Phonétisation statistique adaptable d'énoncés pour le français

Gwénolé Lecorvé Damien Lolive
IRISA, Université de Rennes 1, Lannion, France
gwenole.lecorve@irisa.fr, damien.lolive@irisa.fr

RÉSUMÉ

Les méthodes classiques de phonétisation d'énoncés concatènent les prononciations hors-contexte des mots. Ce type d'approches est trop faible pour certaines langues, comme le français, où les transitions entre les mots impliquent des modifications de prononciation. De plus, cela rend difficile la modélisation de stratégies de prononciation globales, par exemple pour modéliser un locuteur ou un accent particulier. Pour palier ces problèmes, ce papier présente une approche originale pour la phonétisation du français afin de générer des variantes de prononciation dans le cas d'énoncés. Par l'emploi de champs aléatoires conditionnels et de transducteurs finis pondérés, cette approche propose un cadre statistique particulièrement souple et adaptable. Cette approche est évaluée sur un corpus de mots isolés et sur un corpus d'énoncés prononcés.

ABSTRACT

Adaptive statistical utterance phonetization for French *.

Traditional utterance phonetization methods concatenate pronunciations of uncontextualized constituent words. This approach is too weak for some languages, like French, where transitions between words imply pronunciation modifications. Moreover, it makes it difficult to consider global pronunciation strategies, for instance to model a specific speaker or a specific accent. To overcome these problems, this paper presents a new original phonetization approach for French to generate pronunciation variants of utterances. This approach offers a statistical and highly adaptive framework by relying on conditional random fields and weighted finite state transducers. The approach is evaluated on a corpus of isolated words and a corpus of spoken utterances.

MOTS-CLÉS : Phonétisation, variantes de prononciation, treillis de phonèmes, champs aléatoires conditionnels, transducteurs finis pondérés.

KEYWORDS: Utterance phonetization, pronunciation variant modelling, phoneme lattices, conditional random fields, weighted finite state transducers.

1 Introduction

La phonétisation a pour objectif de prédire une séquence de phonèmes à partir d'une séquence de graphèmes. Pour la plupart des langues, cette tâche se limite au cas de mots isolés, réduisant la phonétisation d'un énoncé aux prononciations concaténées de ses mots. Cette approche n'est cependant pas viable pour certaines langues, comme le français, où les transitions entre mots provoquent des modifications de leur prononciation, à moins d'inclure des informations, souvent minimales, sur le contexte phonologique. De plus, cette approche complique la modélisation de stratégies de prononciation globales, par ex. propres à un locuteur ou à un accent particulier. Cette tâche d'adaptation est

*. Cet article reprend un travail présenté par les mêmes auteurs à la conférence ICASSP 2015.

majeure, en particulier en synthèse de la parole (TTS) (Benesty *et al.*, 2008).

Pour palier ces problèmes, ce papier présente une nouvelle méthode pour la phonétisation du français. Cette méthode apporte trois contributions : (i) elle introduit la notion de modèle d'élision pour modéliser les variantes intra-mots ; (ii) elle intègre les contextes phonologiques pour modéliser les variantes inter-mots ; (iii) elle permet de générer des treillis probabilistes de phonèmes à partir d'énoncés, et non seulement de mots isolés. Pour cela, cette méthode repose sur des Champs Aléatoires Conditionnels (CAC) pour estimer les probabilités des phonèmes sur les mots isolés, puis sur des transducteurs finis pondérés (TFP) pour traiter les transitions entre mots. On obtient ainsi des treillis de phonèmes à partir desquels des phonétisations peuvent être dérivées.

Le potentiel de ce phonétiseur est très important. Les treillis de phonèmes générés offrent beaucoup de flexibilité puisque les transitions peuvent être repondérées en utilisant différents modèles de prononciation dédiés à une tâche donnée. Néanmoins, l'objectif de cet article est de présenter la méthode de phonétisation et de ses premiers résultats sans adaptation, et non d'étudier le caractère adaptable des treillis, ce dernier point étant conservé pour de futurs travaux. De manière plus générale, l'objectif de cet article n'est pas la recherche de résultats meilleurs que l'état de l'art mais plutôt de définir un cadre de travail générique et de démontrer son applicabilité pour le français. Ce cadre de travail peut être étendu aisément et complété avec de nouveaux modèles. De plus, il ne repose pas sur des règles expertes et peut donc facilement être porté à de nouvelles langues. Enfin, l'approche proposée peut également tolérer une certaine incertitude dans l'énoncé d'entrée, par exemple pour gérer plusieurs tokénisations.

Dans cet article, la section 2 présente le domaine, la section 3 et 4 introduisent notre méthode de phonétisation, respectivement pour des mots isolés, puis des énoncés. Les expériences y sont présentées sur le lexique de prononciation MHATLex et sur un corpus de parole.

2 État de l'art

La phonétisation est largement étudiée depuis des années, en particulier en reconnaissance automatique de la parole (RAP) et en TTS. La plupart des systèmes reposent principalement sur des lexiques de prononciation construits manuellement pour les mots communs avec une conversion graphème-phonème automatique pour les mots hors vocabulaire, c.-à-d. les mots qui ne sont pas dans le lexique. De nombreuses stratégies ont été proposées pour la conversion graphème-phonème dans la littérature : des méthodes à bases de règles (Béchet, 2001; Claveau, 2009), des approches statistiques (Bisani & Ney, 2008; Illina *et al.*, 2011), ainsi que d'autres approches variées (Bellegarda, 2005; Laurent *et al.*, 2009). Parmi celles-ci, les approches statistiques ont récemment montré des performances intéressantes tout en apportant la possibilité d'interpréter et d'adapter les scores des prononciations générées. Trois principales méthodes s'opposent. D'un côté, les méthodes à n-grammes joints reposent sur des séquences de paires graphème-phonème dont les probabilités sont habituellement obtenues avec des modèles de langage (Bisani & Ney, 2008; Novak *et al.*, 2012; Hahn *et al.*, 2012). Les CAC ont également prouvé leur performance pour traiter le problème de conversion graphème-phonème (Illina *et al.*, 2011; Wang & King, 2011; Hahn *et al.*, 2011; Lehnen *et al.*, 2012). Enfin, des méthodes fondées sur des réseaux de neurones ont également été proposées très récemment et semblent produire les meilleurs résultats (Rao *et al.*, 2015; Yao & Zweig, 2015). Dans cet article, nous nous appuyons sur les CAC car ils sont un moyen simple d'intégrer de multiples connaissances. Le portage de notre méthode à des réseaux de neurones est néanmoins parfaitement envisageable. Dans l'ensemble, nos travaux se rapprochent de (Illina *et al.*, 2011) pour la phonétisation de mots isolés, bien que des différences dans les protocoles expérimentaux empêchent des comparaions directes des résultats.

La prononciation des énoncés, c.-à-d. des séquences de mots, a été étudiée de manière plus partielle. En RAP, l'introduction de TFP comme moyen de décoder les signaux de parole a apporté une nouvelle représentation des alternatives de prononciation des mots (Mohri *et al.*, 2000). En particulier, (Hazen *et al.*, 2005) propose de représenter les énoncés, les prononciations et leurs variations possibles comme des TFP qui peuvent être composés et parcourus pour extraire des variantes de prononciation. Pour la conversion graphème-phonème de mots isolés, des TFP et des treillis de phonèmes (Bodenstab & Fanty, 2007; Polyákova & Bonafonte, 2011) ou encore des CAC (Lehnen *et al.*, 2011) ont également été utilisés pour représenter les alternatives. La philosophie de cet article est très proche en combinant CAC et TFP. Cependant, le travail présenté ici diffère de (Hazen *et al.*, 2005) puisque les mots hors-vocabulaire et les élisions sont traités ici. De plus, (Hazen *et al.*, 2005) se concentre sur l'anglais tandis que notre travail est réalisé sur le français qui est phonologiquement différent. La phonétisation des mots isolés est présentée dans la section 3 avant de passer au niveau énoncé en section 4.

3 Phonétisation des mots isolés

La phonétisation des mots isolés consiste à prédire une séquence de phonèmes à partir d'une séquence de graphèmes. Cette tâche peut être vue comme un problème d'étiquetage automatique. Dans ce travail, cet étiquetage est effectué par deux CAC employés consécutivement, l'un pour prédire une séquence de phonèmes, l'autre pour prédire d'éventuelles élisions sur ceux-ci. Cette section présente l'apprentissage de ces deux modèles, puis les expériences sur la phonétisation de mots isolés.

3.1 Modèle de phonétisation

Le problème de conversion graphème-phonème est traité par l'apprentissage d'un CAC, dit ici *modèle de phonétisation*, sur un corpus aligné de graphèmes et de phonèmes issus d'un lexique de prononciations. Pour de meilleures performances, comme souvent dans la littérature (Jiampojamarn *et al.*, 2007), ces alignements sont effectués entre blocs de graphèmes et phonèmes de taille maximale fixée. Suite à une étude préliminaire, ces blocs sont d'une taille maximale de 2 dans notre travail. Pour rendre possible la tâche d'étiquetage, ces blocs sont ensuite décomposés de telle sorte que tout phonème soit associé à un et un seul graphème et que chaque graphème soit associé à 0, 1 ou 2 phonèmes. Par exemple, les graphèmes « o n » peuvent être alignés au seul phonème /õ/ et « x » au bloc /ks/. Les graphèmes « o », « n » et « x » seraient alors respectivement associés à /õ/, /_/ (c.-à-d. aucun phonème) et /ks/. La littérature montre également qu'il est bon de compléter un graphème par ses voisins (Illina *et al.*, 2011; Wang & King, 2011). Ce voisinage est défini par une fenêtre de $\pm N$ graphèmes dont nous étudierons l'impact à la section 3.3. Enfin, comme le français contient de nombreux homographes aux prononciations différentes [1], la classe grammaticale est une autre information utile. Suivant le choix fait dans (Illina *et al.*, 2011), nous utilisons cette information en la simplifiant à la seule distinction verbe/non verbe. Sur le fond, d'autres caractéristiques comme l'étymologie du mot ou un dialecte à considérer pourraient être utilisées mais ce n'est pas l'objectif du présent travail. Formellement, le CAC que nous entraînons prédit donc chaque phonème p à partir d'un n-gramme de graphèmes $\mathbf{g}$ (le graphème associé et ses voisins) et d'autres caractéristiques $\mathbf{o}$ dérivées du mot à phonétiser. Ce CAC est capable de produire la ou les meilleures hypothèses de phonétisation du mot et de fournir la probabilité *a posteriori* $\phi(p|\mathbf{g}, \mathbf{o})$ de chaque phonème.

1. Par exemple, la graphie *président* se prononce /pʁezidɑ̃/ s'il s'agit du nom ou /pʁezid/ s'il s'agit du verbe *présider*.

3.2 Modèle d'élision

Comme dans d'autres langues, certains phonèmes du français peuvent être élidés. Ces élisions dépendent d'informations variées, comme le contexte phonologique, le type de parole, les règles ou exceptions liées à la grammaires, etc. Le phénomène le plus courant pour illustrer cette variabilité est le cas du schwa (/ə/) qui peut être élidé la plupart du temps. Par exemple, le mot *semaine* peut être prononcé /səmɛn/ ou /smɛn/. De plus, le dernier graphème *e* peut également être prononcé /ə/ lorsqu'il est suivi par une consonne. Ainsi, l'énoncé "*la semaine finit*" peut être prononcé /lasəmɛnfini/ ou /lasəmɛnəfini/. Dans certains contextes, la prononciation de schwas en position finale est une règle, par ex. en poésie. Cependant, toutes les occurrences de schwas ne sont pas optionnelles. Par exemple, le mot *Bretagne* est toujours prononcé /bʁətaɲ/. Des phénomènes similaires existent pour d'autres phonèmes, en particulier les liaisons lorsque l'on considère les liens entre mots consécutifs.

Dans cet article, nous proposons d'entraîner un autre CAC, dit *modèle d'élision*, pour prédire les élisions de phonèmes. Pour chaque phonème dans la prononciation d'un mot, l'étiquette à apprendre est soit *optionnel*, c.-à-d. que le phonème peut être prononcé ou non, soit *obligatoire*. En plus des graphèmes $\mathbf{g}$ et des autres caractéristiques $\mathbf{o}$ utilisées pour apprendre le CAC de phonétisation, des n-grammes de phonèmes sont aussi utilisés ici. Après l'apprentissage, la probabilité d'élision $\varepsilon(p, \mathbf{g}, \mathbf{o})$ d'un phonème donné p est obtenu à partir du CAC d'élision de la manière suivante :

$$\varepsilon(p, \mathbf{g}, \mathbf{o}) = \begin{cases} 0.5 \times \Pr(e|p, \mathbf{g}, \mathbf{o}) & \text{si } e = \textit{optionnel}, \\ 1 - \Pr(e|p, \mathbf{g}, \mathbf{o}) & \text{si } e = \textit{obligatoire}\ , \end{cases} \tag{1}$$

où e représente l'étiquette obtenue par le CAC d'élision pour p et $\Pr(e|p, \mathbf{g}, \mathbf{o})$ est sa probabilité *a posteriori*. Comme seulement deux étiquettes sont possibles et celle retournée est la plus probable, $\Pr(e|p, \mathbf{g}, \mathbf{o})$ est toujours dans l'intervalle $[0, 5; 1]$. D'après (1), $\varepsilon(p, \mathbf{g}, \mathbf{o})$ varie ainsi dans $[0; 0, 5]$. Cette définition permet ainsi d'éviter au modèle d'élision de complètement supprimer les choix faits par le modèle de phonétisation. Rien n'empêche néanmoins de l'adapter, par ex. pour y intégrer de connaissances *a priori*.

En conséquence, la probabilité d'un phonème peut être reformulée de la manière suivante :

$$\Pr(p|\mathbf{g}, \mathbf{o}) = \phi(p|\mathbf{g}, \mathbf{o}) \times (1 - \varepsilon(p, \mathbf{g}, \mathbf{o}))\ , \tag{2}$$

et la probabilité complémentaire d'élider p est :

$$\Pr(\epsilon|p, \mathbf{g}, \mathbf{o}) = \phi(p|\mathbf{g}, \mathbf{o}) \times \varepsilon(p, \mathbf{g}, \mathbf{o})\ , \tag{3}$$

où ϵ signifie l'absence de phonème. En utilisant ces probabilités, un treillis de phonèmes peut être créé pour chaque séquence de phonèmes donnée et le chemin avec la plus grande probabilité est choisi comme la meilleure prononciation. L'architecture d'un tel treillis est illustré par la figure 1a. Les arcs sont étiquetés par un phonème ou ϵ et sont alors respectivement pondérés par les probabilités de (2) ou (3). Ce principe peut être étendu aux meilleures hypothèses retournées par le modèle de phonétisation. Après application du modèle d'élision sur chaque hypothèse, un nouveau treillis peut être construit comme l'union de toutes les séquences alternatives de phonèmes.

3.3 Expériences sur des mots isolés

La méthode de conversion proposée a été appliquée sur le corpus MHATLex (Pérennou & De Calmes, 2000). Ce corpus comprend $450\,000$ mots avec un total de $710\,000$ prononciations. Chaque mot

possède une étiquette grammaticale et chaque prononciation inclut des possibilités d'élision ainsi que les contextes phonologiques pour lesquels chaque prononciation s'applique. Ce corpus est une version plus détaillée du corpus BDLex, utilisé dans (Illina *et al.*, 2011). Les contextes phonologiques sont ignorés pour la première série d'expériences. Ils seront pris en compte dans la section 4. Le corpus a été découpé en trois parties : ensembles d'apprentissage (75 %), de développement (5%), et de test (20%). Les 2 000 mots les plus fréquents du français ont été placés dans l'ensemble d'apprentissage car ces mots ne seront jamais des mots hors-vocabulaire dans des applications réelles, et possèdent de plus des prononciations irrégulières. De plus, les mots issus d'un même lemme ont été regroupés dans le même ensemble afin d'éviter aux différents ensembles d'être morphologiquement trop similaires. Les modèles CAC de phonétisation et d'élision sont appris sur l'ensemble d'apprentissage en utilisant l'ensemble de développement pour définir le critère d'arrêt tandis que les évaluations sont conduites sur l'ensemble de test. Wapiti est utilisé pour entraîner les CAC. Les graphèmes et phonèmes ont été alignés en utilisant un outil d'alignement plusieurs-à-plusieurs [2] et les CAC entraînés grâce à l'outil Wapiti [3] (Lavergne *et al.*, 2010).

Différents jeux de descripteurs ont été testés pour l'apprentissage du CAC de phonétisation. Ceux-ci rassemblent des n-grammes de graphème (pour rappel, un graphème g_i entouré par une fenêtre de $\pm W$ graphèmes) et l'information verbe/non verbe. Différentes tailles de fenêtre W ont été testées, tandis que l'information sur le verbe a toujours été utilisée. En plus de ces descripteurs, le modèle d'élision prend en compte le phonème p_i et ses $\pm W$ phonèmes voisins. W est fixé à la même valeur pour les graphèmes et les phonèmes afin d'éviter l'ajout d'un paramètre supplémentaire.

La table 1 présente les taux d'erreurs au niveau phonème (PER) et au niveau mot (WER) sur l'ensemble de test pour différentes tailles de fenêtre et avec ou sans modèle d'élision. Les résultats sont comparés à ceux obtenus par Liaphon, le système le plus utilisé pour la phonétisation d'énoncés pour le français (Béchet, 2001). Liaphon repose sur des règles manuelles qui couvrent les règles générales de prononciation ainsi que les exceptions. La version utilisée pour les expériences est une version modifiée optimisée pour l'usage en synthèse de parole. Au contraire, les résultats pour l'approche à base de CAC de (Illina *et al.*, 2011) ne sont pas reportés ici car le corpus et la stratégie de partitionnement des données sont différents. Premièrement, il apparaît que, pour $W = 2$, notre approche obtient des résultats proches de ceux de Liaphon, bien que légèrement moins bons. L'accroissement de la taille de la fenêtre des graphèmes apporte un gain. Cependant, après une taille de 2, il est apparu dans nos expériences que la qualité des CAC était dégradée. Cela vient probablement du fait que l'ensemble d'apprentissage contient beaucoup de mots assez proches en raison des contraintes sur les lemmes, ce qui amène à un effet de surapprentissage. Deuxièmement, l'usage d'un modèle d'élision amène de la variabilité dans le treillis de phonèmes sans significativement altérer ou améliorer les résultats.

4 Phonétisation d'énoncés

Dans cette section, nous proposons (i) de modéliser les transitions entre mots en introduisant la notion de contexte phonologique dans le cadre probabiliste posé précédemment et (ii) de calculer l'ensemble des variantes de prononciation d'un énoncé par la composition de TFP. Nous présentons la formalisation de ces contributions, puis leur validation expérimentale sur un corpus de parole.

2. https://code.google.com/p/m2m-aligner/
3. https://wapiti.limsi.fr

Descripteurs	PER (%)	WER (%)
Graphème (sans fenêtre) + verbe/non verbe	$5,8$	$29,9$
+ modèle d'élision	$5,7$	$29,5$
Graphème (± 1) + verbe/non verbe	$2,6$	$11,3$
+ modèle d'élision	$2,4$	$11,6$
Graphème (± 2) + verbe/non verbe	$1,8$	$9,0$
+ modèle d'élision	$1,9$	$9,3$
Liaphon	$1,3$	$6,8$

TABLE 1: PER et WER sur l'ensemble de test de MHATLex.

4.1 Introduction de contextes phonologiques

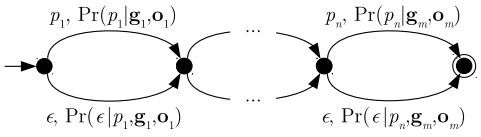
Un mot w_i influence la prononciation des mots précédent et suivant w_{i-1} et w_{i+1}. Réciproquement, la prononciation du mot w_i dépend de celle des mots w_{i-1} et w_{i+1}. L'influence des mots voisins est désignée ici comme le contexte phonologique. Soit l_i l'information transmise par w_i vers la gauche, c.-à-d. à w_{i-1}, et r_i l'information transmise vers la droite à w_{i+1}. De manière symétrique, la prononciation de w_i dépend de r_{i-1} et l_{i+1}. Ainsi, nous proposons d'intégrer r_{i-1} et l_{i+1} comme nouveaux descripteurs dans le processus d'apprentissage des CAC de phonétisation et d'élision.

4.2 Représentation sous forme de transducteurs finis pondérés

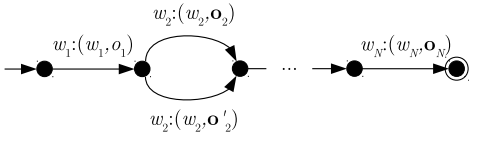
Pour calculer l'ensemble des variantes de prononciation d'un énoncé, nous proposons de construire un treillis de phonèmes en composant deux transducteurs finis pondérés : le premier représentant l'énoncé, le second toutes les prononciations possibles de ses mots.

Comme le montre la figure 1b, la représentation TFP d'un énoncé de N mots consiste simplement en un chaînage de nœuds dont les transitions transposent successivement chaque mot w_i en sa paramétrisation $(w_i, \mathbf{o}_i)$. Ce formalisme accepte d'éventuelles multiples paramétrisations pour un même mot, comme illustré avec le mot w_2 où des chemins alternatifs sont construits dans le transducteur. Par défaut, toutes les transitions ont une probabilité de 1.

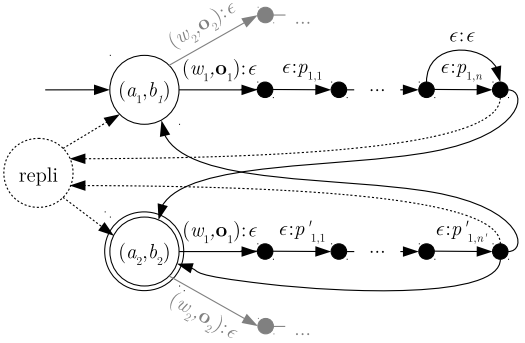
Le TFP du lexique de prononciation est plus complexe car les transitions entre mots doivent être modélisées. La figure 1c illustre son architecture. Pour chaque mot paramétré $(w_i, \mathbf{o}_i)$, plusieurs phonétisations peuvent être acceptables selon le contexte phonologique d'usage du mot. Ces contextes phonologiques sont représentés comme des nœuds (a, b) à partir desquels et vers lesquels chaque phonétisation est reliée. Entre ces nœuds, de même qu'à la figure 1a, chaque séquence de phonèmes est représentée comme une chaîne où $(w_i, \mathbf{o}_i)$ est consommé par le premier arc et les phonèmes $p_{i,j}$ sont les sorties des arcs restants. Les élisions sont représentées par des ϵ-transitions. Finalement, les transitions entre mots sont traitées de la manière suivante : chaque prononciation contextualisée $(r_{i-1}, w_i, \mathbf{o}_i, p_{i,1}, \ldots, p_{i,n}, l_{i+1})$ est liée à tous les nœuds contexte possibles (a, r_{i-1}) et (l_{i+1}, b), pour tous a et b de l'ensemble des contextes respectifs l_i et r_i transmis par w_i à gauche et à droite. Toutes les prononciations sont également liées à un nœud de repli pour autoriser des transitions théoriquement interdites. Les arcs vers les nœuds contexte sont pondérés avec une probabilité de 1 tandis ceux vers le nœud de repli sont pondérés avec une pénalité empirique fixée à e^{-10}. Enfin, en fonction de leur contexte phonologique, certains nœuds contexte sont définis comme terminaux. Les prononciations de chaque mot de l'énoncé sont soit dérivées à partir du dictionnaire, soit, pour les mots hors-vocabulaire, à partir du phonétiseur de mots en contexte présenté à la section 4.1. Pour



(a) Treillis de phonèmes pour un mot isolé.



(b) TFP des mots d'un énoncé.



(c) TFP d'un lexique de prononciation. Les arcs sans étiquette sont des transitions ϵ:ϵ. Les probabilités ne sont pas affichées par souci de clarté.

FIGURE 1

les prononciations dans le vocabulaire, la probabilité de chaque phonème est de 1 s'il est obligatoire et de $0,5$ s'il est optionnel. Dans ce second cas, une ϵ-transition de probabilité $0,5$ est construite en complément parallèlement au phonème. Les probabilités des mots hors-vocabulaire sont, elles, données par le phonétiseur de mots en contexte.

En composant le transducteur de l'énoncé avec celui du lexique, un treillis de phonèmes est obtenu et décodé pour générer la meilleure ou les meilleures prononciations pour l'énoncé.

4.3 Expériences sur des énoncés

Les valeurs possibles des contextes phonologiques sont issues du corpus MHATLex. Deux valeurs sont considérés pour r_{i-1} : l'une indique que le mot précédent se termine par une syllabe ouverte, l'autre par une syllable fermée. Celles pour l_{i+1} sont plus variées : le mot suivant peut débuter par une consonne, une semi-voyelle ou voyelle, un phonème nasal ou non, les liaisons peuvent être interdites ou alors il peut ne pas y avoir de mots suivant (fin de phrase). Ce dernier contexte est le seul cas permettant de définir un nœud contexte comme terminal. Les CAC de phonétisation et d'élision ont été réappris sur l'ensemble d'apprentissage de MHATLex augmenté des informations de contexte.

L'approche proposée a été appliquée sur un corpus de parole, dont la phonétisation a été vérifiée manuellement, d'environ $1\,400$ énoncés pour un total de $12\,000$ mots. Les énoncés ont été phonétisés avec la meilleure configuration de la section 3. Les résultats pour les quatre configuration testées sont mesurés en termes de PER et de taux d'erreurs sur les énoncés (SER, pour *Sentence Error Rate*).

Les résultats sont présentés dans la table 2 et sont comparés à ceux de Liaphon sur le même corpus. Tout d'abord, les PER sont bien plus élevés que sur les mots isolés. Cela montre clairement la difficulté de modéliser la prononciation d'énoncés. Ensuite, nous observons sur les différentes configurations que le modèle d'élision et les contextes phonologiques apportent des améliorations significatives qui, par ailleurs, se complètent en partie. Enfin, notre approche produit de moins bons résultats que Liaphon. Nous pensons que c'est logique car les treillis de phonèmes recensent de nombreux chemins équiprobables du fait de notre stratégie de pondération des prononciations issues du vocabulaire. Ces chemins sont notamment engendrés par des possibilités d'élision ou de liaison. En réalité, cette

Descripteurs et modèles	PER (%)	SER (%)
Graphèmes (± 2) + verbe/non verbe	$22,6$	$88,4$
+ modèle d'élision (sans contextes phonologiques)	$16,8$	$89,2$
+ contextes phonologique (sans modèle d'élision)	$17,7$	$85,6$
+ modèle d'élision + contextes phonologiques	$16,4$	$87,7$
Liaphon	$13,2$	$57,4$

TABLE 2: PER et SER sur le corpus de parole.

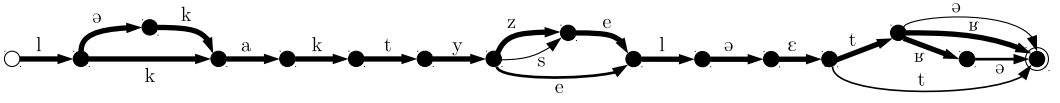


FIGURE 2: Treillis de phonèmes pour l'énoncé « le cactus et le hêtre ». Les transitions les plus probables sont représentées par les lignes les plus épaisses.

hypothèse d'équiprobabilité n'est pas vraie dans la langue. Au contraire, Liaphon fait des hypothèses concernant ces phénomènes. Il serait néanmoins simple de corriger cet effet dans notre méthode.

Un exemple de treillis de phonèmes (élagué) est donné par la figure 2 pour l'énoncé « le cactus et le hêtre », où les mots *cactus* et *hêtre* sont des mots hors-vocabulaire. Des chemins alternatifs apparaissent clairement. De meilleurs résultats pourraient sûrement être obtenus par différentes sophistications que nous prévoyons d'étudier, comme la prise en compte d'informations morphosyntaxiques plus riches (qui permettraient ici de corriger la phonétisation de *cactus*) ou une réévaluation du treillis de phonèmes en post-traitement par un modèle de langage. Néanmoins, cet exemple montre le potentiel de l'approche proposée.

5 Conclusions et perspectives

Cet article présente une nouvelle méthode de phonétisation du français. L'objectif principal de cette approche est de produire des treillis de phonèmes qui peuvent être facilement adaptés à des cas spécifiques, comme un style de parole ou un accent spécifique, en particulier pour la synthèse de parole. Cette méthode repose sur l'utilisation de champs aléatoires conditionnels pour phonétiser les mots isolés, élider certains phonèmes et prendre en compte les contextes phonologiques, ainsi que sur des transducteurs finis pondérés pour étendre la phonétisation à des énoncés.

De nombreuses perspectives sont offertes par cette approche. Tout d'abord, les TFP d'énoncés pourraient intégrer de multiples tokénisations ou prendre en compte des incertitudes sur la paramétrisation, par ex. au niveau des classes grammaticales. Cela peut notamment être utile pour des abréviations ou des acronymes. Ensuite, l'adaptation des treillis de phonèmes peut permettre d'améliorer les applications de la synthèse de parole où une expressivité ou un style de parole particuliers sont nécessaires, par ex. les jeux vidéo, les livres audio ou l'apprentissage de la langue. Enfin, l'utilisation des treillis par un moteur de synthèse de parole offrirait à ce dernier plus de choix et de flexibilité.

Références

BÉCHET F. (2001). LIA_PHON : un système complet de phonétisation de textes. *Traitement Automatique des Langues (TAL)*, (1).

BELLEGARDA J. R. (2005). Unsupervised, language-independent grapheme-to-phoneme conversion by latent analogy. *Speech Communication*, (2).

BENESTY J., SONDHI M. M. & HUANG Y. (2008). *Handbook of speech processing*. Springer.

BISANI M. & NEY H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*.

BODENSTAB N. & FANTY M. (2007). Multi-pass pronunciation adaptation. In *Proc. of ICASSP*.

CLAVEAU V. (2009). Letter-to-phoneme conversion by inference of rewriting rules. In *Proc. of Interspeech*.

HAHN S., LEHNEN P. & NEY H. (2011). Powerful extensions to CRFs for grapheme to phoneme conversion. In *Proc. of ICASSP*.

HAHN S., VOZILA P. & BISANI M. (2012). Comparison of grapheme-to-phoneme methods on large pronunciation dictionaries and LVCSR tasks. In *Proc. of Interspeech*.

HAZEN T. J., HETHERINGTON I. L., SHU H. & LIVESCU K. (2005). Pronunciation modeling using a finite-state transducer representation. *Speech Communication*, **46**(2).

ILLINA I., FOHR D. & JOUVET D. (2011). Grapheme-to-Phoneme Conversion using Conditional Random Fields. In *Proc. of Interspeech*.

JIAMPOJAMARN S., KONDRAK G. & SHERIF T. (2007). Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion. In *Proc. of HLT-NAACL*.

LAURENT A., DELÉGLISE P. & MEIGNIER S. (2009). Grapheme to phoneme conversion using an SMT system. In *Proc. of Interspeech*.

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

LEHNEN P., HAHN S., GUTA V.-A. & NEY H. (2012). Hidden conditional random fields with M-to-N alignments for grapheme-to-phoneme conversion. In *Proc. of Interspeech*.

LEHNEN P., HAHN S. & NEY H. (2011). N-grams for conditional random fields or a failure-transition (φ) posterior for acyclic FSTs. In *Proc. of Interspeech*.

MOHRI M., PEREIRA F. & RILEY M. (2000). Weighted finite-state transducers in speech recognition. In *Proc. of the Intl Workshop on Automatic Speech Recognition : Challenges for the Next Millenium*.

NOVAK J. R., MINEMATSU N. & HIROSE K. (2012). WFST-based grapheme-to-phoneme conversion : open source tools for alignment, model-building and decoding. In *Proc. of the 10th International Workshop on Finite State Methods and Natural Language Processing*.

PÉRENNOU G. & DE CALMES M. (2000). MHATLex : Lexical resources for modelling the French pronunciation. In *Proc. of LREC*.

POLYÁKOVA T. & BONAFONTE A. (2011). Introducing nativization to spanish TTS systems. *Speech Communication*, (8).

RAO K., PENG F., SAK H. & BEAUFAYS F. (2015). Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks. In *Proc. of ICASSP*.

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

YAO K. & ZWEIG G. (2015). Sequence-to-sequence neural net models for grapheme-to-phoneme conversion. *ArXiv Computer Science - Computation and Language*.