Université du Maine
Faculté des Sciences et Techniques

Laboratoire d'Informatique
de l'Université du Maine

MÉMOIRE

présenté en vue d'obtenir

le Diplôme De Master en Informatique

Ingénierie des Systèmes Intelligents – Parcours Recherche

DÉCODAGE HYBRIDE DANS LES SRAP POUR L'INDEXATION AUTOMATIQUE DES DOCUMENTS MULTIMÉDIAS

Mohamed BOUAZIZ

Responsable du stage :
Monsieur Antoine LAURENT

Année Universitaire :
2012 - 2013

Table des matières

Table des figures

Liste des tableaux

Introduction

Durant les dernières décennies, l'information audio a pris une place importante parmi les interfaces de communication. Parallèlement à ce progrès, le traitement de ces grandes quantités de données est devenu indispensable. À cet égard, les Systèmes de Reconnaissance Automatique de la Parole (SRAP) ont réalisé des prouesses théoriques et technologiques considérables.

Selon [Srinivasan et al, 2002], le domaine d'application de la reconnaissance de la parole se base principalement sur deux axes. Le premier axe concerne l'utilisation de la parole comme entrée, notamment, pour les systèmes de dictée vocale, les systèmes de navigation, les applications à but commercial, etc. Le deuxième champ d'application présente la parole comme une source de données ou de connaissances. Les SRAP sont par exemple utilisés pour transcrire les documents multimédias mis à disposition sur le web afin d'améliorer les systèmes d'indexation de ce type de documents.

Dans ce dernier contexte, les SRAP se trouvent face à des données audio avec une grande diversité dans les conditions d'enregistrement, les langues et les accents. Par ailleurs, un des défis majeurs dans ce type de tâche est l'évolutivité du vocabulaire à traiter. En effet, le nombre de mots est manifestement loin d'être fini et ne peut pas être cerné au sein d'un lexique figé. Or, les SRAP, même ceux considérés comme « à large vocabulaire », modélisent une langue par le moyen d'un vocabulaire de taille fixe. Ces systèmes ne peuvent pas, en conséquence, couvrir la totalité des mots prononcés, et surtout dans le cadre d'une telle tâche.

Pour ce qui est de l'impact des mots hors-vocabulaire, ces derniers contribuent considérablement dans la détérioration de la performance d'un SRAP, non seulement parce que ces mots ne sont pas reconnus, mais aussi à cause de leur effet sur la reconnaissance des mots voisins. En effet, parmi les dix mots voisins d'un mot considéré comme hors-vocabulaire, moins de cinq mots, en moyenne, sont correctement reconnus [Dufour 2008]. Par conséquent, cette anomalie aura certainement une influence sur tout traitement potentiel (traduction automatique de la parole, indexation automatique de documents, etc.) [Bisani et al. 2005].

Dans ce qui suit, nous présentons les différents travaux qui s'articulent autour des finalités ci-dessus introduites. Ensuite, nous exposons notre démarche qui consiste à intégrer la solution du « décodage hybride » dans le SRAP développé au sein du Laboratoire d'Informatique de l'Université du Maine (LIUM) afin de remédier à la problématique des mots hors-vocabulaire. L'objectif est, à terme, de proposer des solutions pour réaliser un système utile à l'indexation des documents multimédias disponibles sur le web. Nous envisageons ainsi de nous focaliser sur la reconnaissance des entités nommées, et plus particulièrement, sur celle des noms propres.

Chapitre 1 : ***État de l'art***

1. Les SRAP

Aujourd'hui, la plupart des SRAP sont basés sur des méthodes statistiques permettant la modélisation acoustique et linguistique et la modélisation du vocabulaire de prononciation. La figure 1 schématise l'architecture d'un SRAP statistique.

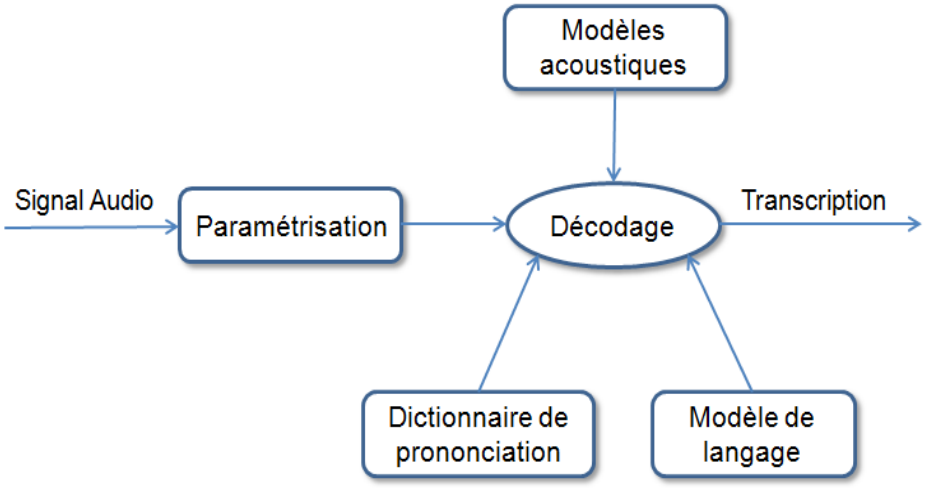

Figure 1 : Architecture d'un SRAP

En se basant sur les composants ci-dessus illustrés, le système de décodage détermine la transcription, autrement dit, la suite de mots, la plus probable en fonction des paramètres acoustiques en entrée. L'approche statistique consiste à chercher la séquence de mots W^* qui, partant d'une séquence d'observations acoustiques X, maximise l'équation suivante :

$$W^* = \arg\max_W P(W|X)$$

P(W|X) représente la probabilité d'émission d'une séquence de mots W sachant X. En appliquant le théorème de Bayes sur P(W|X), l'équation précédente devient :

$$W^* = \arg\max_W \frac{P(X|W)P(W)}{P(X)}$$

P(W) est la probabilité d'occurrence de la séquence de mots W et est estimée à partir du modèle de langage. P(X|W) représente la probabilité fournie par le modèle acoustique qui reflète la vraisemblance du signal de la parole étant donné une séquence de mots. Enfin, P(X) est considérée constante. L'équation peut ainsi être simplifiée et devient :

$$W^* = \arg\max_W P(X|W)P(W)$$

2. Décodage hybride

Dans la partie introductive, nous avons évoqué la cause principale de l'apparition des mots hors-vocabulaire, à savoir, la couverture partielle des mots d'une langue dans le cas des SRAP à large vocabulaire. Les SRAP à vocabulaire ouvert apportent une solution radicale à ce problème. Une des techniques du vocabulaire ouvert consiste à utiliser un modèle de langage basé sur des

unités sous-lexicale (syllabes, morphèmes, phonèmes...) et non pas sur les mots entiers de la langue concernée. En effet, contrairement aux mots qui semblent avoir un nombre infini (vu l'évolution continue des langues), le nombre de lettres est quasiment constant. La combinaison de cette solution avec l'approche classique, qui se base sur des modèles de langage à mots entiers, correspond au « décodage hybride ».

[Bisani et al. 2005] proposent une règle de décision probabiliste s'agissant de déterminer la suite de lettres *g,* de l'ensemble des lettres *G*, qui correspond au mieux à un ensemble de paramètres acoustiques *x*. Cette approche statistique est représentée à travers la formule suivante :

$$\boldsymbol{g}(\boldsymbol{x}) = \underset{\boldsymbol{g}'}{\operatorname{argmax}}\, p(\boldsymbol{g}') \max_{\boldsymbol{\varphi}} p(\boldsymbol{x}|\boldsymbol{\varphi}) p(\boldsymbol{\varphi}|\boldsymbol{g}')$$

Dans cette formule, $p(g)$ correspond au modèle de langage basé sur des unités sous-lexicale. $p(\varphi|\ g)$ correspond au modèle de prononciation qui attribut une suite de phonèmes φ, faisant partie de l'ensemble Φ^* (Φ étant un ensemble fini de phonèmes), à une suite de lettres g. Enfin, $p(x|\varphi)$ correspond au modèle acoustique reliant un ensemble de paramètres acoustiques x à une suite de phonèmes φ. Le modèle de langage et le modèle de prononciation peuvent être combinés au sein du modèle $p(q)$ où l'unité $q=(\varphi,\ g)$, baptisée « graphonème », combine un graphème g avec sa prononciation φ.

Les mots du vocabulaire sont ainsi décomposés en des unités sous-lexicale de taille maximale fixée a priori. Ces unités sont ensuite insérées au sein du même vocabulaire pour former une modélisation linguistique hybride.

En adoptant ces nouveaux concepts sur des données de la langue anglaise, [Bisani et al. 2005] réussissent à atteindre une réduction relative de 30% du taux d'erreur de mots pour des corpus d'évaluation comportant un taux de mots hors-vocabulaire supérieur à 10%. En outre, les auteurs ont montré que le nouveau système reconnaît correctement en moyenne un mot de plus, parmi les mots voisins du mot hors-vocabulaire, par rapport au système de reconnaissance de base (basé sur un modèle de langage à mots entiers).

Les travaux présentés dans [Shaik et al. 2011], représentent une extension du système construit par [Bisani et al. 2005]. En effet, la décomposition des mots n'est plus limitée aux graphonèmes qui doivent avoir un nombre maximal de lettres. Les mots du vocabulaire sont ainsi décomposés en syllabes ou en morphèmes en utilisant des outils spécialisés. Ce système opère sur des données en langue allemande et atteint une réduction relative de 5% en taux d'erreur de mots par rapport au système de base. Ce système reconnaît 40% des mots désignés comme hors-vocabulaire qui représentent 2,3% du corpus de test.

En appliquant des techniques d'apprentissage par unités sous-lexicale sur des données d'entraînement de petite taille de l'amharique, une langue peu-dotée, [Gelas et al. 2012] réussissent à reconnaître jusqu'à 75% des mots hors-vocabulaire en s'appuyant sur une méthode de reconstruction de mots à partir des sorties du décodage en morphèmes. Ce système est testé sur des données de test contenant 9% de mots hors-vocabulaire. Il arrive à réduire considérablement le taux d'erreur de mots de 49%.

La majorité des travaux s'intéressant à la reconnaissance des mots hors-vocabulaire adoptent les syllabes et/ou les morphèmes comme des unités sous-lexicale (par exemple [Zablotskiy et al. 2012] pour le russe, [Rotovnik et al. 2007] pour le slovène, [Gelas et al. 2012] pour le swahili, etc.). En revanche, certains travaux utilisent des unités sous-lexicale de taille plus petite. Par exemple, [Bazzi et al. 2000] présentent une stratégie de construction des mots hors-vocabulaire en partant

d'une suite de phonèmes. Pour ce faire, les auteurs utilisent un modèle de langage en bi-gramme afin de modéliser les contraintes phonotactiques. Avec un léger taux de fausse alarme, la moitié des mots hors-vocabulaire sont correctement reconnus.

Divers travaux s'intéressent ainsi au traitement de l'anomalie des mots hors-vocabulaire par le moyen des techniques du décodage hybride. Malgré le grand nombre de langues concernées, aucune tentative, au moins, à notre connaissance, n'est encore faite pour mettre en œuvre cette problématique dans le cadre d'une tâche de reconnaissance de la parole portant sur la langue française.

3. Indexation automatique des documents multimédias

Nous avons présenté, dans l'introduction, la possibilité d'avoir recours aux SRAP dans le but de raffiner les systèmes d'indexation des documents multimédias. Un des premiers systèmes utilisant un SRAP est SpeechBot. Il s'agit d'un moteur de recherche développé par [Logan et al. 1996] dont le système d'indexation se base sur la transcription automatique des documents audio. Malgré la faible qualité des transcriptions, le système arrive dans [Thong 2002] à satisfaire 77.5% des requêtes effectuées sur des pages web contenant des enregistrements relatifs à un ensemble d'émissions radio.

En ce qui concerne le traitement des mots hors-vocabulaire dans les SRAP appliqués à la tâche de l'indexation des documents multimédias, [Logan et al. 2002] étendent le système d'indexation du moteur de recherche SpeechBot en utilisant la technique du décodage hybride. En premier temps, les auteurs recourent à un SRAP dont le modèle de langage est appris sur une combinaison de mots entiers et d'unités sous-lexicale afin de transcrire les documents audio. En deuxième temps, au moment du test du système d'indexation basé sur les transcriptions automatiques, les mots des requêtes sont découpés selon l'unité sous-lexicale correspondante avant leur exécution. Deux types d'unités sous-lexicale sont utilisés, à savoir, les phonèmes et des unités semblables aux syllabes, appelées particules. L'utilisation de ces techniques, selon les mêmes auteurs, contribue à une légère amélioration dans la précision et le rappel des requêtes comportant des mots hors-vocabulaire, mais cause par contre une augmentation relativement importante dans le taux de fausses alarmes.

Une autre méthode faisant partie des techniques du « vocabulaire ouvert », proposée par [Allauzen et al. 2005], consiste à introduire dynamiquement de nouveaux mots au vocabulaire du système sans être amené à adapter le modèle de langage. Ces mots sont extraits à partir d'un ensemble de métadonnées relatives à des documents multimédias. Le nouveau système réussit non seulement à réduire le taux de mots hors-vocabulaire de 30% mais aussi à reconnaître 84% des entités nommées nouvellement introduites dans le vocabulaire.

*Chapitre 2 : **Approche proposée***

Nous avons exposé, dans la section précédente, un tour d'horizon sur divers travaux traitant de la problématique des mots hors-vocabulaire. L'intégration des techniques du vocabulaire ouvert semble être une solution prometteuse. En revanche, ces différents travaux ne reprennent pas cette solution de la même façon. En effet, les choix sont faits en fonction du domaine d'application, de la langue traitée, des ressources et outils techniques disponibles, etc. Nous présentons ainsi, dans cette section, , dans la conception de notre système, tout en expliquant les motivations correspondantes.

1. Un SRAP hybride

En intégrant la solution du décodage hybride, le dictionnaire de prononciation et la modélisation linguistique ne se restreignent plus sur l'unité « mot ». En effet, comme schématisé dans la figure 2, ces deux modèles se basent désormais sur des unités de deux catégories. La première catégorie représente les mots entiers tandis que des unités de taille plus petite constituent la deuxième catégorie.

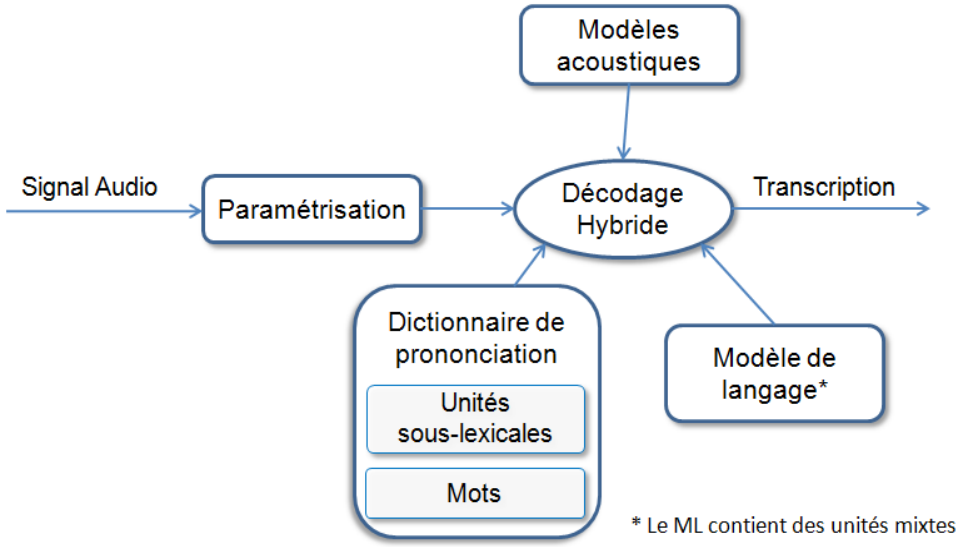

Figure 2 : Architecture d'un SRAP hybride

2. Choix des mots hors-vocabulaire

Nous nous intéressons, dans le cadre de ce travail, à la conception d'une solution remédiant au problème de la reconnaissance des mots hors-vocabulaire, ou au moins en réduisant l'impact. En effet, les mots considérés comme « hors-vocabulaire » peuvent provenir de plusieurs sources. D'une part, [Shaik et al. 2011] affirment que les mots hors-vocabulaire sont plus présents dans les langues morphologiquement riches. Dans des langues fortement flexionnelles, comme le français, l'arabe et l'allemand, la majorité des mots ne sont pas invariables. Ceci provoque une apparition plus fréquente des mots hors-vocabulaire. D'autre part, et particulièrement dans une tâche traitant de la reconnaissance de documents multimédias disponibles sur Internet, les données à traiter peuvent provenir de sources diverses (journaux parlés ou télévisés, débats d'actualité, vidéos d'amateurs...). Dans ce cas, le caractère dynamique du vocabulaire utilisé se manifeste par une émergence quasi-constante de nouveaux mots appartenant à la catégorie des entités nommées (noms propres, noms de villes, etc.). Cette catégorie constitue, selon [Réveil 2013], une grande partie des mots hors-

vocabulaire. Or, ce type de mots est d'une grande importance parmi les index sur lesquels se base un moteur de recherche. Dans une tentative de traitement des mots hors-vocabulaire dans le cadre de l'indexation automatique des documents audiovisuels, [Allauzen et al. 2005] a consacré un intérêt particulier aux entités nommées. Ainsi, nous nous concentrons au sein de ce travail aux entités nommées, et spécifiquement, aux noms propres. Enfin, nous envisageons que cette approche permettra au nouveau système de composer les noms propres hors-vocabulaire en combinant la séquence d'unités sous-lexicale correspondante.

3. Choix de l'unité sous-lexicale

À travers l'étude que nous avons abordée sur l'état de l'art des SRAP incorporant les techniques du vocabulaire ouvert, nous remarquons que le choix de l'unité sous-lexicale représente un élément décisif dans les résultats potentiels du décodage. Cela étant, les morphèmes et les syllabes sont bien les unités les plus utilisées dans la mise en œuvre de ces systèmes. Nous essayons, dans ce qui suit, de démontrer notre décision de choisir les syllabes en tant qu'unité sous-lexicale.

En ce qui concerne les morphèmes, ce choix participe bien à l'amélioration de la reconnaissance des mots hors-vocabulaire, notamment dans [Gelas et al. 2012] et [Shaik et al. 2011]. En revanche, cette solution est mise en œuvre afin de traiter plutôt de la richesse morphologique des langues flexionnelles ou, plus clairement, de permettre au SRAP de synthétiser les mots, qui disposent d'une large variabilité grammaticale, en partant d'un lemme et d'une combinaison de terminaisons grammaticales. La combinaison des différentes unités est assurée par les contraintes spécifiées au sein du modèle de langage appris sur un corpus de texte dont les mots sont découpés. Ainsi, et malgré ce que peut apporter cette stratégie dans le traitement des mots hors-vocabulaire, nous avons plutôt besoin d'une méthode qui sera, autant que possible, adaptée à la reconnaissance des entités nommées.

Pour ce qui est des phonèmes, ce choix s'annonce prometteur dans le sens où chaque langue possède un nombre fini de phonèmes. Cependant, deux défis se manifestent dans ce cas. Premièrement, la transcription de la parole produit une séquence de phonèmes. Pour avoir une transcription d'un signal sonore en mots, une deuxième phase est alors nécessaire. En effet, après le décodage, il faut reconstruire les graphèmes (suite de lettres produisant un phonème) à partir des séquences de phonèmes en sortie. Le deuxième défi qui se pose est le manque de contraintes au sein d'un modèle de langage appris sur des phonèmes vu la petite taille des unités sous-lexicale utilisées.

Partant de cette analyse, nous choisissons d'adopter les syllabes en tant qu'unités sous-lexicale. Nous envisageons ainsi à travers cette stratégie d'avoir des unités dont la combinaison est susceptible de former les mots à retrouver tout en assurant un nombre important de contraintes qui ne diffèrent pas beaucoup de celles dont dispose un modèle de langage appris à partir de mots entiers.

Chapitre 3 : ***Expérimentations***

Dans la section précédente, nous avons présenté notre approche ainsi que les choix qui ont accompagné son élaboration. Afin de mettre en œuvre nos idées, nous nous basons sur le SRAP du LIUM, un système de l'état de l'art qui participe régulièrement à des compagnes d'évaluation internationales. Nous présentons dans ce chapitre ce système et nous exposons plus en détails le protocole expérimental et les résultats relatifs à l'intégration de notre approche.

1. SRAP de base (LIUM'08)

Le SRAP développé par le LIUM [Deléglise et al. 2009] est un système de l'état de l'art basé sur le système CMU Sphinx. Au sein du LIUM, diverses extensions et améliorations y ont été apportées en employant des technologies de pointe qui lui ont permis d'être le meilleur SRAP open-source dans la campagne d'évaluation ESTER 2. Cette dernière a été organisée par la Délégation Générale de l'Armement et l'Association Francophone de la Communication Parlée pendant les années 2007 et 2008. Les étapes de construction du SRAP LIUM'08 est présentée en détail dans la figure 3 [Estève 2009].

1.1. Apprentissage

Concernant le niveau acoustique, en appliquant une paramétrisation de type PLP sur le signal audio, le SRAP du LIUM utilise 39 paramètres acoustiques composés de descripteurs issus de l'analyse PLP, l'énergie, les dérivées premières et les dérivés secondes de ces descripteurs. Les modèles acoustiques sont appris sur des enregistrements audio transcrits dont 240h proviennent d'ESTER 1 et ESTER 2 et 40h d'émissions radio du projet EPAC. Le premier apprentissage acoustique produit des modèles qui dépendent du type du canal et qui sont ensuite adaptés au genre du locuteur par la méthode Maximum A Posteriori (MAP).

Pour ce qui est du niveau linguistique, le modèle de langage est appris à partir de trois types de sources, à savoir, des transcriptions manuelles (des enregistrements utilisés au niveau acoustique et des conversations provenant du corpus PFC [Durand 2002]), des articles de journaux (Le Monde, L'Humanité et les données du French Giga Word Corpus) et des données extraits d'un nombre de sites Internet (L'internaute, Libération, Rue89 et Afrik.com). En se basant sur un vocabulaire de 122k mots, un modèle 3-gram et un modèle 4-gram sont appris pour chacun desdits corpus. L'apprentissage de ces modèles est effectué en recourant à la méthode de lissage par interpolation avec le discounting Kneser-Ney [Kneser et al. 1995, Chen et al. 1999] à l'aide de la boîte à outils SRILM [Stolcke 2002]. Ensuite, tous les modèles sont interpolés suivant des coefficients calculés en fonction du corpus de développement relatif à la tâche concernée. Le SRAP du LIUM dispose ainsi d'un modèle de langage doté de 121k 1-grams. 29M 2-grams, 162M 3-grams et 376M 4-grams.

Assurant le lien entre la modélisation acoustique et lexicale, le dictionnaire phonétique fonctionne avec un ensemble de 35 phonèmes et 5 types de fillers. Les séquences de phonèmes des mots du vocabulaire sont extraits du dictionnaire de phonétisation BDLEX [Perennou et al. 1987]. Les mots absents dudit dictionnaire sont générés par l'outil de phonétisation LIA_PHON [Bechet 2001].

1.2. Segmentation en locuteurs

Le système de segmentation développé par le LIUM se base sur le Critère d'Information Bayésien. Le processus de segmentation adopté consiste à découper le signal acoustique en des segments, regrouper les segments les plus homogènes par le moyen d'une classification hiérarchique et enfin ajuster les frontières qui séparent les différents segments.

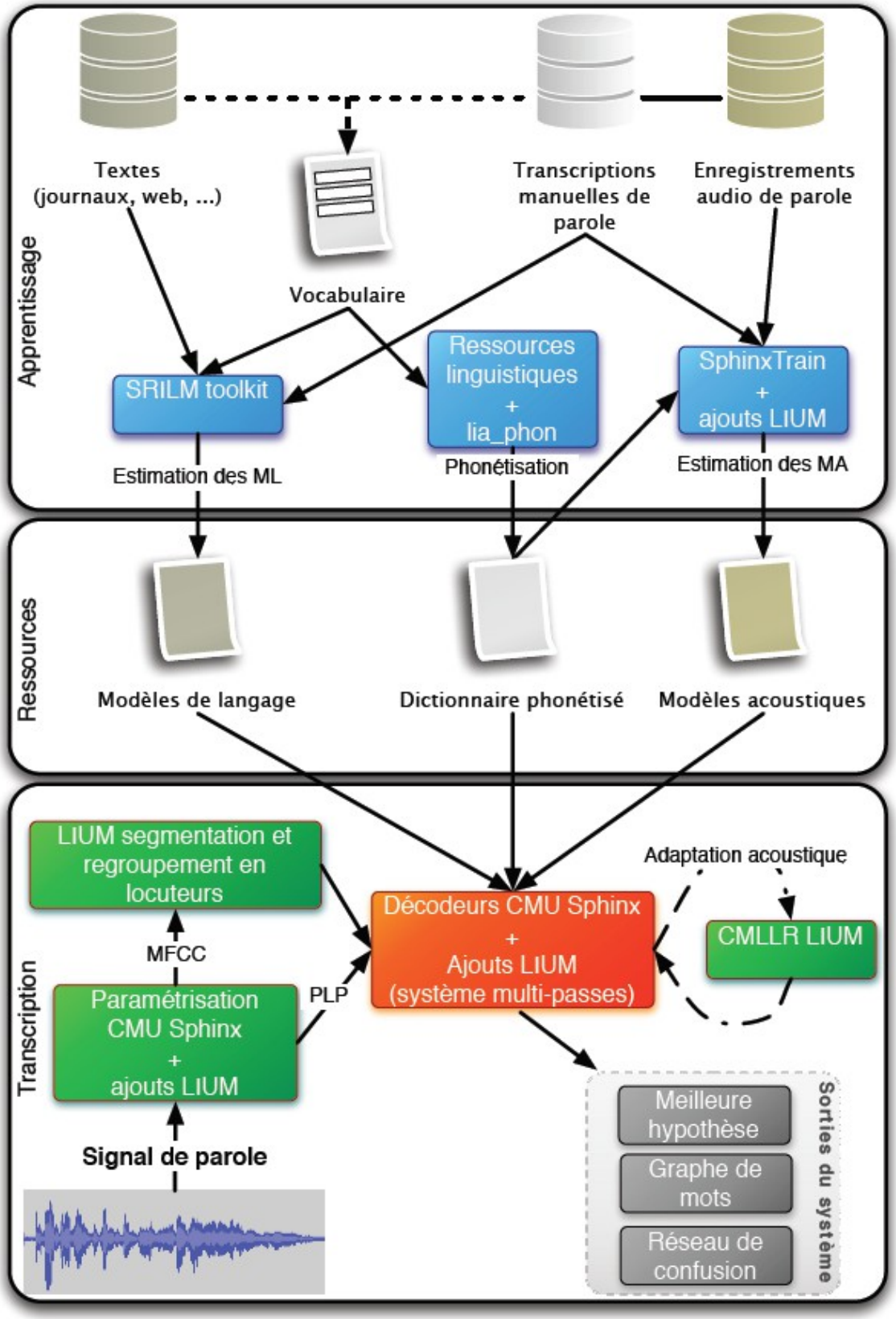

Figure 3 : Architecture détaillée du SRAP LIUM'08

Ce système, détaillé davantage dans [Meignier 2010], a obtenu le meilleur taux d'erreur lors de la compagne d'évaluation ESTER 2.

1.3. Décodage

Le LIUM utilise, pour transcrire le signal de la parole, un processus de décodage sous forme d'une succession de 5 passes. Les trois premières passes se basent sur la version 3.7 du décodeur Sphinx. Une transformation CMLLR est calculée lors de la première passe avec un modèle de langage 3-gram. Un traitement similaire est effectué pour la deuxième passe en se basant sur les méthodes Speaker Adaptive Training (SAT) et Minimum Phone Error (MPE). En partant du graphe de mots généré par la deuxième passe, la troisième passe tend à améliorer la précision acoustique et produit un autre graphe. Les scores linguistiques des mots du graphe sont recalculés à l'aide d'un modèle de langage 4-gram dans la passe 4. Enfin, la dernière passe consiste à transformer le graphe de mots résultant en un réseau de confusions et à appliquer à ce dernier la méthode de consensus [Mangu et al., 2000] afin d'obtenir l'hypothèse finale.

2. Décodage hybride

Après avoir choisi les syllabes en tant qu'unités sous-lexicale, nous passons à la mise en œuvre de notre approche. Notre mission consiste à adapter le SRAP développé au sein du LIUM afin qu'il soit capable de reconnaître les mots qui sont considérés comme une des sources des mots hors vocabulaire, à savoir, les noms propres. Nous présentons, dans cette section, les détails relatifs à la préparation du nouveau système.

2.1. Données d'apprentissage

Tout d'abord, nous cherchons à opérer sur des données qui soient les plus adaptées à notre finalité. Nous choisissons ainsi de tirer profit des données du corpus « REPERE » [Giraudel et al. 2012]. Financé par l'Agence Nationale de la Recherche (ANR) et par la Direction Générale de l'Armement (DGA), le projet REPERE représente un défi ayant comme objectif d'encourager le développement de systèmes automatiques pour la reconnaissance de personnes dans un contexte multimodal. Ce défi se structure en des campagnes d'évaluation annuelles, planifiées entre 2012 et 2014, organisées à la base de 60h de données vidéo de la langue française, collectées et traitées pour ladite fin.

Le corpus REPERE représente des données relatives à des émissions télévisées diffusées sur les chaînes BFM TV et LCP [Site 1]. La répartition de ces données, entre les phases d'apprentissage, d'optimisation et de test est illustrée dans le tableau 1. Le corpus provient de 7 programmes télévisés. Le tableau 2 liste les différentes émissions constituant un extrait de 3 heures de corpus.

Phase	**Volume**
Apprentissage (R_train)	42 h
Développement (R_dev)	9 h
Test (R_test)	9 h

Tableau 1 : Répartition du corpus REPERE entre les différentes phases [site 1]

Étant conçu dans le but d'évaluer les systèmes de reconnaissance des noms de personnes, les segments de parole du corpus REPERE sont fournis avec des transcriptions enrichies par une annotation des noms propres. Afin d'adapter notre système, nous tirons parti de ces annotations pour apprendre, optimiser et évaluer les nouveaux modèles.

Émission	Chaîne	Durée
Ça vous regarde	LCP	15min
Entre les lignes	LCP	15min
Pile et Face	LCP	15min
Top Questions	LCP	30min
LCP INFO	LCP	30min
Planète Showbiz	BFM	15min
BFM Story	BFM	60min

Tableau 2 : Liste des émissions pour 3 heures du corpus REPERE [Giraudel et al. 2012]

2.2. Apprentissage

Pour ce qui est de l'apprentissage du modèle de langage, nous découpons les noms propres en syllabes, au niveau des données d'apprentissage extraites du corpus REPERE, par le moyen de la fonction de syllabation fournie par l'outil LIA_PHON. Le corpus R_train compte 7947 occurrences relatives à 1225 noms propres. Le découpage en syllabes de ces mots produit 1074 syllabes différentes.

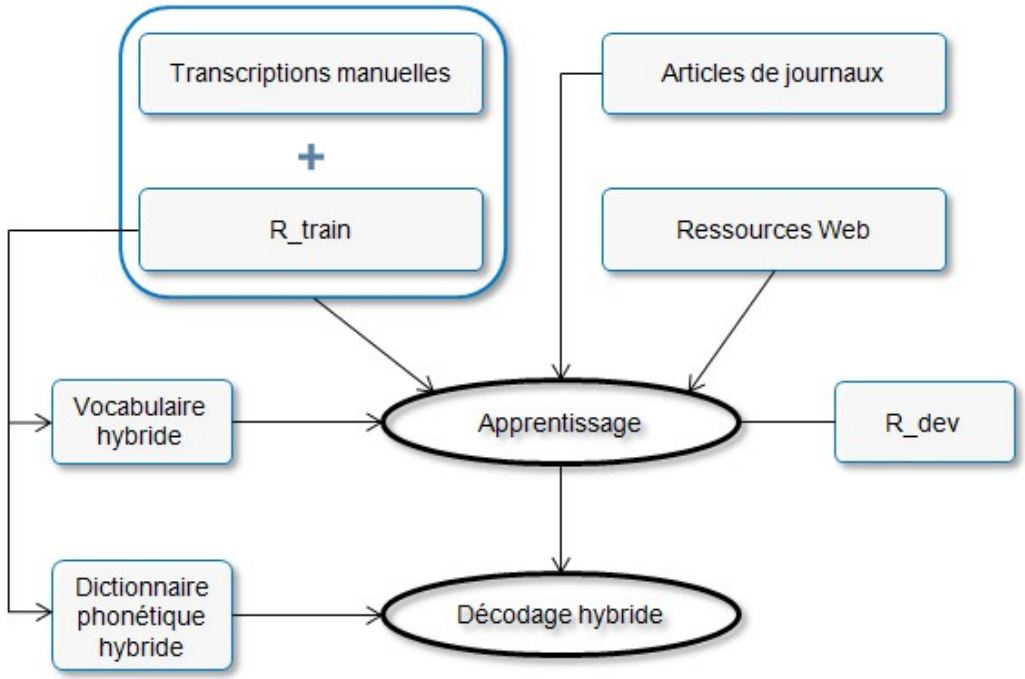

Figure 4 : Préparation du protocole expérimental

Comme schématisé dans la figure 4, nous conservons les mêmes données qui ont servi à apprendre le modèle de langage du système LIUM'08. En outre, nous concaténons le R_train, dont les noms propres sont découpés en syllabes, au corpus qui contient les transcriptions manuelles. Nous introduisons ensuite les nouveaux mots provenant du R_train ainsi que les syllabes générées dans le vocabulaire d'apprentissage du système. Afin d'optimiser le modèle, nous découpons de la même manière le corpus de développement R_dev. Enfin, nous reprenons la même stratégie d'apprentissage du système de base, avec les nouvelles données, et nous procédons à une optimisation du modèle par rapport au corpus R_dev.

Une deuxième expérience, à laquelle nous recourons, consiste à marquer les syllabes résultant du découpage des noms propres, par l'étiquette « SyllEtiqu », dans R_train et R_dev. En effet, nous envisageons par cette technique que notre système soit plus apte à faire une distinction entre les mots et les syllabes qui ont une même orthographe. Ces syllabes sont intégrées de la même manière au sein du vocabulaire. La taille du vocabulaire dans toutes les expériences que nous mettons en œuvre ne dépasse pas 124k mots. Partant du calcul de perplexité des données d'apprentissage sur le corpus R_dev, les poids sont attribués au différents types de corpus d'apprentissage par la boite à outils SRILM. Le tableau 3 illustre les coefficients d'interpolation pour chacune des deux stratégies d'apprentissage abordées.

Apprentissage du modèle de langage	**Transcriptions + R_train**		**Articles de journaux**		**Web**	
	3g	4g	3g	4g	3g	4g
Apprentissage (syllabes non étiquetées)	0,57	0,55	0,42	0,44	0,01	0,01
Apprentissage (syllabes étiquetées)	0,51	0,48	0,48	0,49	0,01	0,01

Tableau 3 : Calcul des poids relatifs aux données servant à l'apprentissage du modèle de langage

Selon le tableau 3, Le corpus comportant les transcriptions manuelles avec le R_train détient, dans la majorité des cas, le coefficient le plus important. Ceci peut s'expliquer par le fait que les données R_train sont les plus semblables aux données de développement R_dev.

2.3. Phonétisation

En ce qui concerne la phonétisation des nouvelles entrées du vocabulaire, nous utilisons l'outil de phonétisation LIA_PHON [Bechet 2001]. Les mots entiers ainsi que les syllabes non étiquetées, c-à-d, celles du premier système (S1), sont phonétisés d'une manière classique. Toutes les variantes de prononciation produites par LIA_PHON sont prises en considération. Les étiquettes des syllabes employées dans le système appris au cours de la deuxième expérience (S2) sont temporairement enlevées pour effectuer la phonétisation automatique. Elles sont ensuite reprises pendant l'intégration des syllabes correspondantes dans le nouveau dictionnaire phonétique.

2.4. Évaluation

2.4.1. Métriques d'évaluation

La dernière étape consiste à concevoir une stratégie afin d'évaluer les sorties du nouveau

système. Ce travail a pour finalité d'améliorer le rendement des systèmes d'indexation automatique des documents multimédias par rapport aux noms propres hors vocabulaire. En effet, nous nous intéressons principalement aux requêtes, saisies par l'utilisateur, qui comportent un ou plusieurs noms propres. En revanche, nous considérons le cas où un moteur de recherche, disposant d'un tel système, n'est pas apte à détecter les noms propres parmi les mots de la requête. Ainsi, nous supposons que le moteur de recherche, avant l'exécution, découpe tous les mots de la requête en syllabes et effectue la recherche sur les transcriptions en sortie du SRAP dédié. Dès lors, nous procédons à une stratégie d'évaluation particulière. D'une part, nous préparons la transcription de référence des données de test (R_test), fournie en format stm[1] (Segment Time Mark), en découpant chaque mot en syllabes. D'autre part, nous procédons à un découpage en syllabes des hypothèses de transcription en tenant compte de la particularité des fichiers d'hypothèse produits en format ctm[1] (Time Marked Conversation). Ces formats, ainsi que les évaluations effectuées, se reposent sur le moteur d'alignement « sclite »[2] (Score-Lite) fourni par NIST (National Institute of Standards and Technology). En adoptant ces considérations, la problématique de reconstruction des mots à partir des séquences de syllabes ne fait donc pas partie de nos préoccupations.

Les données de test comptent 279 noms propres différents. Parmi ces mots, 27 sont originellement hors vocabulaire. Par ailleurs, 13 parmi les noms propres de R_test n'existent pas dans le vocabulaire du nouveau système mais sont présents en syllabes dans le corpus R_train. Le reste des noms propres de R_test sont artificiellement mis hors vocabulaire.

2.4.2. Résultats

Partant du fait que tous les noms propres évaluées n'appartiennent pas au vocabulaire du système, nous considérons que le taux de reconnaissance des noms propres hors-vocabulaire (HV) (nombre de noms propres HV dont toutes les syllabes sont correctement reconnues / nombre de noms propres HV) dans le système de base est nul. Ainsi, nous ne nous intéressons pas non plus aux taux de reconnaissance de syllabes au sein de ces mêmes noms propres (nombre de syllabes correctement reconnues dans l'ensemble des noms propres HV / nombre total de syllabes dans l'ensemble des noms propres HV) pour ledit système. Le tableau 4 récapitule les résultats obtenus en ce qui concerne les taux de reconnaissance ou d'erreur de syllabes pour le système de base ainsi que pour les deux systèmes reposant sur la stratégie du décodage hybride.

Système	TR de noms propres HV	TR de syllabes dans les noms propres HV	SER
Système de base (S0)	0%	-	18,70%
Système avec syllabes (S1)	**31,39%**	42,82%	**19,90%**
Système avec syllabes étiquetées (S2)	31,10%	**42,95%**	24,90%

Tableau 4 : Résultats de reconnaissance (TR : taux de reconnaissance, HV : hors-vocabulaire, SER : Taux d'erreur de syllabes)

Dès lors, comme décrit dans le tableau 1, le système S1, appris sur des noms propres en syllabes, réussit à reconnaître 31,39% des noms propres et 42,82% des syllabes dans l'ensemble de ces noms propres. En revanche, il cause une augmentation, de 1,2% en absolue, du taux d'erreur général de syllabes (SER). Cette légère dégradation de la performance générale du système est

1 http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/infmts.htm
2 http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm

négligeable par rapport au gain que nous obtenons par rapport à la reconnaissance des noms propres, qui sont d'une grande importance pour les systèmes d'indexation.

Quant au système S2, avec une augmentation remarquable dans le SER et une légère diminution du taux de reconnaissance de noms propres, ce système réussit tout de même à obtenir un meilleur taux de reconnaissance de syllabes dans les noms propres (42,95%).

Enfin, nous n'avons pas eu la possibilité de comparer les résultats issus de la mise en œuvre abordée du décodage hybride avec d'autres travaux s'intéressant à la reconnaissance des mots hors-vocabulaire. En effet, à notre connaissance, aucun des travaux de l'état de l'art n'a adopté une stratégie d'évaluation se focalisant sur l'étude de la performance des SRAP par rapport à la reconnaissance des syllabes au sein des noms propres hors-vocabulaire.

Conclusion

Prenant en compte les particularités des systèmes d'indexation automatique des documents multimédias, ce travail constitue, à notre connaissance, un des premiers travaux qui essayent de pallier le problème des mots hors-vocabulaire par le moyen du décodage hybride au sein des SRAP traitant de la langue française. Malgré la diversité et l'importance des choix à faire, nous avons réussi à atteindre un taux considérable de reconnaissance de noms propres hors-vocabulaire.

Ces résultats sont intéressants pour deux raisons principales. D'une part, l'énonciation de mots appartenant à la catégorie des entités nommées, dans les données multimédias disponibles sur le web, est une des causes fondamentales de l'apparition des mots hors-vocabulaire. D'autre part, cette catégorie de mots a bien une grande importance pour les systèmes d'indexation traitant de ces données.

Enfin, l'approche proposée, d'un côté, est en faveur du domaine pris en compte au sein de ce travail, et, d'un autre côté, peut être au profit d'autres champs, notamment, la traduction automatique de la parole. Il est envisagé, dans la continuité de ce travail, d'étendre notre intérêt afin de prendre en considération les autres sous-catégories des entités nommées, et non seulement les noms propres.

Bibliographie & Webographie

[Allauzen et al. 2005] A. Allauzen, J. Gauvain. Open Vocabulary ASR for Audiovisual Document Indexation. Proceedings of the ICASSP, vol 1, p1013-1016, 2005.

[Bazzi et al. 2000] I. Bazzi and J. Glass. Modelling out-of-vocabulary words for robust speech recognition. Proc. ICSLP, Pékin, 2000.

[Bechet 2001] F. Bechet. LIA_PHON - Un système complet de phonétisation de textes. Revue Traitement Automatique des Langues (TAL), vol 42, n°1, 2001.

[Bisani et al. 2005] M. Bisani et H. Ney. Open vocabulary speech recognition with flat hybrid models. Interspeech, Lisbone, Portugal, p725-728, Septembre. 2005.

[Chen et al. 1999] S. F. Chen, J. Goodman. An empirical study of smoothing techniques for language modeling. In Computer Speech and Language. p359-394, 1999.

[Deléglise et al. 2009] P. Deléglise, Y. Estève, S. Meignier, T. Merlin. Improvements to the LIUM French ASR system based on CMU sphinx: what helps to significantly reduce the word error rate? Interspeech, p2123–2126, Brighton UK, 2009.

[Dufour 2008] R. Dufour. From prepared speech to spontaneous speech recognition system: a comparative study applied to French language. IEEE/ACM CSTST, Cergy, France, Octobre 2008.

[Durand 2002] J. Durand, B. Laks, C. Lyche. La phonologie du français contemporain : usages, variétés et structure. Romanistische Korpuslinguistik- Korpora und gesprochene Sprache/Romance Corpus Linguistics - Corpora and Spoken Language, p93–106, 2002.

[Estève 2009] Yannick Estève. Traitement automatique de la parole : contributions. Habilitation à Diriger des Recherches (HDR), LIUM, Université du Maine, 2009.

[Gelas et al. 2012] Hadrien Gelas, Solomon Teferra Abate, Laurent Besacier, François Pellegrino. Analyse des performances de modèles de langage sub-lexicale pour des langues peu-dotées à morphologie riche. JEP-TALN-RECITAL 2012, Atelier TALAf 2012 : Traitement Automatique des Langues Africaines, p53-62, Grenoble, France, juin 2012.

[Giraudel et al. 2012] A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, L. Quintard. The REPERE Corpus: a multimodal corpus for person recognition. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, may 2012.

[Kneser et al. 1995] R. Kneser, H. Ney. Improved backing-off for n-gram language modeling. ICASSP, 1995.

[Logan et al. 1996] B. Logan, P. Moreno, J-M. Van Thong, E. Whittaker. An Experimental Study Of An Audio Indexing System For The Web. Proc. ICSLP, p676-679, 1996.

[Logan et al. 2002] B. Logan, P. Moreno, O. Deshmukh. Word and sub-word indexing approaches for reducing the effects of OOV queries on spoken audio. Proceedings of the second international conference on Human Language Technology Research (HLT '02). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, p31-35. 2002.

[Mangu et al., 2000] H. Mangu, E. Brill, A. Stolcke. Finding consensus in speech recognition:

Word error minimization and other applications of confusion networks. Computer Speech and Language, vol. 14, n°4, p373-400, 2000.

[Meignier et al. 2010] S. Meignier, T. Merlin. LIUM SPKDIARIZATION: AN OPEN SOURCE TOOLKIT FOR DIARIZATION. CMU SPUD Workshop, Dallas, Texas, USA, 2010.

[Perennou 1987] G. Perennou, M. de Calmès. BDLEX lexical data and knowledge base of spoken and written French. European Conference on Speech Technology, 1987.

[Réveil 2013] B. Réveil, K, Demuynck, J.-P. Martens. An improved two-stage mixed language model approach for handling out-of-vocabulary words in large vocabulary continuous speech recognition. Comput. Speech Lang. 2013.

[Rotovnik et al. 2007] T. Rotovnik, M. S. Maučec, Z. Kačič. Large vocabulary continuous speech recognition of an inflected language using stems and endings. Speech Communication, Volume 49, Issue 6, Juin 2007, p437-452.

[Shaik et al. 2011] M. Shaik, A. El-Desoky, R. Schlüter, et H. Ney. Hybrid Language Models Using Mixed Types of Sub-lexical Units for Open Vocabulary German LVCSR. Interspeech, Florence, Italie, Août 2011.

[Srinivasan et al. 2002] S. Srinivasan, E. Brown. Is speech recognition becoming mainstream?. Computer 35, no.4, p38-41, avril 2002.

[Stolcke 2002] Andreas Stolcke. SRILM-An extensible language modeling toolkit. ICSLP, volume 2, p901-904, Denver, Colorado, USA, 2002.

[Thong, 2002] J-M. Van Thong, P. Moreno, B. Logan, B. Fidler, K. Maffey, M. Moores. Speechbot: an experimental speech-based search engine for multimedia content on the web. Multimedia, IEEE Transactions, vol.4, no.1, p88-96, Mars 2002.

[Zablotskiy et al. 2012] S. Zablotskiy, A. Shvets, M. Sidorov, E. Semenkin, Wolfgang Minker. Speech and Language Ressources for LVCSR of Russian. Proceedings of the Eight International Conference on Language Resources and Evaluation, May 2012.

[Site 1] http://www.defi-repere.fr/