Adaptation de la prononciation pour la synthèse de la parole spontanée en utilisant des informations linguistiques

Raheel Qader, Gwénolé Lecorvé, Damien Lolive, Pascale Sébillot

Adaptation de la prononciation pour la synthèse de la parole spontanée en utilisant des informations linguistiques

Raheel Qader[1] Gwénolé Lecorvé[1] Damien Lolive[1] Pascale Sébillot[2]
(1) IRISA/Université de Rennes 1, 6, rue de Kerampont, 22300 Lannion, France
(2) IRISA/INSA Rennes, 263 Avenue Général Leclerc, 35000 Rennes, France
{raheel.qader,gwenole.lecorve,damien.lolive,pascale.sebillot}@irisa.fr

RÉSUMÉ

Cet article présente une nouvelle méthode d'adaptation de la prononciation dont le but est de reproduire le style spontané. Il s'agit d'une tâche-clé en synthèse de la parole car elle permet d'apporter de l'expressivité aux signaux produits, ouvrant ainsi la voie à de nouvelles applications. La force de la méthode proposée est de ne s'appuyer que sur des informations linguistiques et de considérer un cadre probabiliste pour ce faire, précisément les champs aléatoires conditionnels. Dans cet article, nous étudions tout d'abord la pertinence d'un ensemble d'informations pour l'adaptation, puis nous combinons les informations les plus pertinentes lors d'expériences finales. Les évaluations de la méthode sur un corpus de parole conversationnelle en anglais montrent que les prononciations adaptées reflètent significativement mieux un style spontané que les prononciations canoniques.

ABSTRACT

Pronunciation adaptation for spontaneous speech synthesis using linguistic information.

This paper presents a new pronunciation adaptation method which adapts canonical pronunciations to a spontaneous style. This is a key task in text-to-speech as those pronunciation variants bring expressiveness to synthetic speech, thus enabling new potential applications. The strength of the method is to solely rely on linguistic features and to consider a probabilistic machine learning framework, namely conditional random fields, to produce the adapted pronunciations. Features are selected in a first series of experiments, then combined in the backend experiments. Results on the Buckeye conversational English speech corpus show that adapted pronunciations significantly better reflect spontaneous speech than canonical ones.

MOTS-CLÉS : Adaptation de la prononciation, parole spontanée, synthèse de la parole.

KEYWORDS: Pronunciation adaptation, spontaneous speech, speech synthesis.

1 Introduction

Les variantes de prononciation de mots ou énoncés ne sont pas prises en compte par les lexiques et modèles de prononciation utilisés dans les systèmes actuels de synthèse de la parole. Cela limite alors leur capacité à produire des signaux expressifs, notamment pour retranscrire un style spontané. Pour résoudre ce problème, nous proposons une nouvelle méthode d'adaptation de la prononciation dont le but est d'imiter le style spontané de locuteurs individuels et de pouvoir, à terme, être intégrée dans un moteur de synthèse de la parole. La langue étudiée est l'anglais.

Sous l'angle de l'apprentissage automatique, la méthode proposée consiste à prédire pour chaque phonème d'une prononciation canonique si celui-ci doit être supprimé, remplacé, gardé tel quel ou

Phonèmes canoniques	/k ɑ n s ʌ **n** t ɹ eɪ **t ʌ** d · ɪ **n** · oʊ h **aɪ oʊ** /
Phonèmes réalisés	/k ɑ n s n̩ _ t ɹ eɪ _ **ɪ** d · ɪ ɾ̃ · oʊ h **ɑ ʌ** /

TABLE 1 – Prononciations réalisée et canonique pour la séquence de mots « *concentrated in Ohio* ».

complété par des phonèmes à insérer. Pour cela, cette méthode repose sur des Champs Aléatoires Conditionnels (CAC), dont l'utilisation est très répandue pour la phonétisation de mots ou d'énoncés (Wang & King, 2011; Illina *et al.*, 2011; Lecorvé & Lolive, 2015). Les CAC sont très utiles car ils permettent d'intégrer et combiner simplement un très large panel d'informations. Précisément, la force de notre méthode est de ne s'appuyer, en complément des phonèmes canoniques, que sur des informations linguistiques car aucune information acoustique n'est disponible au moment de la phonétisation d'un énoncé en synthèse de la parole.

Les travaux connexes en production de variantes de prononciation peuvent être résumés d'après le type de l'approche retenue et la nature des informations utilisées. Tout d'abord, diverses approches par apprentissage automatique ont déjà été utilisées : des arbres de décision (Fosler-Lussier *et al.*, 1999; Vazirnezhad *et al.*, 2009), des forêts aléatoires (Dilts, 2013), des réseaux de neurones (Chen & Hasegawa-Johnson, 2004; Karanasou *et al.*, 2013), des modèles de Markov cachés (Prahallad *et al.*, 2006) et des CAC (Karanasou *et al.*, 2013). Pour aller plus loin, d'autres travaux ont également proposé de combiner différentes techniques (Vazirnezhad *et al.*, 2009; Kolluru *et al.*, 2014). Il est malheureusement difficile de comparer ces travaux car ceux-ci partagent rarement les mêmes données ou la même tâche exacte. Quant aux informations utilisées, des caractéristiques acoustiques peuvent être extraites à partir de signaux de parole d'un style visé et prises en compte pour l'adaptation de prononciations (fréquence fondamentale, énergie, durée, débit de parole…) (Bates & Ostendorf, 2002; Bell *et al.*, 2009, 2003), tandis que des informations linguistiques peuvent être dérivées de textes (distinction entre mots-outils et mots pleins, probabilité des mots, informations syllabiques, accentuation lexicale dans certaines langues…) (Vazirnezhad *et al.*, 2009; Bell *et al.*, 2009, 2003). Récemment, Dilts (2013) a présenté une étude poussée sur la combinaison de ces deux types d'informations. Ce travail est proche du nôtre mais diffère dans le sens où la technique d'apprentissage automatique est différente et l'objectif était uniquement de réduire les prononciations. En complément, notons également que (Chen & Hasegawa-Johnson, 2004) a montré que les informations d'un phonème canonique doivent être enrichies par celles de leur voisinage pour aboutir à de meilleures adaptations. Enfin, il est important de noter que la plupart des travaux du domaine visent la reconnaissance automatique de la parole alors que les approches pour la synthèse sont encore rares et qu'aucune ne fait un usage aussi intensif que le nôtre des informations linguistiques.

Dans cet article, la section 2 décrit le corpus de parole Buckeye utilisé dans nos travaux ; la section 3 présente la méthode et le protocole expérimental ; la section 4 étudie en préambule des caractéristiques isolées avant que celles-ci ne soient combinées dans les expériences finales de la section 5.

2 Le corpus de parole Buckeye

Nous avons utilisé le corpus de parole conversationnelle Buckeye (Pitt *et al.*, 2005). Ce corpus, en anglais, consiste en 40 entretiens non préparés avec des locuteurs de l'Ohio, aux États-Unis, chaque entretien durant 1 heure. 20 entretiens ont été sélectionnés au hasard, les autres ayant été laissés de côté pour d'éventuels futurs travaux. Les signaux de parole sont fournis avec des transcriptions vérifiées manuellement : une transcription orthographique et deux transcriptions phonétiques, l'une correspondant aux phonèmes canoniques qui auraient dû être prononcés si le style avait été neutre,

Phonèmes

phonème canonique (40) • **position du phonème dans la syllable (20)** • **position inversée du phonème dans la syllabe (22)** • **phonème en début, milieu ou fin de mot (40)**

Syllabes

accentuation lexicale de la syllabe (24) • **partie de la syllabe (24)** • type de syllabe (18) • **position de la syllabe dans le mot (20)**

Mots

mot (40) • graphème (16) • **est-ce un mot vide (d'après une liste) ? (24)** • **fréquence du mot en anglais (22)** • fréquence du mot dans l'entretien (18) • fréquence de la racine en anglais (16) • fréquence de la racine dans l'entretien (19) • position du mot dans l'énoncé (2) • position inverse du mot dans l'énoncé (0) • numéro d'occurrence du mot dans l'entretien (0) • classe grammaticale (17) • longueur en graphèmes (16) • longueur en syllabes (17)

Énoncés

position de l'énoncé dans l'entretien (3) • position inverse de l'énoncé dans l'entretien (4)

TABLE 2 – Liste des caractéristiques (hormis les phonèmes réalisés). En gras, celles qui ont été conservées à l'issue de la phase de sélection. Entre parenthèses, le nombre de votes reçus (cf. section 4).

l'autre aux phonèmes effectivement réalisés par le locuteur dans un cadre de parole spontanée. Chaque locuteur représente environ 7 400 mots et 22 800 phonèmes. Les phonèmes canoniques et réalisés ont été alignés automatiquement. Il en découle que 30 % des phonèmes et 57 % des mots sont prononcés différemment de ce qui était attendu. L'exemple de la table 1 permet de constater à quel point les prononciations réalisées diffèrent généralement des prononciations canoniques.

Ces annotations ont été complétées par de nombreuses autres au moyen d'outils automatiques, conduisant à un total de 23 caractéristiques portant sur les phonèmes, syllabes, mots et énoncés. Leur détail est donné par la table 2. Afin d'être compatible avec l'emploi de CAC, les fréquences ont été catégorisées à masses de probabilité équivalentes en « fréquent », « moyen » et « rare ». Nous présentons maintenant la méthode en elle-même.

3 Présentation de la méthode et du protocole expérimental

Nous formalisons l'adaptation de la prononciation comme la prédiction d'une séquence de phonèmes réalisés à partir d'une séquence de phonèmes canoniques. Expérimentalement, la qualité d'une adaptation se mesure alors à son taux d'erreurs entre les phonèmes prédits, dits *adaptés*, et phonèmes effectivement réalisés dans le corpus.

Dans notre méthode, nous avons choisi d'utiliser des CAC, type de modèles particulièrement adéquat pour l'apprentissage sur des données séquentielles et symboliques. Pour construire au mieux ces modèles, nous avons cherché à ajouter de nouvelles informations en entrée en plus des seuls phonèmes canoniques. Les différentes pistes qui ont été explorées pour cela sont illustrées par la figure 1. Nous les présentons ci-dessous et introduisons les questions qui s'y rattachent.

1. Principalement, chaque phonème p_i à prédire dépend de n caractéristiques $\{c_i^1, \ldots, c_i^n\}$, par exemple le phonème canonique à adapter, sa position dans la syllabe ou la fréquence du mot qui le contient. La question est alors de savoir quelles caractéristiques parmi toutes celles considérées sont pertinentes et quelles autres dégradent l'adaptation.

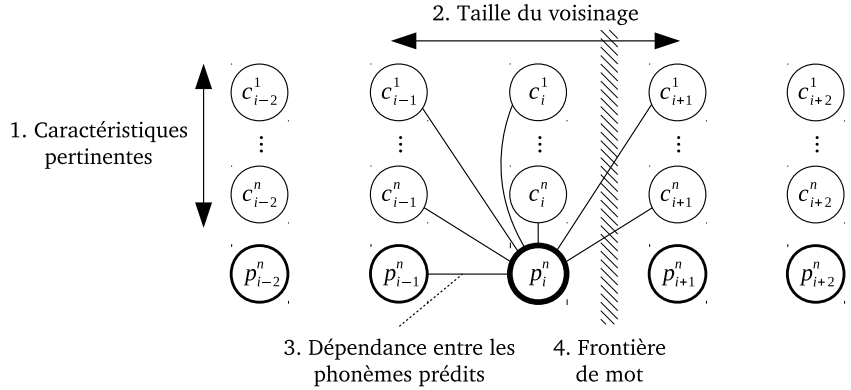

FIGURE 1 – Vue d'ensemble des dépendances et paramètres à traiter pour l'apprentissage des CAC. Les nœuds et arrêtes représentent respectivement différentes informations et leurs liens de dépendance.

2. Ensuite, le panel des informations peut être étendu au voisinage de p_i, par exemple en considérant également le phonème canonique précédent et le suivant, ainsi que les caractéristiques qui leur sont associées. Cette notion de voisinage se définit en pratique par une fenêtre de taille W autour de p_i. Le choix de cette taille est un point que nous avons étudié.
3. De manière analogue, une dépendance entre p_i et la précédente prédiction p_{i-1} peut être considérée. Cela doit notamment permettre d'éviter des enchaînements de phonèmes possiblement non articulables. Nous avons cherché à savoir si cette dépendance est utile.
4. Enfin, il est possible d'interdire ou autoriser la propagation des dépendances par-delà les frontières de mots. Nous avons également étudié ces deux options.

Les CAC permettent naturellement de modéliser tous ces différents types de dépendances (Lafferty *et al.*, 2001) et nous avons pour cela utilisé l'outil Wapiti (Lavergne *et al.*, 2010). Dans le détail, nous avons tout d'abord étudié chaque piste individuellement, puis les avons combinées lors des expériences finales. Avant d'aborder ces travaux, la prochaine section décrit le protocole expérimental.

Les CAC sont entraînés et évalués indépendamment pour chaque locuteur avec l'objectif de déterminer une unique configuration d'apprentissage pour pouvoir généraliser la méthode à tout locuteur. Chaque entretien a été partitionné en ensembles d'entraînement (60 % des énoncés), de développement (20 %) et de test (20 %). L'étude préliminaire des différents paramètres introduits à la section 3 s'est effectuée sur l'ensemble de développement et les expériences finales sur l'ensemble de test.

Pour chaque locuteur, un taux d'erreurs sur les phonèmes (PER, pour *Phoneme Error Rate*) est calculé comme une distance d'édition entre les séquences réalisées et celles canoniques (c.-à-d. non adaptées) ou prédites par l'adaptation. Les résultats reportés dans cet article sont les PER moyens sur la totalité des locuteurs. En complément de ces mesures objectives, des tests d'écoute ont été menés pour mesurer la spontanéité et l'intelligibilité des différentes prononciations étudiées. Ces tests, conduits sur les configurations les plus intéressantes, sont détaillés en section 5.

4 Étude préliminaire des paramètres

Cette section détaille comment la méthode d'adaptation de la prononciation a été réglée sur l'ensemble de développement. L'accent est mis sur les deux aspects majeurs que sont le choix des caractéristiques

	Pas d'adaptation	Phon. canoniques	Car. sélectionnées	Toutes les car.
Unigrammes	30,4	30,4 (0,0)	**24,7 (-5,7)**	26,0 (-4,4)
Uni+bigrammes		25,7 (-4,7)	**24,1 (-6,3)**	26,1 (-4,3)

TABLE 3 – PER (%) sur l'ensemble de développement sans adaptation ou avec adaptation avec différents jeux de caractéristiques.

linguistiques pertinentes et celui de la taille du voisinage à considérer. Néanmoins, l'utilité d'inclure des dépendances entre phonèmes prédits et celle d'adapter à l'échelle de mots isolés ou d'énoncés continus sont également examinées. Le détail de ces études peut être trouvé dans (Qader *et al.*, 2015).

L'entraînement de CAC à partir de trop nombreuses caractéristiques peut conduire à du surapprentissage. Par ailleurs, le temps de calcul et la quantité de mémoire nécessaires à l'entraînement est exponentiel en fonction du nombre de caractéristiques. Ainsi, nous avons effectué une sélection des attributs linguistiques à considérer. Cette sélection s'est faite en recherchant le meilleur ensemble de caractéristiques, c.-à-d. celui qui conduit au plus petit PER, pour chaque locuteur. Nous avons pour cela mis en place un mécanisme de vote. Une caractéristique a reçu un vote par nombre de fois où elle appartenait au meilleur ensemble d'un locuteur. Pour rendre ce processus robuste, deux stratégies de recherche ont été testées pour chaque locuteur : l'une additive, l'autre soustractive. À l'issue des votes, il a arbitrairement été décidé de sélectionner les caractéristiques qui avait reçu au moins $50\,\%$ des votes, soit 20 votes ici. La table 2 reporte entre parenthèses le nombre de votes reçus par chaque caractéristique et en gras celles qui ont finalement été sélectionnées ainsi. Il apparaît que les informations relatives aux syllabes et aux fréquences des mots sont les plus importantes. Ces conclusions sont cohérentes avec de précédents travaux (Adda-Decker *et al.*, 2005; Vazirnezhad *et al.*, 2009; Bell *et al.*, 2009). La table 3 compare les PER obtenus (i) sans adaptation et (ii) avec adaptation sur la base d'aucune caractéristique autre que les phonèmes canoniques, (iii) des caractéristiques sélectionnées et (iv) de toutes les caractéristiques possibles. Ces configurations ont été testées lorsqu'un phonème est prédit soit indépendemment de la prédiction qui le précède (unigrammes), soit en en tenant compte (uni+bigrammes). Il ressort clairement des résultats que la sélection des caractéristiques est une nécessité. Par ailleurs, les dépendances entre prédictions semblent bénéfiques.

Ensuite, comme évoqué en section 3, il peut s'avérer judicieux de considérer les informations provenant du voisinage d'un phonème canonique en cours d'adaptation. Pour déterminer cela, nous avons défini le voisinage comme une fenêtre symétrique [1] de W phonèmes canoniques à gauche et à droite autour du phonème à adapter. Une fenêtre $W{=}0$ signifie ainsi qu'aucun voisinage n'est considéré et $W{=}\pm2$ que 5 phonèmes sont considérés au total (1 au centre, 2 à gauche et 2 à droite). La figure 2 présente les PER obtenus pour différentes valeurs de W. Ces résultats sont présentés soit dans le cas où les fenêtres ne peuvent pas traverser des frontières de mots (mots isolés), soit dans celui où elles le peuvent (énoncés). Les CAC ont été appris à partir des seuls phonèmes canoniques et dans la configuration unigramme (pas de dépendances entre phonèmes prédits). Il apparaît que la prise en compte du voisinage améliore significativement les résultats mais qu'un plateau est vite atteint lorsque W augmente. Ainsi, la valeur $W{=}\pm2$ est retenue pour les expériences finales. Par ailleurs, contrairement à l'intuition, il semblerait que les adaptations mot à mot produisent de meilleurs résultats que lorsque l'adaptation se fait à l'échelle d'un énoncé entier. L'explication de ce phénomène est que les frontières de mots portaient une information utile quant à la position d'un phonème canonique dans son mot. L'utilisation conjointe de fenêtre et d'énoncés supprime cette information. Nous étayerons cette conclusion grâce aux expériences finales.

1. Nous avons également testé des fenêtre dissymétriques mais celles-ci n'ont mené qu'à de moins bons résultats.

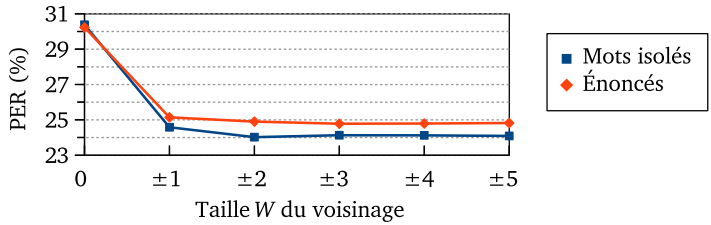

FIGURE 2 – PER en fonction de la taille de la fenêtre, pour les mots isolés et les phrases.

5 Combinaison des paramètres

Cette section présente les expériences finales conduites sur l'ensemble de test de chaque locuteur. Les modèles d'adaptation sont appris en combinant différentes configurations étudiées à la section 4, à savoir :

— à partir des seuls phonèmes canoniques ou en incluant aussi les caractéristiques sélectionnées ;
— sans ou avec prise en compte du voisinage ($W{=}0$ ou $W{=}{\pm}2$) ;
— en incluant ou non une dépendance entre phonèmes prédits (unigrammes ou uni+bigrammes) ;
— en considérant des successions de mots isolés ou, au contraire, des énoncés continus.

La table 4 présente les résultats pour toutes les combinaisons possibles. Les configurations déjà évaluées sur l'ensemble de développement mènent aux mêmes conclusions, à savoir que les prises en compte séparées des caractéristiques linguistiques ou des voisinages produisent des taux d'erreurs plus bas. En outre, ces nouveaux résultats montrent que le croisement de ces deux configurations permet d'obtenir d'encore meilleurs PER. Ces différentes variations sont, certes, faibles mais elles sont statistiquement significatives avec un niveau de confiance de $95\,\%$. Ensuite, les résultats montrent que l'adaptation d'énoncés continus produit de meilleurs résultats absolus que sur des suites de mots isolés. Cet écart, déjà observé sans adaptation, est en réalité dû au fait que des erreurs successives d'insertion et de suppression à la frontière de mots sont considérées comme une seule substitution pour les énoncés. Ces résultats sur les énoncés nous permettent néanmoins de noter que l'inclusion des caractéristiques linguistiques permet de corriger le décalage précédemment observé sur la figure 2 par la réintroduction d'informations sur la position d'un phonème dans son mot. Enfin, nous pouvons noter que la prise en compte des dépendances entre phonèmes prédits devient néfaste lorsque beaucoup de caractéristiques sont prises en compte conjointement. Nous expliquons ces résultats par un phénomène de surapprentissage lié à l'explosion du nombre de paramètres à estimer par le CAC. Pour illustrer ces résultats, la table 5 compare les prononciations réalisées, canoniques et obtenues par deux adaptations sur l'exemple déjà présenté à la table 1. Celui-ci montre que l'adaptation sur la base des caractéristiques linguistiques et des voisinages introduit des variantes de prononciations plus relâchées. Il montre néanmoins également que l'absence de dépendances entre les phonèmes prédits peut conduire à des séquences peu plausibles, comme ici l'enchaînement /n̥n/.

Pour vérifier cette explication et approfondir l'intérêt de tenir compte des dépendances entre phonèmes prédits, nous avons conduit une autre série d'expériences. Nous avons appris un modèle de langage (ML) sur les phonèmes réalisés de l'ensemble des données d'apprentissage [2], puis avons utilisé ce modèle pour réordonner *a posteriori* les meilleures hypothèses d'adaptation fournies par nos CAC. Précisément, chaque hypothèse h est associée à un score $s(h)$ calculé comme une

2. Un seul ML pour les 20 locuteurs a été appris plutôt qu'un par locuteur afin d'estimer des probabilités fiables.

			$W = 0$	$W = \pm 2$
Suites de mots isolés				
Phonèmes canoniques (c.-à-d. pas d'adaptation)			30,5	–
Phonèmes adaptés à partir des	phonèmes canoniques seuls	Unigrammes	30,4 (-0,1)	23,8 (-6,7)
		Uni+bigrammes	25,5 (-5,0)	24,0 (-6,5)
	+ caractéristiques linguistiques	Unigrammes	24,3 (-6,2)	**23,6 (-6,9)**
		Uni+bigrammes	24,1 (-6,4)	24,2 (-6,3)
Énoncés continus				
Phonèmes canoniques (c.-à-d. pas d'adaptation)			30,3	–
Phonèmes adaptés à partir des	phonèmes canoniques seuls	Unigrammes	30,2 (-0,1)	24,9 (-5,4)
		Uni+bigrammes	25,9 (-4,4)	24,2 (-6,1)
	+ caractéristiques linguistiques	Unigrammes	24,1 (-6,2)	**23,4 (-6,9)**
		Uni+bigrammes	23,9 (-6,4)	24,4 (-5,9)

TABLE 4 – PER (%) sur l'ensemble de test. Entre parenthèses, les variations absolues avec les PER des prononciations canoniques seuls (sans adaptation).

Phonèmes réalisés		/k ɑ n s n̩ _ t ɹ eɪ _ ɪ d · ɪ f̃ · oʊ h ɑ ʌ /
Phonèmes canoniques (c.-à-d. pas d'adaptation)		/k ɑ n s **ʌ n** t ɹ eɪ **t ʌ** d · ɪ **n** · oʊ h **aɪ oʊ** / (7 erreurs)
Phonèmes adaptés à partir des	phonèmes canoniques seuls	/k ɑ n s **ʌ n** t ɹ eɪ **t ʌ** d · ɪ **n** · oʊ h **aɪ oʊ** / (7 erreurs)
	+ carac. ling. + voisinage	/k ɑ n s n̩ **n _** ɹ eɪ **ɾ** ɪ d · ɪ **n** · oʊ h **aɪ oʊ** / (6 erreurs)

TABLE 5 – Différentes prononciations de la séquence de mots « *concentrated in Ohio* ». Les prononciations adaptées ont été produites sur la base de mots isolés sans prise en compte des dépendances entre phonèmes prédits. Les erreurs par rapport à la référence sont reportées en gras.

interpolation logarithmique des probabilités fournies par le CAC et par le ML, comme suit :

$$s(h) = \Pr_{\text{CAC}}(h) \times \Pr_{\text{ML}}(h)^{\alpha} \times \beta^{n} \,, \tag{1}$$

où α et β sont deux paramètres à optimiser et n est le nombre de phonèmes dans h. Le facteur β sert à contrebalancer le favoritisme naturel du ML envers les hypothèses les plus courtes. Le réordonnancement consiste alors à sélectionner l'hypothèse de score s le plus élevé. En pratique, le ML est un modèle 5-gramme avec un lissage de Witten-Bell et α et β ont été optimisés de sorte à minimiser le PER sur l'ensemble de développement. L'apprentissage du ML, l'optimisation des paramètres et le réordonnancement ont été effectués grâce à l'outil SRILM (Stolcke *et al.*, 2011). En reprenant les meilleurs résultats de la table 4 (sur la configuration « unigrammes »), l'introduction par le ML des dépendances entre phonèmes prédits permet d'obtenir des améliorations additionnelles significatives du PER sur l'ensemble de test, respectivement de $0,3$ et $0,2$ point pour les mots isolés et les énoncés continus. Ces résultats confortent en outre notre hypothèse sur l'effet négatif d'un trop grand nombre de paramètres lors de l'apprentissage des CAC.

La spontanéité et l'intelligibilité a été évaluée perceptuellement par un test d'écoute AB sur 10 locuteurs anglais natifs. Ce test juge la préférence des testeurs entre des paires de signaux de parole synthésisés sur la base des prononciations non adaptées, adaptées à partir des phonèmes canoniques (C) ou également des informations linguistiques (C+L), ou réalisées. Toutes ces configurations inclus le réordonnancement des hypothèses. Le test contient 40 étapes [3]. À chaque étape, le testeur évalue sa

3. Quelques échantillons peuvent être écoutés sur `http://www-expression.irisa.fr/demos`.

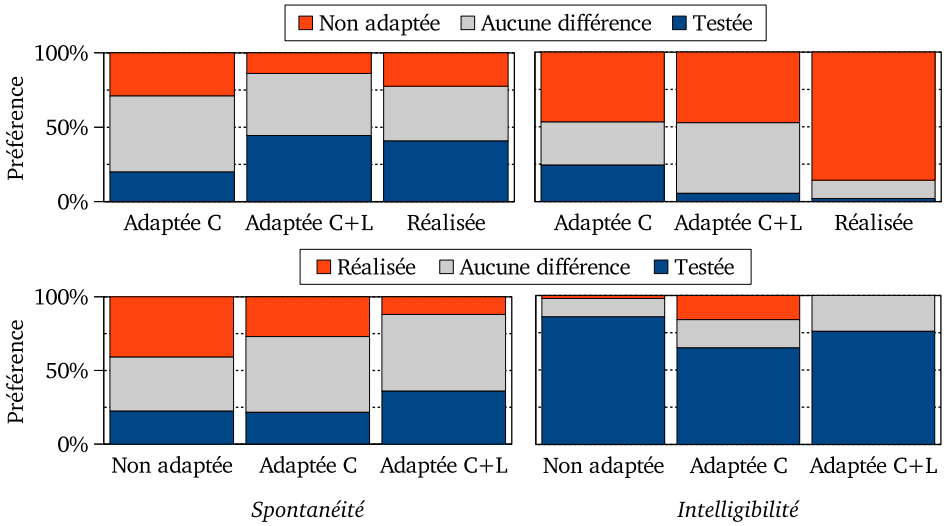

FIGURE 3 – Comparaisons entre les prononciations adaptées et les prononciations non adaptées (en haut) ou réalisées (en bas) en termes de spontanéité et d'intelligibilité.

préférence entre les 2 échantillons proposés en terme de spontanéité et d'intelligibilité. Le système est système HTS classique appris sur les données du challenge Blizzard 2012. Les résultats de ce test sont présentés par la figure 3. Il en ressort que les prononciations adaptées et réalisées sont effectivement jugées plus spontanées que les prononciations non adaptées. Tout particulièrement, il apparaît même que la prise en compte des informations linguistiques pour l'adaptation aboutit à des prononciations qui, une fois synthétisées, sont jugées plus spontanées que les prononciations réalisées. Enfin, les prononciations non adaptées sont les plus intelligibles mais les prononciations adaptées sont, là encore, bien meilleures sur ce point que les prononciations réalisées, notamment lorsque les informations linguistiques sont considérées.

6 Conclusion et perspectives

Dans cet article, nous avons proposé une méthode d'adaptation de la prononciation qui permet d'imiter un style spontané. Les prononciations adaptées sont destinées à améliorer les systèmes de synthèse de la parole. Appliquées à l'anglais, les expériences sur le corpus de parole Buckeye montrent d'ores et déjà que les prononciations adaptées reflètent significativement mieux les prononciations spontanées des locuteurs que les prononciations canoniques d'origine. Ces bons résultats sont en particulier atteints grâce à la prise en compte de caractéristiques linguistiques sélectionnées automatiquement, de voisinages autour de chaque phonème canonique et de dépendances entre phonèmes adaptés.

Parmi les perspectives, ce travail pourrait être complété par une étude sur la prise en compte de caractéristiques articulatoires, prosodiques et acoustiques. Notre méthode pourrait en outre être testée sur d'autres langues ou sur des prononciations fournies par des phonétiseurs automatiques. Ces perspectives sont actuellement en cours d'étude. Notons enfin que, à terme, notre travail pourrait également trouver des applications en reconnaissance automatique de la parole.

Références

ADDA-DECKER M., DE MAREÜIL P. B., ADDA G. & LAMEL L. (2005). Investigating syllabic structures and their variation in spontaneous French. *Speech Communication*, **46**(2).

BATES R. & OSTENDORF M. (2002). Modeling pronunciation variation in conversational speech using prosody. In *ISCA Tutorial and Research Workshop (ITRW) on Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology*.

BELL A., BRENIER J. M., GREGORY M., GIRAND C. & JURAFSKY D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, **60**(1).

BELL A., JURAFSKY D., FOSLER-LUSSIER E., GIRAND C., GREGORY M. & GILDEA D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America*, **113**(2).

CHEN K. & HASEGAWA-JOHNSON M. (2004). Modeling pronunciation variation using artificial neural networks for English spontaneous speech. In *Proc. of Interspeech*.

DILTS P. C. (2013). *Modelling phonetic reduction in a corpus of spoken English using random forests and mixed-effects regression*. PhD thesis, University of Alberta.

FOSLER-LUSSIER E. *et al.* (1999). Multi-level decision trees for static and dynamic pronunciation models. In *Proc. of Eurospeech*.

ILLINA I., FOHR D. & JOUVET D. (2011). Grapheme-to-phoneme conversion using conditional random fields. In *Proc. of Interspeech*.

KARANASOU P., YVON F., LAVERGNE T. & LAMEL L. (2013). Discriminative training of a phoneme confusion model for a dynamic lexicon in ASR. In *Proc. of Interspeech*.

KOLLURU B., WAN V., LATORRE J., YANAGISAWA K. & GALES M. J. F. (2014). Generating multiple-accent pronunciations for TTS using joint sequence model interpolation. In *Proc. of Interspeech*.

LAFFERTY J., MCCALLUM A. & PEREIRA F. C. (2001). Conditional random fields : probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML*.

LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale CRFs. In *Proc. of ACL*.

LECORVÉ G. & LOLIVE D. (2015). Adaptive statistical utterance phonetization for French. In *Proc. of ICASSP*.

PITT M. A., JOHNSON K., HUME E., KIESLING S. & RAYMOND W. (2005). The Buckeye corpus of conversational speech : labeling conventions and a test of transcriber reliability. *Speech Communication*, **45**(1).

PRAHALLAD K., BLACK A. W. & MOSUR R. (2006). Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis. In *Proc. of ICASSP*, volume 1.

QADER R., LECORVÉ G., LOLIVE D. & SÉBILLOT P. (2015). Probabilistic Speaker Pronunciation Adaptation for Spontaneous Speech Synthesis Using Linguistic Features. In *Proc. of SLSP*.

STOLCKE A., ZHENG J., WANG W. & ABRASH V. (2011). Srilm at sixteen : Update and outlook. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, p.5̃.

VAZIRNEZHAD B., ALMASGANJ F. & AHADI S. M. (2009). Hybrid statistical pronunciation models designed to be trained by a medium-size corpus. *Computer Speech & Language*, **23**(1).

WANG D. & KING S. (2011). Letter-to-sound pronunciation prediction using conditional random fields. *IEEE Signal Processing Letters*, **18**(2).