Reconnaissance de parole non native fondée sur l'utilisation de confusion phonétique et de contraintes graphèmiques

Ghazi Bouselmi, Dominique Fohr, Irina Illina, Jean-Paul Haton

Projet Parole, LORIA-CNRS & INRIA, "http ://parole.loria.fr/", BP 239, 54600 Vandoeuvre-lès-Nancy, France
{ bousselm,fohr,illina,jph }@loria.fr

ABSTRACT

This paper presents a fully automated approach for the recognition of non native speech based on acoustic model modification. For a native language (LM) and a spoken language (LP), pronunciation variants of the phones of LP are automatically extracted from an existing non native database. These variants are stored in a confusion matrix between phones of LP and sequences of phones of LM. This confusion concept deals with the problem of non existence of match between some LM and LP phones. The confusion matrix is then used to modify the acoustic models (HMMs) of LP phones by integrating corresponding LM phone models as alternative HMM paths. We introduce graphemic contraints in the confusion extraction process. We claim that prononciation errors may depend on the graphemes related to each phone. The modified ASR system achieved a significant improvement varying between 20.3% and 43.2% (relative) in "*sentence error rate*" and between 26.6% and 50.0% (relative) in "*word error rate*". The introduction of graphemic contraints in the phonetic confusion allowed improvements while using the word-loop grammar.

1. INTRODUCTION

La dégradation des performances des systèmes de reconnaissance automatique de la parole (SRAP) confrontés à un locuteur non natif est un problème bien connu. Une solution pour améliorer les performances des RAPs en présence de parole non native consiste à augmenter leur tolérance face à des variations de prononciation. Le problème consiste à introduire dans ces systèmes des connaissances relatives à ces variantes de pronciations. Différentes approches ont été envisagées et diffèrent selon la manière d'extraire ou d'intégrer ces connaissances dans le système de reconnaissance. Quelques unes de ces approches sont brièvement décrites ci-dessous.

Dans [6], des experts en phonétique ont établi des règles de réécriture phonétique pour quelques couples de langue parlée (LP) et langue maternelle (LM). Ces règles transforment les phonèmes de la LP en phonèmes de la LM. Ainsi, toutes le prononciations alternatives exprimées en termes de phonèmes de la LM sont ajoutées au lexique.

Dans [3], pour chaque phrase du corpus, un alignement forcé de la prononciation canonique et une reconnaissance phonétique (en termes de phonèmes de la LP) sont effectués. Ces deux transcriptions sont ensuite comparés afin d'extraire une confusion phonétique. Enfin, cette dernière sert à ajouter les prononciations alternatives de chaque mot dynamiquement durant la phase de reconnaissance.

Dans [4], une confusion phonétique est extraite d'une façon similaire à la précédente. Toutefois, la prononciation canonique est alignée avec une prononciation phonétique exprimée en termes de phonèmes natifs. Les deux SRAPs de la langue parlée et maternelle sont utilisés à cet effet. Par la suite, les modèles gaussiens des phonèmes natifs sont fusionnés avec ceux des phonèmes natifs avec lesquels ils ont été confondus, et ce pour chaque état des modèles de Markov sous-jacents (HMM).

Suite à une étude de prononciations non native réalisée avec des locuteurs français, nous avons mis en exergue deux principaux problèmes. Nous avons constaté que les locuteurs non natifs ont tendance à prononcer certains phonèmes de la LP comme des phonèmes de leur langue maternelle. Par exemple, certains phonèmes de la LP qui n'existent pas dans la LM sont souvent réalisés comme des phonèmes acoustiquement proches de la LM. C'est le cas de la consonne anglaise '[ð]' (dans le mot *the*) qui est souvent prononcée '[z]' par les locuteurs français. Le second problème, que nous avons constaté, est que la graphie d'un mot influence le locuteur non natif dans la façon de prononcer ce mot. Face à des mots qu'ils ne connaissent pas, ou encore des mots dont la graphie est identique dans les deux langues, ces locuteurs ont souvent tendance à réaliser une prononciation similaire à leur LM. Pour illustrer cela, considérons l'exemple de la table 1 où sont présentées les prononciations canoniques et les prononciations réalisées par des locuteurs français pour les mots anglais "*approach*" et "*position*". La table 1 montre que le phonème anglais '[ə]' est réalisé comme le phonème français '[a]' lorsqu'il correspond au caractère 'a' et comme le phonème français '[ɔ]' lorsqu'il correspond au caractère 'o'.

TAB. 1: *Prononciation du phonème anglais ə*

Mot	Prononciation canonique (phonèmes anglais)	Réalisation acoustique par des français (phonèmes français)
Approach	**[ə]** [p] [r] [əʊ] [tʃ]	**[a]** [p] [r] [ɔ] [t] [ʃ]
Position	[p] **[ə]** [z] [i] [ʃ] **[ə]** [n]	[p] **[ɔ]** [z] [i] [ʃ] **[ɔ]** [n]

Une des difficultés à laquelle sont confrontés les locuteurs non natifs est que certains phonèmes de la LP n'ont pas de correspondant directs dans la LM. C'est le cas de la diphtongue anglaise '[aɪ]' qui peut être réalisée comme la suite de phonèmes italiens '[a] [i]'. Dans notre approche [1], nous avons introduit un nouveau concept de confusion

phonétique qui associe à un phonème de la LP une suite de phonèmes de la LM. Ce concept sera brièvement décrit dans les sections suivantes.

Dans cet article, nous introduisons la contrainte graphémique à la confusion phonétique. Nous supposons que prendre en compte la graphie dans la confusion phonétique peut améliorer les performances de la reconnaissance.

2. NOUVELLE APPROCHE

Nous rappelons brièvement notre méthode déjà décrite dans [1]. La confusion phonétique considérée met en jeu des phonèmes de la langue parlée et maternelle. En effet, les locuteurs non natifs tendent à réaliser les phonèmes comme dans leur langue maternelle. Nous avons donc introduit un nouveau concept de confusion qui associe une suite de phonèmes de la langue maternelle aux phonèmes prononcés.

Pour chaque phrase du corpus d'adaptation, nous effectuons :
- un alignement forcé du signal audio avec la suite de modèles acoustiques de la langue parlée correspondant à la transcription canonique de la phrase,
- une reconnaissance phonétique du signal audio avec des modèles acoustiques de la langue maternelle.

Nous comparons ces deux transcriptions afin d'extraire des règles de confusion phonétique. Ces règles sont ensuite utilisées pour modifier les modèles de Markov (HMM) des phonèmes du SRAP de la langue parlée. Les HMMs correspondant à la séquence de phonèmes natifs sont ajoutés comme chemins alternatifs dans le HMM du phonème de la langue parlée. Nous supposons que cette utilisation de la confusion minimise le surcoût de puissance de calcul pour le nouveau SRAP. En effet, la modification du lexique (ajout de toutes les prononciations possibles) peut induire un surcoût très important ([2]). De plus, la modification des modèles gaussiens, comme dans [5], peut nuire à la cohérence temporelle des modèles acoustiques.

2.1. Extraction des règles de confusion

Deux ensembles de modèles acoustiques, ceux de la LP et de la LM, sont utilisés pour extraire des règles de confusion phonétique. Pour chaque phrase prononcée par un locuteur non natif, nous effectuons :
- un alignement phonétique forcé avec les phonèmes de la LP (prononciation canonique)
- une reconnaissance phonétique avec des phonèmes de la LM. Une simple boucle de phonèmes est utilisée à cet effet.

La comparaison des deux transcriptions permet ensuite de déduire les associations entre les phonèmes de la LP et les séquences de phonèmes de la LM. Un phonème $[K]$ de la langue parlée est associé à la suite de phonèmes $(M_i)_{i\in I}$ si chaque phonème M_i est inclus (pour plus de la moitié) pendant la durée de la prononciation du phonème $[K]$.

La prochaine étape consiste à extraire les règles de confusion à partir de ces associations. Seules les règles les plus pertinentes sont retenues. L'estimation au maximum de vraisemblance de la probabilité des règles ($P(K ==> (M_i)_{i\in I})$) est calculée pour chaque phonème $[K]$ (de la LP). Seules les règles dont la probabilité est supérieure à un seuil arbitraire α sont retenues.

Voici un exemple de règles données par notre système pour la diphtongue anglaise [aɪ] (où la langue maternelle est l'italien) :

[aɪ]==>[a] [i] $\quad P([aɪ]==>[a]\ [i]) = 0.6$
[aɪ]==>[a] [e] $\quad P([aɪ]==>[a]\ [e]) = 0.4$

Les mêmes règles ont été extraites dans le cas où la langue maternelle est l'espagnol et le grec.

2.2. Modification des modèles acoustiques

Les modèles HMMs du SRAP de la LP sont modifiés à l'aide des règles de confusion extraites à l'étape précédente. Pour chaque phone $[K]$ de la langue parlée, un chemin alternatif est ajouté au modèle HMM de $[K]$ (SRAP de la LP). Pour chaque règle $r \in R_K$ (règles sélectionnées pour le phonème $[K]$), un chemin correspondant à la partie droite de la règle est rajouté au modèle HMM de $[K]$. Ce nouveau chemin est la concaténation des modèles HMM (SRAP de la LM) correspondant aux phonèmes de la partie droite de r.

Nous utilisons une pondération entre le modèle acoutique de la LP et ceux de la LM. Ici β correspond au poids du modèle original. Dans le modèle HMM modifié, la transition liant l'état non émetteur de départ au modèle original a une probabilité β. De même, la probabilité liant cet état non émetteur à chaque chemin ajouté (pour une règle $r \in R_K$) est calculée comme suit :

$$P'(r) = (1-\beta)\frac{P(r)}{\Sigma_{x\in R_K}(P(x))} \qquad (1)$$

Étant données les règles de confusion décrite dans le paragraphe 2.1 pour la diphtongue anglaise [aɪ], on obtient le modèle HMM représenté dans la figure 1.

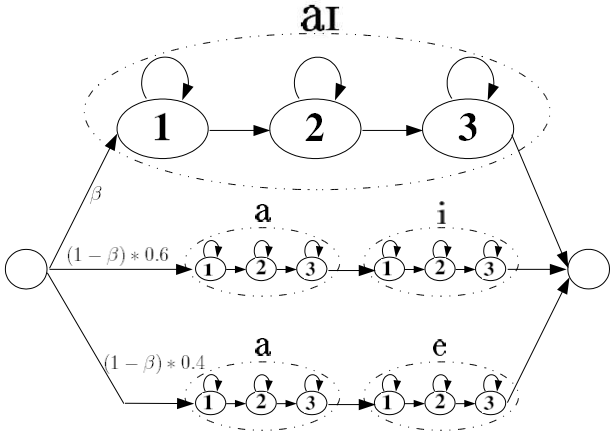

FIG. 1: *HMM modifié pour le phonème anglais [aɪ].*

3. CONTRAINTES GRAPHEMIQUES

Comme nous l'avons expliqué précédemment, les erreurs réalisées par les locuteurs non natifs sont fortement liées à la graphie des mots. Nous supposons donc que la prise en compte des contraintes graphémiques dans la confusion phonétique est susceptible d'apporter une amélioration de performances des SRAP.

La finalité de cette étape est d'associer automatiquement

les phonèmes à leur graphie sous-jacente pour chaque mots du dictionnaire.

3.1. Alignement automatique phonème-graphème

Étant données la graphie et la prononciation canonique d'un mot, le but est d'associer les phonèmes aux graphèmes correspondants. Cependant, l'information recherchée est l'alignement phonème-graphème et non pas une traduction graphème-phonème.

Un système HMM discret a été utilisé pour effectuer cette tâche. Ce système discret a été entraîné sur le dictionnaire phonétique du CMU. Les observations discrètes représentent les graphèmes et les modèles HMMs (mono-état) représentent les phonèmes. Le système HMM discret peut être utilisé pour effectuer l'alignement phonème-graphème sur le dictionnaire du SRAP cible, via l'algorithme de Viterbi.

Mise en place du système HMM discret

L'alignement phonème-graphème est extrait d'une manière automatique à partir d'un grand dictionnaire phonétique. Dans notre système HMM discret, les observations discrètes représentent les graphèmes, les modèles HMM représentent les phonèmes, et le dictionnaire phonétique représente le corpus d'apprentissage.

Les modèles HMM discrets initiaux ont une probabilité d'émission uniforme pour tous les symboles discrets (correspondant chacun à un graphème). Enfin, pour chaque mot du dictionnaire d'apprentissage, un fichier de données discrètes correspondant à la suite de symboles discrets (graphèmes du mot) est créé.

Donc, un alignement *phonétique* est effectué sur le dictionnaire d'apprentissage afin de déterminer les associations entre les phonèmes et les graphèmes. Seules les associations les plus représentées sont retenues. Ceci évite les erreurs dues à un alignement erroné ou encore à une inconsistance dans le dictionnaire d'apprentissage lui même. Une association a_K relative à un phonème $[K]$ n'est retenue que si elle satisfait l'inéquation 2.

$$N(a_K) \geq \gamma \, \Sigma_{a'_K \in A_K} N(a'_K) \qquad (2)$$

où A_K est l'ensemble des associations phonème-graphème pour le phonème $[K]$, $N(a_K)$ le nombre d'apparition de a_K, et γ une constante arbitraire.

Utilisation des contraintes graphémiques

Nous avons adopté une méthode simple pour appliquer les contraintes graphèmiques au SRAP cible. Nous proposons de remplacer les phonèmes dans la prononciation des mots du lexique par le couple (phonème,graphème) pour chacun des mots du dictionnaire (SRAP cible). La prononciation d'un mot n'est plus une suite de phonème, mais une suite de couple (phonème,graphèmes). L'exemple suivant illustre ceci pour le mot anglais "*speech*" :

ancienne prononciation	[s] [p] [i :] [tʃ]
nouvelle prononciation	[s]-S [p]-P [i :]-EE [tʃ]-CH

A cet effet, un alignement forcé sur le dictionnaire du SRAP cible est effectué en utilisant le système HMM discret entraîné (voir section précédente). Ceci fourni les associations phonème-graphème pour chaque mot du dictionnaire cible. Si une association phonème-graphème n'apparaît pas dans la liste des associations retenues (voir section précédente), le phonème reste sans contrainte graphémique dans la prononciation du mot.

La dernière modification consiste à ajouter les modèles HMM correspondant aux nouveaux phonèmes considérés. Pour chaque phonème $[K]$ avec une contrainte graphèmique X, $[K]$ est cloné en un nouveau phonème $[K]$-X.

3.2. Problèmes d'alignement

Dans un mot, un graphème peut être lié à plusieurs phonèmes. C'est le cas du mot anglais "*used*", prononcé [i] [u] [z] [d]. L'application directe de la méthode décrite plus haut donnerait l'alignement unique suivant : [i]-U, [u]-S, [z]-E et [d]-D, qui est bien évidemment faux. En effet, les observations ne peuvent pas être partagées entre les états dans un système HMM. De même, il n'est pas possible de considérer les phonèmes comme étant les observations puisqu'un phonème peut être liés à plusieurs graphèmes. Pour résoudre ce problème, nous avons opté pour la duplication des observations. Par exemple, pour le mot "*used*", la suite de symboles (U, U, S, S, E, E, D, D) sera considérée. Nous obtenons ainsi l'alignement ([i]-U, [u]-U, [z]-SS, [d]-EEDD). Un traitement postérieur donnera l'alignement correct : ([i]-U, [u]-U, [z]-S, [d]-ED).

4. EXPÉRIMENTATIONS

Notre travail a été effectué dans le cadre du projet Européen *HIWIRE* dont le but est d'améliorer les performances des SRAP dans les environnements mobiles et bruités. Le projet *HIWIRE* vise le développement d'un système automatique pour le contrôle vocal d'avions par les pilotes.

4.1. Conditions expérimentales

Notre corpus se compose de 4 parties : française, italienne, grèque et espagnole comportant respectivement 31, 20, 20 et 10 locuteurs. Chacun de ces locuteurs a prononcé 100 phrases en anglais. Nous avons considéré l'union de ces 4 parties comme un corpus international sur lequel nous avons effectué des tests de reconnaissance. La parole a été enregistrée à une fréquence de 16KHz. La paramétrisation MFCC utilisée consiste en 13 paramètres statiques avec leur dérivées premières et secondes. Les 46 monophones anglais ont été entraînés sur la base de donnés *TIMIT*. Les 40 monophones français ont été entraînés sur le corpus *ESTER*, composées de 90 heures de bulletins d'informations radiophoniques. Les modèles acoustiques espagnols, grecs et italiens ont été entraînés sur des corpus de parole espagnol, grec et italien (respectivement). Les modèles gaussiens des HMM possèdent 128 gaussiennes et des matrices de covariances diagonales. Le vocabulaire comporte 134 mots et la grammaire est un langage de commande. Une grammaire libre a également été utilisée (boucle de mots). Nous avons utilisé les 50 premières phrases de chaque locuteur pour l'extraction de la confusion, et les 50 dernières pour les tests.

4.2. Tests et résultats

Nous avons testé le système de référence (sans aucune adaptation) et le système de "confusion phonétique" avec

TAB. 2: *Résultats des tests effectués sur corpus français, italien, espagnol et grec (en %).*

	français		italien		espagnol		grec	
Système	WER	SER	WER	SER	WER	SER	WER	SER
grammaire contrainte :								
- système de référence	6.0	12.8	10.5	19.6	7.0	14.9	5.8	13.2
- "confusion phonétique"	**4.4**	**10.2**	**6.9**	**14.1**	**5.1**	**11.8**	**2.9**	**7.5**
- "confusion phonétique" + contraintes graphèmiques	4.9	11.3	8.2	15.9	6.2	13.6	6.3	15.8
grammaire libre :								
- système de référence	37.7	47.9	45.5	52.0	39.9	53.5	36.7	49.2
- "confusion phonétique"	27.3	42.1	31.3	46.2	**31.3**	**44.5**	**20.2**	**34.9**
- "confusion phonétique" + contraintes graphèmiques	**26.2**	**41.9**	**30.5**	**45.5**	**31.3**	46.5	57.0	76.2

la grammaire contrainte et la grammaire libre. Nous avons effectué des tests séparés sur les corpus français, italien, espagnol, grec et international. Nous avons extrait des règles de confusion entre les phonèmes anglais et les phonèmes de la langue native respective de chaque partie de notre corpus. Toutefois, pour les tests avec le corpus international, nous avons extrait des confusions phonétiques uniquement avec des phonèmes de la langue parlée : confusion entre prononciation canonique (anglais) et phonèmes réellement prononcés (anglais).

Comme le montrent les tableaux 2 et 3, la confusion phonétique apporte une amélioration des taux de reconnaissance sur tous les corpus. Pour les tests utilisant la confusion phonétique et la grammaire contrainte, cette amélioration varie de 26.6% à 50.0% (relatif) en WER (taux d'erreurs en mots) et de 20.3% à 43.2% (relatif) en SER (taux d'erreurs en phrases). Les tests sur la confusion phonétique avec la grammaire libre affichent des améliorations variant de 21.6% à 45.0% en WER et de 11.2% à 29.1% en SER. En revanche, l'ajout des contraintes graphémiques à la confusion phonétique n'a pas eu de répercussions positives sur les taux de reconnaissances (par rapport à la confusion phonétique seule) en ce qui concerne les tests effectués avec la grammaire contrainte. Néanmoins, nous observons une légère amélioration pour les tests impliquant la grammaire libre (par rapport à la confusion phonétique seule) tant en WER qu'en SER. Nous pensons que cette dégradation est dûe à ce que la grammaire contrainte guide très bien la reconnaissance et donc l'ajout des contraintes graphèmiques n'amèliore pas les taux. En effet, il s'agit d'un langage de commande strict composé de seulement 134 mots.

TAB. 3: *Résultats des tests effectués sur le international (en %).*

Système	WER	SER
grammaire contrainte :		
- système de référence	7.1	14.5
- "confusion phonétique"	**5.7**	**12.1**
- "confusion phonétique" + contraintes graphèmiques	5.8	12.4
grammaire libre :		
- système de référence	38.5	49.9
- "confusion phonétique"	31.2	44.8
- "confusion phonétique" + contraintes graphèmiques	**30.2**	**43.7**

5. CONCLUSION

Nous avons présenté une nouvelle approche pour l'amélioration de la reconnaissance automatique de la parole prononcée par des locuteurs non natifs. Elle est basée sur l'utilisation de la confusion phonétique et des contraintes graphèmiques. Dans notre approche, les phonèmes de la langue parlée sont associés à une suite de phonèmes de la langue maternelle du locuteur. Nous avons présenté une nouvelle utilisation de la confusion phonétique qui consiste à ajouter de nouveaux chemins alternatifs dans les modèles HMM des phonèmes de la langue parlée. Nous avons également proposé l'adjonction des contraintes graphèmiques à la confusion phonètique. En effet, les erreurs de prononciations des phonèmes sont fortement corrélées à la graphie des mots. La confusion phonétique a apporté des améliorations significatives dans les taux de reconnaissance sur notre corpus. Toutefois, l'ajout des contraintes graphémiques n'a été favorable que lors de l'utilisation d'une grammaire non contrainte.

6. REMERCIEMENTS

Ce travail a été partiellement financé par le projet Européen *HIWIRE* (Human Input that Works In Real Environments), contrat numéro 507943, "sixth framework programme, information society technologies".

RÉFÉRENCES

[1] G. Bouselmi, D. Fohr, I. Illina, and J.P. Haton. Fully automated non-native speech recognition using confusion-based acoustic model integration. In *In Proc. Eurospeech/Interspeech*, 2005.

[2] S. Goronzy, R. Kompe, and S. Rapp. Generating non-native pronunciation variants for lexicon adaptation. In *Eurospeech*, 2001.

[3] K. Livescu and J. Glass. Lexical modeling of non-native speech for automatic speech recognition. In *ICASSP*, 2000.

[4] J. Morgan. Making a speech recognizer tolerate non-native speech through gaussian mixture merging. In *InSTIL/ICALL*, 2004.

[5] P. Nguyen, P. Gelin, J.-C. Junqua, and J.-T. Chien. N-best based supervised and unsupervised adaptation for native and non-native speakers in cars. In *ICASSP*, 1999.

[6] Stefan Schaden. Generating non-native pronunciation lexicons by phonological rule. In *ICSLP*, 2004.