Caractérisation statistique de la nature des mots Hors-Vocabulaire dans la parole spontanée

Hichem Hamimed et Géraldine Damnati

France Télécom R&D, DIH/IPS
2, avenue Pierre Marzin 22307 Lannion FRANCE
Tél. : +33 (0)2 96 05 21 42 - Fax : +33 (0)2 96 05 35 30
Mél : {hichem.hamimed, geraldine.damnati}@rd.francetelecom.com

ABSTRACT

To improve our knowledge on the acoustic and linguistic characteristics of the Out Of Vocabulary (OOV) words, we present in this article the results of a statistical study on the nature and the contexts of occurrence of OOV words in spontaneous speech. We examined the phonetic and syllabic structure of the OOV words and the other phenomena (false starts, badly pronounced words, truncated words). We also examined the type of utterances containing OOV words, their occurrence rates and their localization in the utterances. We studied the interest of determining different categories for OOV words in the language model. In what follows, we will describe all these analysis and comment the observations that we made.

1. INTRODUCTION

Les serveurs vocaux interactifs sont l'une des principales applications des systèmes de traitement automatique de la parole. Afin de réduire la complexité du système de reconnaissance, la taille du lexique utilisé est limitée et la tâche du serveur est généralement circonscrite à un domaine particulier. Les applications de dialogue oral homme-machine développées à France Télécom R&D permettent de fournir, par exemple, des informations sur les restaurants, sur la météo ou encore sur la bourse. Utiliser un ensemble bien défini de mots (en termes de taille et de sémantique) dans le lexique de l'application permet d'assurer des performances acceptables du système de reconnaissance.
Étant donné un serveur vocal doté d'un vocabulaire de taille limitée, l'apparition de mots Hors-Vocabulaire (HV) lors d'une utilisation non dirigée de ce serveur doit être envisagée pour la simple raison que le mode d'élocution est spontané. La présence de ces mots reste toujours probable même si l'on décide d'augmenter la taille du vocabulaire.
L'emploi de mots HV dans un dialogue peut être source d'erreurs de reconnaissance et par conséquent de mauvaise interprétation de la requête[1]. Dans ce cas, le nombre d'échanges inutiles entre l'utilisateur et le serveur augmente et la probabilité d'échec du dialogue peut augmenter.
Puisque la composition du vocabulaire est une contrainte forte du système de reconnaissance, résoudre les problèmes liés aux mots HV devient une étape importante pour l'amélioration des performances du système de reconnaissance et par conséquent pour l'amélioration de l'interaction avec le serveur. La détection des mots HV fait intervenir trois axes principaux : l'utilisation des modèles de langage pour la prédiction de la présence d'un mot HV dans une requête [5], la modélisation explicite des formes acoustiques des mots HV [1] et la décision vocabulaire/HV par le biais de mesures de confiance [7, 6].
L'étude statistique menée sur nos données expérimentales nous a permis d'examiner certaines propriétés des mots HV. Nous avons comparé dans un premier temps la répartition de la longueur phonétique puis syllabique des mots HV avec celles des mots du vocabulaire, des faux départs, des mots tronqués et des mots mal prononcés. La position d'un mot HV dans un énoncé a également été examinée. Du point de vue du système d'interprétation de la requête, il est intéressant de connaître, lorsqu'elles sont bien prédites, les catégories syntaxiques ou sémantiques des mots HV détectés. Dans la section 3.3, nous présentons à juste titre les catégories déterminées dans le cadre de notre étude et donnons quelques caractéristiques typique à certaines catégorie.
L'étude présentée ici contribue par des statistiques à mieux connaître les caractéristiques des mots HV qui aideront à améliorer leur détection dans la parole spontanée.

2. CORPUS DE DONNÉES

2.1. Description des données

L'étude se fait sur les traces du corpus *PlanResto* de Paris. L'utilisateur communique par le biais du téléphone avec le système, le dialogue est à initiative partagée. L'utilisateur cherche un restaurant parisien en précisant certains critères de choix (prix, lieu, spécialité, ...). Les données ont été enregistrées en condition d'utilisation par des locuteurs externes et aussi par des locuteurs internes qui ont déjà une idée sur le fonctionnement du système. Ces données ont été transcrites par la suite.
Le corpus est découpé en sessions de dialogue, chaque session est constituée d'un ensemble d'enregistrements (énoncés) et chaque enregistrement est un ensemble d'unités (mots ou annotations). Dans notre étude, nous distinguons les phénomènes liés à la parole spontanée (*speech repairs*) des mots HV. En effet, les mots peuvent être mal prononcés, non prononcés jusqu'au bout (faux départs) ou non intelligibles. Ils peuvent être également tronqués au début ou à la fin lorsqu'ils sont en début ou en fin d'énoncés à cause de la détection bruit/parole. Ces phénomènes sont annotés dans les traces de corpus en plus des bruits et des évènements acoustiques.
Nous disposons d'un lexique de 1914 mots et d'un ensemble de données brutes constituées de 8012 enregistrements regroupant, après filtrage des bruits, 30276 mots. Le

[1] Si les mots HV sont omis ou bien remplacés par d'autres mots du vocabulaire

tableau 1 donne la répartition des données en apprentissage et en test.

Corpus	Sessions	Énoncés	Mots		Mots HV	
			Occu.	Diff.	Occu.	Diff.
Appr.	478	6298	26374	833	453	242
Test	58	942	3902	345	67	37

Table 1: Répartition apprentissage / test

Le taux de présence des mots HV dans les traces est raisonnable, si on le compare avec les taux cités dans la littérature, il correspond à 1,72%. C'est la perturbation la plus fréquente si l'on ne tient pas compte des annotations de bruits (5,04%). La détection bruit / parole génère 1,43% de mots tronqués. Les anomalies liées à la parole spontanée représentent 0,68% des éléments du corpus.

2.2. Performances de reconnaissance

Des tests de reconnaissance ont été effectués sur 942 énoncés de ce corpus. Il faut signaler que la version actuelle du système de reconnaissance intègre déjà trois types de modèles de rejet :
<R.GLOBAL> Permet de rejeter les enregistrements contenant des bruits ou des apartés,
<R.UNK_OOV> Permet de rejeter les mots HV,
<R.UNK_SPR> Permet de rejeter les mots tronqués, mal prononcés, les faux départs et les mots non intelligibles.
Ces modèles sont spécifiés acoustiquement par une boucle de modèles de phonèmes hors-contexte et incluent des modèles de bruits. Ces modèles peuvent être soumis à des pondérations qui permettent de favoriser ou de pénaliser le passage par le modèle générique. Dans le cas des rejets <R.UNK_OOV> et <R.UNK_SPR>, cette pondération s'ajoute à la pénalité du modèle de langage qui, elle-même, est différente pour chaque contexte d'apparition. Pour connaître les limites initiales de notre système, nous avons effectué quelques tests de reconnaissance. Sans changement de la structure du modèle générique, nous avons fait varier les pondérations associées aux trois types d'éléments (items) à rejeter et l'heuristique qui permet de contrôler la stratégie d'élagage dans l'espace de recherche. Nous représentons dans le tableau 2 les meilleures performances que nous avons obtenues pour ces trois modèles de rejet et dans le tableau 3 les performances qui correspondent au meilleur taux d'erreurs total sur les mots (21,0%). (FA : Fausses Alarmes, RàT : Rejet à Tort, Ins: Insertion, Omi : Omissions, Prc : *Precision*, Rcl : *Recall*).

	FA	RàT,Ins	Omi	Rcl	Prc
<R.GLOBAL>	12.5%	0.9%	0.0%	87.5%	88.7%
<R.UNK_OOV>	13.4%	5.7%	1.5%	85.1%	20.4%
<R.UNK_SPR>	72.4%	0.3%	24.1%	3.4%	7.8%

Table 2: Performances de reconnaissance sur le corpus *PlanResto*. Taux d'erreurs total sur les mots 24,5%

Nous avons, par ailleurs, mesuré la perplexité du modèle de langage utilisé. Sur un corpus de test de 3902 mots, le bigram donne une perplexité de 14,2. La contribution des mots HV à la perplexité lorsqu'ils sont prédécesseurs est égale à 1,022 et est égale à 1,056 lorsqu'ils sont successeurs. Cela veut dire, que dans notre corpus, il est plus facile de prédire un mot à partir d'un mot HV que de prédire un mot HV à partir d'un autre mot.

	FA	RàT,Ins	Omi	Rcl	Prc
<R.GLOBAL>	25.0%	0.0%	0.0%	75%	100%
<R.UNK_OOV>	73.1%	0.7%	3.0%	23.9%	37.2%
<R.UNK_SPR>	65.5%	0.1%	34.9%	0%	0%

Table 3: Performances de reconnaissance pour un taux d'erreur total sur les mots de 21,0%

3. ANALYSE STATISTIQUE

Afin d'améliorer les performances du système présentées dans la section 2.2, nous avons jugé important d'améliorer d'abord nos connaissances sur les caractéristiques des mots HV et de les comparer avec celles des autres éléments du corpus.
Les mots HV considérés sont des mots complets. À la différence de *Hetherington* dans son étude [4], les phénomènes de la parole spontanée ne sont pas considérés comme des mots HV. En étudiant la longueur des énoncés en nombre de mots, nous avons remarqué que les mots HV apparaissent en moyenne dans des énoncés plus longs que les autres. La moyenne calculée sur des énoncés contenant des mots HV est de 6,2 mots par énoncé alors que la moyenne sur des énoncés sans mots HV est de 3,8 mots par énoncé. Le nombre moyen de mots tous énoncés confondus est de 4,1. Cette dernière constatation mérite d'être développée pour être exploitée lors d'un éventuel post-traitement sur un graphe de mots ou bien dans un module de décision sur le chemin de décodage.

3.1. Étude phonétique

Dans les systèmes de reconnaissance de la parole, un mot HV est généralement représenté acoustiquement par un modèle poubelle. C'est une boucle inconditionnelle qui permet de modéliser toutes les possibilités de suites de phonèmes.
Pour permettre une meilleure sélection des mots HV par le modèle acoustique, des travaux récents tentent de modéliser les séquences entre les phonèmes en appliquant des modèles de langages [2], d'autres moins récents utilisent la contrainte syllabique comme un moyen pour limiter les possibilités de séquences entre phonèmes [5]. Nous ne nous intéressons pas en ce début de section aux contraintes sur les séquences mais plutôt aux contraintes sur la longueur. Intuitivement, et en se basant sur la définition des mots HV dans le cadre de la parole spontanée, on peut supposer qu'un mot HV ne possède pas forcément des propriétés phonétiques différentes des mots du vocabulaire. L'étude qu'a mené *Suhm* [8] sur du texte écrit (WSJ[2]) permet de nuancer cette supposition. Il montre qu'en choisissant des lexiques qui couvrent entre 72.8% et 95.8% du corpus de test, les distributions de la longueur phonétique des ensembles de mots HV issus des différents choix de lexique sont semblables et qu'en prenant un lexique qui couvre 93.9% du corpus de test, les mots HV sont significativement plus longs que tous les mots du vocabulaire. Nous voudrions nous aussi nuancer cette précédente supposition. Nous pensons que la nature de l'application et le bon choix du lexique sont des paramètres qui influ-

[2]La base de données du Wall Street Journal

encent les caractéristiques des mots HV. Nous avons fait l'évaluation sur de la parole spontanée (le corpus *Plan-Resto*). Nous avons calculé la *longueur phonétique* de chaque variante de prononciation déterminée par le phonétiseur. Dans la figure 1, nous donnons les différentes distributions de la longueur phonétique selon la plus courte prononciation des éléments du corpus.

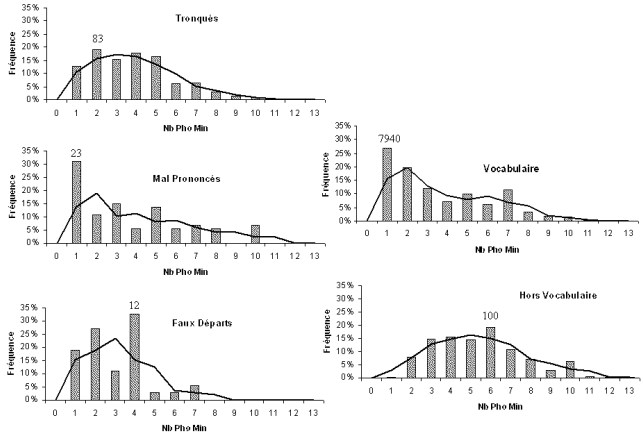

Figure 1: Distributions de la longueur phonétique minimale des éléments du corpus

Les mots HV se distinguent par deux constats :
1- À la différence des mots du vocabulaire, des mots tronqués et des phénomènes de la parole spontanée (faux départs, mots mal prononcés), les mots très courts sont beaucoup moins représentés en proportion dans les mots HV,

$$P(L = 1/w_{\overline{HV}}) \gg P(L = 1/w_{HV}) \qquad (1)$$

2- Le lobe central de la distribution des mots HV se situe entre 4 et 6 phonèmes alors qu'il se situe entre 1 et 4 phonèmes pour les autres distributions.
La longueur phonétique moyenne des w_{HV} est égale à 5,4 alors que les $w_{\overline{HV}}$ ont une longueur phonétique moyenne égale à 3,6.
Les histogrammes de la figure 1 permettent, pour chaque longueur phonétique, de déduire le rapport :

$$R(L) = \frac{P(w_{HV}/L)}{P(w_{\overline{HV}}/L)} = \frac{P(L/w_{HV})P(w_{HV})}{P(L/w_{\overline{HV}})P(w_{\overline{HV}})} \qquad (2)$$

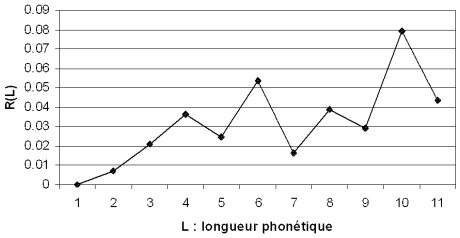

Figure 2: Rapport des probabilités conditionnelles $R(L)$

Les faibles valeurs obtenues résultent de la faible proportion des mots HV dans le corpus. Le rapport $R(L)$ est quasi nul pour des mots de longueur 1 et augmente en fonction de la longueur
Ces observations, faites sur de la parole spontanée, confortent les remarques de *Suhm* [8] et donnent des justifications pour l'introduction d'une restriction sur la longueur phonétique minimale du modèle acoustique générique des mots HV pour notre application.
Nous nous sommes intéressés par la suite au contenu phonétique des mots du vocabulaire et des mots HV. Nous avons calculé la fréquence d'occurrence des quatre catégories de phonèmes : les Plosives-Fricatives-Nasales (PFN), les Liquides (L), les Voyelles (V) et les Semi-Voyelles (SV). Un modèle phonétique bouclé permet équiprobablement le passage par les différents phonèmes, ceci implique que la probabilité de passage par les PFN est légèrement supérieure à celle des V dans la mesure où le nombre de PFN est supérieur à celui des V. Le même raisonnement est tenu pour les SV et les L. Nous avons observé le contraire pour les mots HV et les mots du vocabulaire. En effet, le taux de PFN est inférieur à celui des V et le taux de L est supérieur à celui des SV ($P(V) > P(PFN) > P(L) > P(SV)$). Il s'ensuit qu'un modèle phonétique bouclé sans contraintes caractérise mal les séquences incorrectes de phonèmes. Ceci va jouer en faveur de l'introduction de contraintes de successions sur la suite de phonèmes. Le cas particulier des contraintes syllabiques est détaillé dans la section suivante.

3.2. Étude syllabique

La syllabe est l'un des moyens utilisés pour spécifier les contraintes de successions entre les phonèmes. Le modèle générique syllabique est ainsi plus robuste aux suites de phonèmes improbables [5]. La syllabe est aussi un intermédiaire entre mot et phonème, c'est un compromis entre unités plus longues et moins nombreuses [3].
Outre le fait que les modèles génériques syllabiques intègrent une information supplémentaire déjà présente dans les mots mais non spécifiée par un modèle phonétique, que nous apporte la syllabe si l'on cherche seulement à caractériser la longueur des mots? Pour répondre à cette question, nous avons étudié les *longueurs syllabiques* de chaque élément du corpus [3]. Nous avons déterminé pour chaque mot différents découpages syllabiques correspondants aux différentes variantes de prononciations. Par conséquent, nous avons été en mesure de donner pour chaque mot le nombre de syllabes minimal, maximal et moyen ainsi que les différentes syllabes constituant ce mot. La figure 3 présente les distributions relatives aux éléments du corpus selon le nombre de syllabes moyen calculé sur les différentes prononciations.
Les mots HV se caractérisent par une longueur syllabique moyenne significativement plus grande que les autres éléments du corpus. La tableau 4 récapitule les différentes moyennes des longueurs syllabiques calculées.

	HV	Voc.	Faux départs	Tronqués	Mal Pronon.
Long. syll. moy.	2.6	1.8	1.5	1.9	1.9

Table 4: Longueur syllabique moyenne

50.68% des occurrences des mots du vocabulaire sont monosyllabiques alors que 93.6% des mots HV ne le sont pas. Ceci s'explique par le fait que la majorité des mots très courts sont des mots outils déjà inclus dans le vocabulaire.
Malgré le fait que les mots HV se distinguent par le peu de mots monosyllabiques, il n'est pas simple d'imposer un minimum de deux syllabes dans un modèle syllabique

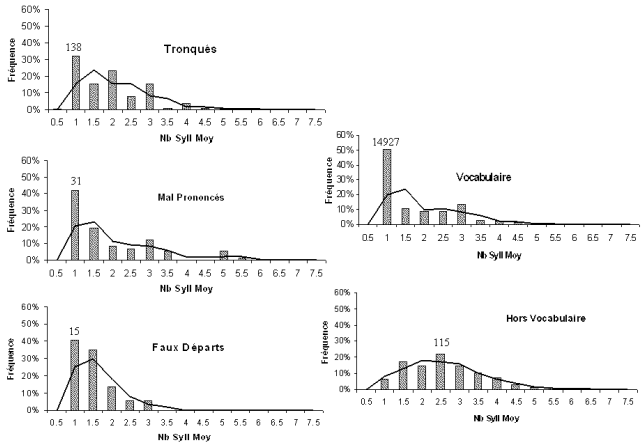

Figure 3: Distribution de la longueur syllabique moyenne des éléments du corpus

générique lorsque la longueur phonétique moyenne d'une syllabe est égale à 2,1 dans notre corpus. Par ailleurs, les mots du vocabulaire contiennent 909 syllabes différentes et les mots HV 436, ce qui laisse penser que le modèle syllabique bouclé serait assez gourmand en temps de calcul. Nous pensons qu'il serait plus intéressant d'utiliser la syllabe comme contrainte sur les séquences de phonèmes ou comme mesure de confiance sur la suite de phonèmes scrutés dans le modèle phonétique que comme unité dans un modèle générique syllabique. Cela peut être fait en exploitant la définition de la syllabe et les règles de sa génération.

3.3. Les catégories de mots Hors-Vocabulaire

Un mot HV bien détecté, c'est sans doute un apport non négligeable pour le système d'interprétation de la requête. Un mot HV bien détecté avec une information en plus sur sa catégorie sémantique ou grammaticale, c'est sans contestation un apport important pour le système d'interprétation. Afin d'aboutir ultérieurement à cet objectif, nous avons essayé de caractériser le contexte d'apparition des mots HV en nous basant sur leur contenu sémantique ou bien grammatical. Nous avons déterminé 7 catégories parmi les 242 mots HV : <Origine-ville-pays> (44) : origine d'une spécialité; <Endroit> (16) : endroit pour situer un restaurant à Paris; <Nom-type-plat> (57) : types de plats et de cuisine; <Nom-restaurant> (16) : noms de restaurants; <Verbe> (34) : verbes conjugués ou non; <Ambiance-restaurant> (3) : caractérise l'ambiance dans les restaurants; <Autres> (72).
Des analyses sur les fréquences d'apparition, de répétition et sur la longueur phonétique des mots de chaque classe ont été faites. Nous avons observé les phénomènes suivants : les <Nom-type-plat> apparaissent dans des énoncés de longueur moyenne, sont fortement répétés dans une même session et sont assez courts par rapport aux autres mots HV. Les <Nom-restaurant> quant à eux sont des mots assez longs, presque quatre phonèmes de plus par rapport à la moyenne. Pour certaines classes (<Nom-restaurant> par exemple), nous avons remarqué une forte présence de ces mots en fin d'énoncé. C'est vraisemblablement dû à la particularité de la langue et au domaine de l'application.
Les modèles de langage peuvent être utilisés pour prédire les catégories des mots HV. Pour cela, il est nécessaire de remplacer chaque occurrence de mot HV dans le corpus par l'étiquette de la catégorie qui lui est associée. Les tests de perplexité que nous avons effectués nous ont poussé à conclure que la quantité et la qualité des données dans chaque catégorie jouent un rôle important. En effet, peu de données dans une catégorie ou bien la non-homogénéité des données au sein d'une même catégorie dégradent les performances de prédiction des classes du modèle de langage. Les premières observations permettent quand même de dégager des comportements propres à certaines catégories et d'envisager leur détection.

4. CONCLUSION ET PERSPECTIVES

Ce travail nous a permis d'examiner les mots HV issus d'un contexte de parole spontanée. Il apparaît que les mots HV sont assez fréquents, qu'ils sont présents dans des énoncés plus longs que les autres et qu'ils sont phonétiquement et syllabiquement plus longs en moyenne que les autres éléments du corpus. Ils se distinguent également par leur rareté parmi les mots courts et par leur forte présence en fin d'énoncé pour certaines catégories d'entre eux. Nous avons pu observer, par ailleurs, une distinction possible entre différentes catégories de mots HV. Dans nos travaux en cours, nous avons commencé la mise en oeuvre des constatations issues de l'analyse phonétique (application des statistiques sur la longueur des mots HV). L'exploration des autres voies s'inscrira dans nos prochaines considérations.

BIBLIOGRAPHIE

[1] A. Asadi, R. Schwartz, and J. Makhoul. Automatic detection of new words in large vocabulary continuous speech recognition system. *Proc ICASSP*, pages 125–128, 1990.

[2] I. Bazzi and J. R. Glass. Modeling Out-Of-Vocabulary words for robust speech recognition. *Proc* 6^{th} *ICSLP*, 2000.

[3] M. El-Bèze. Choix d'unités appropriées et introduction de connaissances dans des modèles probabilistes pour la reconnaissance automatique de la parole. Thèse de doctorat, 1990.

[4] I. L. Hetherington and V. M. Zue. New words: implications for continuous speech recognition. *Eurospeech*, pages 2121–2124, 1993.

[5] T. Kemp and A. Jusek. Modelling unknown words in spontaneous speech. *Proc ICASSP*, pages 530–533, 1996.

[6] N. Moreau and D. Jouvet. Use of a confidence measure based on frame level likelihood ratios for the rejection of incorrect data. *Eurospeech*, pages 291–294, 1999.

[7] T. Schaaf and T. Kemp. Confidence measure for spontaneous speech recognition. *Proc ICASSP*, pages 875–878, 1997.

[8] B. Suhm, M. Woszczyna, and A. Waibel. Detection and transcription of new words. *Eurospeech*, pages 2179–2182, 1993.