

**AVIS DE SOUTENANCE DE THÈSE DE
DOCTORAT**

Monsieur TAZINE Camal soutiendra une thèse
le 21 avril 2005 à 14 heures

**Salle de Réunion – Rez-de-Chaussée
Laboratoire d'Informatique
IUP GMI**

SPÉCIALITÉ : Informatique

Titre de la thèse : Modélisation statistique du langage pour un domaine spécifique en reconnaissance automatique de la parole

Membres du jury :

M. Marc EL BEZE, professeur, Laboratoire d'Informatique (FRE 2487), INRA d'Avignon, Université d'Avignon et des Pays de Vaucluse, Avignon.

M. Renato DE MORI, professeur, Laboratoire d'Informatique (FRE 2487), Université d'Avignon et des Pays de Vaucluse, Avignon.

M. Jacques CHAUCHE, professeur, Laboratoire d'Informatique, de robotique et de micro-électronique, Université de Montpellier 2, Montpellier.

M. Pietro LAFACE, professeur, Laboratoire DAUIN, dipartimento di Automatica e Informatica, III Facolta' Di Ingeneria, Politecnico di Torino, Torino, Italie.

M. Célestin SEDOGBO, directeur de recherche, Laboratoire HIT, Département DAS, THALES Research & Technology France, Orsay.

Résumé de la thèse :

Les modèles de langage des systèmes de reconnaissance de la parole large vocabulaire sont souvent basés sur des méthodes statistiques nécessitant de larges corpus spécifiques au domaine d'application traité. La constitution de ce type de corpus est très coûteuse car elle nécessite un travail humain important. De plus, les corpus influençant énormément les modèles de langages qu'ils génèrent, l'utilisation d'un même corpus pour d'autres domaines d'application est difficile sans perte sévère de précision. Pour cette raison, il est difficile d'utiliser directement un corpus « général » pour traiter des domaines spécifiques. Pourtant, un corpus général de est largement moins coûteux à constituer qu'un corpus spécifique.

Une manière d'obtenir un modèle de langage spécifique en minimisant la main d'œuvre humaine consisterait à l'obtenir en utilisant un corpus « général » de grande taille, non étiqueté. En effet, un tel corpus peut offrir des optiques intéressantes, aussi bien sur le caractère général de la langue que sur le caractère spécifique du domaine d'application à traiter.

Ainsi, dans ce manuscrit, nous défendons l'idée qu'il est possible de rapprocher un corpus général de grande taille non étiqueté, avec un corpus spécifique de petite taille. Pour cela, nous faisons l'hypothèse qu'un corpus du domaine est enfoui dans le corpus général. Nous explorons quelques méthodes de classification automatique de document, et montrons qu'il est possible d'obtenir d'aussi bons résultats que si ce corpus avait été étiqueté manuellement.

De tels procédés permettent la réduction du coût de la collecte de corpus, et ainsi une conception rapide d'un modèle de langage pour un système de reconnaissance de la parole large vocabulaire.