MajecSTIC 2009
Avignon, France, du 16 au 18 novembre 2009

Détection de mots hors-vocabulaire par combinaison de mesures de confiance de haut et bas niveaux

Benjamin Lecouteux[1], *Georges Linarès*[1], *Benoit Favre*[2] *

[1]Laboratoire Informatique d'Avignon (LIA), Université d'Avignon , France
[2] ICSI, 1947 Center St, Suite 600, Berkeley, CA 94704, USA
{benjamin.lecouteux,georges.linares}@univ-avignon.fr; favre@icsi.berkeley.edu

Résumé
Cet article aborde le problème de la détection des mots hors-vocabulaire dans le cadre des Systèmes de Reconnaissance Automatique de la Parole (SRAP) continu grand vocabulaire. Nous proposons une méthode inspirée par les mesures de confiance, qui consiste en l'analyse des sorties du système afin d'y détecter automatiquement les erreurs liées aux mots hors-vocabulaire. Cette méthode combine différents paramètres basés sur l'acoustique, la linguistique, la topologie du graphe de décodage ainsi que sur des paramètres sémantiques. Nous évaluons séparément chacun de ces paramètres et nous estimons leur complémentarité. Les expériences ont été menées sur un corpus d'émissions radio issu de la campagne ESTER. Les résultats montrent des performances intéressantes en condition réelle : nous obtenons un taux de détection de mots hors vocabulaire de 43%-90% pour 2.5%-17.5% de fausse détection.

Abstract
This paper addresses the issue of Out-Of-Vocabulary (OOV) word detection in Large Vocabulary Continuous Speech Recognition (LVCSR) systems. We propose a method inspired by confidence measures, that consists in analyzing the recognition system outputs in order to automatically detect errors due to OOV words. This method combines various features based on acoustic, linguistic, decoding graph and semantics. We evaluate separately each feature and we estimate their complementarity. Experiments are conducted on a large French broadcast news corpus from the ESTER evaluation campaign. Results show good performance in real conditions : the method obtains an OOV word detection rate of 43%-90% with 2.5%-17.5% of false detection.

Mots-clés : Détection des mots hors-vocabulaire, mesures de confiance, reconnaissance automatique de la parole

Keywords: OOV word detection, confidence measures, speech recognition

1. Introduction

Les systèmes de reconnaissance automatique de la parole sont limités par la taille de leurs lexiques. Il en résulte des mots qui ne peuvent être reconnus, alors qu'ils peuvent porter une information potentiellement critique pour la compréhension ou l'indexation.
L'augmentation perpétuelle de la couverture lexicale n'est pas une solution viable, car elle introduirait trop de bruit et demanderait également plus de calculs lors du décodage. Quelques travaux proposent des méthodes dépendantes du domaine pour augmenter le lexique. Ces approches reposent généralement sur des analyses sémantiques du contenu parlé. [13] propose d'utiliser le contexte local pour construire des requêtes Web permettant de retrouver les mots manquants : l'une des principales difficultés liée aux mots hors-vocabulaire est la détection automatique de leur emplacement. Les précédents travaux relatifs à cet aspect proposent d'intégrer des *fillers* dans le modèle de langage. Ces derniers sont supposés absorber les segments non reconnus. Ces méthodes montrent des performances correctes sur de petits vocabulaires, cependant

* Cette recherche est supportée par l'ANR (Agence Nationale de la Recherche), dans le cadre du projet AVISON.

elles nécessitent des réglages fins du modèle de langage et ce, particulièrement sur de grands vocabulaires.
D'autres travaux proposent des approches *a posteriori* où les sorties intermédiaires du SRAP sont analysées afin de localiser les zones de transcription où des erreurs sont dues à des mots hors-vocabulaire (MHV).
Dans cet article, nous présentons une méthode *a posteriori* qui se propose de détecter les MHV en analysant la sortie d'une première passe de reconnaissance automatique. Nous proposons d'étudier différents niveaux de paramètres liés aux mots qui seront évalués indépendamment et combinés via un classifieur statistique. De cette manière, nous classifions chaque mot comme MHV ou non-MHV.
Dans la section suivante, nous faisons un rapide état de l'art relatif à la détection des MHV. Puis dans la section 3, nous présentons notre méthode, dérivée de l'estimation des mesures de confiance : nous nous focalisons sur les différents aspects de la détection des MHV en utilisant des mesures de confiance classiques.
Dans la section 4, nous définissons le cadre expérimental. Nous décrivons le corpus ESTER sur lequel les expériences ont été menées, puis nous détaillons le protocole expérimental.
Dans la section 5, nous proposons différents paramètres pour la détection des MHV. Nous évaluons leur pertinence ainsi que leur complémentarité pour cette tâche. Nous étudions également l'impact du taux d'erreur mot sur notre méthode de détection. Finalement, nous présentons des résultats sur l'ensemble de notre corpus de test dans la section 5. La dernière section présente quelques conclusions et perspectives.

2. État de l'art

La détection des MHV a été approchée de différentes manières ; quelques groupes de recherche [1–3, 5] proposent de modéliser les mots non observés à travers des *fillers* ou des modèles de mots génériques ; le but est de couvrir toutes les prononciations possibles de MHV en les représentant par des sous-unités. Par exemple, la méthode proposée dans [3] permet de retrouver la graphie des mots MHV en utilisant des modèles de séquence de graphèmes. Malheureusement, ces méthodes absorbent souvent une partie du signal correspondant à des mots connus et requièrent d'être très finement réglées.
Les informations de haut niveau (syntaxe, sémantique) sont utilisées à la fois pour la détection de MHV et pour l'estimation de mesures de confiance. [4] propose de détecter les MHV dans un système de dialogue. Les expériences sont effectuées dans des domaines restreints, mais montrent les bénéfices d'un contexte distant pour détecter les entités nommées. [7] propose d'utiliser l'analyse Latent Semantic Analysis (LSA) pour estimer des mesures de confiance dans un SRAP ; ses résultats montrent une bonne précision, mais un faible taux de rappel.
D'autres travaux [8, 9, 15] utilisent une distance d'édition entre le treillis de phonèmes et les mots décodés, les erreurs d'alignement étant supposées plus fréquentes lors de la rencontre de MHV. Ces méthodes obtiennent de bonnes précisions associées à des taux de détection corrects, mais ne prennent pas directement en considération les paramètres sémantiques ou la topologie du graphe de reconnaissance.
Dans d'autres articles [6, 14, 16], les MHV sont identifiés en utilisant des mesures de confiance. Ces approches permettent d'introduire plus d'informations extraites du SRAP. Ces méthodes obtiennent une meilleure précision que les modèles *filler* mais sont limitées aux systèmes de dialogue ou systèmes de reconnaissance de mots isolés. De plus, elles n'utilisent pas de paramètres robustes liés à la linguistique.
Les approches récentes dans le cadre des mesures de confiance, utilisent des information directement issues du processus de décodage : le nombre de mots en compétition à la fin d'un noeud, les vraisemblances normalisées, le comportement du processus de décodage, etc. D'autres travaux liés aux mesures de confiance montrent l'importance des modèles de langage : [10] propose de combiner des paramètres acoustiques et linguistiques tels que le repli du modèle de langage associé aux probabilités *a posteriori*.

3. Détection des mots hors-vocabulaire

3.1. Principe

La méthode proposée se décompose en trois étapes. La première extrait des paramètres de bas niveaux relatifs à l'acoustique et à la topologie du graphe de reconnaissance ainsi que des paramètres de haut niveau liés à la linguistique. A partir de ces paramètres, une première détection des MHV est produite par un classifieur basé sur un algorithme de boosting. Finalement, un module d'analyse sémantique raffine ce processus de détection. La prochaine sous-section détaille ces trois étapes de détection.

3.2. Paramètres extraits

Chaque mot de l'hypothèse est représenté comme un vecteur de 23 paramètres, qui se regroupent en 3 classes.

Les mots hors-vocabulaire induisent des distorsions entre l'hypothèse et la meilleure séquence phonétique. Nous utilisons des **paramètres acoustiques** tels que la log-vraisemblance acoustique du mot, la log-vraisemblance moyenne par trame et la différence entre la log-vraisemblance du mot et la log-vraisemblance du segment correspondant sans contrainte linguistique.

Les **paramètres linguistiques** sont basés sur les probabilités estimées par le modèle de langage 3-gramme utilisé dans le SRAP. Nous utilisons la probabilité 3-gramme, la perplexité du mot dans une fenêtre définie, la probabilité unigramme. Nous ajoutons également l'index proposé par [10], qui représente le repli actuel du mot au niveau du modèle de langage. Cette valeur est de 3 si le mot est un tri-gramme, 2 si c'est un bi-gramme et 1 dans le cas d'un unigramme.

les paramètres liés au graphe se basent sur l'analyse du mot dans le réseau de confusion. L'utilisation de ces paramètres est motivée par l'idée que l'algorithme de recherche tente d'explorer de nombreux chemins alternatifs proposant des scores similaires lorsqu'un MHV apparaît. Le comportement de l'exploration ainsi que la distribution des probabilités *a posteriori* semblent être un bon indicateur pour la détection des mots inconnus. Nous utilisons le nombre de chemins alternatifs ainsi que la probabilité *a posteriori*. Nous incluons également des valeurs relatives à la distribution des probabilités *a posteriori* dans le réseau de confusion : le minimum, le maximum, la moyenne des probabilités *a posteriori*, le nombre de liens nuls avant et après le mot. Un dernier paramètre représente la durée des mots dans une fenêtre de 500ms dont le centre est le mot courant.

3.3. Combinaison des paramètres et classification

Nous utilisons un algorithme de classification de type boosting afin de combiner les paramètres, comme proposé par [11]. Le classifieur est une variante de Adaptive Boosting (Adaboost) : icsiboost [2]. Cet algorithme consiste en une recherche exhaustive de la combinaison linéaire de classifieurs en sur-pondérant les exemples mal appris. Un avantage de ce classifieur est sa capacité à produire des probabilités permettant une interprétation intuitive des résultats.

Les vecteurs destinés au classifieur sont composés de trois mots consécutifs. Pour chaque mot, nous estimons les 23 paramètres précédemment décrits. Finalement, nous obtenons un vecteur de 69 coefficients incluant les mots précédent, courant et suivant.

Le classifieur a été entraîné sur un corpus spécifique qui n'était pas inclus dans l'apprentissage de notre SRAP. Chaque mot de ce corpus a été annoté comme *dans* ou *hors* vocabulaire, en fonction du lexique du SRAP. Les résultats de la classification se séparent en deux classes : MHV ou non-MHV.

3.4. *Latent Semantic Analysis* et 3-grammes issus du Web

Latent Semantic Analysis (LSA) est une technique permettant d'associer des mots qui sont corrélés sémantiquement à travers plusieurs documents. L'hypothèse formulée est que les mots cooccurrents dans un même document sont sémantiquement corrélés.

Dans notre système, une séquence de mot sémantiquement pertinente peut être considérée comme incohérente par le modèle de langage du SRAP en raison de la limite du modèle de langage n-gramme. Pour cette raison, nous ajoutons un estimateur de consistance sémantique qui permet de valider ou rejeter la détection précédemment effectuée par le classifieur.

[2] disponible sur http ://code.google.com/p/icsiboost/

Cette mesure n'a pas été inclue dans le classifieur car elle introduit trop de bruit au niveau des mots qui ne sont pas hors-vocabulaire. Elle est uniquement appliquée sur les mots détectés comme un dernier processus de filtrage.
Ce filtre combine deux mesures basées sur le Web et le corpus Français Gigaword.
La première mesure estime la probabilité de la co-occurrence de mots par un ratio issu des hits issus de google [13]. Cette mesure est calculée sur une fenêtre locale de 3 mots (incluant le MHV) filtrée par une stop-liste.
La seconde mesure utilise LSA pour estimer à quel point le mot détecté comme hors-vocabulaire est sémantiquement proche du segment courant [7] : pour chaque mot détecté, le module LSA sélectionne les 100 mots les plus proches. La cardinalité de cette intersection entre ces 100 mots et le segment courant est normalisée par la taille du segment et la valeur résultante est linéairement combinée avec les scores issus du Web. Un seuil de décision à été empiriquement défini sur le corpus de développement pour obtenir une précision de 100% : sur le corpus de développement 9% des fausses détections sont ainsi supprimées.

4. Cadre expérimental

4.1. Le système de transcription du LIA

Les expériences ont été effectuées en utilisant le système de transcription d'émissions radio utilisé lors de la campagne ESTER. Ce système repose sur un décodeur à base de modèles de Markov cachés (Hidden Markov model : HMM) développé au LIA, Speeral [12]. Speeral est un décodeur asynchrone opérant sur un treillis de phonème ; les modèles acoustiques utilisent des HMM et sont contextuels à base de tri-phones. Le modèle de langage est un modèle 3-gramme classique estimé sur environ 200 millions de mots issus du journal *Le Monde* ainsi que du corpus ESTER (environ 1 million de mots). Le lexique est composé de 67000 mots. Dans ces expériences, une seule passe est effectuée en 3 fois le temps réel.

4.2. Le corpus ESTER

Le corpus ESTER est composé d'heures de radio française. Les corpus d'apprentissage et de développement ont été extraits de la campagne ESTER-2 (100 heures annotées manuellement). Ces données n'ont pas été incluses dans l'apprentissage du SRAP. Le corpus d'apprentissage contient 15000 mots du vocabulaire sélectionnés aléatoirement parmi les 1 million de mots et 15000 MHV extraits de l'ensemble du corpus d'apprentissage.
Nous avons testé notre approche sur 7 heures de parole extraites du test de la campagne ESTER. Le nombre de MHV est de 982, qui pour un corpus de 700011 mots représentent un taux de MHV d'environ 1.4%. Le taux d'erreur mot (Word Error Rate : WER) moyen est de 28.9 : une seule passe a été effectuée.

4.3. Protocole de détection

Les mots hors-vocabulaire ont été manuellement annotés en sélectionnant tous les mots non disponibles dans le lexique. Durant la détection, si un MHV détecté recouvre un vrai MHV, le vrai MHV est considéré comme détecté. Dans toutes les autres situations, nous considérons un mot marqué comme fausse détection.

5. Expériences

Nous avons effectué trois expériences pour estimer la pertinence des paramètres ainsi que leur complémentarité. De plus, nous estimons la robustesse du système en effectuant les expériences pour différentes tranches de WER.

5.1. Pertinence des paramètres

5.1.1. Classification en fonction de chaque paramètre

Dans cette section, nous présentons la capacité de classification pour chaque ensemble de paramètres. Un modèle est entraîné avec icsiboost pour chaque ensemble. Nous utilisons des courbes ROC pour afficher les taux de vraie et fausse détection de MHV. Les courbes sont tracées en faisant varier le seuil de coupure sur les sorties du classifieur.

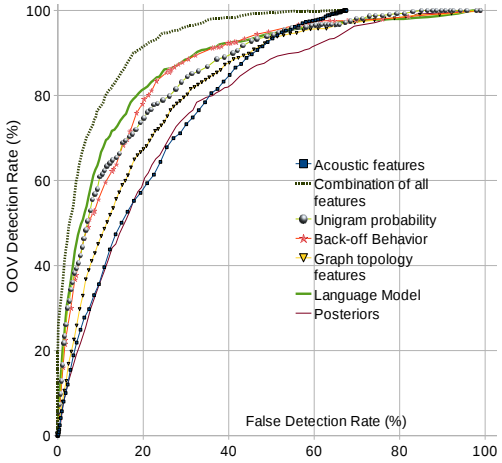


FIG. 1 – *Pertinence de chaque paramètre*

Les courbes ROC, présentées dans la figure 1, montrent la sortie du classifieur pour chaque paramètre comparé à la combinaison de l'ensemble. Les résultats montrent que les paramètres pertinents sont relatifs à la linguistique ainsi qu'à la topologie du graphe, tandis que les paramètres acoustiques et les probabilités *a posteriori* semblent moins intéressants que les autres. Les scores linguistiques et le repli du modèle de langage sont des indicateurs de rupture linguistique (répétition d'unigrammes sans cohérence). Cette rupture est caractéristique lorsque le système rencontre un MHV ou fait des erreurs de reconnaissance. Un point intéressant est la qualité du paramètre de repli qui présente des résultats proches de ceux du modèle de langage.
La topologie du graphe est également un paramètre important, car il est relatif aux difficultés rencontrées lors du processus de décodage. Un résultat plus surprenant est la capacité discriminante de la probabilité unigramme. Ceci est sans doute dû au fait que le système a tendance à se replier sur un certain type de mots lorsqu'il rencontre des MHV : petits et peu fréquents correspondant au signal. Cependant, la combinaison de tous ces paramètres montre une amélioration significative. En dépit de l'hétérogénéité des paramètres, nous observons une bonne complémentarité entre eux.

5.1.2. Contexte linguistique et mots hors-vocabulaire

Les expériences rapportées dans la section précédente montrent la pertinence du contexte lié au mot (Figure 2). Notre topologie inclue les mots précédents et suivants pour chaque MHV examiné. Un résultat inattendu est la prépondérance des mots précédent et suivant qui semblent porter plus d'information que le mot central. Le mot précédent est sans doute plus pertinent par rapport à la rupture linguistique, par ailleurs nous ne devons pas oublier que parmi les paramètres du mot précédent se trouve le score linguistique du mot central.
La combinaison de ces trois ensembles de paramètres permet une amélioration relative de 10%. Par ailleurs, l'utilisation d'une fenêtre plus grande que 3 mots dégrade les résultats.

5.2. Complémentarité des paramètres

Après l'étude de la pertinence des paramètres, nous avons étudié leur complémentarité. Pour cela, nous avons ordonné chaque ensemble de paramètres en fonction de leur apport pour la classification. Nous avons commencé l'apprentissage avec le meilleur ensemble, puis nous avons rajouté un a un chaque ensemble pour observer son incidence.
La courbe ROC (Figure 3) montre la complémentarité de chaque ensemble de paramètres.
Nous mesurons la complémentarité via le EER (Equal Error Rate). Le EER est le point où le taux de détection des MHV et le taux de fausse détection sont égaux. La table de la Figure 3 montre l'évolution du EER pour chaque nouveau paramètre ajouté.
Chaque ensemble de paramètres apporte des informations discriminantes. Un résultat inattendu est la proéminence des paramètres acoustiques. Ils montrent une amélioration significative alors

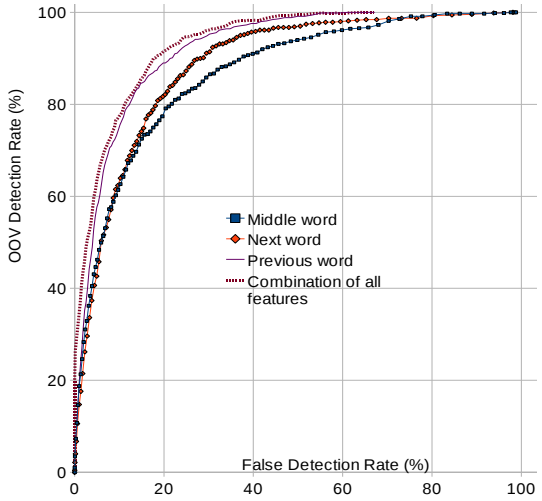
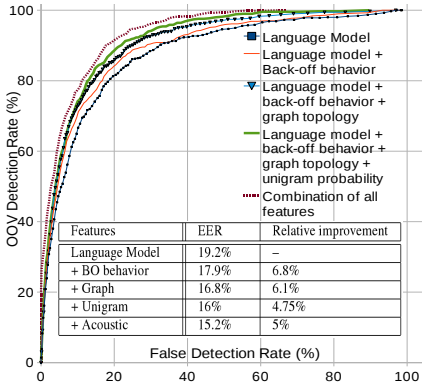


FIG. 2 – *Pertinence des mots dans la fenêtre d'observation*



Features	EER	Relative improvement
Language Model	19.2%	–
+ BO behavior	17.9%	6.8%
+ Graph	16.8%	6.1%
+ Unigram	16%	4.75%
+ Acoustic	15.2%	5%

FIG. 3 – *Complémentarité des paramètres*

qu'ils offrent la plus mauvaise performance individuelle. Cependant, les ensembles de paramètres sont complémentaires et mutuellement dépendants.

5.3. Mots hors-vocabulaire et taux d'erreur mot

Ces dernières expériences proposent d'étudier le comportement de la détection des MHV en fonction du WER contenu dans la transcription. Nous avons trié les segments du corpus de test en 6 tranches de WER (Figure 4).

La figure 4 présente les résultats pour différentes tranches de WER. Nous remarquons une corrélation entre WER et taux de détection de MHV. Les courbes ROC montrent trois tendances : une excellente classification pour un WER compris entre 10% et 30%, une classification stable et correcte pour un WER compris entre 0% et 10% ainsi qu'entre 30% et 50%, et une dégradation au delà de 50%.

La tranche de 0%-10% est surprenante car elle s'éloigne des meilleurs résultats. Elle correspond à des segments avec de très faibles taux de MHV (0.25%), ce qui rend sans doute la détection plus difficile : une bonne couverture lexicale et un bon WER indiquent que les contextes acoustiques et linguistiques sont probablement proches des conditions d'apprentissage pour lesquelles nos paramètres sont moins informatifs.

Cependant, entre 10% et 30%, les taux de classification sont meilleurs que ceux observés dans les autres tranches. Cela correspond sans doute au contexte de décodage le plus fréquent, pour

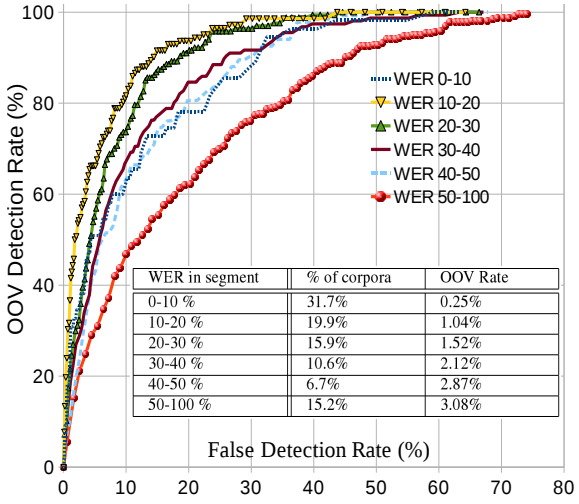


FIG. 4 – *Résultats pour chaque tranche de WER*

lequel les conditions de test ressemblent aux conditions d'apprentissage : les paramètres sont plus pertinents. Au delà de 50%, le bruit détériore la précision.
Globalement, ces expériences montrent une bonne robustesse de notre méthode vis-à-vis du WER, à l'exception des cas de décodage pathologiques.

5.4. Filtrage sémantique des mots hors-vocabulaire

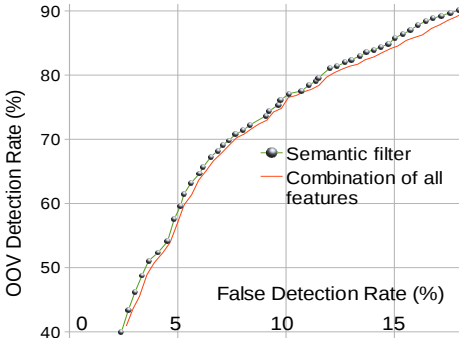


FIG. 5 – *Résultat après filtrage sémantique*

Dans une passe finale, nous proposons d'utiliser l'information sémantique pour filtrer les mots hors-vocabulaire détectés. Nous utilisons un module sémantique uniquement sur les mots détectés, afin de récupérer les mots qui sont cohérents par rapport à leur contexte.
Les courbes ROC sont présentées dans la figure 5. Les résultats montrent que le filtre réduit le EER d'environ 4% relatifs (de 15.2% à 14.6%), tandis que le taux de fausse détection est réduit d'environ 5% relatifs.

6. Conclusion et perspectives

Dans cet article, nous avons présenté une méthode de détection des MHV dans un système de reconnaissance automatique de la parole continue grand vocabulaire. Notre approche consiste à extraire de nombreux paramètres tout au long de la chaîne de décodage : acoustiques, linguistiques, topologie du graphe de reconnaissance. En passe finale, un module sémantique effectue un filtrage pour réduire le taux de fausses détections.

Nos expériences montrent des résultats prometteurs : la détection de MHV semble homogène en dépit d'un WER inconstant. La méthode proposée permet de détecter 43% des mots inconnus avec 2.5% de fausse détection ou 90% pour 17.5% de fausse acceptation. Les expériences montrent que les paramètres linguistiques et basés sur le graphe sont les plus pertinents. Cependant, les paramètres acoustiques associés aux paramètres linguistiques et d'analyse du graphe rendent la détection plus robuste. Finalement, le module sémantique permet une légère mais significative amélioration.

Actuellement, ces résultats reposent uniquement sur la meilleure hypothèse du système ; nous envisageons d'étendre le module sémantique afin de retrouver les chemins alternatifs dans le graphe permettant d'invalider certaines fausses détections de MHV. Nous souhaitons également améliorer notre analyse sémantique, en ajoutant une détection des entités nommées qui représentent une partie non négligeable des mots hors-vocabulaire.

Bibliographie

1. A. Asadi, R. Schwartz, et J. Makhoul. Automatic detection of new words in a large vocabulary continuous speech recognition system. In *Proc. International Conference on Acoustics, Speech, and Signal Processing ICASSP-90*, pages 125–128 vol.1, 1990.
2. Issam Bazzi et James R. Glass. Modeling out-of-vocabularywords for robust speech recognition. In *ICSLP 2000*, 2000.
3. Maximilian Bisani et Hermann Ney. Open vocabulary speech recognition with flat hybrid models. In *Eurospeech*, 2005.
4. Manuela Boros, Maria Aretoulaki, Florian Gallwitz, Elmar Noth, et Heinrich Niemann. Semantic processing of out-of-vocabulary words in a spoken dialogue system. In *Eurospeech*, 1997.
5. G. Boulianne et P. Dumouchel. Out-of-vocabulary word modeling using multiple lexical fillers. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding ASRU '01*, pages 226–229, 2001.
6. Tie Cai et Jie Zhu. Oov rejection algorithm based on class-fusion support vector machine for speech recognition. In *Proc. International Conference on Machine Learning and Cybernetics*, volume 6, pages 3695–3699, 26–29 Aug. 2004.
7. Stephen Cox et Srinandan Dasmahapatra. A semantically-based confidence measure for speech recognition. In *ICSLP*, 2000.
8. Satoru Hayamizu, Katunobu Itou, et Kazuyo Tanaka. Detection of unknown words in large vocabulary speech recognition. In *Eurospeech*, 1993.
9. Hui Lin, J. Bilmes, D. Vergyri, et K. Kirchhoff. Oov detection by joint word/phone lattice alignment. In *Proc. ASRU Automatic Speech Recognition & Understanding IEEE Workshop on*, pages 478–483, 2007.
10. Julie Mauclair, Yannick Estève, Simon Petit-Renaud, et Paul Deléglise. Automatic detection of well recognizedwords in automatic speech transcriptions. In *LREC*, 2006.
11. Pedro J. Moreno, Beth Logan, et Bhiksha Raj. A boosting approach fo confidence scoring. In *Europseech*, 2001.
12. P. Nocera, C. Fredouille, G. Linares, D. Matrouf, S. Meignier, J.-F. Bonastre, D. Massonié, et F. Béchet. The lia's french broadcast news transcription system. In *SWIM : Lectures by Masters in Speech Processing*, 2004.
13. Stanislas Oger, Georges Linarès, et Frédéric Béchet. Local methods for on-demand out-of-vocabulary word retrieval. In *LREC*, 2008.
14. Hui Sun, Guoliang Zhang, Fang Zheng, et Mingxing Xu. Using word confidence measure for oov words detection in a spontaneous spoken dialog system. In *Eurospeech*, 2003.
15. C. White, G. Zweig, L. Burget, P. Schwarz, et H. Hermansky. Confidence estimation, oov detection and language id using phone-to-word transduction and phone-level alignments. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2008*, pages 4085–4088, 2008.
16. Sheryl R. Young. Recognition confidence measures :detection of misrecognitions and out-of-vocabulary words. Technical report, Carnegie Mellon University, 1994.