique-symbolique, présente l'intérêt d'avoir expérimenté l'apprentissage symbolique en reconnaissance. Nos travaux liés à une autre problématique de la reconnaissance, celle du locuteur, sont présentés. Enfin, les travaux les plus récents proposent une extension d'un espace de représentation quelconque par une multi-résolution temporelle qui, munie de la mesure d'inertie et du calcul de centroïde, permet d'estimer la stationnarité du signal vocal.

Introduction au chapitre 1

Si je savais l'évaluation très importante dans notre méthodologie de recherche en parole, c'est surtout après notre participation à la première campagne de dictée vocale francophone que le paradigme d'évaluation m'est apparu fondamental. A l'origine de cette campagne : l'agence francophone pour l'enseignement supérieur et la recherche[1]. En 1994, deux des actions de recherche lancées par l'AUPELF-UREF concernent la linguistique informatique écrite et orale. Les thèmes retenus pour ces actions montrent l'ampleur du projet. Pour l'écrit : -accès à l'information en langage naturel, -alignement de corpus multilingues, -construction automatique de terminologie et de relations sémantiques entre les termes, -compréhension de texte. Pour l'oral : -dictée vocale, -dialogue oral, -synthèse vocale. L'idée maîtresse est que pour maintenir l'importance de la langue française au plan international, il est indispensable d'arriver à une haute technologie en ingénierie de la langue. En conséquence, un objectif est d'accélérer la progression de la recherche dans le traitement automatique de la langue écrite et orale. Pour ce faire, les ressources essentielles au développement des systèmes de traitement du langage doivent être disponibles, et des systèmes de mesure doivent être élaborés pour comparer objectivement les différentes approches du domaine et assurer une synergie de la recherche. L'idée n'est pas nouvelle, et, dans son appel d'offres, l'AUPELF-UREF souligne l'avance considérable des pays anglophones qui ont su mettre en œuvre ce paradigme d'évaluation. En effet, depuis bien des années, aux Etats-Unis, des campagnes de tests comparatifs sont menées sur des bases de données communes avec des ressources partagées et les mêmes outils d'évaluation. Les organismes précurseurs sont, pour l'écrit, NIST[2], avec ses évaluations menées depuis 1978 sur la recherche textuelle (TREC, Text Retrieval Evaluation Conference), et, pour l'oral, ARPA[3], avec ses campagnes de tests menées depuis 1980 sur la reconnaissance vocale. Enfin, une référence : sponsorisé par la NSF[4], la Communauté Européenne[5] et OGI[6], « Survey of the State of the Art in Human Language Technology »[7], livre impressionnant par sa qualité et son volume, est le résultat d'une collaboration internationale sur l'état de l'art dans les technologies du langage. Le treizième et dernier chapitre de ce livre traite de l'évaluation « plurielle » et de sa mise en œuvre dans les différents champs du traitement du langage ; c'est l'une de nos principales références sur l'évaluation.

---

[1] AUPELF-UREF.
[2] National Institute for Science and Technology.
[3] Advance Research Project Agency.
[4] National Science Foundation.
[5] Direction Générale XIII-E de la Commission des Communautés Européennes.
[6] Oregon Graduate Institute - Center for Spoken Language Understanding.
[7] *Survey of the State of the Art in Human Language Technology*, Editeurs : R. Cole, J. Mariani, H. Uszkoreit, N. Varile, A. Zaenen, A. Zampolli, V. Zue, Publié par Giardini Ed. et Cambridge University Press, 1995.

Chapitre 1

Evaluation et paradigme d'évaluation

L'évaluation est au cœur des progrès dans tous les domaines de l'ingénierie. Bien que les méthodologies d'évaluation soient en constante évolution, leur succès est déjà tel que l'évaluation pourrait être un domaine de « recherche propre ». Les compétitions se généralisent, elles sont déjà ouvertes à de nombreux domaines technologiques. Par exemple, l'ATP[8] de NIST finance et organise des compétitions dans des technologies aussi diverses que l'agriculture, la biotechnologie, la chimie, les machines-outils, l'automobile, la micro-électronique, l'électronique, l'architecture parallèle, la réalité virtuelle, l'information et la communication, entre autres... Toutes ces compétitions ne peuvent être menées qu'avec des méthodologies d'évaluation adéquates.

1.1. Modes d'évaluation

Depuis la conception d'un système et jusqu'à son transfert technologique, l'évaluation est présente. Si la méthodologie d'évaluation varie, elle est un moteur indéniable dans les phases -de conception, -de développement, -de maintenance, et -d'utilisation finale. L'évaluation est « plurielle » ; en effet, différents modes d'évaluation sont identifiés relativement à leurs buts. Il s'agit principalement de l'adéquation, du diagnostic, du quantitatif et du qualitatif.

- **Evaluation d'adéquation**
  L'adéquation dont il s'agit est celle du système à la tâche assignée, avec la prise en compte de la qualité et du coût. Ce type d'évaluation, couplé avec l'identification des besoins potentiels des utilisateurs, est particulièrement intéressant dans l'information aux « consommateurs ». Le but n'est pas tant

[8] Advanced Technology Program.

d'identifier le « meilleur système », que de donner une information comparative qui permette à l'utilisateur de faire un choix éclairé du produit relativement à ses besoins et à ses contraintes.

- **Evaluation de diagnostic**
  C'est ici un maillage de l'espace des entrées possibles, qui permet de dresser un profil du comportement attendu du système. L'une des difficultés est bien évidemment de construire une suite de tests représentative. Typiquement, les tests de programme par jeux d'essai relèvent de l'évaluation de diagnostic. Tout programmeur connaît bien les limites d'une telle approche : les jeux d'essai ne peuvent servir de preuve de programme, mais seulement de contre-exemples. Un autre exemple de ce mode d'évaluation est le développement des compilateurs. C'est à la suite d'une évaluation de diagnostic très poussée que la norme peut être créée et que les compilateurs peuvent être agréés en fonction de la qualité de leurs réponses aux tests de normalisation (*benchmarks*). Les tests de diagnostic permettent la maintenance d'un système évolutif par des tests de régression sur les versions consécutives du système. Un dernier exemple concerne les systèmes d'interaction Homme-Machine pour lesquels l'information de diagnostic est utile : ce que « réussit » un système et ce qui lui reste « à acquérir » sont autant d'informations importantes pour la conception et le développement de ces systèmes.

- **Evaluation de performance quantitative**
  Le triptique critère/mesure/méthode permet l'évaluation de performance quantitative. En reconnaissance de la parole, où la problématique est de retrouver, à partir du signal acoustique, les mots organisés en phrases, le distinguo est clair. Un classique du triptique est, -pour le critère, la « bonne » reconnaissance, -pour la mesure, le taux d'erreur en mots, et -pour la méthode, la comparaison de la transcription automatique du système et de la transcription de référence (dans le « meilleur des cas », faite par un expert humain) du signal acoustique. Ce mode d'évaluation, adapté à la comparaison d'implémentations alternatives, est utile pour le développement et l'amélioration des systèmes. Pour une même tâche de reconnaissance, les tests contrastifs permettent d'évaluer la contribution d'un composant sur les performances globales du système.

- **Evaluation de performance qualitative**
  Cette évaluation vise une meilleure connaissance de ce qui est prédominant dans l'évaluation quantitative du système. Comment et pourquoi le système fonctionne-t-il ? L'un des intérêts évident est que, généralement, pour un système très sophistiqué, cette connaissance est indispensable afin de simplifier ou d'améliorer ce système. Un exemple en est le projet d'évaluation MultiKron de NIST pour les applications parallèles et distribuées. La principale préoccupation est ici d'améliorer les applications ; l'observation du système, des goulots d'étranglement par exemple, peut conduire à des améliorations. L'instrumentation nécessaire à la capture de cette information ne doit cependant pas trop perturber le système sous peine de devenir non significative. La solution de NIST au problème a été la conception et le développement d'une puce « espionne ». Cette puce

mesure les performances des processeurs parallèles et des stations de travail sur des réseaux à haut débit, en enregistrant des événements comme par exemple les écritures-mémoire et les interruptions. Les mesures d'évaluation ainsi obtenues ont permis aux chercheurs de comprendre la source des goulots d'étranglement et d'apprendre comment améliorer les applications. Pour nombre de nos travaux, nous reviendrons sur l'évaluation qualitative.

A travers les méthodologies que nous venons de voir, il devient évident que l'évaluation joue un rôle capital pour le bon développement des systèmes, pour les intégrateurs de systèmes et enfin pour les consommateurs. Si l'évaluation n'est pas une politique scientifique, elle est au moins un guide précieux pour la recherche.

1.2. Paradigme d'évaluation en reconnaissance vocale

Le paradigme d'évaluation a pour but d'accélérer le développement de la recherche et de la technologie par des campagnes d'évaluation, selon un principe qui peut être résumé par « tâche commune, données communes, évaluation commune ». Bien que l'on retrouve ce paradigme dans bien des domaines, entre autres, celui de l'ingénierie de la langue, c'est dans sa mise en œuvre en reconnaissance de la parole que nous présentons le principe du paradigme défini par ARPA et utilisé dans les plans d'évaluation, aux Etats-Unis et en Europe (SQALE, AUPELF-UREF) :

- Diffusion d'un appel d'offres sur une **tâche spécifiée**.
- **Distribution** aux participants des **bases de données** nécessaires au développement des modèles et des systèmes.
- **Test** des systèmes participants sur des **données similaires** en un temps limité.
- Appel à une **infrastructure** pour la **mesure** et **la comparaison des performances** des systèmes.

Les principes du paradigme d'évaluation vont faire l'objet d'un développement pour en apprécier tout l'intérêt en reconnaissance de la parole. Enfin, l'historique des plans d'évaluation ARPA menés aux Etats-Unis depuis 1987 montre les progrès manifestes réalisés, dans le cadre de leur mise en œuvre, en reconnaissance de la parole.

1.2.1. « Tâche commune » Spécification de la tâche de reconnaissance

Spécifier une tâche de reconnaissance, c'est également fixer la complexité de la reconnaissance. De nombreux paramètres sont liés à la tâche de reconnaissance, leur spécification induit des facteurs de complexité qui ont une influence sur les performances « attendues » du système et sur les techniques à employer.

- **Spécification sur la dépendance du système aux locuteurs**
  Si le(s) locuteur(s) de test est (resp., sont) le(s) même(s) que le(s) locuteur(s) d'apprentissage, la reconnaissance est monolocuteur (resp., multilocuteur). Dans le cas où les locuteurs de test sont distincts des locuteurs d'apprentissage, la reconnaissance est indépendante du locuteur. La complexité est ici croissante ; elle s'explique par la grande variabilité du signal vocal dont on sait toute l'importance pour un même locuteur mais également pour des locuteurs différents. La variabilité a un effet immédiat sur les techniques à mettre en œuvre pour l'apprentissage des modèles acoustiques. Par exemple, la reconnaissance indépendamment du locuteur fait *a priori* appel à des techniques d'adaptation des modèles (indépendants du locuteur) au locuteur de test.
- **Spécification sur la nature du document audio traité**
  Indépendamment de l'environnement lors de l'enregistrement audio, il existe une grande variété de documents audio. Ainsi, une grande différence existe entre la parole lue, préparée, spontanée et conversationnelle. Cette spécification a une influence sur les modèles de langage requis, mais également sur la modélisation acoustique. Dans le cas de la parole lue ou préparée, les modèles de langage sous-jacents sont les modèles de l'écrit ou des modèles apparentés. Le langage de l'écrit est généralement bien structuré, et obéit à des règles de grammaire établies. En fait, la spécialisation ou non-spécialisation des documents traités est un autre facteur de complexité à considérer. Toute spécialisation (e.g., dictée de lettres commerciales, dictée de rapports médicaux, lecture de textes journalistiques) permet *a priori* une meilleure connaissance et une intégration possible des modèles de langage associés. La parole spontanée relève plus du langage naturel pour lequel le vocabulaire, la variété des expressions et des thèmes abordés sont bien plus importants, les modèles plus difficiles à apprendre. La parole spontanée est, de plus, sujette à des respirations, des hésitations, des erreurs de prononciation, des corrections ; autant de spécificités difficiles à modéliser. Pour la parole conversationnelle, deux paramètres de complexité s'ajoutent : le choix de l'échelle d'articulation allant de l'hypo-articulation à l'hyper-articulation, et les recouvrements possibles, sur le document audio, des signaux de parole des protagonistes. Enfin, sur des documents de type audio-vidéo, après le mixage des bandes sonores, on pourra trouver des superpositions de signaux (e.g., parole/parole, parole/musique, parole/bruit, bruit/musique). Dans ce cas, les différentes zones devront être caractérisées pour une reconnaissance adaptée, voire une décorrélation des signaux, avant de procéder à la reconnaissance vocale.
- **Spécification des conditions d'enregistrement**
  Lors de l'enregistrement des données audio, l'environnement peut-être non bruité ou bruité. Les bruits peuvent être additifs ou convolutifs. Le microphone peut être connu ou inconnu. Cette complexité est croissante dans le sens où les techniques mises en œuvre doivent être robustes au bruit, ce qui est loin d'être facile. Là encore, des techniques d'adaptation au bruit peuvent s'avérer nécessaires, ainsi que des primitives de pré-traitement du signal. Le canal de transmission typique pour la parole est le téléphone. Il est caractérisé par sa bande passante [300-

3400 Hz] qui n'est pas sans conséquence sur le signal vocal reçu ; la qualité de la transmission (courte ou longue distance) est variable et peut être responsable de distorsions et d'altérations du signal.

- **Spécification du vocabulaire de reconnaissance**
  La spécification de la taille du vocabulaire a une influence sur les performances et la complexité des systèmes de reconnaissance. Le passage d'un petit vocabulaire à un grand vocabulaire ou à un très grand vocabulaire présente un coût de calcul et un coût-mémoire. A architecture matérielle équivalente, l'optimisation peut devenir cruciale, sans parler de l'implémentation de la reconnaissance en temps réel. Une autre considération porte sur le vocabulaire fermé ou non fermé. Dans le premier cas, seuls les mots du vocabulaire considéré peuvent avoir une occurrence dans les corpus de test. Dans le deuxième cas, des mots n'appartenant pas au vocabulaire de reconnaissance peuvent avoir une occurrence. Ces mots appelés mots « hors vocabulaire » ont une influence sur la majoration des erreurs estimée à 2.5 fois leur nombre. L'un des grands principes de Reconnaissance des Formes demeure : « seuls les mots du vocabulaire de reconnaissance peuvent être reconnus », et l'on conçoit bien toute l'importance de la couverture du vocabulaire de reconnaissance relativement aux tests répondant à la spécialisation de la tâche.

Compte tenu de la multiplicité des paramètres de spécification de tâche et de la complexité qu'ils peuvent engendrer à la reconnaissance, toute comparaison de systèmes ne peut *a priori* être établie que sur une « tâche commune » de reconnaissance.

1.2.2. « Données communes » Corpus de développement et corpus de test

Pour les besoins stricts de la méthodologie d'évaluation, les « données communes » se résument à la distribution aux participants des corpus acoustiques de développement et de test. Les données de développement ont pour but de permettre l'adaptation des systèmes à la tâche spécifiée pour le test. Le corpus de développement doit être représentatif du corpus de test. Par conséquent, les données doivent être homogènes aussi bien par leur nature (e.g., texte dont le test est issu) que par leur condition d'enregistrement (e.g., matériel d'enregistrement identique). Dans le cas d'une reconnaissance indépendamment du locuteur, les locuteurs de développement et de test sont différents. Les données de développement sont antérieures aux données de test par leur date d'enregistrement et, par exemple, par les textes dont elles proviennent. La contrainte pour le test est que les données n'aient servi à aucun apprentissage et à aucun développement, les locuteurs étant évidemment différents. De plus, les données de développement et de test doivent être annotées afin de permettre l'évaluation de performance des systèmes suivant la méthodologie retenue. Dans le cas d'une évaluation par taux d'erreur en mots, la transcription orthographique des corpus est requise. Les données de développement et de test peuvent correspondre à plusieurs heures de parole. Si l'on conçoit bien l'intérêt d'une infrastructure, organisatrice du test, pour la mise en œuvre et la validation de la méthodologie, l'historique des plans d'évaluation met en évidence un autre apport du paradigme d'évaluation, celui du « partage de données » qui concerne également les ressources linguistiques.

1.2.3. « Partage de données » Ressources linguistiques

Les systèmes de reconnaissance sont en grande partie conditionnés dans leur performance par l'adéquation et la qualité des ressources linguistiques qu'ils utilisent. Dans un système de reconnaissance vocale, trois composants sont distingués suivant leur niveau de dépendance à la langue et à la tâche de reconnaissance : -le « cœur » du système dont l'implémentation est indépendante de la langue, -la représentation acoustique de la langue, bien évidemment dépendante de la langue, et -le modèle de langage du domaine, dépendant du langage et de la tâche de reconnaissance. Les ressources linguistiques utiles aux deux derniers composants désignent principalement -des corpus oraux (acoustiques), -des corpus écrits (textuels), ainsi que -des bases de données structurées à partir des connaissances académiques de la langue. Pour être utilisés par des systèmes de l'état de l'art généralement fondés sur l'observation et sur des techniques de modélisation statistique, les **corpus** doivent être de « **très larges** » corpus, disponibles sur **support électronique**, avec plusieurs niveaux d'**annotation** dont certains sont absolument nécessaires.

- **Corpus acoustiques**
  Dans leur meilleure adéquation à l'apprentissage des **modèles acoustiques des phonèmes**, les corpus acoustiques doivent être étiquetés par les phonèmes, les mots, les phrases, chacune de ces unités devant être référencée au signal vocal. Au minimum, la transcription orthographique des phrases correspondant effectivement au signal vocal est requise. Le **contrôle** des annotations est fondamental pour un bon apprentissage des modèles, mais il est fastidieux et coûteux lorsqu'il est assuré par un humain. Un contrôle automatique de qualité ne sera possible que lorsque les systèmes de reconnaissance auront atteint leur haute technologie.
- **Corpus textuels**
  Pour leur part, les corpus textuels doivent être « nettoyés », compte tenu de la variété de leur mise en forme et de leur contenu. Ils sont, entre autres, homogénéisés, corrigés, voire lemmatisés, et ceci afin de calculer des **modèles de langage**. Les modèles de langage les plus couramment utilisés sont des modèles statistiques estimés sur de grands corpus textuels (des centaines de millions de mots). Les modèles usuels sont les n-grammes et les n-classes ; il s'agit de l'estimation probabiliste de l'observation de n entités consécutives ; dans le premier cas, il s'agit des mots ; dans le second cas, il s'agit d'étiquettes de classes (e.g., classes grammaticales) associées aux mots.

Les corpus acoustiques utilisés peuvent atteindre des centaines d'heures de parole, les corpus textuels des centaines de millions de mots. La collecte sur support électronique et l'annotation des corpus représentent un travail considérable et qui pour être mené à bien nécessite un investissement important en temps et en moyens humains. Dans son appel d'offre de 1994, l'AUPELF-UREF remarquait que la quasi-absence des corpus écrits ou oraux en langue française était un très gros handicap pour l'élaboration des systèmes de traitement du langage ; et soulignait le retard des pays francophones sur les pays anglophones.

- **Lexique phonétisé**
  Dans l'hypothèse de reconnaissance du système, seuls les mots du vocabulaire de reconnaissance pourront être émis. Pour couvrir, dans une langue, l'étendue de la communication parlée, ce vocabulaire doit tendre vers le lexique, au sens linguistique, c'est-à-dire à l'ensemble des mots de la langue. En l'état actuel des techniques de reconnaissance, la **transcription phonétique** d'un **lexique** assurant la **couverture** des **corpus de test** et des **corpus acoustiques d'apprentissage** est **indispensable** pour le décodage et l'apprentissage acoustique des mots. L'une des sources de la connaissance du **lexique phonétisé** est académique, il s'agit du dictionnaire de la langue. Pour sa meilleure adéquation aux performances des systèmes, le lexique phonétisé utilisé doit être représentatif des élisions possibles, des variantes de prononciation ; il doit couvrir les flexions en genre, en nombre, ainsi que les flexions verbales qui ne sont pas décrites *in extenso* dans un dictionnaire ; les liaisons ou variantes de prononciation dues aux liaisons doivent également être représentées. De plus, dans le cas de l'utilisation de modèles de langage n-classes, tout mot du lexique phonétisé doit être *taggé* c'est-à-dire **étiqueté**, par exemple par sa classe grammmaticale (e.g., nom masculin singulier, adjectif cardinal, verbe à la première personne du singulier, adverbe), sa classe sémantique (e.g., nom de famille, nom de pays, entité de lieu, entité économique, entité politique), voire sa classe fonctionnelle (e.g., sujet, complément d'objet direct ou indirect).

L'évolution de la tâche de reconnaissance, telle que l'indexation des documents audio, montre qu'il est impératif, pour les indexer dans leur diversité, de disposer des ressources linguistiques les plus larges possibles. Des laboratoires de linguistique écrite ou orale ont consacré nombre de leurs efforts de recherche et de développement à l'informatisation des ressources académiques. De nombreuses autres entrées du dictionnaire sont à considérer, comme, par exemple, les noms propres (e.g., prénoms, noms de famille, pays, villes) ; le travail est considérable. De plus, en raison de l'évolution de la langue, ces bases de données nécessitent une maintenance continuelle, l'évolution se traduisant par l'enrichissement de nouveaux mots dans le dictionnaire et dans la langue courante.

L'importance de la quantité et de la qualité des corpus et bases de connaissance expliquerait à elle seule l'intérêt du « partage de données » pour les chercheurs. Ce partage devrait faire partie à part entière du paradigme d'évaluation.

1.2.4. « Evaluation commune » Evaluation de performance des systèmes

Par « évaluation commune », on entend outils d'évaluation communs mis en œuvre, validés et utilisés par une infrastructure organisatrice des tests. L'évaluation de performance la plus « classique » est celle de l'évaluation de la « boîte noire » ; elle consiste à comparer l'hypothèse de reconnaissance du système et la transcription de référence du signal acoustique d'entrée. La principale mesure de performance est le taux d'erreur en mots, défini par le rapport du nombre de mots erronés de l'hypothèse de reconnaissance au nombre de mots de la transcription de référence. Les différentes erreurs sont la substitution, l'insertion et l'omission. L'identification de ces erreurs résulte de l'alignement des mots de l'hypothèse et des mots de la référence. L'alignement est obtenu par une variante de l'algorithme de Wagner et Fischer habituellement utilisé pour la comparaison de chaînes de caractères. Une substitution a lieu lorsque les mots alignés diffèrent ; une insertion (resp. omission) a une occurrence lorsqu'un mot de l'hypothèse (resp. référence) n'a pas de correspondant dans la référence (resp. hypothèse). Pour des langues, comme par exemple le Mandarin, la mesure de performance est le taux d'erreur en caractères dont le principe est similaire à la mesure déjà exposée, mais transposée au niveau du caractère.

1.2.5. Historique des plans d'évaluation ARPA

Comme nous l'avons vu, les campagnes d'évaluation sont originaires des Etats-Unis. L'historique des plans d'évaluation d'ARPA[9] des dix dernières années montre les progrès manifestes en reconnaissance de la parole, imputables au paradigme d'évaluation. Les plans d'évaluation ont été très progressifs à leur début, période que l'on peut estimer de 1987 à 1992. Cette période correspond à la mise en œuvre du paradigme : au développement des bases de données, des ressources, des outils d'évaluation et des systèmes. A partir de 1992, la tâche de reconnaissance est de complexité continuellement croissante. Les bases de données et les ressources sont en constante évolution.

- **De 1987 à 1992**
  Durant cette période, la spécification des tâches de reconnaissance peut se résumer de la façon suivante : reconnaissance multilocuteur de la parole lue, avec un vocabulaire fermé de 1000 mots. La base de données acoustique *Resource Management* est utilisée pour l'apprentissage, le développement et le test. L'utilisation des ressources est contrainte : le corpus acoustique d'apprentissage des modèles est fixé, le lexique de reconnaissance est donné ainsi que les modèles de langage statistiques. L'objectif était alors de focaliser le développement des

[9] J.-L. Gauvain, « *Avancée en Reconnaissance de la Parole : de Resource Management à Broadcast News* », 1ères JST FRANCIL, Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la Langue de l'AUPELF-UREF, pp. 11-12, Avignon, 1997.

systèmes sur la modélisation acoustique. Sur des tâches comparables, les taux d'erreur en mots passent de 20% en 1987 à 4% en 1992. En 1992, les modèles de Markov cachés sont déjà à la base des systèmes de l'état de l'art.

- **De 1992 à 1997**
  De nombreux corpus sont utilisés pendant cette période : *Wall Street Journal* (1992-1993), *North American Business News* (1994-1995), *Broadcast News*, *Switchboard* (1996). De 1992 à 1993, la taille du vocabulaire progresse de 5 k-mots à 20 k-mots ; l'utilisation des ressources est contrainte pour l'évaluation de tests contrastifs. Par exemple, en 1993, évaluation des performances quantitatives d'un modèle de langage bigramme 5 k versus un modèle de langage trigramme 5 k. On peut également remarquer les premières évaluations comparatives avec enregistrement téléphonique « courte distance ».
  A partir de 1994, l'utilisation des ressources n'est plus contrainte, les participants ont acquis, au cours des campagnes annuelles, les ressources et toute une expérience qui leur permet de mettre en œuvre les stratégies les plus performantes. La taille du vocabulaire de reconnaissance atteint les 65 k-mots. En 1995, l'influence des conditions d'enregistrement est évaluée (enregistrement téléphonique, microphone connu ou inconnu) dans le but de développer des systèmes plus robustes à la variété des environnements acoustiques. En 1996, l'évaluation porte sur *Broadcast News :* la tâche est complexifiée par la nature même des documents audio (cent heures d'émissions radiophoniques et télévisuelles). Le corpus acoustique inclut la combinaison de parole lue, de parole spontanée, de divers environnements d'enregistrement, de musique, de fond musical. Les taux d'erreur varient de 27 à 54% suivant les systèmes testés. Une autre campagne d'évaluation, sur *Switchboard*, visant la reconnaissance de la parole conversationnelle en anglais via le téléphone donne 45% de taux d'erreur. Les derniers plans d'évaluation sont étendus à des langues non anglophones ; en l'occurrence l'Espagnol et le Mandarin.

Quelques résultats sont donnés à titre indicatif, ils permettent par exemple pour la première période d'apprécier la progression de performance des systèmes. Les résultats donnés pour *Broadcast News*, pour la seconde période, ne peuvent être appréciés qu'après une meilleure connaissance de la tâche (i.e., écoute des corpus de test) et du mode d'évaluation utilisé (i.e., tolérance d'erreur sur les segments autres que ceux de parole « propre »).

1.3. Succès de l'évaluation

En **reconnaissance de la parole**, les **succès de l'évaluation** sont manifestes surtout si l'on se réfère au paradigme d'évaluation et à son application aux Etats-Unis. Leur historique sur les dix dernières années met en évidence une première période de cinq années nécessaires à la consolidation des systèmes. L'aspect pluridisciplinaire de la Parole est pris en compte, et les efforts des « concepteurs » de systèmes de reconnaissance sont focalisés sur le développement même des systèmes, et non sur l'acquisition des ressources linguistiques sur lesquelles la communauté de recherche n'est pas spécifiquement formée. Le « partage des données » est effectif, les corpus acoustiques sont donnés ainsi que le lexique et les modèles de langage statistiques. Les cinq dernières années montrent une évolution très rapide de la tâche de reconnaissance ; les systèmes traitent alors des vocabulaires de reconnaissance de 65 k-mots, et ils commencent à s'adapter à la transcription d'émissions radio et télédiffusées dont la difficulté de décodage est bien plus importante que la « parole de laboratoire ». Les tâches de reconnaissance sont plus réalistes, et s'orientent déjà vers l'indexation des documents audio. En 1997, deux systèmes de dictée vocale temps réel, dépendant du locuteur et fonctionnant sur un matériel « grand public », sont commercialisés (Via Voice d'IBM et Naturally Speaking de Dragon System), un troisième système devrait bientôt l'être (Voice Xpress de Lernout et Hauspie) ; on peut remarquer qu'IBM et Dragon participent régulièrement aux plans d'évaluation ARPA. L'influence de l'évaluation sur l'évolution technologique est visible, comme en témoigne **l'augmentation des performances** des systèmes sur des **tâches** qui **se complexifient** de plus en plus et parfois très rapidement.

L'**efficacité** du principe de **partage de données** (voire ressources) et d'**évaluation commune** se retrouve dans bien d'autres domaines de l'ingénierie de la langue : la recherche de documents, l'extraction d'information, l'analyse du langage naturel, la traduction automatique, la reconnaissance du locuteur, la synthèse. En recherche de documents dans de grandes bases de données, les requêtes, à l'évaluation, sont variées, et les corpus utilisés peuvent atteindre des dizaines de milliers de documents. En extraction d'information à partir du texte, plus axée sur la compréhension, les corpus textuels d'apprentissage sont passés, en dix ans, de quelques centaines de textes très courts à plusieurs milliers de textes plus longs ; les corpus de test, eux, sont passés de l'ordre des dizaines à celui des centaines. L'évaluation de type « boîte noire » compare la sortie des systèmes de recherche ou d'extraction d'information à une référence provenant d'une expertise humaine, évidemment coûteuse lorsque l'on considère l'importance des corpus traités, et pas toujours fiable en raison de l'ambiguïté inhérente aux textes en langage naturel. Pour des raisons de faisabilité, en recherche d'information, seules les premières hypothèses (quelques centaines) de plus forte vraisemblance sont évaluées. Des mesures standard de performance des systèmes ont été définies, que l'on peut interpréter comme une évaluation de diagnostic des systèmes, couplée à l'attente de l'utilisateur. Les systèmes sont évalués, par les mesures de rappel et de précision, sur leur capacité -à retrouver toute l'information pertinente et -à retrouver seulement l'information pertinente. L'infrastructure, mise à la disposition des chercheurs dans le cadre de l'évaluation, -permet une **évaluation statistique** des performances des systèmes, -**évite** la **duplication des efforts** de mise en œuvre d'une évaluation, et -**focalise la recherche** sur des **problèmes difficiles**.

Les **plans d'évaluation**, nationaux ou internationaux, aboutissent à la **visibilité** des domaines et à la **dissémination** de la recherche via des conférences ou des workshops dédiés à l'évaluation de performance. En ingénierie de la langue, on peut citer, aux

Etats-Unis, SLTW[10], TREC[11], MUC[12], MTEW[13], et, en France, les premières JST Francil[14] en 1997 ; en Europe LREC.

En ingénierie de la langue, comme dans tout domaine, il est capital d'arriver très vite à la **haute technologie** ; par son transfert industriel, elle est porteuse de richesse et d'emplois, et l'enjeu est d'envergure. Tous les spécialistes savent l'importance du marché qui s'ouvrira à l'industrie de la langue. Les corpus, les ressources linguistiques et de nombreux outils de traitement automatique sont communs aux chercheurs des deux composantes « écrite » et « orale » de la linguistique informatique. Dans ces conditions, on conçoit bien toute l'importance de la **mise en œuvre** du **paradigme d'évaluation** pour le développement de la recherche et de la technologie. De nombreux pays lancent des programmes de **collectes de données** (orales ou écrites) relatives à leur **langue** ; des **infrastructures** ont été créées comme LDC[15] aux Etats-Unis, et ELRA[16] en Europe. De nombreux **projets** ont été financés pour la **création de bases de connaissance** complémentaires aux **ressources linguistiques** déjà acquises. Parmi ces projets, on peut citer des projets européens ESPRIT ; ONOMASTICA, par exemple, est un projet dont le but est la création de larges dictionnaires phonétisés des noms propres et des endroits connus pour onze langues européennes.

1.4. Limites et danger de l'évaluation

Tous les domaines ne sont pas « égaux » devant l'évaluation, et les méthodologies existantes admettent certaines **limites**. En reconnaissance de la parole, les méthodes d'évaluation sont simples à élaborer ; mais en dialogue oral, par exemple, l'évaluation est bien plus difficile. Si la performance de compréhension peut être évaluée, comme par exemple l'extraction d'information, il n'en va pas de même pour la qualité du dialogue et sa portabilité à la diversité des applications. D'une manière générale, il n'existe pas actuellement de méthodologie d'évaluation « incontestée » pour les **systèmes interactifs**. Une autre évaluation « difficile » est celle, dont est chargé le VVRG[17], sur les avancées technologiques de la visualisation tridimensionnelle. Sa mission est d'évaluer le rôle de la visualisation tridimensionnelle dans l'accès, la manipulation et l'échange d'information. Il s'agit de systèmes interactifs de recherche et d'accès aux documents dans ce que l'on peut appeler les « archives » du multimédia. La complexité de l'évaluation de tels systèmes s'explique entre autres par un protagoniste dont on parle peu : l'**utilisateur** et son **interaction avec le système**. En effet, l'évaluation du système doit tenir compte du confort pour l'utilisateur, et du « plus » réel apporté par le système

[10] Spoken Language Technology Workshops.
[11] Text Retrieval Conferences.
[12] Message Understanding Conferences.
[13] Machine Translation Evaluation Workshops.
[14] Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la Langue.
[15] Linguistic Data Consortium.
[16] European Language Ressources Association.
[17] Visualization and Virtual Reality Group.

lorsque l'utilisateur aura à sa disposition les outils sophistiqués de visualisation des documents à un niveau global mais également analytique, et qu'il pourra ainsi organiser, filtrer et inspecter les documents. Il n'existe pas non plus de méthodologie pour estimer la portabilité des systèmes sur de nouvelles applications.

L'un des **dangers** que les experts craignent dans l'évaluation tient à la **périodicité** même des **campagnes d'évaluation** et des **temps de développement et de test** qu'elles induisent. En effet, un risque peut exister de **détournement des efforts de recherche**. On sent effectivement très bien le danger de l'interaction devenant prépondérante sur l'action : « dix lignes de commentaires, une ligne de programme ». Les campagnes d'évaluation semblent être à consommer avec modération pour préserver l'innovation dans la recherche. On peut citer en exemple le dernier concours en reconnaissance du locuteur[18], concours auquel nous avons participé en 1997. Sur dix systèmes participants, neuf étaient fondés sur une technique déjà reconnue à l'état de l'art : les combinaisons de gaussiennes (GMM - Gaussian Model Mixture) ; seul un système était fondé sur la modélisation Auto-Régressive Vectorielle (ARV). Dans son « information en retour du gouvernement »[19], le Département de la Défense notifiait -le « lissage » des systèmes, -le comportement asymptotique des systèmes GMM avec des performances groupées, -son intérêt de voir émerger d'autres modélisations dont on pourrait espérer une amélioration des performances par fusion, et enfin -l'importance des moyens matériels utilisés par les systèmes GMM pour supporter une modélisation de plus en plus lourde. Nul doute que ces remarques seront suivies d'effet sur les prochains plans d'évaluation en reconnaissance du locuteur.

[18] Speaker Recognition Workshop – 31 March - 1 April 1998 - University of Maryland.
[19] Government feedback.

Introduction au chapitre 2

Avant d'introduire les travaux sur la reconnaissance markovienne regroupés dans ce chapitre, il me semble important de revenir sur le paradigme d'évaluation qui vient d'être présenté. C'est relativement à notre expérience de la première campagne de dictée vocale francophone de l'AUPELF-UREF que les principes du paradigme seront approfondis.

Il faut rappeler que le but premier du paradigme est d'accélérer le développement de la recherche et de la technologie. Nous avons vu comment une politique scientifique pouvait créer une infrastructure pour aider à l'un des facteurs de cette accélération : la mise en commun d'une tâche, de données, voire de ressources, et d'outils. Mais un autre facteur est indispensable à l'amélioration et à la maîtrise des systèmes (dans notre cas) : celui d'apprendre objectivement les raisons de l'amélioration ou de la dégradation des systèmes. Dans le cadre du paradigme d'évaluation, les tests contrastifs sont très indiqués. Pour être efficaces, les plans d'évaluation doivent être établis en conséquence et évoluer au fil des campagnes et des objectifs atteints. La complexité des systèmes est telle que les tests contrastifs d'un système ont peu de signification pour les autres participants ; ce n'est que sur la base de **références communes** que les **tests contrastifs** sont des éléments de **connaissance objective** pour tous les participants et utiles à l'amélioration des systèmes. Sa mise en œuvre n'est pas simple, elle consiste à établir -le cahier des charges, -le protocole d'évaluation des tests, -le calendrier (distribution, développement et test), et à veiller à la conformité au plan ; elle se doit également d'être rigoureuse et adaptée aux niveaux des systèmes.

En 1994, la dictée vocale en langue française est l'un des thèmes de l'action de recherche concertée pour le traitement de la parole, lancée par l'AUPELF-UREF. En répondant à l'appel d'offres, les participants s'engagent sur quatre ans à tester, au cours de deux campagnes d'évaluation, « *un système de dictée vocale capable de reconnaître de la parole continue, mono ou multilocuteur, pour des vocabulaires variant d'un millier à quelques dizaines de milliers de mots au cours du programme* ». La tâche de reconnaissance pressentie est la dictée de textes de journaux de langue française. L'un des objectifs est la mise en œuvre du paradigme d'évaluation s'étendant à l'ensemble des actions de recherche en linguistique écrite et orale, les protocoles étant voisins des protocoles ARPA. Deux catégories de systèmes seront finalement évaluées lors de la première campagne : les systèmes 20 k-mots et les systèmes « sans contrainte de vocabulaire ».

A l'échelle internationale, on peut replacer ce plan relativement aux évaluations faites aux Etats-Unis et en Europe. Aux Etats-Unis, les plans d'évaluation ARPA continuent leur croisière : en 1994, les systèmes traitent déjà un vocabulaire de reconnaissance de 65 k-mots, et en 1996, la tâche de reconnaissance proposée est la transcription d'émissions de radio et de télévision. En Europe, les premiers résultats du projet SQALE[20] sont publiés en 1995. Le but est alors d'expérimenter le paradigme

[20] Speech Recognition Quality Assessment for Linguistic Engineering.

d'évaluation dans un environnement multilingue (allemand, anglais, français), afin d'évaluer, relativement à la langue, -les difficultés de reconnaissance et -le degré d'indépendance des techniques à la langue. On peut remarquer que les données d'apprentissage (orales et textuelles), procurées par les participants, sont comparables en taille à celles de l'évaluation ARPA de 1992. Le vocabulaire de reconnaissance est de 20 k-mots pour l'anglais et le français, et de 65 k-mots pour l'allemand.

Le LIMSI, laboratoire français, a développé un système plusieurs fois couronné à l'état de l'art[21], entre autres dans les évaluations ARPA. Dans le contexte national et international, l'appel d'offres de l'AUPELF de 1994 résume bien la situation. Les laboratoires, dans leur ensemble, n'ont pas fait de la dictée vocale une priorité de recherche. Dans ces conditions, il ne fait aucun doute qu'une mise en œuvre efficace du paradigme, dont l'enjeu n'est plus à démontrer, est capitale pour amener les laboratoires à cette compétence.

Cette première campagne d'évaluation nous a montré que la mise en œuvre du paradigme n'est pas simple, le niveau hétérogène aussi bien des systèmes que de l'expérience des participants dans le domaine n'étant pas facile à gérer. Avec le recul, je vais maintenant essayer de décrire ce qui serait à mes yeux un paradigme d'évaluation « idéal ». Je me placerai dans le cas où les participants disposent d'un système de dictée vocale dans la langue utilisée, mais inadapté à la tâche de reconnaissance choisie et aux conditions d'acquisition. Dans ce cas, la mise en œuvre d'un paradigme d'évaluation idéal demande un certain nombre de phases (i.e., apprentissage ou adaptation des modèles acoustiques, apprentissage du modèle de langage, ajustement des paramètres du système) avant la phase de test proprement dite. Toutes ces phases doivent être parfaitement coordonnées car la chronologie des mises à disposition des ressources indispensables est alors cruciale. Les phases d'apprentissage et de développement sont analysées relativement au calendrier d'une campagne d'évaluation et au besoin de références communes. Par la suite, la durée de chaque phase sera estimée en prenant comme hypothèse des moyens matériels 2 PC à 200 MHz, correspondant au budget alloué à chaque équipe.

- **Apprentissage des modèles acoustiques**
  Cette phase n'est pas nécessaire quand les modèles acoustiques sont fournis à chaque équipe. De façon évidente, la distribution de cette ressource est intéressante pour les tests contrastifs concernant les méthodes d'adaptation des modèles acoustiques et l'intégration des modèles de langage.
  Pour une reconnaissance à grand ou à très grand vocabulaire, indépendante du locuteur, il n'est pas envisageable de modéliser chaque mot par un modèle de Markov caché distinct. Cela demanderait à chaque locuteur d'apprentissage de prononcer au moins une fois chaque mot du lexique de test. Les mots doivent être modélisés par une concaténation d'un nombre réduit d'unités. Les unités

J. M. Steeneken, D. A. Van Leuwen, « *Multi-Lingual Assessment of Speaker Independant Large Vocabulary Speech Recognition Systems : the SQALE Project* », European Conference of Speech Communication and Technology, pp. 1271-1274, Madrid, 1995.

[21] A notre connaissance, à l'état de l'art dans trois langues (-*anglais*, évaluation ARPA ; -*allemand*, évaluation SQALE ; -*français*, évaluation AUPELF-UREF).

privilégiées sont les phonèmes, en raison de la simplicité de la modélisation des mots pouvant être déduite de leur transcription phonétique. Un mot est alors modélisé par la concaténation des modèles de Markov cachés des phonèmes[22] de sa transcription phonétique. En conséquence, la ressource essentielle à l'apprentissage des modèles acoustiques est une base de données de parole annotée phonétiquement de manière à apprendre les modèles phonétiques.

L'**annotation idéale** doit être l'**étiquetage** « **fin** » **en phonèmes** ; dans ce cas, le début et la fin de chaque phonème sont référencés au signal. L'apprentissage des paramètres du modèle phonétique se fait à partir de l'observation des segments d'une même classe phonétique, en utilisant l'algorithme de Viterbi ou de Baum-Welch. L'annotation de la base de données est, comme nous l'avons vu avec son contrôle de qualité, très coûteuse, mais c'est une **référence commune** dont l'absence rend déjà plus difficile la comparaison des méthodes d'apprentissage.

En effet sans cet étiquetage fin, une autre référence commune est le **réseau phonétique des phrases** du corpus acoustique d'apprentissage. Ce réseau permet, par une succession de décodage et de réestimation de modèles acoustiques simplifiés, d'obtenir, après contrôle, un étiquetage fin du corpus d'apprentissage indispensable à l'apprentissage des modèles phonétiques. Cependant cet étiquetage n'est plus une référence commune car il peut différer pour chaque équipe.

Enfin, la référence commune la plus simple est la distribution du **lexique phonétisé** couvrant le **corpus d'apprentissage**. Dans ce cas, le réseau de décodage doit être reconstruit à partir des variantes phonétiques correspondant à un mot, de règles de liaison entre les mots, ainsi que de règles contextuelles qui peuvent différer suivant les participants. L'étiquetage final est encore plus variable selon les options du système. Dans ces conditions, que peut-on conclure de tests contrastifs sur les méthodes d'apprentissage des systèmes ?

En prenant par exemple BREF-TOTAL[23], corpus acoustique de 100 heures de parole, distribué pour la campagne, la taille du lexique phonétisé nécessaire à l'apprentissage des modèles acoustiques est de 30 854 mots. Les réseaux phonétiques correspondant aux 66 559 phrases de cette base sont composés en moyenne de 70 nœuds phonétiques et de 95 arcs. L'étiquetage phonétique comporte environ 4 millions de segments phonétiques. Sur cet exemple, la durée de la construction de l'étiquetage fin peut être estimée à ***deux semaines*** si le lexique phonétisé est fourni.

Dans le cas où les modèles acoustiques ne sont pas obtenus par adaptation mais initialisés à partir du corpus d'apprentissage, le **temps d'apprentissage** des **modèles acoustiques** peut être excessivement long : ***trois mois*** dans le cas de BREF-TOTAL. Ces temps doivent naturellement être pris en compte dans le calendrier établi pour la campagne, dans le cas où l'étiquetage fin n'est pas fourni et que l'adaptation des modèles est estimée peu sûre.

---

[22] Modèles phonétiques.

[23] J.-L. Gauvain, L. F. Lamel, M. Eskénazi, « *Design Considerations and Text Selection for BREF, a Large French Read-Speech Corpus* », International Conference on Spoken Language Processing, pp. 2359-2362, Kobe, Japan, 1990.

- **Apprentissage des modèles de langage**
  L'adaptation d'un système de dictée vocale à une tâche nécessite la fourniture d'un modèle de langage spécifique à cette tâche. Dans le cas où cette ressource n'est pas fournie, le temps de construction d'un modèle de langage standard de type n-grammes peut être estimé à ***une semaine*** si l'on suppose « nettoyés » les textes utilisés. Le nettoyage des textes fixe les conventions de l'évaluation du décodage et a une influence directe sur les modèles de langage. Il est également à considérer comme une référence commune nécessaire à la comparaison objective des systèmes.

- **Phase de développement**
  La phase de développement consiste entre autres à vérifier la pondération entre les modèles acoustiques et les modèles de langage. Elle détermine, par une série de tests dans les conditions réelles (i.e., chaque test de développement doit *a priori* être de même durée que le test final), le poids optimal (pour cette tâche) des modèles de langage. Les ressources essentielles pour la phase de développement sont -**les modèles acoustiques**, -**les modèles de langage**, -**le lexique phonétisé de test** (*a priori* différent de celui d'apprentissage). La durée de cette phase doit être un multiple (i.e., entre 5 et 10 fois) de celle de la phase de tests. Dans notre exemple où la **durée du test** était d'***un mois et demi***, la durée minimale de la phase de développement doit être de ***huit mois***.

Pour conclure sur mon avis d'un **paradigme d'évaluation « idéal »** et sur sa mise en œuvre, je dirais que la base de données de parole annotée phonétiquement au sens de l'étiquetage « fin » est une référence commune à distribuer en premier (au temps ***T0***). Les références communes au calcul des modèles de langage et le lexique phonétisé du test sont ensuite à fournir (au temps ***T0 + 3 mois***). La phase de test peut alors intervenir au temps ***T0 + 11 mois***. Il ne s'agit que de la durée minimale, car une nouvelle phase d'apprentissage des modèles acoustiques peut être nécessaire quand ceux-ci se révèlent inadéquats lors de la phase de développement. Ces durées sont naturellement fonction du parc-machine utilisé qui, lui aussi, doit être pris en compte dans le paradigme d'évaluation. On peut remarquer que les plans d'évaluation ARPA sont annuels et que les corpus de développement considérés sont les tests de l'année précédente lorsque les tâches de reconnaissance sont conservées d'une année sur l'autre. Seules une bonne maîtrise des techniques markoviennes, une grande expérience, et une haute performance des systèmes, permettent de s'affranchir du paradigme d'évaluation idéal.

Chapitre 2

Reconnaissance markovienne

Introduits en 1913 par Markov, les modèles de Markov cachés n'ont été appliqués à la reconnaissance de la parole qu'au début des années 70 par J. K. Baker[24] et F. Jelinek[25] ; ces modèles sont aujourd'hui (encore) à la base des systèmes de l'état de l'art en reconnaissance vocale[26]. La description des systèmes participant aux campagnes d'évaluation nationales ou internationales montre qu'ils sont tous à dominante markovienne. Plusieurs facteurs expliquent le succès et l'utilisation, qui tend à se généraliser, de la modélisation markovienne : leur cadre mathématique, le développement et la mise à disposition de grandes bases de données nécessaires à l'estimation statistique, une architecture adéquate à l'intégration de connaissances linguistiques stochastiques, et enfin l'évolution technologique matérielle.

Rappelons tout d'abord le cadre de la modélisation markovienne en reconnaissance de la parole continue. Le continuum de parole est représenté par une séquence d'observations (vecteurs acoustiques) calculées à partir d'analyses (synchrones/asynchrones à la période fondamentale) à court terme. Les modèles de Markov cachés sont des automates d'états finis probabilistes, qui modélisent en reconnaissance de la parole des unités linguistiques (e.g., phonèmes), voire des unités acoustiques (e.g., fénones). Dans le cas simple de la modélisation des phonèmes, la modélisation acoustique des mots est simplement déduite de leur transcription phonétique par la concaténation des modèles phonétiques correspondants. La topologie des modèles est définie par le nombre des états considérés de l'automate et les transitions autorisées entre ces états. Aux

---

[24] J. K. Baker, « *Stochastic Modeling for Automatic Speech Understanding* », in D. R. Reddy editor, *Speech Recognition*, pp. 521-542, Academic Press, New York, 1975.

[25] F. Jelinek, « *Continuous Speech Recognition by Statistical Methods* », IEEE Transaction on Acoustic, Speech, Signal Processing, vol. 64, n° 4, pp. 532-556, 1976.

[26] R. De Mori, *Spoken Dialogues with Computers*, Signal Processing and Its Application Series, Academic press, 1998.

F. Jelinek, *Statistical Methods for Speech Recognition (Language, Speech, and Communication),* MIT Press, 1998.

transitions entre états sont associées d'une part leur probabilité de transition et d'autre part leur probabilité d'émission d'une observation : ces probabilités, liées à une topologie donnée, constituent les paramètres du modèle markovien. Connaissant la topologie des modèles ainsi que le type de distribution d'émission des transitions, peuvent alors être utilisés des algorithmes -d'apprentissage des paramètres du modèle à partir d'un ensemble d'observations, et -de décodage d'une séquence d'observations. De plus, l'architecture markovienne présente l'avantage d'intégrer des connaissances linguistiques fondées, pour les plus simples, sur l'estimation statistique de la contiguïté d'entités linguistiques, ou pour les plus sophistiquées, sur des grammaires non contextuelles[27]. Les systèmes markoviens sont ainsi auto-organisateurs, à partir d'observations suffisantes en quantité et en qualité ; il est théoriquement possible d'apprendre les paramètres du modèle pour un décodage optimal, même si en pratique un élagage le rend sous-optimal.

L'évolution technologique des ordinateurs n'est pas sans influence sur les possibilités et sur la maîtrise des systèmes markoviens. La reconnaissance markovienne est puissante mais coûteuse en temps de calcul et en mémoire. Elle doit établir un compromis entre les solutions algorithmiques (recherche et optimisation) et les contraintes matérielles (rapidité et mémoire). Les ordinateurs « grand public » sont devenus bon marché, et avec une configuration raisonnable, ils permettent déjà une dictée vocale temps-réel de 65 k-mots dépendant du locuteur, ce qui était inimaginable il y a encore quelques années.

Ce chapitre regroupe, en deux parties, les travaux de notre équipe sur la modélisation markovienne. La première partie est consacrée à la description de notre **système de dictée vocale D-DAL**, des options que nous avons choisies, de ses composants linguistiques, et de ses résultats à la première campagne d'évaluation de dictée vocale francophone. La seconde partie présente deux **systèmes hybrides** fondés sur la modélisation markovienne. Le premier système se caractérise par son contrôle temporel dans les modèles de Markov cachés. Le second système utilise la règle de décision des K-plus proches voisins comme alternative à l'estimation gaussienne des probabilités d'émission.

2.1. D-DAL : le dictaphone-Dactylographe du LIP6

Le moteur de reconnaissance de D-DAL a été conçu pour pouvoir utiliser tout modèle de langage de type n-grammes (i.e., probabilité d'un mot connaissant les *n-1* précédents). Le processus de décodage est alors divisé en *n-1* passes synchrones. La première passe utilise l'algorithme du *one pass*[28] et un modèle de langage bigramme pour créer un graphe de mots centiseconde de profondeur P_1 ($P_1 < V$) où V est la taille du vocabulaire

[27] A. Corazza, R. De Mori, R. Gretter, G. Satta, « *Language Modeling using Stochastic Context-Free Grammars* », Speech Communication, vol. 13, pp. 163-170, 1993.

[28] H. Ney, « The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition », IEEE Transactions on Acoustic, Speech, and Signal Processing, vol. 32, n° 2, pp. 263-271, 1984.

de reconnaissance. La deuxième passe réorganise ce graphe de mots en utilisant un modèle de langage trigramme pour ne garder qu'un graphe de mots de profondeur P_2 ($P_2 < P_1$). Enfin la dernière passe (n-$1^{ème}$ passe) engendre l'hypothèse textuelle de D-DAL. La structure d'un graphe de mots centiseconde de profondeur P et de longueur T (nombre d'observations résultant de l'analyse centiseconde du signal à décoder) est une matrice G de taille PxT dont chaque élément $g_{ij\ (i=1,...,P\ ;\ j=1,...,T)}$ repère les informations du décodage au temps j suivantes : l'indice du mot de l'hypothèse de décodage, l'indice du début de ce mot associé au décodage, la probabilité acoustique de ce mot calculée au décodage.

Cette division en *n-1* passes permet de diminuer fortement les coûts de calcul car la réorganisation d'un graphe de mots centiseconde de profondeur P par un modèle de langage de type n-grammes a un coût de calcul en $O(P^n)$. Ce coût s'explique par le nombre de n-grammes estimés à chaque trame : chaque mot pouvant être précédé par P mots. L'ensemble des passes a pour but de faire remonter peu à peu la bonne solution, avec l'espoir de la retrouver en première position après la dernière passe. Le coût de calcul de la première passe, en $O(V^2)$, est le plus important ; dans le cas d'une reconnaissance grand vocabulaire, V est de l'ordre de 65 k-mots. Pour obtenir une reconnaissance dans un temps raisonnable, il est indispensable d'optimiser le nombre de n-grammes scrutés ainsi que leur temps d'estimation.

Des méthodes de coupure ont d'abord été développées pour réduire fortement le nombre de n-grammes scrutés. Elles ont nécessité le calcul préalable des minima et maxima des n-grammes, des n-grammes en fonction du dernier mot, des n-grammes en fonction de deux derniers mots, etc... A elles seules ces techniques ne nous ont pas permis de réduire suffisamment le nombre de n-grammes estimés. Des méthodes de recherche en faisceau ont dû être développées pour limiter arbitrairement le nombre de solutions explorées.

L'estimation d'un n-gramme à l'aide de la méthode *backing off*[29], l'une des plus rapides, peut nécessiter plusieurs étapes. La première consiste à chercher si ce n-gramme est apparu dans les corpus utilisés pour apprendre les modèles de langage, et à accéder à son compte (nombre d'occurrences). Dans le cas où ce n-gramme n'est pas apparu, ce qui est très probable, le n-gramme doit être calculé à l'aide du *backing off* du (n-1)-gramme ; une deuxième étape cherche le (n-1)-gramme qui peut lui-même ne pas être apparu, etc... Chaque recherche d'un n-gramme est longue (quelques dizaines de comparaisons) car le nombre de n-grammes ($n > 1$) est en général supérieur à quelques dizaines de millions. Une recherche en trois étapes a permis d'optimiser ces recherches : l'accès direct en fonction du dernier mot du n-gramme, l'adressage calculé, et la recherche dichotomique dans les collisions.

Ce moteur de reconnaissance a été implémenté par Fabrice Lefèvre en modifiant entre autres des parties du code source de HTK[30], la plate-forme de développement des systèmes de reconnaissance markoviens, de l'université de Cambridge.

[29] S. M. Katz, « Estimation of Probabilities for the Language Model Component of a Speech Recognizer », IEEE Transactions on Acoustic, Speech and Signal Processing, vol. 3, pp. 400-401, 1987.
[30] Hidden Markov model ToolKit.
S. J. Young, *HTK Version 1.4: Reference Manual and User Manual*, Cambridge University Engineering Department - Speech Group, 1992.

2.1.1. D-DAL96

Dans le cadre de la première campagne de dictée vocale francophone, D-DAL a concouru dans la catégorie « sans contrainte de vocabulaire », indépendamment du locuteur. Je suis responsable de ce contrat, d'une durée de quatre ans, financé par l'AUPELF-UREF.

Les ressources utilisées sont les archives électroniques des dix dernières années *du Monde* (du 01/01/87 au 02/11/96), le corpus acoustique BREF-TOTAL, le DP-L (dictionnaire phonétisé du LIP6) et deux PC Pentium-Pro 200 MHz. Les modèles de langage dynamiques sont de type *backing off* avec un seuil de 2 (estimation des n-grammes et des n-classes ayant au moins deux occurrences). 74 classes grammaticales sont considérées pour les n-classes. Les vecteurs acoustiques sont composés des 15 coefficients MFCC, de l'énergie et de leurs delta-coefficients.

Les travaux de Claude Barras ont été une contribution importante à l'apprentissage des modèles acoustiques. Leur originalité est l'utilisation d'un test du chi-2 pour le dédoublement des gaussiennes. Cela permet d'adapter le nombre de gaussiennes de chaque état de chaque machine à la distribution des données d'apprentissage. Les modèles acoustiques sont les suivants : -les monophones (37 modèles, dont le silence, représentés par 1 758 gaussiennes), -les biphones gauches (1 104 modèles représentés par 9 545 gaussiennes), -les biphones droits (1 104 modèles représentés par 11 436 gaussiennes), -les triphones (10 005 modèles représentés par 10 879 gaussiennes).

Fondés sur les travaux de Claude Montacié en reconnaissance du locuteur indépendamment du texte, deux types d'adaptation au locuteur des modèles acoustiques sont disponibles dans D-DAL. Les modèles des locuteurs d'apprentissage, ainsi que les modèles « homme » et « femme », sont appris à partir de la modélisation auto-régressive vectorielle des signaux d'élocution. Une mesure inter-locuteurs est définie entre deux modèles de locuteurs. Le premier type d'adaptation consiste en la détection du genre (féminin et masculin) du locuteur de test, à partir des modèles « homme » et « femme ». Les modèles acoustiques spécialisés par genre, préalablement calculés, sont alors sélectionnés. Le second type d'adaptation, plus novateur, consiste à étudier la typologie du locuteur de test et à adapter les modèles acoustiques indépendants du locuteur. Le locuteur le plus proche du locuteur-test, parmi les locuteurs d'apprentissage et les locuteurs appris depuis le début du test, est tout d'abord recherché. Dans le cas où le locuteur-test reste inconnu, les modèles acoustiques sont adaptés sur les signaux de parole du locuteur d'apprentissage le plus proche. Dans le cas contraire, les modèles acoustiques précédemment adaptés pour ce locuteur sont utilisés.

Dans sa version 96, D-DAL est organisé en quatre phases de reconnaissance que nous décrivons ici brièvement :

- Typologie par genre du locuteur de test et sélection des modèles acoustiques correspondants.
- Sélection d'un lexique dynamique de 40 k-formes (lexicalement, acoustiquement

et grammaticalement différentes) par décodage acoustico-phonétique du signal de test et analyse du treillis phonétique de profondeur 3.

- Décodage par l'algorithme *one pass* engendrant un graphe de mots centiseconde de profondeur 60 utilisant le modèle de langage des bigrammes et des biclasses avec le traitement des liaisons et des pauses. Le graphe de mots comporte en moyenne 600 000 arcs et permet d'engendrer potentiellement 600^{60} hypothèses différentes de phrases pour une durée type de 10 secondes.
- Réorganisation du graphe de mots en utilisant le modèle de langage des trigrammes et des triclasses pour donner l'hypothèse textuelle de D-DAL.

Dans ces conditions, la durée de décodage d'une phrase de dix secondes peut être estimée à 2h30 sur un PC Pentium-Pro 200 MHz.

Collaborations : Claude Montacié, Claude Barras (thésard - Equipe Parole), Fabrice Lefèvre (thésard - Equipe Parole).

Publications liées : 4 articles dont 2 internationaux [16, 17] et 2 nationaux [31, 33], 1 workshop national [42], 2 rapports de fin de contrat [48, 50].

Publications jointes : Recueil des publications page 1 [33] et page 5 [17].

2.1.2. Composants linguistiques

Les composants linguistiques nécessaires à un système de dictée vocale, comme nous l'avons déjà vu, regroupent les lexiques phonétisés et les modèles de langage.

En reconnaissance vocale, le vocabulaire de reconnaissance est l'ensemble des mots pouvant apparaître dans l'hypothèse de décodage du système. Ce vocabulaire doit permettre de construire le réseau de décodage des mots qui est utilisé par le moteur de reconnaissance. Rappelons que dans le cas de la reconnaissance à grand vocabulaire, les unités de reconnaissance sont les phonèmes, et les modèles acoustiques sont les modèles de Markov cachés de ces unités. Le réseau de décodage est la description des alternatives de mots de ce vocabulaire ; chaque mot y est représenté comme la concaténation des modèles acoustiques des unités phonétiques le constituant. La connaissance du vocabulaire de reconnaissance est par conséquent indispensable dans sa version phonétisée pour construire ce réseau. La nature de l'annotation des bases de données acoustiques utilisées pour l'apprentissage des modèles acoustiques conduit à distinguer un autre lexique : le lexique d'apprentissage. Comme il a été dit auparavant, la meilleure adéquation des bases acoustiques au problème de la reconnaissance vocale est leur étiquetage fin. S'il existe des « astuces » pour parvenir à un apprentissage sans cet étiquetage, c'est par une procédure itérative d'obtention de l'étiquetage fin mais avec peu d'assurance sur sa qualité. Dans ce cas, les procédures font appel à la connaissance même imparfaite des modèles acoustiques, et pour chaque phrase d'apprentissage elles permettent d'aligner sa transcription phonétique sur le signal vocal correspondant ; les modèles acoustiques peuvent alors être réestimés. Ainsi, un lexique phonétisé (de qualité) couvrant les corpus d'apprentissage est indispensable. C'est afin de répondre aux exigences de la reconnaissance markovienne « sans contrainte de vocabulaire » et à la meilleure couverture des corpus, tant en reconnaissance qu'en apprentissage, que nous avons développé le **DP-L**, le **dictionnaire phonétisé du LIP6**.

Le vocabulaire de reconnaissance n'est *a priori* pas le même que le lexique d'apprentissage. Sans modèle prédictif de ce qui peut être dit, on choisit usuellement un vocabulaire de reconnaissance de taille fixe, composé des mots les plus fréquents observés pour une tâche. Nous avons introduit le principe du **lexique dynamique** qui est une stratégie de décodage pour un système de reconnaissance utilisant un très large vocabulaire, et dont le but est de prédire au mieux la couverture du test.

Dans le cadre probabiliste des systèmes de dictée vocale, les **modèles de langage** les plus classiques parce que les plus « simples à intégrer » sont les modélisations statistiques des **n-grammes** (probabilité d'un mot connaissant les *n-1* précédents) et/ou des **n-classes** (probabilité d'une classe de mots connaissant les *n-1* précédentes). Leur puissance est par conséquent limitée à un contexte de proximité sans apport réel de la sémantique ; dans la pratique, 2 ou 3 sont des valeurs usuelles pour *n*, bien que l'on trouve des quintigrammes dans la littérature. Ces modèles sont essentiels aux performances des systèmes de dictée vocale. Leurs comptes sont calculés sur de grandes bases de données textuelles, les modèles de langage sont ainsi spécialisés dans le domaine des bases textuelles apprises.

DP-L : le Dictionnaire Phonétisé du LIP6

Pour couvrir toute l'étendue de la communication parlée, il serait évidemment souhaitable que le vocabulaire de reconnaissance couvre toute l'expression de la langue. Les dictionnaires de la langue et les dictionnaires spécialisés forment la base d'une telle ressource. Quelques ordres de grandeur : *Larousse* répertorie -76 000 entrées dans son Lexis, -30 600 entrées dans son dictionnaire des noms de famille et prénoms de France ; *Le Robert* répertorie -80 000 entrées lexicales, -42 000 entrées dans son dictionnaire des noms propres en histoire, géographie, arts, littérature, et sciences. On peut remarquer que seuls les dictionnaires de la langue incluent systématiquement, dans la structure des articles, la phonétique normative (décrite à partir de l'API[31]) et la caractérisation grammaticale du mot.

Pour les besoins de la dictée vocale « sans contrainte de vocabulaire », nous avons entièrement développé le DP-L (Dictionnaire Phonétisé du LIP6). Le corpus d'apprentissage considéré pour la campagne d'évaluation est BREF-TOTAL. Cette base n'est pas annotée par son étiquetage fin, seule la transcription orthographique des phrases est donnée. Le lexique d'apprentissage phonétisé est, comme nous l'avons vu, indispensable à la modélisation acoustique : c'est l'une des premières exigences de couverture du DP-L. La base de données comprend 66 559 phrases représentant 30 854 mots différents, parmi lesquels environ 9 500 mots sont des noms propres ou des mots étrangers.

[31] Alphabet Phonétique International.

Le DP-L a été construit à partir des 80 k-mots phonétisés du dictionnaire *Le Grand Robert*. Il a par la suite été corrigé, en particulier avec l'ajout de plus de 4 000 élisions manquantes correspondant aux *e muets*. Par exemple, la transcription phonétique du mot *appelé*, que l'on trouve dans le dictionnaire, considère un *e muet* [a p l e] (le *e* ne se prononce pas). Nous avons corrigé cette transcription par [a p ə* l e], où l'étoile indique la réalisation optionnelle du phonème /ə/, et qui correspond effectivement à une variante de prononciation que l'on peut retrouver dans les signaux de parole. Le DP-L a ensuite été étendu par l'adjonction des flexions qui ne sont pas décrites *in extenso* dans le dictionnaire. Parmi ces flexions, -les flexions en genre, en nombre, -les flexions verbales (en temps et en conjugaison), soit plus de 500 000 formes lexicalement différentes. Les variantes phonétiques engendrées par les liaisons ont été traitées, en moyenne 2,15 variantes phonétiques par mot. le DP-L répertorie ainsi plus d'un million de formes phonétiquement différentes.

Au nettoyage et à l'analyse des textes des dix dernières années, puisés dans les archives *du Monde*, on estime à 600 000 les éléments présentant une emphase et pouvant correspondre à des noms propres ou à des sigles. Les 15 000 noms propres et sigles les plus fréquents ont été phonétisés par analogie et contrôlés à l'écoute en utilisant le logiciel de synthèse MBR-PSOLA[32]. Parmi les noms propres (e.g., prénoms, pays, villes, lieux « connus »), les noms de famille, par leur nombre et leur caractère international, montrent l'étendue du travail d'acquisition et le problème de leur phonétisation. En effet, il existe souvent au moins deux phonétisations issues de la prononciation, l'une francisée et l'autre de la prononciation d'origine : par exemple, les trois variantes phonétiques pour *Maastricht* [m a s t r i k], [m a s t r i t* ʃ][33] ont été retenues. La phonétisation des mots étrangers entrant dans la langue française présente le même ordre de difficulté. Pour les sigles, la transcription de l'acronyme, lorsqu'elle existe, est également considérée (e.g., *ONU* : /o ɛ n y/, /o n y/). Ces ressources en langue française sont pour certaines déjà développées, pour d'autres en cours de développement, mais elles ne sont pas encore distribuées par les infrastructures prévues à cet effet.

Collaborations : Claude Montacié, Mathias Heidl (stagiaire de MST)

Lexique dynamique

Le lexique dynamique est une stratégie de décodage pour un système de reconnaissance à très large vocabulaire, utile pour réduire le coût du décodage lexical et prédire au mieux le vocabulaire de reconnaissance. Tout mot hors du vocabulaire de reconnaissance ne sera pas reconnu ; une hypothèse sera néanmoins émise par le système pouvant conduire à une substitution de plusieurs mots au mot inconnu. On estime que les mots hors vocabulaire induisent 2,5 fois leur nombre d'erreurs. Compte tenu des architectures matérielles utilisées, la taille maximum du vocabulaire de

[32] T. Dutoit, H. Leich, « *MBR-PSOLA : Text-To-Speech Synthesis based on an MBE Re-Synthesis of the Segments Database* », Speech Communication, vol. 13, n° 3-4, 1993.
[33] /t*/ : t optionnel.

reconnaissance est de 65 k-mots. Il est par conséquent stratégique qu'un vocabulaire de reconnaissance de cette taille puisse couvrir au mieux la tâche de reconnaissance considérée.

Le lexique dynamique vise à sélectionner un sous-lexique du vocabulaire de reconnaissance, couvrant le « mieux possible » le signal de test à décoder. Son principe est fondé sur le calcul des dissimilarités entre le treillis phonétique à *n* choix résultant du décodage acoustico-phonétique (DAP) du signal de test, et la transcription phonétique de chaque mot du vocabulaire originel. Les **mesures de dissimilarité acoustico-lexicales** sont basées sur l'algorithme de Wagner et Fisher développé, à l'origine, pour comparer des chaînes de caractères. Il a été modifié pour trouver la position optimale d'une sous-chaîne dans une chaîne. Trois opérations sont définies pour la comparaison des symboles : la substitution, l'omission et l'insertion. Un coût est affecté à chaque opération. Le but de cet algorithme de programmation dynamique est de minimiser la somme des coûts de ces opérations. La fonction de coût est apprise pour prendre en compte les erreurs du DAP. Les coûts des opérations définies entre deux phonèmes sont choisis inversement proportionnels aux pourcentages donnés par la matrice de confusion calculée d'après le DAP d'un ensemble d'apprentissage. Pour intégrer la fréquence d'occurrence d'un mot, la dissimilarité obtenue par programmation dynamique est multipliée par une fonction inverse de cette probabilité *p* d'apparition (e.g., *-log(p)*).

Le lexique dynamique que nous avons choisi de sélectionner pour D-DAL96 a une taille de 40 k-mots ; il est composé des 20 k-mots les plus fréquents des dix ans *du Monde*, et des 20 k-mots les plus proches du signal de test, dont 10 k-mots sans intégrer la fréquence d'occurrence des mots. L'intérêt de cette sélection est -pour les 20 k-premiers mots, de ne pas écarter les mots très fréquents même en cas de très mauvais décodage, -pour les 10 k-mots suivants qui prennent en compte la fréquence des mots, de ne pas écarter les mots fréquents en cas de mauvais décodage, et -pour les derniers 10 k-mots, de sélectionner les mots peu fréquents bien décodés. L'apport du principe du lexique dynamique relativement au vocabulaire de reconnaissance peut être évalué par sa couverture du corpus d'un test de reconnaissance. L'intérêt du lexique dynamique a été montré *a posteriori* de la première dictée vocale de l'AUPELF : pour un même taux de couverture, le lexique dynamique 40 k-mots que nous avions sélectionné est à comparer au lexique des 100 k-mots les plus fréquents *du Monde*.

Collaboration : Claude Montacié.

Publications liées : 2 articles dont 1 international [17] et 1 national [36].

Publication jointe : Recueil des publications page 5 [17].

Modèles de langage

Les modèles de langage sont une composante dépendante de la tâche de reconnaissance. Les modèles que nous avons utilisés pour la campagne AUPELF sont spécialisés sur la reconnaissance de textes journalistiques. Les bigrammes, les trigrammes, les biclasses et les triclasses, ont été appris à partir des dix dernières années des archives du journal *Le*

Monde[34]. Le corpus textuel correspondant est composé de 9 millions de phrases et de plus de 250 millions de mots. Un travail préliminaire au calcul des modèles de langage est le « nettoyage » des corpus textuels. Nettoyer ces textes a consisté à s'affranchir des balises de mise en forme, à corriger dans la mesure du possible les erreurs (coquilles, accents, etc...), à transcrire en mots les chiffres et les abréviations (Mme, Mm : Madame, Me : Maître), à homogénéiser les termes. Les conventions définies au nettoyage, telles les règles de réécriture, l'éclatement ou non des sigles, le traitement des emphases, doivent être prises en compte à l'évaluation puisqu'elles sont également partagées par le vocabulaire de reconnaissance.

L'utilisation des modèles de langage n-classes nécessite le *taggage* des corpus textuels, c'est-à-dire l'étiquetage des mots en classes. Ainsi, le vocabulaire de reconnaissance et le lexique de couverture des textes d'apprentissage des modèles de langage doivent évidemment être *taggés*. Les classes considérées pour les modèles n-classes peuvent être de natures diverses. Néanmoins, l'importance de la grammaire dans l'expression écrite nous a conduits à considérer les classes grammaticales[35] caractérisées par leur recouvrement, contrairement à des classifications automatiques consistant généralement à partitionner l'ensemble des mots. D'environ 400 classes grammaticales initiales, celles-ci ont été réduites à 74 de façon à *tagger* automatiquement les textes par un algorithme de type avant-arrière initialisé par exemples. La liste[36] des classes grammaticales que nous avons utilisées est la suivante :

1. Adjectifs cardinaux 2. Adjectifs féminins pluriels 3. Adjectifs féminins singuliers 4. Adjectifs masculins pluriels 5. Adjectifs masculins singuliers 6. Adverbes 7. Articles féminins singuliers 8. Articles masculins singuliers 9. Articles pluriels 10. Conjonctions 11. Débuts de phrases 12. Fins de phrases 13. Interjections 14. Locutions 15. Locutions verbales 16. Noms féminins pluriels 17. Noms féminins singuliers 18. Noms de lettres 19. Noms masculins pluriels 20. Noms masculins singuliers 21. Noms de mois 22. Noms propres, Chefs d'Etat et de Gouvernement 23. Noms propres, entités « économiques » 24. Noms propres, entités objets « marques » 25. Noms propres, entités « politiques » 26. Noms propres, noms de famille 27. Noms propres, entités « emphasées » 28. Noms propres, entités « géographie étrangère » 29. Noms propres, entités « géographie française » 30. Noms propres, hommes politiques 31. Noms propres, entités « inconnues » 32. Noms propres, lieux connus 33. Noms propres, mots étrangers 34. Noms propres, pays 35. Noms propres, prénoms 36. Noms propres, villes étrangères 37. Noms propres, villes françaises 38. Onomatopées 39. Verbes première personne du singulier 40. Verbe *avoir* première personne du singulier 41. Verbe *être* première personne du singulier 42. Verbes deuxième personne du singulier 43. Verbe *avoir* deuxième personne du singulier 44. Verbe *être* deuxième personne du singulier 45. Verbes troisième personne du singulier 46. Verbe *avoir* troisième personne du singulier 47. Verbe *être* troisième personne du singulier 48. Verbes première personne du pluriel 49. Verbe *avoir* première personne du pluriel 50. Verbe *être* première personne du pluriel 51. Verbes deuxième personne du pluriel 52. Verbe *avoir* deuxième personne du pluriel 53. Verbe *être* deuxième personne du pluriel 54. Verbes troisième personne du pluriel 55. Verbe *avoir* troisième personne du pluriel 56. Verbe *être* troisième personne du pluriel 57. *ayant* 58. *Etant* 59. Préfixe 60. Préposition 61. Pronom neutre 62. Pronom première personne du pluriel 63. Pronom première personne du singulier 64. Pronom deuxième personne du pluriel 65. Pronom deuxième personne du singulier 66. Pronom féminin troisième personne du pluriel 67. Pronom féminin troisième personne du singulier 68. Pronom indéfini 69. Pronom masculin troisième personne du pluriel 70. Pronom masculin troisième personne du singulier 71. Sigles 72. Verbes à l'infinitif 73. *Avoir* 74. *Etre*

---

[34] Je remercie Madame Foliot de l'OCD de m'avoir introduite auprès de Monsieur Deshaies, responsable technique *au Monde*.

Je suis profondément reconnaissante à Monsieur Deshaies pour la compréhension et la bienveillance qu'il a bien voulu nous accorder.

[35] M. Grevisse, *Le bon usage - Grammaire française* , 8ème édition, Duculot-Gembloux-Hatier Paris, 1964.

[36] Cette liste nous a souvent été demandée.

L’utilisation du lexique dynamique implique un modèle de langage dynamique pour les n-grammes. Un modèle n-grammes est précalculé à partir du lexique des 64 k-mots les plus fréquents ; les corpus textuels sont filtrés par ce lexique, les n-grammes manquants seront calculés sur ce corpus ou estimés par la méthode classique du *backing off*. Sur les corpus *du Monde*, les modèles de langage précalculés correspondent à plus de 10 millions de bigrammes (sur 250 milliards possibles), près de 40 millions de trigrammes (sur 125×10^{15} possibles), 3 000 biclasses (sur 5 500 possibles) et 60 000 triclasses (sur 400 000 possibles).

Les modèles de langage sont essentiels aux performances des systèmes de dictée vocale, l’exemple suivant permet de s’en convaincre. Dans l’ordre, nous donnons la référence lue (1), sa transcription phonétique (2), l’hypothèse acoustique de D-DAL96 obtenue sans la prise en compte de ses modèles de langage (3), sa transcription phonétique (4) et enfin l’hypothèse de D-DAL96 avec l’intégration de ses modèles de langage (5) avec sa transcription phonétique (6).

1. Louis Viannet poursuit une stratégie qui consiste à avoir deux fers au feu
2. l w i v j a n e p u r s ɥ i y n s t r a t e ʒ i k i k ɔ̃ s i s t a a v w a r d ø f ɛ r o f ø
3. louent Ui Viannet et Poor’s j’ sût in estrades thé gis tique consistait avouât hardes œufs air Rauf feux
4. l u w i v j a n e e p u r s ʒ s y i n ɛ s t r a d t e ʒ i t i k k ɔ̃ s i s t ɛ a v w a a r d œ f ɛ r r o f f ø
5. Louis Viannet et poursuit une stratégie qui consiste avoir deux fers au feu
6. l w i v j a n e e p u r s ɥ i y n s t r a t e ʒ i k i k ɔ̃ s i s t a v w a r d ø f ɛ r o f ø

Collaboration : Claude Montacié.

Publications liées : 2 articles dont 1 international [17] et 1 national [36].

Publication jointe : Recueil des publications page 5 [17].

2.1.3. Première campagne de dictée vocale en langue française

La première campagne d’évaluation de dictée vocale[37] en langue française a eu lieu du 3 février 1997 au 17 mars 1997. Deux catégories de systèmes de dictée vocale avaient été définies : les systèmes 20 k-mots et les systèmes sans « contrainte de vocabulaire ». Cinq laboratoires ont participé à ces tests : le CRIM, le CRIN, l’INRS, le LIMSI et le LIP6. Les trois premiers laboratoires ont concouru dans la première catégorie 20 k-mots, le LIMSI dans les deux catégories, et le LIP6 dans la dernière catégorie. Cette campagne a été financée et organisée par l’AUPELF-UREF. Grâce à ce financement, les ressources suivantes ont pu être distribuées : les corpus acoustiques BREF-TOTAL (corpus d’apprentissage) et ILORB1_DEV (corpus de développement) en octobre 1996, le corpus textuel des années 87-88 du journal *Le Monde* et les modèles de langage associés (bigrammes, trigrammes) à la même date, un lexique phonétisé des 20 k-mots les plus fréquents (années 87-88) en novembre 1996.

[37] J.-M. Dolmazon, F. Bimbot, G. Adda,, J.-C. Caerou, J. Zeiliger, M. Adda-Decker, « *Première campagne AUPELF d’évaluation des systèmes de Dictée Vocale, organisation et résultats* », à paraître.

Les systèmes[38] utilisent tous une recherche multi-passe, qui comprend au minimum la génération d'un graphe de mots avec un premier modèle de langage, suivi d'un décodage avec un modèle de langage plus élaboré. Le premier modèle de langage est un bigramme ou un bigramme/biclasse ; le second, un modèle trigramme ou trigramme/triclasse. Certains systèmes utilisent d'autres passes comme l'adaptation non supervisée, l'utilisation d'un modèle grammatical formel, la sélection d'un lexique dynamique. Deux algorithmes de décodage sont employés : l'algorithme de Viterbi-bloc ou l'algorithme *one pass*. Les modèles acoustiques sont basés sur des modèles de Markov d'ordre 1 ou d'ordre 2. Les unités acoustiques utilisées sont les biphones, les triphones ou les triphones dépendant de leur position dans le mot.

Pour les systèmes « sans contrainte de vocabulaire », la tâche de reconnaissance a porté sur 655 phrases (structurées en paragraphes), lues par vingt locuteurs, du journal *Le Monde* (extraites du mois de novembre 1996), ce qui correspond à deux heures de parole. Le système utilisé par le LIP6 était D-DAL96. Après la phase d'adjudication, sur les 655 phrases du test officiel, le taux de mots reconnus était de 69% avec un taux d'insertion de 8%. Les erreurs sont dues à des problèmes acoustiques, linguistiques ou grammaticaux.

Les résultats présentés ci-dessous sont représentatifs de la complexité de la tâche. On trouvera la référence de la phrase lue (REF), l'hypothèse de D-DAL (HYP), les erreurs de substitution (S), d'omission (O) et d'insertion (I), et les scores finaux.

Ex 1) Cause de l'erreur : dédoublement ou fusion de deux phonèmes identiques
REF: *Louis Viannet ** poursuit une stratégie qui consiste à avoir deux fers au feu*
HYP: *Louis Viannet et poursuit une stratégie qui consiste * avoir deux fers au feu*
Eval : I O
Scores : (#Correct 12 #Substitution 0 #Omission 1 #Insertion 1)

Ex 2) Cause de l'erreur : faute d'accord
REF: *les nationalistes eux mêmes disent leur inquiétude face à cette situation*
HYP: *les nationalistes eux mêmes disent leurs inquiétudes face à cette situation*
Eval : S S
Scores : (#Correct 9 #Substitution 2 #Omission 0 #Insertion 0)

Ex 3) Cause de l'erreur : poids trop faible du modèle de langage
REF: *le législateur n'a jamais envisagé un tel cas*
HYP: *le législateur n'a jamais envisagé un tel écart*
Eval : S
Scores : (#Correct 8 #Substitution 1 #Omission 0 #Insertion 0)

---

[38] G. Adda, M. Adda-Decker, J.-L. Gauvain, L. F. Lamel, « *Le système de dictée vocale du LIMSI pour l'évaluation AUPELF'97* », 1ères JST FRANCIL, Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la Langue de l'AUPELF-UREF, pp. 35-40, Avignon, 1997.

D. Fohr, J.-P. Haton, J.-F. Mari, K. Smaïli, I. Zitouni, « *Maud : un prototype de machine à dicter vocale* », 1ères JST FRANCIL, Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la Langue de l'AUPELF-UREF, pp. 25-30, Avignon, 1997.

C. Savariaux, A. Farhat, M. Héon, D. O'Shaughnessy, C. Lee, « *Nouvelles avancées en reconnaissance de la parole continue grand vocabulaire du français basées sur le système de reconnaissance de l'INRS-Télécommunications* », 1ères JST FRANCIL, Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la Langue de l'AUPELF-UREF, pp. 31-34, Avignon, 1997.

Ex 4) Cause de l'erreur : mot hors du vocabulaire de reconnaissance
REF: *son premier film en mille neuf cent cinquante huit après avoir rejoint ***** la Nikkatsu*
HYP: *son premier film en mille neuf cent cinquante six après avoir rejoint année quatre sous*
Eval : S I S S
Scores : (#Correct 11 #Substitution 3 #Omission 0 #Insertion 1)

Ex 5) Cause de l'erreur : l'emphase
REF: *selon Peter Kessler du Haut Commissariat des Nations Unies pour les réfugiés*
HYP: *selon Peter Kessler du haut commissariat des nations unies pour les réfugiés*
Eval : S S S S
Scores : (#Correct 8 #Substitution 4 #Omission 0 #Insertion 0)

Ex 6) Cause de l'erreur : erreurs acoustiques et linguistiques
REF: *il n' existe pas d'autres***** motif à leur assassinat*
HYP: *il * existait pas d'autres motifs alors à Saint Cyr*
Eval : O S I S S S
Scores : (#Correct 5 #Substitution 4 #Omission 1 #Insertion 1)

La mesure pour évaluer les performances du système est liée à la meilleure solution au sens de la suite de mots : c'est le calcul du pourcentage de reconnaissance des mots à l'aide des pourcentages d'identification, de substitution, d'insertion et d'omission. Cette évaluation fondée sur la meilleure hypothèse a le mérite d'être simple à implémenter mais il est à notre avis trop réducteur pour donner une bonne information de la capacité du système de dictée vocale. D'autres indices peuvent être utilisés tout en gardant les mots comme unités d'évaluation. L'évaluation pourrait porter sur la qualité du treillis de mots à N choix, N étant fixé à un certain pourcentage (e.g., 1%) de la taille du vocabulaire. Les indices de l'évaluation pourraient alors être les suivants : -le pourcentage de présence des mots de la phrase-test dans le treillis, -la profondeur moyenne et maximum d'apparition des mots de la phrase-test dans le treillis, -le pourcentage de mots de la phrase-test présents dans le treillis, -le rapport des probabilités entre la phrase-test et la meilleure phrase. Ces indices d'évaluation originaux seraient à calculer sur : l'acoustique pur, et l'acoustique contraint par un modèle de langage standard.

2.2. Systèmes hybrides

Nous avons développé ce que nous appelons deux systèmes hybrides : deux processus de décision inhabituels, indépendants de la théorie markovienne, sont étudiés dans le cadre de la reconnaissance par modèles de Markov cachés. Le premier processus de décision est le contrôle du décodage markovien par l'analyse du signal de test. Le contrôle temporel dans les modèles de Markov cachés a fait l'objet de la thèse de Claude Barras[39]. Le second consiste à utiliser la famille des règles de décision des plus proches voisins (ppv) comme alternative possible à l'estimation gaussienne des probabilités d'émission des observations par les états. Fabrice Lefèvre doit prochainement soutenir sa thèse sur cet axe de recherche.

[39] C. Barras, « *Reconnaissance de la parole continue : adaptation au locuteur et contrôle temporel dans le modèles de Markov cachés* », thèse de Docteur de l'Université Pierre et Marie Curie, 29 mai 1996.

2.2.1. Contrôle temporel dans la modélisation markovienne

L'une des faiblesses reconnue aux modèles de Markov cachés concerne la modélisation de la durée inhérente à leur topologie. Certains auteurs ont introduit des modèles de durée qui corrigent les probabilités des modèles phonétiques, par exemple avec une loi de Poisson ou une loi Gamma. Ces modélisations explicites proviennent de statistiques sur la durée des phonèmes et ne se réfèrent pas au signal acoustique à décoder. Un contrôle temporel des transitions phonétiques peut également être obtenu à partir de l'analyse du signal à décoder, les changements acoustiques permettant par exemple de guider l'enchaînement des modèles phonétiques au cours du décodage. Un contrôle des transitions phonétiques dans le système de décodage markovien a ainsi été couplé aux discontinuités du signal détectées par des méthodes de traitement de signal. La méthode des Ruptures de Modèles[40] est l'une des méthodes qui a été testée. Son application à la segmentation du signal de parole donne un nombre de ruptures supérieur au nombre de phonèmes à reconnaître, certaines ruptures correspondant *a priori* à des non-stationnarités infra-phonémiques. Cette segmentation engendrée par les ruptures ne peut donc être utilisée directement pour le contrôle du décodage. Cependant, l'étude statistique de la distribution des ruptures sur des bases de données de référence a montré une forte corrélation entre la position des ruptures et les instants de transition entre phonèmes référencés par l'étiquetage fin. Les ruptures détectées sont interprétées, dans le système de décodage, comme des instants de transition possibles entre les modèles représentant les unités phonétiques. Ainsi, lors de l'étape d'identification utilisant l'algorithme de Viterbi, la probabilité $p(M_j / M_i)$ de transition entre les modèles M_i et M_j dépend de la distance de la rupture la plus proche. Si cette distance est supérieure à un seuil fixé *a priori*, la probabilité de transition inter-modèles est pénalisée par un facteur multiplicatif α ($0 < \alpha < 1$). La valeur retenue pour α est celle qui optimise les résultats sur l'ensemble d'apprentissage. Ce protocole est indépendant du type de modèle de Markov utilisé. Dans les expériences menées par Claude Barras, le contrôle temporel est à l'origine d'une amélioration de 2% des performances du décodage acoustico-phonétique.

Collaborations : Claude Barras (thésard - Equipe Parole), Claude Montacié.

Publications liées : Thèse de l'Université Paris 6 de Claude Barras, 6 articles dont 1 ouvrage de synthèse [3], 2 internationaux [12, 16], 3 nationaux [28, 30, 31] et 1 workshop [41].

Publications jointes : Recueil des publications page 9 [3], page 13 [12], page 17 [16].

2.2.2. Système HMM-Kppv

Plusieurs méthodes ont été utilisées pour estimer la probabilité d'émission d'un vecteur acoustique par un état d'un modèle de Markov caché. Une des premières méthodes

[40] R. André-Obrecht, « *A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals* », IEEE Transaction on Acoustic, Speech and Signal Processing, vol. 36, n° 1, pp. 29-40, 1988.

employées a été fondée sur les techniques de quantification vectorielle. A chaque partition de cette quantification est associée une table des probabilités de tous les états. L'estimation d'une probabilité d'émission revient à la recherche de la partition associée au vecteur acoustique, puis à une recherche dans une table. L'avantage de cette méthode est son très faible coût de calcul permettant une implémentation temps réel. Son principal inconvénient est d'associer la même probabilité à tous les vecteurs acoustiques d'une partition, ce qui est irréaliste. La recherche d'une meilleure estimation de la probabilité d'émission a conduit à l'utilisation d'une somme de multigaussiennes. Cette méthode donne indiscutablement les meilleurs résultats, mais au prix d'un nombre toujours plus élevé de multigaussiennes (e.g., plusieurs centaines de milliers). En dehors des problèmes classiques d'apprenabilité (apprentissage par cœur ou modélisation), cette inflation des paramètres d'un modèle de Markov pose deux problèmes majeurs : -le premier est l'impossibilité d'interpréter cette énorme masse de données, de lui donner une cohérence pour pouvoir, par exemple, comprendre la raison d'une erreur, -le second est le coût de calcul nécessaire, toujours croissant pour des augmentations négligeables de performance. Cette augmentation des coûts provoque une course plus technologique que scientifique qui peut empêcher le développement d'autres modélisations peut-être bien moins performantes à l'origine mais, à notre avis, pleines d'avenir lorsqu'elles seront adaptées au cadre markovien.

Plusieurs modélisations alternatives ont été étudiées, comme d'autres familles de fonctions de densité de probabilité (e.g., laplaciennes), l'analyse linéaire ou les perceptrons multicouches. Dans le premier cas, nous retrouvons les mêmes problèmes que pour les multigaussiennes, avec un coût de calcul encore supérieur, tandis que les méthodes d'analyse linéaire (composantes principales ou méthodes discriminantes) ne sont en fait qu'un changement de l'espace de représentation. Le principal avantage des perceptrons multicouches est de pouvoir approximer n'importe quelle fonction continue : la précision de cette approximation est fonction du nombre de nœuds sur la couche cachée. Même si ce nombre de nœuds peut être élevé (e.g., plus de 1000 nœuds), le nombre de paramètres du réseau de neurones reste cependant, pour des performances égales[41], plus faible que ceux d'un système à base de multigaussiennes. Un autre avantage des perceptrons multicouches est leur caractère discriminant : cependant l'utilisation des techniques GPD (*Generalized Probabilistic Descent*) permet d'obtenir cette propriété avec pratiquement n'importe quel type de modélisation. Des méthodes spécifiques d'apprentissage ont dû être développées pour utiliser au mieux le perceptron multicouches comme estimateur de la probabilité d'émission d'un vecteur acoustique. On peut citer la méthode REMAP[42] (*Recursive Estimation and Maximization of A Posteriori probabilities*) dérivée de la méthode MAP. Comme dans toutes les méthodes à base de réseaux de neurones, la phase d'apprentissage peut être extrêmement longue.

[41] H. Bourlard, C. J. Wellekens, « *Links Between Markov Models and Multilayer Perceptrons* », IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 12, n° 12, pp. 1167-1178, 1990.

T. Robinson, « *An Application of Recurrent Nets to Phone Probability Estimation* », IEEE Transaction on neural Networks, vol. 5, n° 3, 1994.

[42] H. Bourlard, Y. Konig, N. Morgan, « *REMAP : Recursive Estimation and Maximization of A Posteriori Probabilities in Connectionist Speech Recognition* », Eurospeech, European Conference of Speech Communication and Technology, pp. 1663-1666, Madrid, 1995.

Cependant, d'autres d'estimateurs eux aussi très efficaces peuvent être employés. Il s'agit des estimateurs dits « non-paramétriques » qui ne sont pas basés sur un modèle. Ces estimateurs utilisent les méthodes de Parzen ou la règle des K plus proches voisins (K-ppv). Dans cette approche, la probabilité d'émission d'un vecteur acoustique x est fonction des vecteurs acoustiques d'apprentissage proches de x. Un tel estimateur présente des propriétés théoriques très intéressantes dans le cas où le nombre de vecteurs d'apprentissage est suffisamment grand : -estimation d'une densité de probabilité quelconque, -pas de problème de convergence à l'apprentissage. La probabilité d'erreur de cet estimateur décroît quand K augmente. De plus, P.A. Devijver et J. Kittler[43] ont démontré qu'elle était bornée par $\left(1+2/\sqrt{\pi K/2}\right)E^*$, où E^* est l'erreur théorique minimale (i.e., risque de Bayes). Par exemple, la probabilité d'erreur de la règle de décision d'un estimateur 50-ppv est inférieure à 120% de l'erreur minimale. Tous ces avantages ont naturellement une contrepartie : la nécessité d'un ensemble d'apprentissage de grande taille, qui entraîne des difficultés pour déterminer les vecteurs acoustiques d'apprentissage proches d'une observation x. Avec Claude Montacié, nous avons développé un algorithme rapide de K-ppv combinant les heuristiques de Kittler-Richetin[44], de Friedman[45] et de Fukunaga[46]. Le temps de calcul des 50-ppv d'un vecteur acoustique (12 MFCC + Energie) est inférieur à 100 ms (Pentium 450 MHz) en prenant plus d'un million de vecteurs d'apprentissage. Le développement d'un système hybride HMM-Kppv fait actuellement l'objet de la thèse de Fabrice Lefèvre. Les premiers résultats sur la base TIMIT sont encourageants mais ils montrent, comme dans le cas des réseaux de neurones, la nécessité d'adapter les algorithmes d'apprentissage à un nouvel estimateur. Fabrice Lefèvre essaye actuellement de déterminer pourquoi le système hybride HMM-Kppv tire moins bien parti des delta-paramètres que les systèmes HMM classiques, pour lesquels l'augmentation de performance est de l'ordre de 10%. Nous mettrons prochainement à la disposition de la communauté, les fichiers des 50-ppv déjà calculés sur la base de données TIMIT.

Collaborations : Fabrice Lefèvre (thésard - Equipe Parole), Claude Montacié.

Publications liées : 3 articles dont 1 ouvrage de synthèse [5] et 2 internationaux [20, 22].

Publication jointe : Recueil des publications page 21 [20].

[43] P. A. Devijver, J. Kittler, *Pattern Recognition. A statistical approach*, Prentice Hall International, 1982.

[44] J. Kittler, « *A Method for Determining k-Nearest Neighbors* », Kibernetes, vol. 7, pp. 313-315, 1978.

M. Richetin, G. Rives, M. Naranjo, « *Algorithme rapide pour la détermination des k plus proches voisins* », RAIRO Informatique, vol. 14, n° 4, pp. 369-378, 1980.

[45] J. H. Friedman, J. L. Bentley, R. A. Finkel, « *An Algorithm for Finding Nearest Neighbors* », IEEE Transaction on Computer, vol. 24, pp.1000-1006, 1975.

[46] K. Fukunaga, P. M. Narendra, « *A Branch and Bound Algorithm for Computing k-Nearest Neighbors* », IEEE Transaction on Computers, vol. 24, pp. 750-753, 1975.

Introduction au chapitre 3

En reconnaissance des formes, l'un des problèmes tient au choix même de l'espace de représentation des formes, car il n'existe aucune démarche algorithmique qui permette de le formaliser. C'est l'**adéquation** du couple **espace de représentation** et **processus de décision** qui conditionne les performances de la reconnaissance. Actuellement, les systèmes markoviens ou les systèmes hybrides sont « encore » incontournables pour la reconnaissance de la parole continue grand vocabulaire, indépendamment du locuteur.

Avant l'avènement des systèmes auto-organisateurs, les recherches en reconnaissance de la parole trouvent principalement leur application dans la reconnaissance de mots isolés ou de mots connectés, et dans le décodage acoustico-phonétique. Le « grand » processus de décision pour un continuum acoustique est alors la comparaison dynamique, pour laquelle d'innombrables travaux ont porté sur les optimisations du temps de calcul. Les espaces métriques, couples de l'espace de représentation et de la mesure de dissimilarité, y sont indispensables et trouvent de plus leur utilité dans des applications aussi variées que -la segmentation, -la quantification vectorielle, -le codage par compression d'états stables, -le test de qualité objectif des vocodeurs, -la reconnaissance du locuteur. Ainsi, l'analyse du signal en un espace de représentation, et le choix ou le développement d'une métrique adaptée, ont été étudiés, évalués et comparés par de nombreux laboratoires. Des orientations de la recherche ont été guidées par les rapports d'analyse de grands projets nationaux ou internationaux comme ARPA, incitant par exemple à développer, dans les systèmes de reconnaissance, l'intégration de la perception ou de la prosodie. Les approches novatrices, entre autres de l'intelligence artificielle, sont assez systématiquement expérimentées en parole. On peut citer deux approches remarquées par les sessions qui leur ont été consacrées dans les congrès de parole, -les systèmes experts et -les réseaux de neurones. Si la modélisation markovienne a déplacé les centres d'intérêt, elle n'a cependant pas résolu tous les problèmes, des problématiques anciennes demeurent.

Outre les thèmes spécifiques à sa modélisation, comme -la topologie des modèles, -les modèles de durée, -la modélisation des densités d'émission, des thèmes tels que -l'apprentissage des modèles acoustiques, -l'adaptation au locuteur des modèles acoustiques, sont encore problématiques même dans la reconnaissance markovienne. Et il est un choix que l'on remet peu en question lorsque l'on s'attelle à ces travaux, il s'agit du choix de l'espace de représentation et de son analyse. Ou du moins, est-ce surtout dans le but d'avoir une représentation robuste au bruit que l'on reconsidère ce choix.

Nos travaux n'ont pas échappé à « l'air du temps », et c'est dans ce contexte que l'on peut situer les travaux résumés dans cette partie de façon chronologique, des plus anciens aux plus récents. Ces travaux concernent -l'**évolution de l'évaluation** des composants de systèmes, -le développement d'un **espace métrique perceptivement consistant** et ses modes d'**évaluation**, -une approche **numérique-symbolique** de la reconnaissance, -la **reconnaissance du locuteur** qui est un autre thème de recherche de la reconnaissance, et enfin -l'analyse par **multi-résolution temporelle** permettant une rétroaction sur tout processus de décision.

Chapitre 3

Représentation et décision

Jusque dans les années 80, de nombreux travaux ont été consacrés aux espaces métriques. L'intérêt des espaces métriques, qu'ils soient associés aux normes L_p ou spécifiques à la parole, s'explique par -leur utilité en reconnaissance analytique où il s'agit d'associer un point de R^p à un prototype de classe, et -leur nécessité en reconnaissance globale avec le succès de l'algorithme de comparaison dynamique qui calcule la distance optimale entre deux continuums de parole. De nombreuses études ont été menées pour leur évaluation et leur comparaison dans l'espoir de voir « émerger » un espace métrique dans le cadre d'une application considérée. Néanmoins, on ne peut que constater la difficulté de leur comparaison, comme en témoigne la publication d'études comparatives parfois contradictoires et pour la plupart inexploitables. Deux solutions ont été proposées pour résoudre le problème posé par les études comparatives. Il s'agit en fait de **deux paradigmes d'évaluation**, plus généraux, des **composants de systèmes** de reconnaissance et qui sont discutés dans ce chapitre. Après ma thèse en 1987, mes travaux sur le développement et l'évaluation d'une **mesure de dissimilarité** perceptivement consistante ont été poursuivis. Des **évaluations qualitatives** des espaces métriques sont menées dans les domaines de la perception et de la phonétique.

Un facteur de complexité de la parole est sa grande variabilité inter-locuteurs mais également intra-locuteur. Dans bien des approches, comme la reconnaissance phonétique et le système expert, l'invariance a été recherchée. En reconnaissance phonétique, la problématique est -l'estimation des indices caractérisant les traits acoustiques ou articulatoires, -les règles de décision reliant les indices et les traits et -la structure des règles de décision. Dans le système expert, la difficulté est d'élaborer voire d'organiser les règles d'une base de connaissance permettant de déduire par inférence l'invariance symbolique. L'approche numérique consiste à passer au niveau symbolique de la reconnaissance, à partir d'observations et de processus de décision numériques. Contrairement à celle-ci, notre approche **numérique-symbolique**, nous a conduits à une transformation numérique-symbolique des observations avant d'appliquer un processus de décision symbolique. Les observations du signal vocal sont représentées dans un formalisme logique adapté à un système d'apprentissage symbolique. Ce système est capable de détecter automatiquement les régularités logiques d'un ensemble d'apprentissage, et d'engendrer le système de règles reflétant ces régularités. C'est sur ce système de règles qu'est alors fondé le processus de décision.

Si en reconnaissance de la parole, la problématique est liée à l'invariance des unités linguistiques relativement à la variabilité intra-locuteur et inter-locuteurs, celle de la **reconnaissance du locuteur** s'intéresse au problème dual, c'est-à-dire à reconnaître des caractéristiques du locuteur indépendamment des unités linguistiques employées à l'élocution.

A l'analyse du signal de parole, des artefacts peuvent apparaître, et s'il y a plusieurs raisons à cela, la non-stationnarité du signal en est vraisemblablement une. *A priori*, la présence d'artefacts ne facilite pas le processus de décision et on peut même penser qu'ils en sont un biais. C'est dans le but de détecter les non-stationnarités du signal vocal que nous avons introduit la **multi-résolution temporelle** qui peut être généralisée à toute technique d'analyse, et exploitée dans tout processus de décision.

3.1. Paradigmes d'évaluation des composants de système

Le mode d'évaluation des composants de système de reconnaissance le plus fréquemment utilisé repose sur l'évaluation de performance sur une tâche spécifiée et sur une base de données considérée. L'évaluation est alors quantitative, les performances de reconnaissance (taux d'erreur ou de reconnaissance) sont les mesures de l'évaluation. Un exemple de composant de système est l'espace métrique, c'est un composant fondamental aux performances des systèmes, au moins par la représentation qu'il donne du signal de parole. L'évaluation de ce composant est très importante car « tout est choix » dans un espace métrique, et la justification des choix relève souvent de l'expérience.

- **Choix de la représentation**
  Le choix de l'espace de représentation comprend -les paramètres d'analyse du signal qui peuvent être fondés sur la théorie du signal (e.g., FFT[47]) ou sur une modélisation (e.g., LPC[48]), -le prétraitement du signal, -la durée des fenêtres d'observation, -le fenêtrage, -la dimension considérée de l'espace de représentation. Dans le cas de l'utilisation des connaissances en perception, leur mode d'intégration -en modélisation (e.g., PLP[49]) ou -en échelle d'intégration fréquentielle (Klatt[50], Zwicker[51]).

[47] FFT (Fast Fourier Transform), transformée de Fourier discrète.
[48] LPC (Linear Predictive Coding), prédiction linéaire.
[49] PLP (Perceptive Linear Prediction), prédiction linéaire perceptive.
[50] D. H. Klatt, « *A Digital Filter Bank for Spectral Matching* », IEEE International Conference on Acoustic, Speech and Signal Processing, pp. 573-576, 1979.

- **Choix de la métrique**
  L'espace de représentation est muni d'une distance ou d'un indice de distance. Dans les deux cas, les options de choix sont multiples. Si la distance euclidienne est classiquement utilisée pour les vecteurs de représentation, c'est bien plus pour une raison de rapidité de calcul que d'adéquation à la représentation. En perception, par exemple, l'inégalité triangulaire n'a pas de sens dans la perception des entités sonores.

Si de nombreuses publications ont été consacrées à l'étude comparative des espaces métriques, elles n'ont pourtant pas apporté de réponse au choix du « meilleur » espace métrique. En effet, d'une étude à l'autre, -la tâche varie (reconnaissance analytique ou globale de la parole, reconnaissance du locuteur), -les bases de données sont développées le plus souvent par les laboratoires eux-mêmes ; elles ont une complexité intrinsèque inconnue pour les lecteurs, et leur influence n'est *a priori* pas quantifiable sur les performances de reconnaissance. Du fait des études comparatives, les efforts de programmation, nécessaires à l'implémentation d'espaces métriques souvent complexes, se trouvent dupliqués dans de nombreux laboratoires. Les choix qui y sont faits ne sont pas toujours conformes à la théorie ou aux choix des auteurs. Les conditions d'analyse (e.g., durée des fenêtres, type du fenêtrage) et la dimension de l'espace de représentation sont autant de paramètres qui peuvent biaiser les évaluations. Pour ces raisons, on trouve dans la littérature des études comparatives contradictoires. Ainsi, nombreux sont les paramètres qui, lorsqu'ils diffèrent, rendent difficile l'appréciation des diverses études comparatives. Dans les années 80, deux solutions ont été proposées pour pallier ce problème.

Une première proposition consiste en la définition d'un **système de référence** (J. K. Baker[52], G. Chollet & al.[53]) définissant une mesure « étalon » commune à tous les chercheurs. Le système de référence est défini comme un système de reconnaissance reflétant l'état de l'art, relativement aux algorithmes employés et aux performances de reconnaissance. Il doit inclure les représentations du signal de parole usuelles ainsi que les distances interspectrales usuelles. Plusieurs systèmes de référence, utilisant la comparaison dynamique, ont été proposés pour la **reconnaissance** de **mots isolés** ou de **mots connectés** (parole continue à vocabulaire limité) (SAM[54], EAGLES[55]). Le but de tels systèmes est également d'évaluer les bases de données. Le système de référence est un appareil de mesure qui étalonne leur complexité intrinsèque, et dont l'intérêt pourrait être de prédire les performances d'un système sur une autre base de données. L'évaluation par un système de référence n'a pourtant jamais été expérimentée à

---

[51] E. Zwicker, R. Feldtkeller, *Psychoacoustique. L'oreille récepteur d'information*, Traduction française par C. Sorin, Masson, Paris, 1981.

[52] J. Baker, « *Assessment of using a Reference System for the Evaluation of Speech Recognizers and Databases* », report to National Bureau of Standards, 1982.

[53] G. Chollet, C. Gagnoulet, « *On the Evaluation of Speech Recognizers and Databases using a Reference System* », IEEE IEEE International Conference on Acoustic, Speech and Signal Processing, pp. 2026-2028, 1982.

[54] SAM 84-85.

[55] Eagles.

l'échelle nationale ou internationale. Le principe de sa propagation permet peut-être d'expliquer cet état de fait. Dans le cas de sa seule spécification, le système doit être simple à implémenter pour que les laboratoires l'utilisent effectivement. La duplication de code est alors inévitable, et cette perte d'efficacité favorise le principe de sa distribution, et par conséquent l'existence d'une infrastructure chargée de sa maintenance. Pour être au niveau de l'état de l'art, une maintenance logicielle est nécessaire compte tenu de l'évolution des techniques, et chaque version doit être suivie de la mise à jour des étalonnages des diverses bases de données antérieurement traitées. Pour la reconnaissance de la parole continue à grand vocabulaire, si l'on considère la technologie actuelle et la complexité des systèmes de l'état de l'art, envisager la maintenance d'un tel système devient irréaliste.

La deuxième solution consiste en la **diffusion** de **larges bases de données** considérées comme des bases de référence. Avec cette solution, c'est la qualité de la conception de la base de données qui permet effectivement à un paradigme d'évaluation de se développer. La base de données DARPA-TIMIT[56] en est un parfait exemple : elle est annotée jusqu'à l'étiquetage fin, et contrôlée à l'écoute, de façon à aider au mieux l'apprentissage et l'évaluation de tests. Les corpus d'apprentissage et de test sont spécifiés dans leur version réduite ou complète. Si l'on revient au principe des « données communes », on peut remarquer que les chercheurs ont trouvé, dans le **décodage acoustico-phonétique** (DAP), un consensus sur la « tâche commune ». Le DAP, passage du signal acoustique en un treillis de phonèmes, est sur ces bases de données l'une des tâches de reconnaissance usuelles. Les potentiels et les contraintes des systèmes de reconnaissance classifiés selon leur approche lexicale ou sub-lexicale expliquent l'intérêt des chercheurs pour le DAP. L'approche lexicale est fondée sur la reconnaissance acoustique de mots ou de groupes de mots, et nécessite l'apprentissage acoustique des entités lexicales qui la contraignent à un vocabulaire de reconnaissance limité. A l'opposé, l'approche sub-lexicale présente l'intérêt d'un vocabulaire de reconnaissance théoriquement illimité ; elle est fondée sur une reconnaissance phonétique, nécessitant un apprentissage acoustique des unités phonétiques. Ainsi, pour des systèmes de reconnaissance à grand vocabulaire, on admettait obligatoire l'étape du DAP caractérisée par la localisation et l'identification des phonèmes. Le DAP est devenu lui-même un sujet de recherche pour lequel tous les composants de systèmes ont pu être testés pour améliorer ses performances. Les chercheurs se sont organisés pour comparer leurs travaux ; par exemple, les regroupements allophoniques ont fait l'objet d'un consensus pour l'évaluation quantitative du DAP sur TIMIT. Si les performances du DAP ont peu à peu augmenté, elles ont actuellement dépassé un taux de 70% sur TIMIT ; le taux minimum de 80% estimé[57] pour espérer une bonne performance de reconnaissance n'est pas encore atteint. Les évaluations de composants sur les performances du DAP ne semblent toutefois pas être le meilleur paradigme. En effet, des améliorations de performance du DAP ne garantissent pas, surtout si elles sont mineures, une amélioration des performances d'un système de reconnaissance

[56] DARPA-TIMIT, base de données anglo-américaine de parole continue de 630 locuteurs représentant les huit dialectes américains.

[57] J.-S. Liénard, J. Mariani, G. Renard, *« Intelligibilité de phrases phonétiquement altérées : application à la transmission de la parole »*, 9th ICA, I-95, 1977.

sophistiqué pour lequel, par exemple, des connaissances linguistiques sont utilisées. Peaufiner le DAP sur une base de données, à travers les composants qu'il utilise, a moins d'intérêt que travailler sur sa robustesse à la grande variété des bases de données audio. Une illustration de la « fragilité » des systèmes est le passage de la base de données TIMIT à NTIMIT[58], base de données pour laquelle on observe un effondrement des performances de reconnaissance, et entre autres de celles du DAP.

3.2. Espace métrique : développement et évaluations

L'une de mes motivations dans le développement d'une mesure de dissimilarité perceptivement consistante s'insère dans les perspectives de recherche de l'époque : l'intégration de la perception. On peut remarquer, à cet effet, le regain d'intérêt pour la perception dans le codage[59]. La norme MPEG couche 3, concernant le codage du signal audio pour un stockage et une restitution répondant aux besoins du multimédia, en est un exemple. Son codage comporte une composante psycho-acoustique, qui permet le codage transparent d'un signal Haute-Fidélité stéréo.

L'intégration de connaissances sur la perception sonore peut-elle constituer un apport aux performances de reconnaissance des systèmes ? Si une mesure de dissimilarité s'avérait perceptivement consistante, et même si elle ne donnait aucune amélioration en reconnaissance, elle présenterait déjà l'intérêt de pouvoir être utilisée comme une mesure subjective en synthèse. Mes travaux sur la mesure par appariement de pics spectraux (APS), fondée sur des critères perceptifs, ont été poursuivis après ma thèse dans le but d'optimiser le coût de calcul et d'en évaluer le comportement relativement à des connaissances phonétiques. Des études comparatives d'espaces métriques ont été menées sur ce dernier principe d'évaluation qualitative.

3.2.1. Perception et espace métrique en reconnaissance

Les recherches en synthèse de Xavier Rodet, mon directeur de thèse, ont influencé le choix d'une représentation des observations fondée sur l'enveloppe spectrale[60]. Plus précisément, l'enveloppe spectrale est représentée par ses maxima/pics spectraux dont on sait toute l'importance en production et en perception.

Un spectre à court-terme S est alors représenté par l'ensemble de ses pics spectraux $\{P_k\}_{(k=1,..,K)}$ caractérisés, par analogie aux formants, par leur fréquence centrale f_k (Hz),

[58] NTIMIT, version bruitée de TIMIT par simulation de lignes téléphoniques courtes et longues distances.

[59] M. P. Guimaraes, N. Moreau, M. Bonnet, « *Perceptual Filter Comparisons for Wideband and FM Bandwith Audio Coder* », Eurospeech, European Conference of Speech Communication and Technology, pp. 1503-1506, Rhodes, Grèce, 1997.

[60] La synthèse par morphing des enveloppes spectrales de deux voies, l'une masculine et l'autre féminine, est à l'origine de la synthèse de la voie de Castrat pour la bande audio du film « *Farinelli* » de Gérard Corbiau. Cette synthèse a été réalisée par l'équipe d'Analyse et Synthèse de l'IRCAM dirigée par Xavier Rodet.

leur largeur de bande l_k (Hz) et leur amplitude a_k (dB). Pour paramétriser le signal de parole en cet espace de représentation, j'ai développé un algorithme pour détecter directement les paramètres des pics spectraux sur l'enveloppe LPC logarithmique.

$$S = \{P_k (f_k, l_k, a_k)\}_{(k=1,..,K)}$$

La mesure de dissimilarité APS procède par appariement optimal des pics spectraux des formes spectrales à comparer. Cet appariement est fondé sur une **mesure de distorsion locale inter-pics** $\delta(P_i, P_j)$ que l'on peut interpréter comme une déformation spectrale locale, et où sont intégrées les **connaissances perceptives**.

$$\delta(P_i, P_j) = \omega_f(f_i).e_f(P_i, P_j) + \omega_l(f_i).e_l(P_i, P_j) + \omega_a(f_i).e_a(P_i, P_j)$$

$$\text{où,}\quad e_f(P_i,P_j) = \frac{|f_i\text{-}f_j|}{f_i+f_j}\ ;\quad e_l(P_i,P_j) = \frac{|f_i\text{-}f_j|}{f_i+f_j}\ ;\quad e_a(P_i,P_j) = |a_i\text{-}a_j|$$

et où $\omega_f(f_i)$, $\omega_l(f_i)$ et $\omega_a(f_i)$ sont des coefficients de pondération de la distorsion observée respectivement -en fréquence, -en largeur de bande et -en amplitude. La mesure est calculée à partir de la contribution à la distorsion spectrale des trois paramètres, estimée à partir de leur écart relatif ou absolu, et pondérée par des valeurs reflétant leur importance en perception. L'une des difficultés majeures est évidemment la quantification de ces pondérations. Les **seuils différentiels perceptifs** (*sdp*) des formants sont les principales connaissances de la perception prises en compte. Les *sdp* sont estimés pour les trois paramètres (*f*, *l*, *a*) d'après des expériences en psycho-acoustique[61] sur la perception de voyelles synthétiques. Ces seuils sont relatifs aux classes des formants, et bien que leur connaissance soit incomplète, les divers *sdp* ont été extrapolés à des valeurs discrètes correspondant aux quatre premiers formants. Ces pondérations discrètes ont ensuite été normalisées de façon à ce qu'une variation égale au *sdp*, qu'elle soit -en fréquence, -en largeur de bande ou -en amplitude, apporte la même contribution à la mesure inter-pics. La version originelle d'APS présentait l'inconvénient d'une classification des pics spectraux en formants pour le calcul d'une distorsion locale. Pour pallier la pondération discrète et optimiser le calcul d'APS, des tables de pondération fréquentielle sont utilisées. Elles sont estimées à partir des valeurs discrètes $\{\Omega_f(F_i), \Omega_l(F_i), \Omega_a(F_i)\}_{(i=1,..,4)}$ pour les trois paramètres ($\{\omega_f(f_i)\}$, $\{\omega_l(f_i)\}$, $\{\omega_a(f_i)\}$) en prenant en compte les statistiques des formants d'un système vocalique de référence du français.

Il n'y a pas de bi-univocité entre les formants (resp., antiformants) et les maxima (resp., minima) détectés sur une représentation spectrale, et entre autres sur l'enveloppe spectrale LPC. Dans une séquence d'observations, le suivi de la trajectoire des pics spectraux met en évidence des « apparitions » ou des « disparitions » de pics spectraux qui s'expliquent soit -par des phénomènes physiques d'échange de résonateurs avant/arrière et de points focaux, soit -par des artefacts de mesure. Et d'une manière générale, les spectres à comparer n'ont pas le même nombre de maxima spectraux. Pour la **mesure de distorsion globale inter-spectres** $D_{APS}(S^R, S^T)$, il faut résoudre le problème des appariements des pics spectraux et décider des pénalisations à donner aux

[61] J. L. Flanagan, *Speech analysis, synthesis and perception*, Springer-Verlag, Berlin, 1972.

différents types de distorsion globale. Nos connaissances sur la perception -des phénomènes d'intégration à l'échelle fréquentielle (Bark/Mel), -de la théorie de l'intégration à 3,5 Barks qui définit une distance critique entre les formants pour leur perception, ou -de la théorie du F'_2 qui estime un centre de gravité pour la perception des formants au-delà du premier, nous ont conduits à choisir un principe d'appariement optimal des spectres considérés. La distance globale est calculée par comparaison dynamique des spectres avec les contraintes locales de Sakoe et Chiba pour lesquelles les pondérations ont été adaptées à l'asynchronie fréquentielle des maxima spectraux.

La proposition d'une nouvelle mesure de dissimilarité ne peut se concevoir sans **évaluation** ni sans **étude comparative**. Des études, spécifiques à la mesure APS, ont été menées -quant à la validité des pondérations et -quant à leur optimisation, sur la base d'une **évaluation quantitative**. Cette évaluation a porté sur la **reconnaissance** des **noyaux vocaliques** d'un corpus de parole continue. Dans une étude comparative incluant douze protocoles de reconnaissance par mesure de dissimilarité, les scores d'identification d'**APS** se sont avérés significativement plus élevés que ceux obtenus par deux autres mesures classiques (**LLR**[62] et **MFCC**[63]) choisies pour leurs performances reconnues. Les noyaux vocaliques, dont l'identification est fiable à plus de 95%, ont permis d'amorcer l'identification de cycles vocaliques /VCV/ par comparaison dynamique. L'identification des continuums est fondée sur -une analyse pseudo-synchrone à la période fondamentale, et -une comparaison dynamique avec relâche aux frontières sur les deux axes pour le meilleur alignement des débuts et fins de cycles vocaliques. Si, dans le domaine de la reconnaissance, APS a montré une qualité certaine, l'analyse de son comportement en phonétique et en perception peut conduire à des améliorations, ne serait-ce que par la mise en évidence de comportements inadéquats relativement aux connaissances de ces domaines.

Collaborations : Xavier Rodet

Publications liées : ma thèse de l'Université Paris 6 [45], 3 articles nationaux [24, 25, 26], 1 workshop international [37], 2 workshops nationaux [38, 39].

3.2.2. Evaluations qualitatives

Dans le cas d'APS, une évaluation de la consistance perceptive de la mesure s'impose. L'évaluation, décrite ici, a été possible grâce à une étude de psycho-acoustique de D. Tuffelli et H. Ye de l'ICP[64]. Je leur suis très reconnaissante d'avoir intégré APS dans cette étude sur la cohérence entre les mesures de dissimilarité et la perception humaine de sons synthétiques.

- L'évaluation est fondée sur la perception catégorielle des sons. Entre deux références, deux voyelles dans le cas de cette étude, onze sons intermédiaires sont

[62] LLR (*Log Likelihood Ratio*), distance spécifique à la modélisation LPC.
[63] MFCC (*Mel Frequency Cepstrum Coefficients*), coefficients cepstraux en échelle Mel, munis de la distance euclidienne.
[64] Institut de la Communication Parlée de Grenoble.

synthétisés par pas successifs d'interpolation linéaire des formants. Lors des tests psycho-acoustiques, il est demandé aux auditeurs de catégoriser les sons intermédiaires par l'une des références, en leur donnant auparavant -les références phonétiques (test X) ou -les références acoustiques (test ABX). Pour chaque son intermédiaire, une mesure du jugement catégoriel humain est estimée par la différence du nombre des auditeurs jugeant de l'appartenance du son aux références. Le jugement catégoriel simulé par les distances de dissimilarité est simplement calculé comme la différence des distances du son aux références. Pour chaque couple de références, à partir des statistiques (jugement humain et jugement simulé) sur les sons intermédiaires, on calcule le coefficient de corrélation. Douze paires de voyelles du système vocalique français sont considérées ; le coefficient de corrélation moyen est alors choisi comme critère d'évaluation de la cohérence entre la mesure de dissimilarité et la perception humaine. Six mesures de dissimilarité ont été comparées suivant ce critère. Dans le cas d'APS, le coefficient moyen calculé est de 0,89 (le meilleur de l'étude comparative) dans les tests X, et de 0,90 dans les tests ABX, ce qui montre une « bonne » corrélation linéaire à la perception humaine. Dans les tests ABX, les plus « fins » au sens de l'étude perceptive, les références pour lesquelles on constate le coefficient le plus faible sont les couples de voyelles (/ɛ/, /œ/ ; r=0.77) et (/e/, /ø/ ; r=0,81). Pour ces voyelles, la différence porte essentiellement sur le deuxième formant (différence respective de 400 Hz et 600 Hz) et semble indiquer que les pondérations dans la zone fréquentielle du second formant sont insuffisantes. Ces résultats ne se retrouvent pas sur les tests X où la catégorisation est phonétique.

Collaborations : Denis Tuffelli et Hayan Ye de l'Institut de la Communication Parlée.

Travaux liés : 1 article[65] international de Denis Tuffelli et Hayan Ye où la mesure de dissimilarité APS est référencée.

Mes connaissances en phonétique ont été extrêmement enrichies par mes collaborations avec l'ICP en les personnes de L.-J. Boë, D. Tuffelli et H. Ye, et je les en remercie. L.-J. Boë est à l'origine du protocole d'évaluation du comportement des mesures de dissimilarité en phonétique. Le principe repose sur les techniques d'analyse multi-dimensionnelle qui permettent de reconstruire un espace de représentation (à faible dimension) à partir d'un tableau de distances ou de dissimilarités entre objets. C'est en comparant cet espace « reconstruit » à des espaces de référence bien connus que l'on peut obtenir une évaluation qualitative. Pour cette étude, c'est dans les domaines -de l'acoustique, -de l'articulatoire, et -du perceptif, que sont considérés les espaces de référence.

- Une ***première étude*** du comportement des mesures de dissimilarité est fondée sur des critères de **bonne représentation**. Les mesures sont évaluées sur leur capacité à **respecter** la **structure du système vocalique**. On connaît toute l'importance en phonétique des deux premiers formants (F1 et F2) caractéristiques de la

[65] D. Tuffelli, H. Ye, « *Distortion Measure Evaluation using Synthetic Sounds and Human Perception* », 12th ICA, Montreal Symposium on Speech Recognition, pp. 30-33, 1986.

production articulatoire. C'est dans ce système, le plus discriminant, d'axes fréquentiels privilégiés (plan F1/F2) que l'on représente usuellement les systèmes vocaliques. Les travaux de synthèse des voyelles à deux formants confirment l'importance de ces deux paramètres en perception. Considérant des voyelles naturelles ou des stimuli de référence du système vocalique étudié, on peut calculer la matrice des dissimilarités inter-voyelles à partir de laquelle les techniques d'analyse multidimensionnelle sont appliquées. Ce que l'on attend *a priori*, c'est que l'espace reconstruit sur deux dimensions conserve les relations structurelles entre les voyelles de la représentation F1/F2. Pour la représentation des mesures de dissimilarité, la qualité est évaluée par la sommation des scores attribués par rapport au respect des relations structurelles du système vocalique. Les scores (-1 : mauvais, 0 : médiocre, 1 : bon) sont attribués suivant le respect des oppositions des voyelles -antérieur/postérieur, -ouvert/fermé, -labialisé/non labialisé, -extrême/non extrême. Sur ce mode d'évaluation, six mesures de dissimilarité originelles ont été comparées avec, quand la mesure le permet, leurs variantes en représentation (FFT ou LPC) et en intégration perceptive (Klatt ou Zwicker) : soit au total 17 dissimilarités. Les références utilisées proviennent -des stimuli synthétiques de onze voyelles de référence du système vocalique français, et -de stimuli de parole naturelle de la langue africaine Bejonde. L'algorithme de Kruskal, implémenté à l'ICP, est choisi pour l'analyse multidimensionnelle. Dans cette étude, APS obtient le meilleur respect structurel du système vocalique du français et de la langue africaine Bejonde dont le système est caractérisé par sa répartition vocalique équilibrée et deux voyelles centrales très proches. Les projections ont également été comparées à l'espace perceptif[66] issu de l'analyse multidimensionnelle INDSCAL de données perceptives.

- La ***deuxième étude*** est fondée sur le **pouvoir de discrimination** des mesures de dissimilarité, évalué au **voisinage des points focaux** pour les voyelles [i, a, u, y]. Ces points focaux correspondent à de fortes perturbations des formants (en fréquence, en amplitude et en largeur de bande) F1/F2 ou F2/F3, dues à l'échange des affiliations des cavités avant/arrière, et que l'on peut observer lorsque le lieu de constriction varie dans les limites autorisées par l'articulatoire. Pour chacune de ces 4 voyelles, 11 stimuli synthétiques sont engendrés au voisinage des points focaux. Le critère retenu pour évaluer la discrimination est le rapport des distances inter-classes et intra-classe calculé par voyelle. Ce rapport peut être calculé directement à partir des mesures de dissimilarité, ou bien à partir de l'espace reconstruit par Kruskal ; plus il est élevé, meilleur est le pouvoir discriminant. L'étude comparative des 17 dissimilarités conclut sur une meilleure discrimination des distances dérivées de la LPC relativement à leur homologue FFT, et à l'efficacité des MFCC et d'APS.

Collaborations : Louis-Jean Boë, Denis Tuffelli, Hayan Ye de l'ICP.

Publications liées : 2 articles dont 1 international [6] et 1 national [27].

Publication jointe : Recueil des publications page 33 [27].

[66] F. Lonchamp, « *Recherche sur les indices perceptifs des voyelles orales et nasales* », Thèse de 3ème cycle, Laboratoire de Phonétique, Université de Nancy II, 1978.

Avec le recul, une meilleure analyse multidimensionnelle aurait pu être choisie. L'algorithme **de Kruskal** est couramment utilisé en psychologie expérimentale ; il présente néanmoins le désavantage d'être sous-optimal. L'algorithme est fondé sur la méthode du gradient avec un ajustement heuristique. Après initialisation aléatoire, les objets sont reclassés itérativement pour obtenir une distorsion de configuration (*stress*) minimale. Ce principe ne garantit pas l'obtention d'une configuration de *stress* minimum minimorum. Plusieurs tirages aléatoires initiaux sont alors effectués et seule la configuration de *stress* minimale est retenue. Dans ces conditions, l'**analyse factorielle** semble préférable. Sous le nom d'analyse en composantes principales, elle est appliquée à un tableau individus-variables pour permettre de visualiser l'information contenue dans le tableau. Mais l'analyse factorielle peut également être appliquée à un tableau de distances ou d'indices de distance inter-individus. Son but est alors de représenter les individus dans un système d'axes, de façon à ce que les distances originelles du tableau soient conservées le mieux possible. L'analyse factorielle ne se limite pas à la recherche d'une représentation euclidienne quelconque de l'ensemble des individus, mais à la recherche de la meilleure. On démontre que s'il existe une représentation euclidienne des individus correspondant au tableau inter-individus donné, les produits scalaires entre individus sont indépendants de la représentation euclidienne et se déduisent de ce tableau. Enfin, les composantes principales f_k calculées à l'aide du principe de dualité de l'analyse factorielle sont, s'il en existe une, une représentation euclidienne d'un tableau inter-individus ; les k premières permettent de conserver au mieux par projection orthogonale dans un espace de dimension k, les distances entre individus. Il suffit ainsi de calculer les valeurs propres $\{\lambda_k\}$ et les vecteurs propres $\{v_k\}$ de la matrice des produits scalaires des individus pour calculer les composantes principales $f_k = \sqrt{\lambda_k}\ v_k$, coordonnées des individus sur le $k^{ème}$ axe. Pour un tableau donné d'indices de distance, il n'existe pas toujours de représentation euclidienne ; dans ce cas l'analyse factorielle en donnera une image approchée, et l'on pourra juger de la qualité de représentation des individus dans le système d'axes.

Les bases de données de référence annotées sont le support idéal pour tester les données naturelles de la parole continue, et l'évaluation qualitative des espaces métriques sur ces données devrait être riche en enseignement. Une adaptation à la première étude consisterait, par exemple, -à estimer les indices de distances inter-individus, correspondant à deux classes vocaliques, par leur distance inter-classes ou leur distance inter-centroïdes, et -à automatiser les critères d'évaluation.

3.3. Numérique-Symbolique

Le problème en Reconnaissance des Formes est de définir à quelle forme-type ressemble le plus la forme-test observée. Quelle que soit l'application considérée, des connaissances sont *a priori* nécessaires pour mettre en œuvre les algorithmes de Reconnaissance des Formes : une connaissance sémantique du domaine, c'est-à-dire la connaissance exhaustive des formes à reconnaître, encore appelées « classes ». Cette connaissance est indissociable de l'apprentissage : « ne sera reconnu que ce qui a été appris et bien appris ». Pour des formes appartenant à une même classe, des distorsions peuvent être observées. Autant de domaines étudiés que de types de distorsion observables. Dans la parole par exemple, les distorsions peuvent provenir -des conditions d'enregistrement, -d'erreurs du système de mesure (transformations mathématiques du continu au discret), -des variantes allophoniques (variantes de réalisation d'un phonème suivant le contexte phonétique), etc... C'est dans la complexité des données captées du « monde réel » et de leurs distorsions observables que la Reconnaissance des Formes trouve à relever son plus grand défi.

Si des traits ou des combinaisons de traits demeuraient invariants quelles que soient les formes observables d'une classe, non seulement ils constitueraient des fonctions idéales de reconnaissance, mais de plus, ils pourraient être validés par synthèse. Les techniques statistiques ne permettent pas de construire ces traits ou ces combinaisons de traits, car elles demandent l'utilisation de « distances », de mesures de dissimilarité, de densité de probabilité, autant de fonctions numériques qui biaisent de manière implicite la reconnaissance. D'où notre grand intérêt pour les systèmes d'apprentissage symbolique et d'acquisition des connaissances capables de « construire » automatiquement ces ensembles de traits. C'est le système CHARADE, conçu par Jean-Gabriel Ganascia[67], qui nous a permis d'expérimenter l'approche de l'apprentissage symbolique en reconnaissance de la parole. Ce système est conçu pour détecter des régularités logiques ou statistiques existant dans un ensemble d'exemples, et engendrer l'ensemble des règles de production reflétant ces régularités. Le principe d'apprentissage repose sur l'utilisation de treillis distributifs, d'une correspondance de Galois et d'heuristiques fondées sur la prise en considération d'un modèle du domaine (langage de description) et d'un modèle de la tâche (induction des règles).

Deux représentations logiques ont été testées. Elles consistent en une transformation numérique-symbolique des observations, et sont toutes deux choisies pour leur possibilité d'interprétation en termes de formants. L'intérêt est que de nombreuses études sur la quantification des formants prototypes de classe nous permettent d'interpréter les règles et ainsi d'évaluer la qualité des systèmes de règles engendrés.

- La première représentation est fondée sur des **descripteurs binaires** de **balances d'énergie spectrale** calculées au moyen d'un banc de filtres triangulaires suivant l'échelle de fréquence Mel. Quatre niveaux d'analyse ont été retenus pour la précision des descripteurs. A un niveau d'analyse i, la bande de fréquence est divisée en 2^i canaux de fréquence. A chaque paire de canaux consécutifs correspond un descripteur dont la valeur est celle de l'expression logique ($E_k > E_{k+1}$), où E_k est l'énergie du $k^{ème}$ canal. On peut remarquer qu'au quatrième niveau les canaux sont suffisamment étroits pour permettre la localisation de masses spectrales telles que les formants.
- La seconde représentation est fondée sur le **codage binaire** de l'ensemble des **maxima spectraux** du **spectre à court terme**. 30 descripteurs binaires sont calculés sur 30 bandes de fréquence contiguës de 63 Mels. Ces descripteurs codent les masses (m) et vallées (v) spectrales. Le descripteur d_k prend la valeur m si la bande B_k correspondante inclut tout ou partie de la largeur de bande d'un pic

[67] Nous remercions Jean-Gabriel Ganascia, Professeur responsable du groupe ACASA du thème APA au LIP6, pour toute l'aide et la confiance qu'il nous a témoignées durant ces recherches.

spectral. L'intérêt est ici de coder la fréquence centrale et la largeur de bande des pics spectraux. Pour prendre en compte l'évolution temporelle du signal de parole, nous avons choisi la Décomposition Temporelle[68], Généralisée[69] (DTG) à un espace non euclidien grâce à la notion de géodésique. Cette méthode étudie les déformations de la trajectoire spectrale pour localiser les principaux événements du signal de parole. APS est la mesure de dissimilarité utilisée pour représenter la trajectoire spectrale. La coopération du système CHARADE et de la technique de DTG munie de la mesure APS permet ainsi d'intégrer trois structurations *a priori* importantes pour la parole : la structuration logique (i.e., les règles de production), la structuration temporelle (i.e., les événements acoustiques), et la structuration fréquentielle (i.e., les masses spectrales).

Le couplage numérique-symbolique a été évalué de manière quantitative et qualitative. Les tâches considérées pour l'évaluation (e.g., séparation de macro-classes, identification de voyelles), les études comparatives (e.g., classifieurs par plus proches voisins, classifieur par mémoires associatives de Hopfield), et les performances obtenues, sont décrites dans les articles référencés. Après quelques années de recul, nous donnerons ici les conclusions qui nous semblent importantes.

- La première représentation pourrait être qualifiée de **multi-résolution fréquentielle**. En effet, les descripteurs organisés sur quatre niveaux correspondent à des intégrations de l'énergie par bandes de fréquence variables. Du fait de la détection même des régularités logiques, CHARADE apparaissait comme le système idéal pour mettre en évidence l'intérêt de la multi-résolution fréquentielle par l'occurrence de combinaisons de descripteurs de niveaux différents dans le système de règles engendré. Mais sur l'identification des voyelles, l'analyse qualitative du système de règles de production n'a pas permis cette mise en évidence.
- Pour l'**identification des voyelles**, le numérique-symbolique, tel qu'il a été implémenté dans la seconde représentation, a donné des résultats d'identification comparables voire meilleurs que des opérateurs de reconnaissance fondés sur plusieurs paramétrisations usuelles dérivées de la prédiction linéaire. Ce résultat est déjà excellent, compte tenu de la description binaire des observations. Des travaux d'Olivier Desté[70] sur l'apprentissage par quantification vectorielle des pics spectraux ont encore permis d'améliorer ces performances.
- Un intérêt certain de la base des règles engendrée par CHARADE se trouve dans son évaluation qualitative qui peut être utile à une rétroaction sur la représentation des observations. Les **règles** peuvent être **analysées**, **interprétées**, et permettent la simulation des prototypes de classe. L'analyse qualitative des règles engendrées

[68] B. S. Atal, " Efficient Coding of LPC Parameters by Temporal Decomposition ", IEEE International Conference on Acoustic Signal and Speech Processing, pp. 81-84, 1983.

[69] C. Montacié, « Décodage acoustico-phonétique : apport de la décomposition temporelle généralisée et de transformations spectrales non linéaires », Thèse d'Université, Ecole Nationale Supérieure des Télécommunications, 1991.

[70] O. Desté, « *CHARADE-LVQ : le numérique au service du symbolique pour un apprentissage meilleur* », 3èmes journées Symboliques Numériques, Paris, 1992.

a ainsi permis de vérifier l'importance de la répartition des masses spectrales pour la caractérisation des voyelles. Cette analyse des règles engendrées pour les voyelles cardinales extrêmes ([i], [a], [u]), donne par exemple des masques cohérents avec nos connaissances phonétiques.

- Dans le même ordre d'idée et sans résoudre, sur le plan algorithmique, le choix de l'espace de représentation des observations, une approche de recherche en Apprentissage Symbolique du thème APA[71] s'avère novatrice dans le sens où elle explique la partie de l'exemple sur laquelle portent l'apprentissage et la généralisation. Il s'agit d'une information importante, même si elle doit être supervisée, et qui peut être également exploitée pour une **rétroaction sur la représentation**. La phase d'explication de l'apprentissage est une technique « espionne » qui peut guider très utilement la recherche d'une meilleure représentation.

Collaborations : Jean-Gabriel Ganascia, Claude Montacié, Olivier Desté (thésard-Equipe ACASA), Patrick Chazé (stagiaire du DEA IARFA).

Publications liées : 5 articles dont 1 ouvrage de synthèse [2], 3 internationaux [7, 8, 9] et 1 national [29]; 1 workshop [40], 1 rapport de fin de contrat [46].

Publications jointes : Recueil des publications page 39 [2] et page 45 [9].

3.4. Reconnaissance du locuteur

La reconnaissance du locuteur s'attache à extraire du signal vocal une information sur les caractéristiques du locuteur. Cette information est *a priori* complémentaire de l'information qui véhicule les unités linguistiques et que l'on s'attache à extraire en reconnaissance de la parole. Claude Montacié est à l'origine de l'utilisation de la modélisation Auto-Régressive Vectorielle (ARV) et d'une mesure de distorsion inter-locuteurs appelée distance d'Itakura vectorielle (DIV). Le composant de vérification du locuteur, qu'il a conçu et développé sur cette approche, s'est avéré extrêmement performant ; la vérification est indépendante du texte prononcé et de la langue employée, la phase d'apprentissage (i.e., 30 secondes de parole) est courte et ne nécessite que l'estimation d'un modèle ARV. Des expériences, menées sur la base de données DARPA-TIMIT, montrent qu'il est possible de reconnaître sans erreur un locuteur parmi 630 avec quelques secondes de parole, et 5% d'erreurs avec une seconde de parole. C'est sur ce composant, impressionnant par ses performances en mode d'élocution coopératif, que j'ai eu le plaisir de participer aux études et à la valorisation, puisque cette technologie a été brevetée puis licenciée, pour le contrôle d'accès, à la société CDVI-International.

Dans le cadre du projet AMIBE[72], j'ai participé à l'étude des deux axes de recherche suivants : -le premier pour obtenir une décision de vérification en continu sans attendre

[71] Groupe ACASA - Apprentissage Symbolique.
[72] Application Multimodales pour Interfaces et Bornes Evoluées.

la fin d'une élocution, -le second pour déterminer la fiabilité de cette vérification en fonction du type de segments de parole. Deux types de méthodes de vérification en continu ont été développées : la vérification analytique et la vérification segmentale. Dans une vérification « analytique », tout ensemble de vecteurs cepstraux consécutifs permet une vérification du locuteur. La décision est prise en comparant le dernier vecteur avec la prédiction réalisée par les modèles ARV d'apprentissage sur les vecteurs précédents. Dans une vérification « segmentale », un segment phonétique permet une vérification du locuteur à partir de la distance DIV entre le modèle ARV du segment et les modèles ARV appris. Les résultats sur la base de données TIMIT de vérification analytique du locuteur montrent que les phonèmes les plus discriminants, par ordre décroissant, sont : les diphtongues, les voyelles, les nasales/liquides, les glissées, les plosives/fricatives/affriquées. Les résultats de vérification segmentale sont faibles pour les phonèmes d'une durée de réalisation inférieure à une dizaine de centisecondes.

Les études précédentes ont porté sur de la parole dite « propre ». Nous nous sommes également intéressés à la reconnaissance du locuteur au téléphone dont les problèmes spécifiques réduisent fortement les performances des systèmes classiques. Des expériences sur la base de données DARPA-NTIMIT (630 locuteurs) montrent la dégradation des performances en reconnaissance du locuteur sur du signal de qualité téléphonique (99,2% vs 78,1% pour 168 locuteurs de TIMIT et NTIMIT). Dans le but d'améliorer ces performance, Jean-Luc Le Floch a développé des méthodes de coopération et de compétition de modèles. Deux types de modèles ont été utilisés : des modèles ARV appris sur des segments phonétiques choisis et le modèle GMM (*Gaussian Mixture Model*) développé par D. Reynolds[73]. La coopération consiste en une recherche d'ensemble de pondérations entre les vraisemblances des résultats de différents modèles. Ces pondérations sont calculées à partir des distances DIV. La compétition de modèles consiste à déterminer le modèle le plus vraisemblable pour un segment phonétique donné. Les résultats obtenus pour la méthode par compétition (82,6% pour 168 locuteurs de NTIMIT) ont montré l'intérêt de cette approche.

Collaborations : Claude Montacié, Jean-Luc Le Floch (thésard - Equipe Parole).

Publications liées : 1 brevet [1], 6 articles dont 4 internationaux [10, 11, 13, 15] et 2 nationaux [32, 34], 1 workshop international [44].

Publications jointes : Recueil des publications page 25 [10], page 29 [11].

3.5. Multi-résolution temporelle

Nos études sur l'analyse de la parole en « formants », ou plus exactement des maxima spectraux, nous ont conduits à la problématique de l'analyse du signal de parole. La première évidence est, nous l'avons déjà vu, qu'il n'existe pas de relation bi-univoque entre les maxima spectraux analysés et les formants. Les « apparitions » ou les

[73] D Reynolds, C. Rose, « *Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models* », IEEE Transactions on Speech and Audio Processing, vol. 3, pp. 72-83, 1995.

« disparitions » de maxima ne sont pas tous imputables aux phénomènes physiques des points focaux. Des artefacts apparaissent à l'analyse, et sont observés sur des bases de données aussi bien françaises qu'anglaises. A partir d'une enveloppe LPC, des artefacts peuvent être mis en évidence suivant le choix de -la méthode d'estimation des coefficients de prédiction (méthode par autocorrélation ou par covariance), -de la durée de la trame d'analyse, -du nombre de coefficients LPC, etc... La présence d'artefacts est *a priori* un facteur de dégradation de la reconnaissance. De façon à traiter les artefacts, nous proposons d'étendre l'espace de représentation du continuum de parole, en considérant non plus une séquence d'observations mais une séquence d'ensembles d'observations, calculée par une analyse que nous appelons multi-résolution temporelle.

La parole est essentiellement un signal non stationnaire. Sa production s'accompagne de changements de configuration articulatoire de l'ordre de 10 fois par seconde (estimation du débit phonétique). En raison de ces changements, la transformée de Fourier à court terme est l'une des principales mesures du signal de parole. Pour être significative, cette mesure doit être faite sur un intervalle de temps fini (trame) pour lequel l'appareil articulatoire peut être considéré immobile. La durée de la trame est généralement choisie par expérience et s'étend typiquement de 20 à 30 ms. Le continuum de parole est usuellement représenté par une séquence d'observations (i.e., vecteurs issus d'analyse à court terme). Les observations sont calculées par analyse asynchrone à la période fondamentale (*pitch*), typiquement toutes les 10 ms. En raison de son coût de calcul, moins usuelle est l'analyse synchrone à la période fondamentale ; en effet, si elle ne nécessite pas obligatoirement le calcul de la période fondamentale, elle nécessite un recalage temporel pour une analyse à portion de glotte équivalente ouverte. A notre connaissance, la durée de la trame est une alternative inusitée pour l'analyse synchrone ou asynchrone. La multi-résolution est adaptable quelles que soient les techniques d'analyse utilisées et quel que soit le mode d'analyse (synchrone ou asynchrone).

Le principe de la multi-résolution temporelle est de prendre en compte deux paramètres de choix dans l'analyse : le centre de la trame d'analyse et la durée de la trame analysée. L'analyse multi-résolution résultera d'un ensemble d'analyses au voisinage d'un point considéré du signal, chacune de ces analyses étant faite sur un ensemble de trames de durée variable.

L'intérêt de l'ensemble d'observations multi-résolution est de pouvoir estimer la stationnarité du signal de parole par une mesure telle que l'inertie. Plus l'inertie est faible plus le signal est stationnaire ; dans le cas « idéal », le continuum au voisinage du point d'analyse peut être représenté par une observation quelconque de l'ensemble des observations de multi-résolution. Plus l'inertie est forte, plus le signal est non-stationnaire ; dans le cas « extrême », le continuum ne peut être représenté par aucune de ces observations. Dans la plupart des cas, la mesure sera intermédiaire, mais il sera possible par des techniques de centroïdes ou de centroïdes généralisés d'obtenir la meilleure observation-prototype du continuum à partir de l'ensemble de multi-résolution. Si l'on considère la stabilité comme l'adéquation entre -l'application considérée, -la technique d'analyse utilisée et -le continuum analysé, on conçoit bien que si le couple technique d'analyse et continuum analysé est évalué « hautement » non-stationnaire, l'adéquation a peu de chance d'être réalisée. La prise de décision sur ce continuum peut s'avérer délicate, et son influence est *a priori* néfaste sur les performances de l'application considérée.

L'analyse par multi-résolution temporelle a été faite suivant un mode asynchrone centi-seconde sur la base de données anglo-américaine TIMIT et sur la base de données française BREF80. L'espace des pics spectraux muni de la mesure de distorsion APS est choisi pour visualiser l'évolution spectrale du signal analysé par multi-résolution. Les trajectoires de l'inertie et des centroïdes, obtenues par multi-résolution, ont été analysées. Les conclusions à nos travaux préliminaires ainsi que les perspectives de l'analyse par multi-résolution sont les suivantes :

- L'analyse des trajectoires de l'inertie et des maxima spectraux mettent bien en évidence les zones de haute stationnarité infra-phonémique avec des évolutions types, et de forte non-stationnarité à la transition entre les phonèmes. Des **prototypes de phonèmes** peuvent être calculés par le couplage des notions de **centroïde** et d'**inertie**.
- L'analyse par multi-résolution temporelle met en évidence le problème bien connu de la difficulté d'étiquetage de certaines parties du signal vocal. Dans la cas d'un étiquetage « fin », délimitant de façon temporelle le début et la fin d'un phonème, l'étiquetage des transitions est reconnu incertain par les phonéticiens eux-mêmes. L'une des applications de la multi-résolution est l'aide à la **correction automatique** de l'**étiquetage « fin » des bases de données**. Si un étiquetage peut être déduit à partir du décodage markovien par l'alignement des corpus acoustiques et des transcriptions phonétiques, son contrôle de qualité à l'écoute est indispensable tant que la reconnaissance n'est pas parfaite. Les fenêtres de forte inertie indiquent *a priori* les transitions entre phonèmes, « fin du phonème précédent (P), début du phonème suivant (S) ». Au voisinage immédiat des transitions détectées par leur forte inertie, les trames contiguës d'inertie plus faible témoignent de la réalisation -soit du phonème P si la trame est en amont de la transition, -soit du phonème S dans le cas d'une trame en aval. Au voisinage d'une transition, l'analyse des valeurs des distances inter-trames et des étiquettes phonétiques peut révéler une incohérence dans l'étiquetage. On peut envisager que ces informations combinées à l'inertie permettent de définir un critère de détection et de correction automatique d'erreur. Cette technique pourrait s'avérer utile, au cours des phases de décodage et de réestimation, pour l'**apprentissage** des modèles acoustiques lorsque l'étiquetage « fin » des corpus d'apprentissage n'est pas donné. Quelle que soit la tâche de reconnaissance considérée, le processus de décision devrait tenir compte et traiter au mieux cette incertitude que la multi-résolution permet d'estimer.

Collaboration : Claude Montacié.

Publication liée : 1 article international [21].

Publication jointe : Recueil des publications page 49 [21].

Conclusions et perspectives

Ce sont mes contributions à dix années de recherche post-doctorale dans le domaine du traitement informatique de la parole qui sont résumées dans ce mémoire. Ces travaux s'intègrent dans le thème « Apprentissage et Acquisition des Connaissances » du LIP6, dont le groupe « Parole » est dirigé depuis octobre 1991 par Claude Montacié et moi-même. Mes activités de recherche se sont situées principalement dans l'approche statistique de la Reconnaissance des Formes. Dans le cadre de la modélisation markovienne, mes travaux ont contribué à la réalisation collective d'un système de dictée vocale ; dans celui de la modélisation auto-régressive vectorielle, j'ai participé à des études et à la valorisation d'un système de vérification du locuteur. Dans nombre de mes travaux (espaces métriques, numérique-symbolique, dictée vocale), je me suis attachée à travailler dans le cadre de paradigmes d'évaluation qu'ils soient quantitatifs ou qualitatifs.

Pendant ces dix années, j'ai co-encadré quatre thèses dont l'une est déjà soutenue, j'ai encadré ou co-encadré plus de dix stagiaires de DEA. J'ai assuré, la responsabilité pour l'université Pierre et Marie Curie, de projets de recherche locaux, nationaux ou internationaux qui nous ont permis d'assurer une partie du financement indispensable aux activités de recherche de notre équipe. Ces activités m'ont permis d'être titulaire d'un contrat d'encadrement doctoral depuis 1994, renouvelé en 1998. Dans le cadre de l'administration de la recherche, j'ai également fait partie pendant quatre ans du Conseil Scientifique de l'Université Pierre et Marie Curie, chargée de l'expertise des conventions de recherche.

Dans le domaine de la reconnaissance de la parole, consacré au traitement informatique d'un signal biologique particulièrement complexe par son codage et son décodage, les cinq dernières années couvrent une évolution, rapide et spectaculaire des technologies vocales. En effet, c'est près d'un demi-siècle qu'il a fallu, pour voir émerger la dictée vocale des laboratoires de recherche publics ou privés. En reconnaissance de la parole, c'est un succès indéniable de la recherche. Pour expliquer cette évolution, il y a évidemment d'une part la technologie matérielle qui a accompagné l'évolution des recherches, et d'autre part, à mon avis, les deux facteurs déterminants que sont la diffusion de larges bases de données annotées et les paradigmes d'évaluation tel qu'ils ont été développés aux Etats-Unis. Tellement attendue pour ses multiples applications dans la communication Homme-Machine, la dictée vocale fait déjà l'objet d'une exploitation commerciale. Tout observateur extérieur pourrait considérer cet état de fait comme un aboutissement des recherches en reconnaissance de la parole. Mais, que la reconnaissance soit mono-locuteur ou indépendante du locuteur, l'amélioration spectaculaire des performances provient de ce que l'on a bien cerné et bien modélisé les domaines de l'application de reconnaissance. En particulier, c'est l'intégration des connaissances linguistiques, en l'occurrence des modèles de langage de mieux en mieux adaptés à une tâche bien spécifique, qui peut donner cette impression d'une problématique résolue. Cependant, lorsque ces contraintes linguistiques se relâchent et deviennent plus difficilement modélisables, soit parce ce que la parole est plus spontanée (conversation) ou soit que l'on dispose d'une quantité insuffisante de textes d'apprentissage, les performances se dégradent fortement. Cette détérioration s'explique également en partie par une qualité insuffisante du décodage acoustique « pur » qui doit encore faire l'objet d'efforts de recherche. De plus, la reconnaissance de la parole peut désormais s'inscrire dans des problématiques plus larges comme la recherche d'information ou la compréhension de la parole. Cette évolution se fera en interaction avec d'autres communautés de recherche et nous permettra d'appréhender toute la complexité de la communication parlée.

Mes perspectives à court terme sont de développer et d'utiliser la notion de multi-résolution généralisée pour prendre en compte les phénomènes de non-stationnarité dans le signal de parole. Deux axes de recherche peuvent être explorés dans le cadre de la reconnaissance markovienne. Le premier consiste à utiliser le coefficient de stationnarité d'une trame dans le calcul de ses probabilités d'émission. Le second est plus complexe : il vise à développer une paramétrisation spécifique pour les zones de non-stationnarité. L'une des difficultés est alors la gestion de deux types de paramétrisation dans un modèle de Markov caché.

La recherche d'information est l'une de mes perspectives à moyen terme. En effet, l'acquis et le savoir-faire de notre équipe en matière de dictée vocale et de reconnaissance du locuteur, nous a permis de travailler sur un problème plus large que la reconnaissance de la parole : l'indexation. Ces recherches ont commencé avec le projet TRANSDOC, coordonné par Pascal Faudemay du LIP6 et Lemin Chen du laboratoire Heudiasyc de l'Université de Technologie de Compiègne. Ce projet d'indexation de documents multimédia a permis à notre équipe de développer des outils préliminaires[74] à l'indexation comme l'alignement de script[75] ou la détection silence/bruit/musique/parole[76]. Ces recherches se poursuivent et s'approfondissent dans le cadre du projet AVIR auquel nous participons. Il s'agit d'un projet ESPRIT dont le but est l'indexation et la recherche d'informations multimédia pour utilisateurs non-experts. Il réunit Philips

[74] Publication jointe : Recueil des publications page 53 [14].
[75] Publication jointe : Recueil des publications page 64 [19].
[76] Publication jointe : Recueil des publications page 68 [23].

(Hollande), Techmath (Allemagne), le LEP (France), La RAI (Italie), TV Spielfilm (Allemagne), et les Universités de Brescia (Italie) et de Paris 6 (France). Son but est d'apporter de nouvelles solutions permettant à des utilisateurs non-experts d'accéder, de stocker et de rechercher facilement une information audiovisuelle. Ces solutions sont basées sur l'utilisation conjointe d'agents intelligents et d'un profil-utilisateur permettant de sélectionner et d'enregistrer automatiquement les programmes correspondant à l'intérêt personnel de l'utilisateur. Ceci ne peut seulement être réalisé que si conjointement aux données audio-visuelles sont fournies des méta-données décrivant le contenu des canaux, avec un niveau de détail suffisant pour un fonctionnement automatique. La production et la normalisation (norme MPEG-7) de ces méta-données est l'un des objectifs de ce projet. Nous nous intéressons surtout aux méta-données obtenues à partir des sources audio. Notre contribution principale sera l'étude et le développement d'un serveur de dictée vocale adapté.

Enfin je terminerai par une problématique qui m'intéresse, celle de l'évolution d'un système de dictée vocale qu'il s'agisse de la prise en compte de nouvelles avancées scientifiques ou de l'adaptation rapide à une application spécifique. Cette problématique est due au nombre toujours plus important de contributions scientifiques sur le sujet. La difficulté d'une évaluation comparative de ces contributions ralentit à mon avis les progrès dans ce domaine car elles doivent être intégrées dans un système complexe. Le développement dans le cadre du génie logiciel, avec -ses buts, réutilisabilité, maintenance, portabilité, et -ses moyens, spécifications, contrôle de flux de données, peut être utilisé pour parvenir à une solution à l'évolution des systèmes de dictée vocale. Il s'agirait alors de développer la notion de composants logiciels réutilisables correspondant à une contribution scientifique. En considérant tout ce que la recherche pourrait y gagner, on peut se demander pourquoi une telle notion n'a pu se concrétiser qu'en partie[77] ? L'un des éléments de réponse est que l'intérêt de la recherche, aussi important soit-il, n'occulte cependant pas l'intérêt propre des organismes de recherche et des chercheurs. Ces intérêts particuliers relèvent de la propriété intellectuelle et industrielle. C'est en ménageant ces droits et en établissant les devoirs des individus utilisant les composants logiciels faits par autrui que ce type de diffusion pourra se développer. Un devoir des utilisateurs et une garantie que doit offrir la plate-forme est alors un usage des composants logiciels dans l'unique but de la recherche en excluant leur utilisation dans des projets donnant lieu à un financement. Pour préserver la propriété industrielle, seul les codes-objets des composants logiciels peuvent être accessibles. Toutes ces contraintes doivent nous obliger à identifier, à définir et à normaliser avec précision les composants logiciels utilisés dans un système de dictée vocale, à définir les spécifications pour que ces composants puissent s'intégrer sans perte d'efficacité ou de rapidité. Je pourrais donner comme exemple l'intégration d'un composant de type algorithme de décodage avec celui d'accès rapide à un modèle de langage. Tous ces efforts de normalisation doivent naturellement être sous-tendus par des actions de coopération entre les différentes équipes travaillant dans ce domaine.

[77] Speech Training and Recognition Unified Tool - http ://tcts.fpms.ac.be/speech/strut.html.
Intel Recognition Primitives Library - Référence Manual.

Bibliographie personnelle

Brevet

1. C. Montacié, **M.-J. Caraty**, J.-L. Le Floch, « Procédé de reconnaissance vocale d'un locuteur mettant en œuvre un modèle prédictif, notamment pour les applications de contrôle d'accès », Brevet français 9605566, déposé par l'Université Pierre et Marie Curie, le 3 mai **1996.** Extension internationale PCT/FR 9700753 déposé le 25 Avril **1997.**

Publications dans des ouvrages de synthèse

2. **M.-J. Caraty**, C. Montacié, X. Rodet, « Automatic Learning of a Production Rule System for Acoustic-Phonetic Decoding » in P. Laface and R. De Mori (Editor), *Speech Recognition and Coding, New Advances and Trends*, NATO ASI Series, Vol. F 75, pp. 311-316, Springer-Verlag, Berlin, **1992**.

3. C. Barras, **M.-J. Caraty,** C. Montacié, « HMM based Acoustic-Phonetic Decoding with Constrained Transitions and Speaker Topology » in A. Ayuso and R. De Mori (Editor) *Speech Recognition and Coding, New Advances and Trends*, NATO ASI Series, Vol. F 147, pp. 72-75, Springer-Verlag, Berlin, **1995**.

4. C. Montacié, **M.-J. Caraty**, R. Andre-Obrecht, L.-J. Boë, P. Deléglise, M. El-Bèze, I. Herlin, P. Jourlin, T. Lallouache, B. Leroy, H. Méloni, « Applications Multimodales pour Interfaces et Bornes évoluées (AMIBE) », Ecole d'été, Fondements et perspectives en traitement de la parole, pp. 155-164, Luminy, France, **1995**.

5. F. Lefèvre, C. Montacié, **M.-J. Caraty**, « K-Nearest Neighbours Estimator in a HMM-Based System » *in* Pointing (Editor) *Computational Models of Speech Patterns Processing*, NATO ASI Series, Springer-Verlag, Berlin, **1998** (à paraître).

Congrès internationaux avec comité de lecture

6. H. Ye, **M.-J. Caraty**, L.-J. Boë, D. Tuffelli, « Structural (Phonetic) Evaluation of Dissimilarities Functions Used in Speech Recognition », Eurospeech, European Conference on Speech Communication and Technology, pp. 404-407, Paris, France, **1989**.

7. **M.-J. Caraty**, C. Montacié, P. Chazé, X. Rodet, « Application of an Automatic Learning System for Acoustic-Phonetic Decoding », Cognitiva, pp. 847-850, Madrid, Espagne, **1990**.

8. C. Montacié, **M.-J. Caraty**, X. Rodet, « Experiments in the Use of an Automatic Learning System for Acoustic-Phonetic Decoding », Actes ICSLP, International Conference on Spoken Language Processing, pp. 357-360, Kobe, Japon, **1990**.

9. **M.-J. Caraty,** C. Montacié, C. Barras, **«** Integration of Temporal and Frequential Structurations in a Symbolic Learning System », IEEE-ICSLP, International Conference on Spoken Language Processing, pp. 475-478, Banff, Canada, **1992**.

10. C. Montacié, P. Deléglise, F. Bimbot, **M.-J. Caraty**, « Kinematic Techniques for Speech Processing : Temporal Decomposition and Multivariate Linear Prediction », IEEE-ICASSP, International Conference on Acoustic Signal and Speech Processing, pp. 153-156, San Francisco, Etats-Unis, **1992**.

11. J.-L. Le Floch, C. Montacié, **M.-J. Caraty, «** Investigations on Speaker Characterization from Orphée System Technics », Actes IEEE-ICASSP, International Conference on Acoustic Signal and Speech Processing, pp. 149-152, Adelaide, Australie, **1994**.

12. C. Barras, **M.-J. Caraty**, C. Montacié, « Temporal Control and Training Selection for HMM-Based System », Eurospeech, European Conference on Speech Communication and Technology, pp. 27-30, Madrid, Espagne, **1995**.

13. J.-L. Le Floch, C. Montacié, **M.-J. Caraty**, « Speaker Recognition Experiments on the NTIMIT Database », Eurospeech, European Conference on Speech Communication and Technology, pp. 379-382, Madrid, Espagne, **1995**.

14. P. Faudemay, L. Cheng, C. Montacié, **M.-J. Caraty**, C. Maloigne, X. Tu, M. Ardebilian, J.-L. Le Floch, « Multi-channel Video Segmentation », International Conference on Multimedia Storage and Archiving Systems, SPIE Symposium, pp. 252-264, Boston, Etats-Unis, **1996**.

15. J.-L. Le Floch, C. Montacié, **M.-J. Caraty**, « GMM and ARVM Cooperation and Competition for Text-Independent Speaker Recognition on Phone Speech », Actes IEEE-International Conference on Spoken Language Processing, pp. 2411-2414, Philadelphie, Etats-Unis, **1996**.

16. C. Montacié, **M.-J. Caraty**, C. Barras, « Mixture Splitting Technic and Temporal Control in a HMM-Based Recognition System », Actes IEEE-International Conference on Spoken Language Processing, pp. 977-980, Philadelphie, Etats-Unis, **1996**.

17. **M.-J. Caraty**, C. Montacié C., F. Lefèvre, « Dynamic Lexicon for a Very Large Vocabulary Vocal Dictation », Eurospeech, European Conference on Speech Communication and Technology, pp. 2691-2694, Rhodes, Grèce, **1997**.

18. P. Faudemay, C. Montacié, **M.-J. Caraty**, « Video Indexing Based on Image and Sound », International Conference on Multimedia Storage and Archiving System, SPIE Symposium, pp. 57-69, Dallas, Etats-Unis, **1997**.

19. C. Montacié, **M.-J. Caraty**, « Sound Channel Video Indexing », Eurospeech, European Conference on Speech Communication and Technology, pp. 2359-2362, Rhodes, Grèce, **1997**.

20. C. Montacié, **M.-J. Caraty**, F. Lefèvre, « KNN versus Gaussian in a HMM-based Recognition System », Eurospeech, European Conference on Speech Communication and Technology, pp. 529-532, Rhodes, Grèce, **1997**.

21. **M.-J. Caraty**, C. Montacié, « Multiresolution Analysis for Speech Analysis », Actes IEEE-International Conference on Spoken Language Processing, pp. 955-958, Sydney, Australie, **1998**.

22. F. Lefèvre, C. Montacié, **M.-J. Caraty**, « Influence of the Probability Estimator in a HMM-based System », Actes IEEE-International Conference on Spoken Language Processing, Sydney, Australie, **1998**.

23. C. Montacié, **M.-J. Caraty**, « A Silence/Noise/Music/Speech Splitting Algorithm », Actes IEEE-International Conference on Spoken Language Processing, pp. 1579-1582 Sydney, Australie, **1998**.

Congrès nationaux avec comité de lecture

24. **M.-J. Caraty,** X. Rodet, « Distance interspectrale à critères perceptifs », 14èmes JEP, Journées d'Etude sur la Parole, pp. 87-90, Paris, France, **1985**.

25. **M.-J. Caraty,** X. Rodet, « Etude comparative de mesures de distorsion spectrale », 15èmes JEP, Journées d'Etude sur la Parole, pp. 297-300, Aix-en-Provence, France, **1986**.

26. **M.-J. Caraty,** X. Rodet, « Reconnaissance de /VCV/ (cycles vocaliques) en parole continue par comparaison dynamique », 16èmes JEP, Journées d'Etude sur la Parole, pp. Etats-Unis 96-99, Hammamet, Tunisie, **1987**.

27. H. Yé , **M.-J. Caraty**, L.-J. Boë, D. Tuffelli, « Distances interspectrales et macro-sensibilité des voyelles focales », 17èmes JEP, Journées d'Etude sur la Parole, pp. 20-25, Nancy, France, **1988**.

28. C. Barras, **M.-J. Caraty**, P. Deléglise, C. Montacié, R. André-Obrecht, X. Rodet, « Décomposition temporelle et ruptures de modèles pour le décodage acoustico-phonétique », 19èmes JEP, Journées d'Etude sur la Parole, pp. 335-340, Bruxelles, Belgique, **1992**.

29. **M.-J. Caraty,** C. Montacié, « Intégration de la décomposition temporelle généralisée dans un système d'apprentissage symbolique. Application à la reconnaissance des voyelles », 19èmes JEP, Journées d'Etude sur la Parole, pp. 387-392, Bruxelles, Belgique, **1992**.

30. C. Barras, **M.-J. Caraty**, C. Montacié, « Contrôle temporel et sélection de l'apprentissage appliqués aux modèles de Markov Cachés », RFIA, Reconnaissance des Formes et Intelligence Artificielle, pp. 391-396, Paris, France, **1994**.

31. C. Barras, **M.-J. Caraty**, C. Montacié, « Topologie des locuteurs et pré-segmentation pour les modèles de Markov Cachés », 20èmes JEP, Journées d'Etude sur la Parole, pp. 491-494, Lannion, France, **1994**.

32. J.-L. Le Floch, C. Montacié, **M.-J. Caraty**, « Investigation de protocoles pour la reconnaissance des locuteurs fondée sur le système Orphée », 20èmes JEP, Journées d'Etude sur la Parole, pp. 255-258, Lannion, France, **1994**.

33. **M.-J. Caraty**, C. Barras, F. Lefèvre & C. Montacié, « D-DAL : un système de dictée vocale développé sous l'environnement HTK », 21èmes JEP, Journées d'Etude sur la Parole, pp. 289-292, Avignon, France, **1996.**

34. J.-L. Le Floch, C. Montacié, **M.-J. Caraty**, « Coopération et compétition de modèles en reconnaissance du locuteur », 21èmes JEP, Journées d'Etude sur la Parole, pp. 395-398, Avignon, France, **1996**.

35. P. Faudemay, L. Chen, C. Montacié, **M.-J. Caraty,** X. Tu, "Segmentation multi-canaux de vidéos en séquences, 3èmes journées CORESA, Journées d'études et d'échanges "Compresion et Représentation des Signaux Audiovisuels", Issy-les-Moulineaux, France, **1997**.

36. C. Montacié, **M.-J. Caraty,** F. Lefèvre, « Lexique dynamique dans un système de reconnaissance de la parole », 2ème CFA, Congrès Français d'Acoustique, pp. 413-416, Marseille, France, **1997**.

Workshops

37. **M.-J. Caraty**, « Interspectral Distorsion Measure and Perceptual Differential Limens », Second Franco-British Meeting on Speech, Sussex, Grande-Bretagne, **1988**.

38. **M.-J. Caraty**, J.-C. Richard, X. Rodet, « Décodage Acoustico-Phonétique : reconnaissance de formes spectro-temporelles », 1ères Journées du GRECO-PRC Communication Homme-Machine, pp. 117-122, Toulouse, France, **1988**.

39. **M.-J. Caraty**, X. Rodet, « Reconnaissance de formes spectro-temporelles, Actes du Séminaire Action Décodage Acoustico-Phonétique », pp. 47-52, Nancy, France, **1988**.

40. **M.-J. Caraty**, C. Montacié, P. Chazé, X. Rodet, « Application d'un système de règles au décodage acoustico-phonétique », 2èmes Journées du GRECO-PRC Communication Homme-Machine, pp. 255-265, Toulouse, France, **1991**.

41. C. Barras, **M.-J. Caraty**, C. Montacié, P. Deléglise, R. André-Obrecht, F. Bimbot, J.-L. Le Floch, « Approche événementielle par la reconnaissance de la parole », 3èmes Journées du GRECO-PRC Communication Homme-Machine, pp. 24-34, Nancy, France, **1994**.

42. **M.-J. Caraty**, C. Montacié, F. Lefèvre, "D-DAL le Dictaphone-Dactylographe du LIP6", Atelier concours de dictée vocale de l'AUPELF-UREF, Avignon, France, **1997**.

43 P. Faudemay, **M.-J. Caraty**, C. Montacié, « Benchmarking Methods for Audio », MPEG Requirements Group, MPEG Meeting, Document ISO/MPEG N3313, Tokyo, Japon**, 1998**.

44. C. Montacié, **M.-J. Caraty,** « AR-Vector Models for Speaker Recognition », Speaker Recognition Workshop, National Institute of Standard and Technology, Maryland, Etats-Unis, **1998**.

Thèse

45. **M.-J. Caraty, «** Contribution au décodage acoustico-phonétique : études de distances interspectrales et reconnaissance de cycles vocaliques », Thèse d'Université, Université Pierre et Marie Curie, 2 juillet **1987**.

Rapports de fin de contrat

46. C. Barras, **M.-J. Caraty**, O. Deste, J.-G. Ganascia, C. Montacié, « Acquisition de Connaissance en Traitement de la Parole », Actes de la Journée IBP, Journée de l'Institut Blaise Pascal, pp. 32-37, Paris, **1992**.

47. C. Montacié, **M.-J. Caraty**, « Projet AMIBE : rapport préliminaire », LAFORIA94/12, **1994**.

48. **M.-J. Caraty**, C. Montacié, Rapport d'Activité 1995 (arrêté à la date du 31 mars 1996) ARC « Linguistique, Informatique et corpus oraux : Dictée vocale », « ILOR B1 », Convention ARC n° X/7.10.04/ILOR.B1t.rec/ Paris6.LAFORIA.MJC, **1996**.

49. C. Montacié, **M.-J Caraty**, T. Lallouache, P. Deléglise, P. Jourlin, R. André-Obrecht, I. Herlin, T. Guiard, M. Alissali, M. El-Bèze, B. Jacob, B. Leroy, A. Adjoudani, A. Rogozan, H. Méloni, N. Parlangeau, F. Pellegrino, C. Cénac, Projet du GDR-PRC « AMIBE : Applications Multimodales pour Interfaces et Bornes évoluées », 90 pages, **1996**.

50. **M.-J. Caraty**, C. Montacié, Rapport Biennal d'Activité Scientifique 1995-1996 (arrêté à la date du 25 mars 1997) ARC « Linguistique, Informatique et corpus oraux : Dictée vocale », « ILOR B1 », Convention ARC n° X/7.10.04/ILOR.B1t.rec/ Paris6.LAFORIA. MJC, **1997**.

51. C. Montacié, **M.-J. Caraty**, « Reconnaissance du Locuteur en Indexation », Rapport d'Activité Scientifique 1996 (arrêté à la date du 21 mai 1997), projet du Ministère de l'Industrie TRANSDOC, **1997**.

Curriculum Vitae

	Marie-José Caraty
Date et lieu de naissance :	née le **25 décembre 1957** à PARIS 10ème
Nationalité :	**française**
Situation de famille :	**mariée**, **un enfant** né le 1er août 1992
Adresse personnelle :	7, rue Emile Gœury 94140 Alfortville
Adresse professionnelle :	**LIP6 - Laboratoire d'Informatique de Paris 6** Université Pierre et Marie Curie 4, place Jussieu 75232 Paris Cedex 05 Tél : 01 44 27 47 22 E-mail : Marie-Jose.Caraty@lip6.fr
Situation actuelle :	**Maître de Conférences** à l'**UFR d'Informatique** de l'université Pierre et Marie Curie (**Paris 6**)
Date d'installation :	le 1er octobre 1989

Diplômes

Licence d'Informatique	Université Pierre et Marie Curie, mention *bien*, 1980
Maîtrise d'Informatique	Université Pierre et Marie Curie, mention *bien*, 1981
DEA TAI	Université Pierre et Marie Curie, mention *bien*, 1982 TAI (Traitement Algorithmique de l'Information) M1-Reconnaissance des Formes, M2-Intelligence Artificielle

Doctorat de l'Université Pierre et Marie Curie, mention *très honorable*, 1987

Spécialité : Informatique

Titre : « *Contribution au décodage acoustico-phonétique : études de distances interspectrales et reconnaissance de cycles vocaliques* »

Jury : J.-C. SIMON (président et rapporteur), G. CHOLLET (rapporteur), X. RODET (directeur de thèse), L.-J. BOE, J. MARIANI, G. PERENNOU

Fonctions assurées dans le cadre de l'enseignement et de la recherche

Depuis 1989 — **Maître de Conférences** à l'UFR d'Informatique de Paris 6

- **1993** promue à la **1ère classe** des Maîtres de Conférences
- **1990-1994** titulaire d'un **contrat pédagogiqu**e
- **1994-2002** titulaire d'un **contrat d'encadrement doctoral et de recherche** (contrat renouvelé en 1998)

1985-1988 — **Assistante Associée** à l'UFR d'Informatique de Paris 6

1983-1984 — **Vacataire** à l'UFR d'Informatique de Paris 6

1981-1982 — **Conseil** et **formatrice** en Micro-Informatique à la Société Mégalpha International

1981-1982 — **Maître Auxiliaire** de Mathématiques au Lycée Saint-Vincent à Senlis

Participation à l'administration des structures universitaires

1989-1993 — Membre élu au **Conseil Scientifique de l'Université Pierre et Marie Curie**
Chargée de l'expertise scientifique des conventions de recherche de l'UFR d'Informatique

1989-1993 — Membre élu au **Conseil de l'UFR d'Informatique**

- **1991-1992** Membre nommé au **bureau de l'UFR**
  Chargée de mission pour le développement des relations européennes

1994-1998 — Membre élu de la **Commission de Spécialistes**, **27ème section**

1990-1998 — Membre nommé du **Conseil Scientifique** du **LAFORIA**, puis du **LIP6**

Participation à l'administration de la recherche

1998-2000 — Membre élu du bureau de la **section GFCP** (Groupe Francophone de la Communication Parlée) de la **SFA** (Société Française d'Acoustique)

Participation aux programmes de recherche nationaux

GRECO PRC, Action « **Décodage Acoustico-Phonétique** », **1987-1989**
Sujet : « *Reconnaissance de noyaux vocaliques dans la parole continue* »
Responsable : M.-J. Caraty. Participants : J.-C. Richard et R. Sorokine

GRECO PRC, Action « **Reconnaissance de la Parole** », **1990-1992**
Partenaires : ENST-Paris, IRISA-Rennes
Sujet : « *Intégration et modélisations complémentaires du signal vocal* »
Responsables : M.-J. Caraty et C. Montacié. Participant : C. Barras

Projet fédérateur IBP, **1990-1992**
Partenaires : Equipe ACASA (LAFORIA)
Sujet : « *Apprentissage symbolique et décodage acoustico-phonétique* »
Responsables: J.-G. Ganascia(Equipe ACASA), M.-J. Caraty et C. Montacié (Equipe Parole). Participants : C. Barras et O. Desté

GRECO PRC, Action « **Reconnaissance de la Parole** », **1992-1993**
Partenaires : CERFIA-Toulouse, ENST-Paris, LIA-Avignon
Sujet : « *Détection de mots dans la parole continue* »
Responsables LAFORIA : M.-J. Caraty et C. Montacié. Participant : C. Barras

GDR-PRC « **Communication Homme Machine** », **1993-1996**
Sujet : « *AMIBE : Applications Multimodales pour Interfaces et Bornes Evoluées* »
Coordinateur : C. Montacié. Partenaires : six laboratoires (20 chercheurs)

Projet du ministère de l'Industrie, **TRANSDOC**, **1996-1998**
Sujet : « *Indexation audio* »
Responsable Thème APA : C. Montacié. Participant : M.-J.Caraty

Participation aux programmes de recherche internationaux

Projet ESPRIT 28798, **AVIR**, 1998-2000
Objectif : Conception et réalisation d'un serveur d'indexation audio
Responsables : **P. Faudemay** (pour Paris 6), C. Montacié (Equipe Parole)

Participation à la première campagne de test des systèmes de dictée vocale francophone, 1997
Projet de **l'AUPELF-UREF** ARC-B1, 1996-2000
Responsable pour Paris 6 : **M.-J. Caraty**, participants : C. Montacié, F. Lefèvre

Participation aux évaluations des systèmes de vérification du locuteur, 1998
Projet du **National Institute of Standard and Technology**, SRE, 1998
Responsable pour Paris 6 : C. Montacié. Participant : **M.-J. Caraty**

Co-encadrement de thèses

Notre équipe est constituée de deux permanents, Claude Montacié et moi-même. Nous co-encadrons sur une base de 100% les thésards de notre équipe. Les temps d'encadrement, estimés en pourcentage, sont précisés.

Claude Barras [60%]
Titre : « *Reconnaissance de la parole continue : adaptation au locuteur et contrôle temporel dans les modèles de Markov cachés* »
Doctorat de l'Université Pierre et Marie Curie, spécialité Informatique
Mention très honorable avec les félicitations du jury

Thèse soutenue le 29 mai 1996
Jury : R. de Mori (président et rapporteur), R. André-Obrecht (rapporteur), X. Rodet (directeur de thèse), M.-J. Caraty, M. El-Bèze, P. Gallinari, C. Montacié

Claude Barras a développé et testé différentes approches -pour l'apprentissage (classes de locuteurs, test du khi-carré multidimensionnel) des modèles de Markov cachés et -pour le décodage (contrôle temporel fondé sur les ruptures de modèles) en reconnaissance markovienne. Il les a implémenté et testé à l'aide du logiciel HTK sur la base de données TIMIT.

Jean-Luc Le Floch [40%]
Sujet : « *Identification du locuteur en milieu bruité indépendamment du texte* »
Doctorat de l'Université Pierre et Marie Curie, spécialité Informatique

Jean Luc Le Floch a développé des méthodes de reconnaissance du locuteur par coopération/ compétition de modèles AR-vectoriels et de sommes pondérées de multigaussiennes. Il a testé ces méthodes sur la base de données N-TIMIT. Il a arrêté sa thèse, pour l'instant, pour développer une serrure vocale chez un industriel à partir de notre brevet.

Fabrice Lefèvre [40%]
Sujet : « *Etude et amélioration des modèles de Markov cachés discriminants par des méthodes non paramétriques* »
Doctorat de l'Université Pierre et Marie Curie en cours, spécialité Informatique

Date de soutenance prévue : début 1999

Fabrice Lefèvre a développé des méthodes d'apprentissage des modèles de Markov cachés en utilisant un estimateur non-paramétrique (K-ppv). Il a également participé à la première campagne de dictée vocale en implémentant le moteur de reconnaissance de notre système.

Sergios Pétridis [60%]
Sujet : « *Indexation sonore : parole et musique* »
Doctorat de l'Université Pierre et Marie Curie en cours, spécialité Informatique

Soutenance prévue : en 2000

Sergios Pétridis est en thèse depuis deux an. Il a développé un algorithme de segmentation silence/bruit/parole.

Participation à des jurys de thèse

Mohamed Hermak, Docteur Ingénieur - Université Paris 6, spécialité Informatique, 15/12/89
Titre de la thèse : « *Représentation et édition de paramètres spectro-temporels pour la synthèse de la parole* »
Jury : X. Rodet (Président), **M.-J. Caraty** (Rapporteur), F. Aussedat (Rapporteur), A. Fromentel

Chantal Laura, Docteur de l'Université Paris 6, spécialité Informatique, 18/12/90
Titre de la thèse : « *Entre segments et règles, une approche perceptive pour la synthèse automatique de la parole* »
Jury : M. Baudry (Président et Rapporteur), M. Bonnet (Rapporteur), X. Rodet (Directeur de thèse), G. Boulakia, **M.-J. Caraty**

Claude Barras, Docteur de l'Université Paris 6, spécialité Informatique, 29/05/96
Titre : « *Reconnaissance de la parole continue : adaptation au locuteur et contrôle temporel dans les modèles de Markov cachés* »
Jury : R. de Mori Président et Rapporteur), R. André-Obrecht (Rapporteur), X. Rodet (Directeur de thèse), **M.-J. Caraty**, M. El-Bèze, P. Gallinari, C. Montacié

Marcos Perreau Guimaraes, Docteur de l'Université Paris 5, spécialité Informatique, 23/06/98
Titre de la thèse : « *Optimisation des ressources binaires et modélisation psycho-acoustique pour le codage audio* »
Jury : P. Deléglise (Président), H. Leich (Rapporteur), J. Menez (Rapporteur), M. Bonnet (Directeur de thèse), N. Moreau (Directeur de thèse), **M.-J. Caraty**, J. Soumagne

Liste des publications jointes

Chapitre 2. D-DAL - Lexique dynamique

p. 1 **M.-J. Caraty**, C. Barras, F. Lefèvre & C. Montacié, « D-DAL : un système de dictée vocale développé sous l'environnement HTK », 21èmes JEP, Journées d'Etude sur la Parole, pp. 289-292, Avignon, France, **1996**. **[33]**

p. 5 **M.-J. Caraty**, C. Montacié C., F. Lefèvre, « Dynamic Lexicon for a Very Large Vocabulary Vocal Dictation », Eurospeech, European Conference on Speech Communication and Technology, pp. 2691-2694, Rhodes, Grèce, **1997**. **[17]**

Chapitre 2. Systèmes hybrides

p. 9 C. Barras, **M.-J. Caraty,** C. Montacié, « HMM based Acoustic-Phonetic Decoding with Constrained Transitions and Speaker Topology » in A. Ayuso and R. De Mori (Editor) *Speech Recognition and Coding, New Advances and Trends*, NATO ASI Series, Vol. F 147, pp. 72-75, Springer-Verlag, Berlin, **1995**. **[3]**

p. 13 C. Barras, **M.-J. Caraty**, C. Montacié, « Temporal Control and Training Selection for HMM-Based System », Eurospeech, European Conference on Speech Communication and Technology, pp. 27-30, Madrid, Espagne, **1995**. **[12]**

p. 17 C. Montacié, **M.-J. Caraty**, C. Barras, « Mixture Splitting Technic and Temporal Control in a HMM-Based Recognition System », Actes IEEE-International Conference on Spoken Language Processing, pp. 977-980, Philadelphie, Etats-Unis, **1996**. **[16]**

p. 21 C. Montacié, **M.-J. Caraty**, F. Lefèvre, « KNN versus Gaussian in a HMM-based Recognition System », Eurospeech, European Conference on Speech Communication and Technology, pp. 529-532, Rhodes, Grèce, **1997**. **[20]**

Chapitre 3. Evaluations qualitatives d'espaces métriques

p. 33 H. Yé , **M.-J. Caraty**, L.-J. Boë, D. Tuffelli, « Distances interspectrales et macro-sensibilité des voyelles focales », 17èmes JEP, Journées d'Etude de la Parole, pp.20-25, Nancy, France **1988**. **[27]**

Chapitre 3. Numérique-symbolique

p. 39 **M.-J. Caraty**, C. Montacié, X. Rodet, « Automatic Learning of a Production Rule System for Acoustic-Phonetic Decoding » in P. Laface and R. De Mori (Editor), *Speech Recognition and Coding, New Advances and Trends*, NATO ASI Series, Vol. F 75, pp. 311-316, Springer-Verlag, Berlin, **1992**. **[2]**

p. 45 **M.-J. Caraty,** C. Montacié, C. Barras, **«** Integration of Temporal and Frequential Structurations in a Symbolic Learning System », IEEE-ICSLP, International Conference on Spoken Language Processing, pp. 475-478, Banff, Canada, **1992**. **[9]**

Chapitre 3. Reconnaissance du locuteur

p. 25 C. Montacié, P. Deléglise, F. Bimbot, **M.-J. Caraty**, « Kinematic Techniques for Speech Processing : Temporal Decomposition and Multivariate Linear Prediction », IEEE-ICASSP, International Conference on Acoustic Signal and Speech Processing, pp. 153-156, San Francisco, Etats-Unis, **1992**. **[10]**

p. 29 J.-L. Le Floch, C. Montacié, **M.-J. Caraty**, « Investigations on Speaker Characterization from Orphée System Technics », Actes IEEE-ICASSP, International Conference on Acoustic Signal and Speech Processing, pp. 149-152, Adelaide, Australie, **1994**. **[11]**

Chapitre 3. Multi-résolution

p. 49 **M.-J. Caraty**, C. Montacié, « Multiresolution Analysis for Speech Analysis », Actes IEEE-International Conference on Spoken Language Processing, pp. 955-958, Sydney, Australie, **1998**. **[21]**

Conclusions Indexation multimédia

p. 53 P. Faudemay, L. Cheng, C. Montacié, **M.-J. Caraty**, C. Maloigne, X. Tu, M. Ardebilian, J.-L. Le Floch, « Multi-channel Video Segmentation », International Conference on Multimedia Storage and Archiving Systems, SPIE Symposium, pp. 252-264, Boston, Etats-Unis, **1996**. **[14]**

p. 64 C. Montacié, **M.-J. Caraty**, « Sound Channel Video Indexing », Eurospeech, European Conference on Speech Communication and Technology, pp. 2359-2362, Rhodes, Grèce, **1997**. **[19]**

p. 68 C. Montacié, **M.-J. Caraty**, « A Silence/Noise/Music/Speech Splitting Algorithm », Actes IEEE-International Conference on Spoken Language Processing, pp. 1579-1582, Sydney, Australie, **1998**. **[23]**