Combinaison de systèmes pour la phonétisation automatique de noms propres

Antoine Laurent, Sylvain Meignier, Yannick Estève, Paul Deléglise

Combinaison de systèmes pour la phonétisation automatique de noms propres

Antoine Laurent[†§], *Sylvain Meignier*[†], *Yannick Estève*[†], *Paul Deléglise*[†]

LIUM[†] – Laboratoire d'Informatique de l'Université du Maine – Le Mans
prenom.nom@lium.univ-lemans.fr
Spécinov[§] – Trélazé
a.laurent@specinov.fr

ABSTRACT

Large vocabulary automatic speech recognition (ASR) technologies perform well in known, controlled contexts. However recognition of proper nouns is commonly considered as a difficult task. Accurate phonetic transcription of a proper noun is difficult to obtain, although it can be one of the most important resources for a recognition system. In this article, we propose methods of automatic phonetic transcription applied to proper nouns. The methods are based on combinations of the rule-based phonetic transcription generator LIA_PHON and an acoustic-phonetic decoding system. On the ESTER corpus, we observed that the combined system obtained better results than our reference system (LIA_PHON). The WER decreased on segments of speech containing proper nouns, without affecting negatively the results on the rest of the corpus.

1. Introduction

Les systèmes de reconnaissance vocale à grand vocabulaire ont des performances correctes dans des contextes d'utilisation connus et contrôlés. Cependant, les noms propres sont fréquemment des mots hors vocabulaire et leur reconnaissance est généralement considérée comme une tâche difficile.

De nombreuses situations nécessitent de transcrire correctement les noms propres. Il est généralement intéressant de savoir qui parle et quand, en particulier lors de tâches comme l'indexation multimédia, la recherche documentaire, la transcription et le compte-rendu de réunions. Bien que la phonétisation des mots soit l'une des principales ressources nécessaires au bon fonctionnement du système de reconnaissance vocale, la phonétisation des noms propres est difficile à obtenir. En effet, un nom propre écrit de la même manière sera prononcé différemment selon l'origine de ce nom et selon l'origine du locuteur.

La prononciation des noms propres est moins normalisée que la prononciation des autres mots. Une suite de lettres dans un mot quelconque sera généralement prononcée de la même manière quel que soit le mot dans lequel cette suite de lettres apparaîtra. Ce n'est pas le cas pour les noms propres. Il est difficile, lors de la lecture de certains noms propres, de déterminer la façon dont ils doivent être prononcés.

Deux approches communes au problème de la phonétisation automatique sont proposées dans la littérature : l'approche par système à base de règles [2], et l'approche utilisant des statistiques comme les arbres de classification [5] ou les HMM [4, 1]. Pour le cas particulier des noms propres, une étude sur la génération dynamique des déformations plausibles des formes canoniques de noms propres est proposée dans [3]. Cette étude a été menée dans le cadre d'une application d'annuaire téléphonique développée par France Télécom R&D. La méthode utilisée consiste à réévaluer les n meilleures hypothèses de reconnaissance générées lors d'une passe de décodage dans laquelle les déformations dépendent de la nature des hypothèses en compétition.

La méthode que nous proposons repose sur la combinaison d'un générateur utilisant des règles de transcription phonétique et d'un système de décodage acoustico-phonétique. Ce dernier permet l'extraction d'un nombre élevé de transcriptions phonétiques, y compris quelques prononciations peu communes. Les transcriptions phonétiques sont obtenues à partir du décodage des portions de signal contenant le mot cible. Ces portions sont détectées et extraites automatiquement d'enregistrements préalablement transcrits en mots. Le générateur utilisant des règles, quant à lui, tend à produire des transcriptions phonétiques respectant les principes généraux de conversion graphèmes vers phonèmes.

La méthode proposée est appliquée, comme dans [3], à la phonétisation automatique de noms propres en vue d'améliorer la transcription automatique de journaux radiophoniques. Les nouvelles phonétisations générées seront évaluées en terme de taux d'erreur mot (Word Error Rate — WER), et en terme de taux d'erreur nom propre (Proper Noun Error Rate — PNER). Ces taux seront évalués sur le corpus de la campagne d'évaluation ESTER [8].

Tout d'abord, nous allons présenter les avantages et les inconvénients du générateur utilisant des règles ainsi que du système de décodage acoustico-phonétique. Ensuite, nous présenterons une combinaison de ces deux méthodes. En dernier lieu, nous commenterons et présenterons les résultats obtenus.

2. Systèmes de phonétisation automatique

2.1. À base de règles

LIA_PHON est un système de phonétisation automatique à base de règles [2]. Il utilise la graphie des mots pour déterminer les suites de phonèmes correspondantes. L'un des atouts de ce système est d'être capable de phonétiser des mots sans nécessiter le signal sonore correspondant.

LIA_PHON a participé à la campagne de tests des phonétiseurs français connue sous le nom d'ARC B3 (Action de Recherche Concertée B3). Les phonétisations générées par le système ont été comparées à des textes phonétisés par des experts. Le taux d'erreur phonème a été calculé sur la même base que le taux d'erreur mot utilisé en Reconnaissance Automatique de la Parole. Sur 86938 phonèmes, 99,3% des transcriptions phonétiques générées par LIA_PHON ont été identifiées comme correctes. Cependant, les résultats présentés dans [2] montrent une répartition non uniforme des erreurs selon les classes de mots. 25,6% des erreurs générées par LIA_PHON sont issues de la phonétisation des noms propres qui ne représentent pourtant que 5,8% des mots du corpus de test.

En effet, la phonétisation des noms propres présente un haut niveau de variabilité difficile à prédire. Par exemple, dans le corpus de développement ESTER, le prénom du chanteur "Joey Starr" est prononcé de quatre manières différentes ("dZoe", "dZoj", "Zoe", ou "Zoj" en format Sampa), bien que tous les locuteurs parlent français et connaissent ce chanteur. Cette variabilité illustre la très grande difficulté pour générer l'ensemble des règles permettant de produire l'ensemble des variantes de phonétisation.

Idéalement, le système devrait être capable de détecter à la fois l'origine du nom propre et la façon dont les gens, en fonction de leurs origines socio-culturelles, pourraient prononcer ce nom. Malheureusement, ces deux tâches sont très complexes, voire impossibles, car le système ne dispose pas d'informations *a priori* sur l'origine du locuteur.

2.2. Décodage acoustico-phonétique

Un système de décodage acoustico-phonétique (DAP) permet de générer la suite de phonèmes la plus probable correspondant à un signal sonore. Pour obtenir les phonétisations automatiques des noms propres, les portions du signal correspondant aux mots à phonétiser sont extraites automatiquement des transcriptions manuelles. Ces portions sont ensuite décodées en utilisant le système de DAP. Les noms propres qui sont présents plusieurs fois dans le corpus peuvent donc être phonétisés différemment, permettant ainsi d'obtenir des variantes de prononciations.

Les auteurs de [4] indiquent qu'un décodage non contraint ne permet pas d'obtenir un décodage phonétique fiable. Nous sommes arrivés expérimentalement à la même conclusion.

L'utilisation d'un modèle de langage permet de guider le système de reconnaissance vocale en minimisant le risque de voir apparaître des séquences de phonèmes très peu probables. Le décodage a été contraint en utilisant des triphones à états partagés et un modèle de langage 3-gram pour générer la meilleure hypothèse de transcription de phonèmes.

Ce système est très proche d'un système de décodage utilisé en reconnaissance de la parole, mis à part que son lexique et son modèle de langage ne contiennent que des phonèmes à la place de mots. Ce modèle de langage 3-gram a été appris à partir du dictionnaire de phonétisations utilisé pour la tâche de transcription lors de la campagne ESTER 2005. Il contient environ 65000 variantes de mots créés à l'aide de BDLEX [6] et de LIA_PHON. Seuls les mots absent de BDLEX ont été phonétisés automatiquement avec LIA_PHON. Les mots identifiés comme étant des noms propres ont été supprimés de ce dictionnaire avant d'apprendre le modèle de langage 3-gram pour les phonèmes.

Comme expliqué plus haut, la première étape consiste à isoler les portions du signal qui correspondent aux noms propres en utilisant la transcription manuelle à notre disposition. Les mots de la transcription manuelle ne sont pas alignés sur le signal sonore : les instants de début et de fin de chaque mot ne sont pas disponibles. Seuls les instants de début et de fin de chaque segment ("phrase" contenant plusieurs mots) sont à notre disposition. Les frontières temporelles de chaque mot des segments sont déterminées en les alignant sur le signal à l'aide d'un système d'alignement forcé. La phonétisation des noms propres utilisée pour cette tâche est obtenue avec LIA_PHON. Les transcriptions phonétiques des noms propres générés de cette manière sont par hypothèse erronées. Par conséquent, la détection des frontières peut être imprécise. Les portions du signal détectées peuvent ainsi parfois chevaucher les mots voisins du nom propre. Lorsque le système de DAP est appliqué à ces portions de signal, il peut générer des phonèmes erronés en début et/ou fin de nom propre ce qui introduit une cause d'erreur lors de l'utilisation ultérieure de cette variante en transcription.

3. Combinaison

La combinaison des deux méthodes de transcription phonétique vise à tirer partie de chacune tout en ne dégradant pas le reste du processus de reconnaissance vocale.

3.1. Union

La première méthode de combinaison suit la plus simple des stratégies, en construisant un dictionnaire de phonétisations contenant l'union des phonétisations générées par LIA_PHON et par le DAP. Cette méthode génère un nombre important de variantes de phonétisations (voir section 4.2).

3.2. Sélection

Un dictionnaire de phonétisations contenant un nombre excessif de variantes augmente le taux d'erreur mot lors de son utilisation en transcription. Par-

tant de cette constatation, nous proposons de conserver uniquement les variantes de phonétisation utilisées par le système de transcription lors d'un décodage du corpus de développement.

Pour chaque variante de phonétisation de chaque nom propre, un dictionnaire temporaire est construit. Ce dictionnaire contient uniquement la forme de phonétisation du nom propre à évaluer et l'ensemble des autres mots qui ne sont pas des noms propres. Les phrases qui contiennent ce nom propre sont transcrites en utilisant le dictionnaire temporaire. La phonétisation du nom propre est considérée comme valide si le nom propre apparaît au moins une fois dans le résultat de la transcription des phrases. Dans ce processus, les autres mots du dictionnaire temporaire jouent le rôle d'un modèle de rejet quand nous essayons de reconnaître le nom propre évalué.

4. Expériences

4.1. Corpus

Les expériences ont été menées sur le corpus ESTER. ESTER est une campagne d'évaluation française de transcription d'émissions radiophoniques qui s'est déroulée en janvier 2005 [8]. Le corpus ESTER a été divisé en trois parties : entraînement, développement et évaluation.

Le corpus d'entraînement est composé de 81 heures d'émissions radiophoniques provenant de quatre stations (France Inter, France Info, RFI et RTM). Ce corpus a été utilisé pour l'entraînement du système de reconnaissance vocale. Le corpus de développement est composé de 12,5 heures d'émissions radiophoniques provenant des mêmes radios. Ce corpus a été utilisé pour générer et valider les phonétisations du DAP. Le corpus de test, utilisé pour évaluer les méthodes proposées, contient 10 heures d'enregistrement provenant des mêmes stations de radio et de deux autres stations, enregistrées 15 mois après le corpus de développement.

En plus de la transcription orthographique, les entités nommées, dont les noms de personnes, sont annotées dans l'ensemble des corpus.

4.2. Modèles acoustiques et linguistiques

Le système de décodage utilise le logiciel CMU Sphinx 3.6 de la Canergie Mellon University. Nos expériences ont été menées en décodant le signal caractérisé par 12 paramètres acoustiques MFCC avec l'énergie complétés de leurs dérivées première et seconde. Les modèles acoustiques ont été appris sur le corpus d'entraînement d'ESTER, ils dépendent du genre du locuteur. Le modèle de langage a été entraîné sur les transcriptions manuelles du corpus comptant 1,35 millions de mots. Des articles du journal français "Le Monde" ont été ajoutés, menant le nombre de mots à 319 millions. Tous les noms propres sont présents dans le modèle de langage. Lors des expériences, le même modèle de langage a été utilisé aussi bien lors du développement que lors de l'évaluation des dictionnaires créés. Tous les dictionnaires contiennent les même noms propres, seules leurs phonétisations changent.

Variantes de phonétisation La figure 1 présente le nombre de variantes de phonétisation générées pour les noms propres présents dans le corpus de développement en fonction de la méthode utilisée. Le corpus de développement ESTER contient 1098 noms propres différents, présent 4791 fois.

Le système de phonétisation automatique à base de règles LIA_PHON génère 1443 phonétisations différentes sur ce corpus, soit une moyenne d'environ 1,31 variantes de phonétisation par nom propre.

Sur le même corpus, le système de DAP génère 3881 phonétisations, soit approximativement 3,53 variantes par nom propre. Le nombre de variantes est plus de 2,5 fois supérieur au nombre de variantes générées par LIA_PHON.

L'union des variantes de phonétisations générées par LIA_PHON et par le DAP monte ce nombre de phonétisations à 3984, soit ≈3,63 variantes par nom propre. La technique de sélection visant à éliminer les variantes de prononciation superflues générées par le DAP ramène le nombre de variantes à 3523, soit ≈3,21 variantes par nom propre.

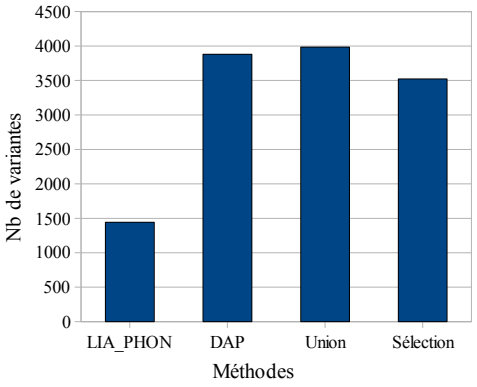

Fig. 1: Nombre de variantes de phonétisation générées selon la méthode utilisée.

4.3. Métrique

Nous proposons d'évaluer la qualité des phonétisations des noms propres créés en terme de taux d'erreur mot (Word Error Rate — WER) et de taux d'erreur nom propre (PNER — Proper Noun Error Rate). Le PNER est calculé de la même manière que le taux d'erreur mot classique utilisé en Reconnaissance Automatique de la Parole à la différence qu'il est appliqué non pas à l'ensemble des mots mais uniquement aux noms propres :

$$PNER = \frac{I + S + E}{N} \quad (1)$$

avec I le nombre d'insertions erronées de noms propres, S le nombre de substitutions de noms propres par un autre mot (ou un autre nom propre), E le nombre d'élisions de noms propres (c'est-à-dire le nombre de noms propres "supprimés" dans la transcription) et N le nombre total de noms propres.

Le WER permet d'évaluer l'impact du dictionnaire sur l'ensemble du corpus de test, alors que le PNER

permet d'évaluer la qualité de la détection des noms propres.

4.4. Résultats

Le tableau 1 présente les résultats de décodage obtenus en utilisant les différents ensembles de phonétisations de noms propres générés.

	LIA_PHON	**Union**	**DAP**	**Sélection**
PNER	**26,0%**	21,5% (-4,5%)*	32,3% (+6,3%)*	**22,1% (-3,9%)***
WER	26,8%	26,9%	27,2%	26,8%

Tab. 1: Résultats du décodage sur le corpus de test ESTER.
* : Différence entre la méthode évaluée et le système de référence LIA_PHON

Le système utilisant LIA_PHON uniquement obtient 26,8% de WER et un taux de 26,0% de PNER. Ce système est notre système de référence.

Le système utilisant les phonétisations obtenues à partir d'un DAP seul obtient les plus mauvais WER et PNER : respectivement 27,2% et 32,3%.

L'union de LIA_PHON avec le système de DAP obtient les meilleures performances en terme de PNER. Cependant, le WER est légèrement plus élevé (0,1 %) que dans le système de référence. D'autres mots sont substitués par des noms propres faisant augmenter le WER.

La stratégie de sélection sur les phonétisations générées par le système de DAP permet de réduire le nombre de variantes de phonétisation. L'union de ces phonétisations filtrées avec les phonétisations générées par LIA_PHON est référencée en tant que "Sélection" dans le tableau 1. Avec ce système, nous observons un gain de 3,9 % en terme de PNER sans dégrader le WER. Bien que cette méthode soit moins performante sur les noms propres que l'union, elle permet de ne pas dégrader les performances de la transcription.

Le taux d'erreur mot est peu affecté par le changement de dictionnaire, car les noms propres ne représentent qu'une petite partie des mots du corpus. Le corpus de test compte 1840 noms propres sur 113918 mots, c'est à dire approximativement 1,6% des mots. Pour observer l'influence des différentes méthodes proposées en terme de WER, nous proposons d'évaluer séparément les segments qui contiennent des noms propres. Le tableau 2 montre les résultats pour les segments contenant ou ne contenant pas de noms propres.

Les résultats les plus significatifs sont les résultats obtenus avec la méthode de "Sélection" : elle permet un gain de 0,5 point en terme de WER par rapport à LIA_PHON sur les segments contenant des noms propres sans affecter le WER des autres segments.

Segments	**LIA_PHON**	**Union**	**DAP**	**Sélection**
avec noms propres	**26,9%**	26,4% (-0,5%)*	28,4% (+1,5%)*	**26,4% (-0,5%)***
sans noms propres	26,8%	27,0%	27,0%	26,8%

Tab. 2: Taux d'erreur mot (WER) sur le corpus de test d'ESTER pour les segments contenant des noms propres, et pour ceux n'en contenant pas.
* : Différence entre la méthode évaluée et le système de référence LIA_PHON

5. Conclusion

Cet article présente une méthode de génération automatique de phonétisations de noms propres. Nous avons proposé différentes façons de combiner un système de phonétisation automatique à base de règles (LIA_PHON) avec un système de décodage acoustico-phonétique.

Sur le corpus ESTER, le système de combinaison proposé obtient de meilleurs résultats que le système de référence (LIA_PHON). Avec cette méthode de combinaison, le taux d'erreur sur les mots diminue de 0,5 point sur les segments de parole contenant des noms propres sans dégrader le reste du corpus. La combinaison permet également d'observer un gain de 3,9 points en terme de taux d'erreur sur les noms propres.

La méthode proposée pourra être utilisée en identification nommée [7]. Cette tâche consiste à trouver les identités des locuteurs (nom et prénom) à partir de la transcription. Les nouvelles phonétisations des noms propres permettant d'obtenir un meilleur décodage, elles devraient faciliter la détection des noms de locuteurs.

Les développements futurs pourront s'attacher à la généralisation de notre méthode sur les autres mots du système identifiés comme étant des mots fréquemment mal décodés.

Références

[1] L. R. Bahl, S. Das, P. V. deSouza, M. Epstein, R. L. Mercer, B. Merialdo, D. Nahamoo, M. A. Picheny, and J. Powell. Automatic phonetic baseform determination. In *ICASSP*, pages 173–176, December 1991.

[2] F. Béchet. LIA_PHON : un système complet de phonétisation de textes. In *TAL*, pages 47–67, 2001.

[3] F. Béchet, R. de Mori, and Subsol G. Dynamic generation of proper name pronunciations for directory assistance. In *ICASSP*, pages 745–748, 2002.

[4] M. Bisani and H. Ney. Breadth-first for finding the optimal phonetic transcription from multiple utterances. In *Eurospeech*, 2001.

[5] R. I. Damper, Y. Marchand, M. J. Adamson, and K. Gustafson. Automatic phonetic baseform determination. In *ESCA*, pages 53–58, 1998.

[6] M. De Calmes and G. Perennou. BDLEX : a lexicon for spoken and written french. In *LREC*, pages 1129–1136, 1998.

[7] Y. Estève, S. Meignier, Deléglise P., and J. Mauclair. Extracting true speaker identities from transcriptions. In *ICSLP*, 2007.

[8] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J. F. Bonastre, and G. Gravier. The ESTER phase II evaluation campaign for the rich transcription of french broadcast news. In *Eurospeech*, Lisbon, Portugal, September 2005.