Décodage hybride dans les SRAP pour l'indexation automatique des documents multimédia

Mohamed BOUAZIZ[1, 2] Antoine LAURENT[1, 3] Yannick ESTEVE[1]
(1) LIUM, Le Mans, France
(2) LIA, Avignon, France
(3) LIMSI-CNRS, Orsay, France
mohamed.bouaziz@alumni.univ-avignon.fr, antoine.laurent@limsi.fr,
yannick.esteve@lium.univ-lemans.fr

RÉSUMÉ
Certains Systèmes de Reconnaissance Automatique de la Parole (SRAP) atteignent des taux d'erreur de l'ordre de 10%. Toutefois, notamment dans le cadre de l'indexation automatique des documents multimédia sur le web, les SRAP se trouvent face à la problématique des mots hors-vocabulaire. En effet, les entités nommées en constituent une grande partie et sont remarquablement importantes pour les tâches d'indexation. Nous mettons en œuvre, dans ce travail, la solution du décodage hybride en utilisant les syllabes comme unités sous-lexicales. Cette méthode est intégrée au sein du SRAP LIUM'08 développé par le Laboratoire d'Informatique de l'Université du Maine. Avec une légère dégradation de la performance générale du système, environ 31% des noms de personne hors vocabulaire sont correctement reconnus.

ABSTRACT
Hybrid decoding in ASR systems for multimedia documents auto-indexing

Some state of the art Automatic Speech Recognition (ASR) systems reach even about 10% in word error rate. Yet, these systems suffer from out-of-vocabulary (OOV) words particularly in the task of auto-indexing of multimedia documents on the web. Indeed, not only do named entities take a big part of OOVs, but they are also of significance in multimedia documents indexing. In this work, we implement a syllable-based hybrid decoding approach within the LIUM'08 ASR system. Despite a small decrease in the system's general performance, about 31% of proper nouns are correctly recognized.

MOTS-CLÉS : mot hors-vocabulaire, Indexation automatique, vocabulaire ouvert, décodage hybride, unité sous-lexicale.

KEYWORDS: out-of-vocabulary, auto-indexing, open vocabulary, hybrid decoding, sub-lexical unit.

1 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. Selon (Brown, 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 multimédia 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 au problème des mots hors-vocabulaire à cause du changement continu du vocabulaire à traiter. En effet, les SRAP 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.

Les mots hors-vocabulaire sont en partie responsables de la dégradation de la performance d'un SRAP. Non seulement ces mots ne sont pas reconnus, mais aussi ils influent négativement sur la reconnaissance des mots voisins. En effet, parmi les dix mots voisins d'un mot hors-vocabulaire, moins de cinq mots, en moyenne, sont correctement reconnus (Dufour, 2008). Par conséquent, cette anomalie aura une influence sur tout traitement potentiel (traduction automatique de la parole, indexation automatique de documents, etc.) (Bisani et Ney, 2005).

Dans les deux sections suivantes, nous présentons les différents travaux qui s'articulent autour des finalités ci-dessus introduites. Ensuite, nous exposons, dans les sections 4 et 5, notre contribution 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). Enfin, la section 6 est réservée pour la présentation des résultats de la mise en œuvre de notre contribution visant à remédier à la problématique des mots hors-vocabulaire.

2 Décodage hybride

Les SRAP à vocabulaire ouvert permettent de pallier la couverture partielle des mots d'une langue. Une des techniques du vocabulaire ouvert consiste à utiliser un modèle de langage basé sur des unités sous-lexicales (syllabes, morphèmes, phonèmes,...) et non pas sur des mots entiers. La combinaison de cette solution avec l'approche classique correspond au « décodage hybride ».

(Bisani et Ney, 2005) décomposent les mots du vocabulaire en des unités sous-lexicales dénommées « graphonèmes » qui combinent un graphème avec sa prononciation. 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, les auteurs 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, (Bisani et Ney, 2005) 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. Les travaux présentés dans (Shaik *et al.*, 2011) représentent une extension du système construit par (Bisani et Ney, 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-lexicales 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. Ce système est testé sur des données de test contenant 9% de mots hors-vocabulaire et arrive à réduire de 50,3% le taux d'erreur de mots.

La majorité des travaux s'intéressant à la reconnaissance des mots hors-vocabulaire adoptent les syllabes et/ou les morphèmes comme des unités sous-lexicales (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-lexicales de taille plus petite. Par exemple, (Bazzi et Glass, 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 des mots hors-vocabulaire par le moyen du décodage hybride. Malgré le grand nombre de langues concernées, aucune tentative, au moins à notre connaissance, n'a encore mis en œuvre cette solution dans le cadre de la langue française.

3 Indexation automatique des documents multimédia

Nous avons évoqué, dans l'introduction, la possibilité d'avoir recours aux SRAP dans le but de raffiner les systèmes d'indexation des documents multimédia. Un des premiers essais dans ce contexte est SpeechBot, 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 selon (Van Thong *et al.*, 2002) à satisfaire 77.5% des requêtes effectuées sur des pages web contenant des émissions radio. En ce qui concerne le traitement des mots hors-vocabulaire, (Logan *et al.*, 2002) étendent leur système en utilisant la technique du décodage hybride. Dans un premier temps, les documents audio sont transcrits par un SRAP dont le modèle de langage est appris sur une combinaison de mots entiers et d'unités sous-lexicales. Ensuite, les mots des requêtes sont découpés avant leur exécution. Les phonèmes et des unités semblables aux syllabes, dénommés particules, sont utilisés comme unités sous-lexicales. 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. En revanche, elle cause une augmentation relativement importante dans le taux de fausses alarmes.

Dans un contexte similaire, une autre tentative de traitement des mots hors-vocabulaire, décrite dans (Allauzen et Gauvain, 2005), consiste à introduire dynamiquement de nouveaux mots au vocabulaire du système sans adapter le modèle de langage. Ces mots sont extraits à partir d'un ensemble de métadonnées relatives à des documents multimédia. 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.

4 Méthodologie

Les travaux exposés ci-dessus ne reprennent pas la solution du décodage hybride de la même façon. Nous présentons, dans cette section, les principales décisions que nous étions amenés à prendre dans la conception de notre système.

4.1 Choix des mots hors-vocabulaire

L'apparition des mots hors-vocabulaire peut provenir de plusieurs facteurs. D'une part, (Shaik *et al.*, 2011) affirment que ces mots apparaissent plus souvent dans le traitement des langues morphologiquement riches. Par exemple, dans des langues comme le français, l'arabe et l'allemand, la majorité des mots possèdent une grande variabilité. D'autre part, et particulièrement dans une tâche traitant de la reconnaissance de documents multimédia du web, les données à traiter peuvent provenir de diverses sources (journaux parlés ou télévisés, débats d'actualité, vidéos d'amateurs,...). Dans ce cas, le caractère dynamique du vocabulaire utilisé se manifeste par une émergence continue de nouveaux mots appartenant à la catégorie des entités nommées (noms propres, noms de villes, etc.). Cette catégorie constitue, selon (Réveil *et al.*, 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. Ainsi, nous nous concentrons au sein de ce travail aux entités nommées, et spécifiquement, aux noms de personne. Enfin, nous envisageons que cette approche permettra au nouveau système de composer les noms de personne hors-vocabulaire en combinant la séquence d'unités sous-lexicales correspondante.

4.2 Choix de l'unité sous-lexicale

À travers l'étude de l'état de l'art que nous avons abordée sur l'utilisation du décodage hybride, 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 les unités les plus utilisées dans ce contexte.

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. En effet, elle permet au SRAP de synthétiser plus efficacement les mots qui disposent d'une large variabilité grammaticale en partant d'un lemme et d'une combinaison de terminaisons grammaticales. Ainsi, 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 problèmes 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. En outre, un deuxième souci se manifeste à cause du manque de contraintes au sein d'un modèle de langage appris sur des phonèmes vu la petite taille des unités sous-lexicales utilisées.

Partant de cette analyse, nous choisissons d'adopter les syllabes en tant qu'unités sous-lexicales. 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 suffisant de contraintes.

5 Protocole expérimental

Afin de mettre en œuvre nos idées, nous partons du SRAP du LIUM (Deléglise *et al.*, 2009), le meilleur SRAP open-source dans la campagne d'évaluation ESTER 2, basé sur le système CMU Sphinx. Les différents modèles du LIUM'08 sont appris à partir de données provenant de plusieurs origines (ESTER1, ESTER2, EPAC, articles de journaux, sites internet). Ce système utilise 39 paramètres acoustiques issus d'une paramétrisation de type PLP. Les modèles acoustiques dépendent du type du canal et sont adaptés au genre du locuteur par la méthode Maximum A Posteriori (MAP). En se basant sur un vocabulaire de 122k mots, le modèle de langage est appris à l'aide de la boîte à outils SRILM (Stolcke *et al.*, 2002) et en recourant à la méthode de discounting Kneser-Ney (Kneser et Ney, 1995). Avec 35 types de phonèmes et 5 types de fillers, le dictionnaire phonétique est construit à l'aide du dictionnaire de phonétisation BDLEX (Perennou et Calmes, 1987) et l'outil de phonétisation LIA_PHON (Béchet, 2001). Le système de segmentation (Meignier et Merlin, 2010), basé sur le Critère d'Information Bayésien, a obtenu le meilleur taux d'erreur lors de la compagne d'évaluation ESTER 2. Enfin, un processus de décodage en 5 passes est utilisé pour transcrire le signal de la parole.

Après avoir choisi les syllabes en tant qu'unités sous-lexicales, nous passons à la mise en œuvre de notre approche. Étant conçu dans le but d'évaluer les systèmes de reconnaissance des noms de personne, les transcriptions du corpus REPERE [1] sont enrichies par une annotation de ces noms. Ainsi, nous tirons parti de ces annotations pour apprendre, optimiser et évaluer les nouveaux modèles. Les données correspondent à 7 programmes télévisés diffusés sur les chaînes BFM TV et LCP. 42 heures du corpus REPERE servent à l'apprentissage (R_train), 9 heures à l'optimisation (R_dev) et 9 heures au test (R_test).

Concernant le système de base, nous faisons rejoindre le R_train au corpus d'apprentissage et nous introduisons les nouveaux mots qu'il possède dans le vocabulaire dudit système. En revanche, pour ce qui est de l'apprentissage du modèle de langage hybride, nous découpons les noms de personne en syllabes au niveau de R_train, par le moyen de la fonction de syllabation fournie par l'outil LIA_PHON. Ce corpus compte 7947 occurrences relatives à 1225 noms de personne et le découpage en syllabes de ces mots produit 1074 syllabes différentes. Nous désignons par "nom de personne", dans notre protocole expérimental, un nom ou un prénom d'une personne. Ensuite, nous faisons rejoindre le nouveau R_train au corpus d'apprentissage et nous introduisons les syllabes générées, dans le vocabulaire du système de base. Afin d'optimiser le modèle, nous utilisons le corpus de développement R_dev après y avoir découpé, de la même manière, les noms de personne. Une deuxième expérience, à laquelle nous recourons, consiste à marquer les syllabes résultant du découpage des noms de personne, 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. La taille du vocabulaire dans toutes les expériences que nous mettons en œuvre ne dépasse pas 124k mots.

La phonétisation des nouvelles entrées du vocabulaire est effectuée en utilisant l'outil de phonéti-

1. www.defi-repere.fr

sation LIA_PHON. 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.

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 multimédia par rapport aux noms de personne. Nous nous intéressons ainsi aux requêtes, saisies par l'utilisateur, qui comportent un ou plusieurs noms de personne considérés comme hors-vocabulaire pour le SRAP utilisé. En revanche, nous considérons le cas où un moteur de recherche, disposant d'un tel système d'indexation, n'est pas apte à repérer les noms de personne parmi les mots de la requête. Ainsi, nous supposons que ledit moteur découpe tous les mots d'une requête en syllabes avant d'effectuer la recherche dans les index basés sur les transcriptions automatique. Dès lors, nous procédons à une stratégie d'évaluation particulière. D'une part, nous découpons en syllabes les mots des hypothèses de transcription en tenant compte de la particularité des fichiers d'hypothèse produits en format ctm[2] (Time Marked Conversation). D'autre part, nous préparons la transcription de référence des données de test (R_test), fournie en format stm [2] (Segment Time Mark), en découpant chaque mot en syllabes. Ces formats, ainsi que les évaluations effectuées, se reposent sur le moteur d'alignement « sclite » [3] (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 de personne différents. Parmi ces mots, 27 sont originellement hors-vocabulaire. Par ailleurs, 13 parmi les noms de personne 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 de personne de R_test sont artificiellement mis hors vocabulaire.

6 Résultats

Le tableau 1 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 NP HV	TR de syllabes dans les NP HV	SER
Système de base (S0)	0%	21,05%	18,70%
Système avec syllabes (S1)	**31,39%**	42,82%	**19,90%**
Système avec syllabes étiquetées (S2)	31,10%	**42,95%**	24,90%

TABLE 1: Résultats de reconnaissance (TR : taux de reconnaissance, NP : nom de personne, HV : hors-vocabulaire, SER : Taux d'erreur de syllabes)

2. www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/infmts.htm
3. www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm

Partant du fait que les noms de personne évalués n'appartiennent pas au vocabulaire, le taux de reconnaissance des noms de personne hors-vocabulaire (HV) (nombre de noms de personne HV dont toutes les syllabes sont correctement reconnues / nombre de noms de personne HV) dans le système de base est nul. Comme décrit dans le tableau 1, ce système reconnaît tout de même 21,05% des syllabes au sein de ces mêmes noms de personne. Ce taux correspond au nombre de syllabes correctement reconnues dans l'ensemble des noms de personne HV divisé par le nombre total de syllabes dans l'ensemble des noms de personne HV.
Le système S1, appris sur des noms de personne en syllabes, réussit à reconnaître 31,39% des noms de personne et 42,82% des syllabes dans l'ensemble de ces noms de personne. 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 négligeable par rapport au gain que nous obtenons par rapport à la reconnaissance des noms de personne, 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 de personne, ce système réussit tout de même à obtenir un meilleur taux de reconnaissance de syllabes dans les noms de personne (42,95%).

Enfin, nous n'avons pas eu la possibilité de comparer les résultats de cette mise en œuvre 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 de personne hors-vocabulaire.

7 Conclusion

Prenant en compte les particularités des systèmes d'indexation automatique des documents multimédia, 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, la mise en oeuvre de l'approche proposée réussit à reconnaître environ un tiers des noms de personne hors-vocabulaire.
L'intérêt de ces résultats se manifeste, entre autres, dans deux aspects. D'une part, l'énonciation de mots appartenant à la catégorie des entités nommées, dans les données multimédia 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, 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 de personne. Par ailleurs, le traitement des mots hors-vocabulaire peut être au profit d'autres axes. Notamment dans le cadre de la traduction automatique de la parole, le système de traduction s'appuie sur la sortie des SRAP. Ainsi, un système de traduction de la parole profiterait de la réduction de l'influence des mots hors-vocabulaire dans la performance du SRAP.

Références

ALLAUZEN, A. et GAUVAIN, J.-L. (2005). Open vocabulary asr for audiovisual document indexation. *In Proceedings of the ICASSP*, volume 1, pages 1013–1016.

BAZZI, I. et GLASS, J. (2000). Modeling out-of-vocabulary words for robust speech recognition. *In The Proceedings of the sixth International Conference on Spoken Language Processing*, volume 1.

BÉCHET, F. (2001). Lia phon : un systeme complet de phonétisation de textes. *Traitement automatique des langues*, 42(1):47–67.

BISANI, M. et NEY, H. (2005). Open vocabulary speech recognition with flat hybrid models. *In INTERSPEECH*, pages 725–728.

BROWN, E. (2002). Is speech recognition becoming mainstream ?

DELÉGLISE, P., ESTEVE, Y., MEIGNIER, S. et MERLIN, T. (2009). Improvements to the lium french asr system based on cmu sphinx : what helps to significantly reduce the word error rate ? *In INTERSPEECH*, pages 2123–2126.

DUFOUR, R. (2008). From prepared speech to spontaneous speech recognition system : a comparative study applied to french language. *In Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology*, pages 595–599. ACM.

GELAS, H., ABATE, S. T., BESACIER, L. et PELLEGRINO, F. (2012). Analyse des performances de modèles de langage sub-lexicale pour des langues peu-dotées à morphologie riche. *JEP-TALN-RECITAL*, page 53.

KNESER, R. et NEY, H. (1995). Improved backing-off for n-gram language modeling. *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, page 181–184.

LOGAN, B., MORENO, P. et DESHMUKH, O. (2002). Word and sub-word indexing approaches for reducing the effects of oov queries on spoken audio. *In Proceedings of the second international conference on Human Language Technology Research*, pages 31–35.

LOGAN, B., MORENO, P., VAN THONG, J.-M. *et al.* (1996). An experimental study of an audio indexing system for the web. *In in Proc. ICSLP*. Citeseer.

MEIGNIER, S. et MERLIN, T. (2010). Lium spkdiarization : an open source toolkit for diarization. *In CMU SPUD Workshop*, volume 2010.

PERENNOU, G. et CALMES, M. d. (1987). Bdlex lexical data and knowledge base of spoken and written french. *In European conference on Speech Technology*.

RÉVEIL, B., DEMUYNCK, K. et MARTENS, J.-P. (2013). An improved two-stage mixed language model approach for handling out-of-vocabulary words in large vocabulary continuous speech recognition. *Computer Speech & Language*.

ROTOVNIK, T., MAUČEC, M. S. et KAČIČ, Z. (2007). Large vocabulary continuous speech recognition of an inflected language using stems and endings. *Speech communication*, 49(6):437–452.

SHAIK, M. A. B., MOUSA, A. E.-D., SCHLÜTER, R. et NEY, H. (2011). Hybrid language models using mixed types of sub-lexical units for open vocabulary german lvcsr. *In INTERSPEECH*, pages 1441–1444.

STOLCKE, A. *et al.* (2002). Srilm-an extensible language modeling toolkit. *In INTERSPEECH*.

VAN THONG, J.-M., MORENO, P. J., LOGAN, B., FIDLER, B., MAFFEY, K. et MOORES, M. (2002). Speechbot : an experimental speech-based search engine for multimedia content on the web. *Multimedia, IEEE Transactions on*, 4(1):88–96.

ZABLOTSKIY, S., SHVETS, A., SIDOROV, M., SEMENKIN, E. et MINKER, W. (2012). Speech and language resources for lvcsr of russian. *In LREC*, pages 3374–3377.