Automates lexico-phonétiques pour l'indexation et la recherche de segments de parole

Julien Fayolle[1,5] Fabienne Moreau[2,5]
Christian Raymond[3,5] Guillaume Gravier[4,5]
(1) INRIA Rennes (2) Université de Rennes 2 (3) INSA Rennes (4) CNRS
(5) IRISA, Campus de Beaulieu, 35042 Rennes Cedex
Prenom.Nom@irisa.fr

RÉSUMÉ

Ce papier[1] présente une méthode d'indexation de segments de parole qui combine des hypothèses lexicales et phonétiques au sein d'un index hybride à base d'automates. La recherche se fait via un appariement lexico-phonétique semi-imparfait qui tolère certaines imperfections pour améliorer le rappel. Un vecteur de descripteurs, contenant des scores d'édition et une mesure de confiance, pondère chaque transition permettant de caractériser la pertinence des segments candidats pour une recherche plus précise. Les expériences montrent la complémentarité des représentations lexicales et phonétiques et leur intérêt pour rechercher des requêtes d'entités nommées.

ABSTRACT

Lexical-phonetic automata for spoken utterance indexing and retrieval

This paper presents a method for indexing spoken utterances which combines lexical and phonetic hypotheses in a hybrid index built from automata. The retrieval is realised by a lexical-phonetic and semi-imperfect matching whose aim is to improve the recall. A feature vector, containing edit distance scores and a confidence measure, weights each transition to help the filtering of the candidate utterance list for a more precise search. Experiment results show that the lexical and phonetic representations are complementary and we compare the hybrid search with the state-of-the-art cascaded search to retrieve named entity queries.

MOTS-CLÉS : recherche d'information, indexation de parole, représentations lexico-phonétiques, automates et transducteurs, mesures de confiance, distances d'édition, apprentissage supervisé.

KEYWORDS: information retrieval, speech indexing, lexical-phonetic representations, automata and transducers, confidence measures, edit distances, supervised learning.

1 Introduction

La recherche de contenus parlés (Chelba *et al.*, 2008) fait appel aux domaines de la reconnaissance automatique de la parole (RAP) et de la recherche d'information (RI). Seulement les outils de RI textuelle ne sont pas adaptés aux transcriptions automatiques qui sont particulièrement bruitées de par leur nature incomplète et incertaine. En effet, ces transcriptions contiennent de nombreuses erreurs de reconnaissance touchant notamment les mots hors vocabulaire (OOV pour

[1] Travaux réalisés dans le cadre du programme QUAERO, financé par OSEO, agence française pour l'innovation.

out-of-vocabulary) absents des lexiques de transcription et les entités nommées qui véhiculent les informations essentielles du discours (*e.g.*, noms de personnes, de lieux ou d'organisations) nécessaires à la RI. On distingue deux types d'approches pour pallier ces défauts. On peut, d'une part, améliorer le rappel en faisant appel à une représentation de plus bas niveau composée de sous-mots (*i.e.*, subdivisions du mot comme les syllabes ou les phonèmes) qui permet de représenter les mots OOV et plus généralement tous types d'erreurs lexicales. Il est aussi possible d'utiliser des représentations plus denses qu'une simple transcription telles que le graphe, le réseau de confusion ou la liste des N meilleures hypothèses. D'autre part, on peut améliorer la précision en estimant des mesures de confiance qui indiquent le degré de fiabilité de la reconnaissance permettant ainsi de filtrer le bruit. On s'intéresse ici à combiner ces deux approches pour une tâche de recherche de segments de parole.

Cette tâche consiste à retrouver, dans un ensemble de contenus parlés, tous les segments de parole contenant une requête textuelle donnée. On distingue deux stratégies dans l'état de l'art pour combiner efficacement deux niveaux de représentations lexicales et phonétiques. La première considère deux index séparés utilisés en "cascade", *i.e.*, la recherche utilise par défaut l'index lexical et se replie sur l'index phonétique que si nécessaire (Saraclar et Sproat, 2004), ce qui permet d'éviter le bruit de la recherche phonétique dans la plupart des cas. La seconde stratégie modélise les deux niveaux au sein d'un index hybride (Hori *et al.*, 2007; Yu et Seide, 2004), offrant l'avantage d'un possible appariement lexico-phonétique entre la requête et l'index. La méthode proposée reprend l'idée d'un index hybride car il permet des appariements lexico-phonétiques impossibles avec deux index séparés. La structure de l'index est basée sur les automates car ils peuvent représenter tous types de sorties de RAP. L'originalité de la méthode est qu'elle pondère les transitions des automates par un vecteur de descripteurs qui permet de caractériser la pertinence des segments candidats à la requête donnée. Les descripteurs utilisés comprennent : des scores d'édition calculés par un transducteur d'appariement semi-imparfait qui tolére certaines imperfections des représentations ; et une mesure de confiance indiquant la fiabilité des symboles reconnus. Les expériences comparent les performances des combinaisons hybride et cascadée pour rechercher des requêtes d'entités nommées. On présentera tout d'abord la méthode (section 2) puis les expériences (section 3) pour enfin conclure (section 4).

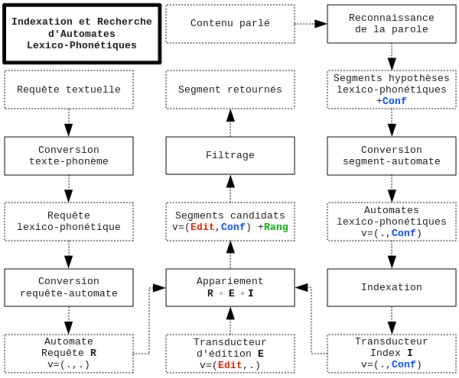

FIG. 1 – Vue générale de la méthode proposée.

2 Méthode proposée

La méthode proposée reprend le cadre général d'indexation d'automates pour la recherche de segments de parole présenté par (Allauzen *et al.*, 2004) en l'adaptant aux automates lexico-phonétiques. La figure 1 donne une vue générale de la méthode. À partir des sorties de RAP, on construit des automates lexico-phonétiques qui constituent l'index (section 2.1). La requête textuelle est phonétisée et aussi représentée par un automate lexico-phonétique. Un appariement plus ou moins imparfait est rendu possible en composant successivement la requête, un transducteur d'édition et l'index (section 2.2). Cette opération renvoie une liste de segments candidats qui peut être filtrée à l'aide d'un vecteur de descripteurs qui pondère chaque segment (section 2.3).

2.1 Automates lexico-phonétiques

Dans cet article, un automate lexico-phonétique désigne simplement un automate à états finis dont les symboles appartiennent soit à un alphabet lexical Σ^{lex} soit à un alphabet phonétique Σ^{ph}, et dont les poids sont multi-dimensionnels. L'automate peut ainsi avoir des chemins lexicaux et phonétiques concurrents pondérés par un vecteur de descripteurs variés (*e.g.*, voir figure 2). On définit l'automate sur le semi-anneau tropical de sorte que le poids d'un chemin soit la somme des poids de ses transitions et que le chemin le plus court soit celui de poids minimal. On peut toujours déterminer ce chemin le plus court si les poids sont toujours comparables, *i.e.*, s'ils sont totallement ordonnés. C'est précisément le cas lorsqu'on considère l'ordre lexicographique (aussi appelé ordre alphabétique) comme dans (Can et Saraclar, 2011). Chaque transition correspond à un symbole s (lexical ou phonétique) reconnu entre les temps de début t_d et de fin t_f avec une mesure de confiance associé c. Le poids de la transition est le suivant :

$$v = (0, 0, 0, 0, 0, w_{conf}^{lex+ph} = -(t_f - t_d).log(c))$$

où w_{conf}^{lex+ph} est un score de confiance commun aux niveaux lexical et phonétique. Il est proportionnel à la durée du symbole s pour que les chemins lexico-phonétique concurrents ayant des nombres différents de symboles soient comparables.

L'automate ainsi construit est ensuite converti en un transducteur de facteurs acceptant toutes les sous-séquences de l'automate en entrée et donnant l'identifiant du segment de parole en sortie. L'index est constitué de l'union de tous les transducteurs de facteurs (Allauzen *et al.*, 2004).

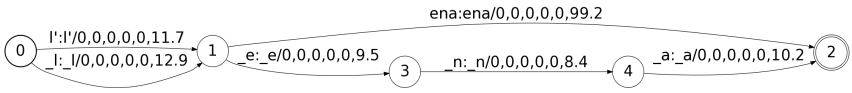

FIG. 2 – Exemple d'automate lexico-phonétique acceptant 4 chemins : 1 lexical ("l' ena"), 1 phonétique ("_l _E _n _a"), et 2 lexico-phonétiques ("l' _E _n _a" et "_l ena"). Les pondérations sont de la forme $(0, 0, 0, 0, 0, w_{conf}^{lex+ph})$.

2.2 Appariements lexico-phonétiques

L'appariement entre la requête R et l'index I peut être réalisé par la simple composition automate-transducteur $R \circ I$. Il est cependant possible d'obtenir un appariement plus flexible en utilisant un transducteur d'édition E par la composition successive $R \circ E \circ I$ (Mohri, 2002). Nous présentons trois types de transducteurs d'édition lexico-phonétiques correspondant à des appariements parfait, imparfait et semi-imparfait et calculant les scores d'édition du vecteur

$$v = (w_{cor}^{lex}, w_{cor}^{ph}, w_{sup}^{ph}, w_{ins}^{ph}, w_{sub}^{ph}, 0)$$

qui comprend les nombres de mots corrects, de phonèmes corrects, et d'erreurs phonétiques (suppressions, insertions et substitutions).
Le transducteur d'appariement parfait n'a pour but que de compter les mots et phonèmes corrects. Le compte des mots corrects vient en premier dans l'ordre lexicographique afin de privilégier les appariements lexicaux plutôt que phonétiques. Les imperfections ne sont pas tolérées, ce qui rend ce transducteur particulièrement restrictif.
Le transducteur d'appariement imparfait permet de compter non seulement les mots et phonèmes corrects mais aussi les erreurs phonétiques. Le problème est que l'appariement se fait sans aucune contrainte. Ainsi toutes les imperfections sont tolérées (*e.g.*, chemins ne comptant aucun symbole correct), ce qui le rend particulièrement gourmand.
Un bon compromis entre entre ces deux approches extrêmes peut être de compter les imperfections sous certaines contraintes. Le transducteur d'appariement semi-imparfait proposé tient compte de la variabilité phonétique connue a priori afin de limiter les possibilités d'imperfection : "sur une fenêtre glissante de ϕ phonèmes, le taux de phonèmes corrects doit être supérieur ou égal à τ". Les paramètres sont ici fixés arbitrairement à $\phi = 2$ et $\tau = 1/2$ en guise d'expérience préliminaire. De plus amples recherches seront nécessaires pour les fixer correctement.
La figure 3 illustre ces trois types de transducteurs pour un alphabet lexico-phonétique restreint.

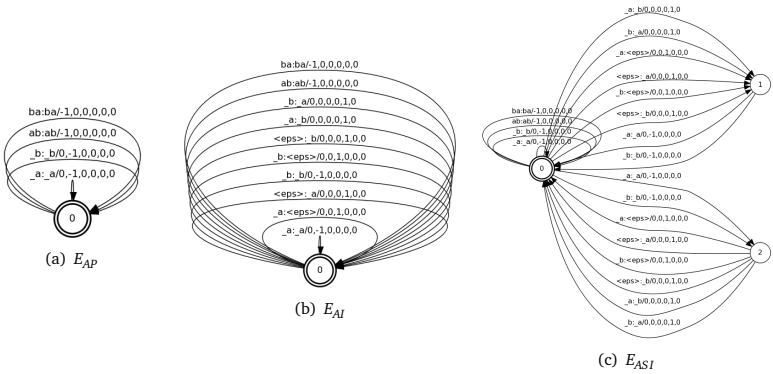

FIG. 3 – Transducteurs d'édition pour appariements lexico-phonétiques parfait (a), imparfait (b) et semi-imparfait (c) dans le cas où $\Sigma^{lex} = \{ab, ba\}$ et $\Sigma^{ph} = \{\_a, \_b\}$. Les pondérations sont de la forme $(w_{cor}^{lex}, w_{cor}^{ph}, w_{sup}^{ph}, w_{ins}^{ph}, w_{sub}^{ph}, 0)$.

2.3 Filtrage des segments candidats

Après appariement et projection sur l'étiquette de sortie, on obtient une liste de segments pondérés et ordonnés suivant l'ordre lexicographique. Chaque segment candidat est ainsi associé à un vecteur de 7 descripteurs :

$$(rang, w_{cor}^{lex}, w_{cor}^{ph}, w_{sup}^{ph}, w_{ins}^{ph}, w_{sub}^{ph}, w_{conf}^{lex+ph})$$

Déterminer si un segment est pertinent ou non à partir de ces descripteurs se ramène à un problème de classification binaire qui peut se résoudre par une méthode d'apprentissage quelconque (*e.g.*, arbre de décision). La probabilité estimée qu'un segment soit pertinent peut ensuite être seuillée suivant le compromis rappel-précision recherché.

3 Expériences

Dans cette partie, nous détaillons le protocole expérimental (section 3.1) permettant de mettre en œuvre la méthode proposée à travers deux expériences sur la complémentarité des représentations lexicales et phonétiques (section 3.2) et la recherche de segments de parole (section 3.3).

3.1 Protocole expérimental

Les données audio utilisées pour les expériences rassemblent 6h d'émissions radiophoniques francophones (2h africa1, 2h tvme, 2h rfi) issues du corpus ESTER2 (Galliano *et al.*, 2009) dont les transcriptions de référence sont annotées manuellement en entités nommées. La RAP est réalisée par un système de transcription à large vocabulaire (65k mots) dont les taux d'erreurs par mot sur ce corpus varient de 16.0% à 42.2%. Les données sont automatiquement décomposées en 3447 segments de parole. La liste des N meilleures hypothèses est réordonnancée grâce à un étiquetage morpho-syntaxique (Huet *et al.*, 2010). Le niveau lexical n'est constitué que de la meilleure hypothèse de transcription. Le niveau phonétique est obtenue en forçant l'alignement entre le signal audio et la prononciation du niveau lexical. Les mesures de confiance lexicales et phonétiques sont calculées à partir des probabilités a posteriori et de l'entropie entre les différentes hypothèses (Chen *et al.*, 2006). Pour éviter les problèmes d'appariement dûs aux flexions morphologiques, les mots sont lemmatisés par l'outil TreeTagger[2].
Les automates ont été implémentés avec OpenFST[3]. Les tailles respectives des index lexical, phonétique et hybride sont de 9.9, 32.8 et 47.6 Mo.
Pour estimer la probabilité qu'un segment candidat soit pertinent étant donné l'ensemble des descripteurs, on a utilisé un bagging sur 20 arbres de décision (Bonzaiboost[4]). L'évaluation se fait suivant une validation croisée sur 5 ensembles d'échantillons : 80% pour l'apprentissage et 20% pour le test.
Les requêtes sont exclusivement composées d'entités nommées extraites des transcriptions de référence. Elles sont phonétisées grâce au lexique phonétique ILPho[5]. Si le mot ne se trouve pas

[2] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[3] http://www.openfst.org/
[4] http://bonzaiboost.gforge.inria.fr/
[5] http://catalog.elra.info/product_info.php?products_id=760

dans le lexique, de multiples prononciations sont générées via le phonétiseur Lia_phon[6]. En plus des deux jeux de requêtes IV (pour in-vocabulary) et OOV habituels, nous proposons un troisième jeu de requêtes composées à la fois de mots IV et OOV (*e.g.*, prénom IV suivi du nom OOV). Ces requêtes IV/OOV sont intéressantes car elles représentent un niveau de difficulté intermédiaire (a priori plus difficile que les requêtes IV mais moins que celles OOV) et sont plus fréquentes que les requêtes OOV. Le tableau 1 montre la répartition des requêtes utilisées. La recherche de segments de parole est évaluée en terme de MAP (mean average precision) correspondant à l'aire sous la courbe rappel/précision.

3.2 Complémentarité des représentations lexicales et phonétiques

Cette expérience préliminaire consiste à mesurer la qualité des représentations lexicales et phonétiques ainsi que leur complémentarité. En alignant, pour chaque segment de parole, les automates lexico-phonétiques d'hypothèse et de référence à l'aide d'un transducteur d'édition imparfait, il est possible d'obtenir le tableau 2 qui donnent les taux de symboles corrects pour les termes IV et OOV composant les entités nommées. On utilise, d'une part, le niveau lexical sur les zones correctement reconnues et, d'autre part, le niveau phonétique sur les zones erronées. On constate que 73.89% des lemmes sont bien reconnus. Pour les lemmes mal reconnus, on peut heureusement se replier sur le niveau phonétique dont 67.73% des phonèmes sont corrects. Cela montre bien que les niveaux lexical et phonétique sont complémentaires et justifie donc leur combinaison pour rechercher des entités nommées.

3.3 Recherche de segments de parole

Le but de cette expérience est de comparer les recherches de segments de parole pour des index, des requêtes, des appariements et des filtrages différents. On distingue les recherches utilisant un index "lexical", un index "phonétique", deux index "cascadés" (méthode de l'état de l'art qui consiste à ne chercher dans l'index phonétique que si la recherche lexicale n'a rien donné) et un index "hybride" lexico-phonétique. Les requêtes peuvent être IV, OOV et IV/OOV. L'appariement est parfait ou semi-imparfait. L'appariement imparfait a été mis de côté car il est trop gourmand en temps de calcul. Deux filtrages sont considérés utilisant de simples seuillages soit sur le score de confiance lexico-phonétique (f-conf) soit sur la probabilité estimée de façon supervisée par les arbres de décisions qui combinent les 7 descripteurs présentés précédemment (f-super). La méthode de référence correspond à une recherche cascadée dont l'appariement est parfait et dont le filtrage est basé sur un seuillage du score de confiance. Le tableau 3 rapporte les résultats obtenus.

[6] `http://www.atala.org/LIA-PHON`

#mots	1	2	3	4	5	6	7	8+	total
IV	209	276	125	73	29	24	16	18	770 (68%)
OOV	76	43	1	.	.	.	.	.	120 (10%)
IV/OOV	.	120	73	29	11	8	4	2	247 (22%)

TAB. 1 – Répartition des requêtes en fonction du type et de la longueur en nombre de mots.

terme d'entité nommée	% lemmes dans la référence	% lemmes correct sur les zones correctes	% phonèmes corrects sur les zones erronées
IV	93.57	78.97	67.34
OOV	6.43	0.00	68.54
Tous	100.00	73.89	67.73

TAB. 2 – Complémentarité des représentations lexicales et phonétiques pour les entités nommées.

Appariement	Index	Parfait				Semi-Imparfait			
		lex	**ph**	**cas**	**hyb**	**lex**	**ph**	**cas**	**hyb**
IV	**f-conf**	.634	.577	.673	.577	.634	.015	.047	.013
	f-super	.631	.646	*.677*	*.681*	.629	*.693*	*.713*	***.729***
OOV	**f-conf**	.000	.036	.036	.036	.000	.001	.001	.001
	f-super	.000	*.053*	*.053*	*.053*	.000	***.139***	***.139***	***.139***
IV/OOV	**f-conf**	.000	.024	.024	*.029*	.000	.001	.001	.001
	f-super	.000	*.024*	*.024*	*.024*	.000	***.256***	***.256***	*.250*
Global	**f-conf**	.523	.479	.556	.478	.523	.009	.015	.008
	f-super	.520	.540	*.568*	*.570*	.519	*.610*	*.637*	***.650***

TAB. 3 – Evaluation en MAP de la recherche de segments de parole : méthode de référence, *meilleur que la référence*, **meilleur(s) résultat(s)**.

De manière générale, on remarque tout d'abord que la méthode de référence peut facilement être améliorée pour tous les types de requêtes en utilisant un appariement semi-imparfait et un filtrage supervisé (le filtrage sur le score de confiance n'est pas suffisant). Deuxièmement, la recherche hybride (accompagnée d'un filtrage supervisé) obtient des performances supérieures ou équivalentes aux recherches lexicale et phonétique, ce qui justifie la combinaison hybride.

Plus spécifiquement, la recherche hybride obtient les meilleurs résultats pour les requêtes IV. Pour les requêtes OOV, les recherches phonétiques, cascadées et hybrides sont équivalentes puisqu'elle ne font appel qu'au niveau phonétique. Pour les requêtes mixtes IV/OOV, il est surprenant de constater que la recherche phonétique soit meilleure que celle hybride. Cela est dû au fait que le rang donne trop d'importance aux appariements lexicaux même lorsque ceux-ci ne sont pas pertinents (mots mal reconnus ou mots très fréquents). Nous pensons que l'ajout d'un score tf*idf dans le vecteur de pondération et l'utilisation de meilleurs mesures de confiance pourront aider à mieux gérer ces cas.

Finalement, la recherche hybride (avec un appariement semi-imparfait et un filtrage supervisé) offre les meilleures performances globales.

4 Conclusion

Nous avons présenté une méthode d'indexation et de recherche de segments de parole représentés sous forme d'automates lexico-phonétiques. Les résultats montrent la complémentarité des niveaux lexical et phonétique (extraits de la meilleure hypothèse de reconnaissance de la parole) et l'avantage d'un index hybride. L'utilisation d'un appariement semi-imparfait et d'un filtrage supervisé (combinant des scores d'édition et un score de confiance) permet d'améliorer significativement la recherche en terme de MAP.

En perspective, de nombreux aspects de la méthode sont encore à améliorer. On peut envisager une amélioration du rappel par une meilleure adaptation des transducteurs d'appariement semi-imparfait et l'utilisation de représentations plus denses (*e.g.*, N meilleures hypothèses) ; mais aussi une amélioration de la précision en utilisant des mesures de confiance de meilleure qualité (Fayolle *et al.*, 2010) et en enrichissant le vecteur de descripteurs avec d'autres types d'informations (*e.g.*, scores tf*idf).

Références

Allauzen, C., Mohri, M. et Saraclar, M. (2004). General indexation of weighted automata - application to spoken utterance retrieval. *In HLT/NAACL'04*, pages 33–40.

Can, D. et Saraclar, M. (2011). Lattice indexing for spoken term detection. *IEEE Transactions on Audio, Speech & Language Processing*, 19(8):2338–2347.

Chelba, C., Hazen, T. J. et Saraçlar, M. (2008). Retrieval and browsing of spoken content. *Signal Processing Magazine, IEEE*, 25(3):39–49.

Chen, T.-H., Chen, B. et Wang, H.-M. (2006). On using entropy information to improve posterior probability-based confidence measures. *In ISCSLP'06*, pages 454–463.

Fayolle, J., Moreau, F., Raymond, C., Gravier, G. et Gros, P. (2010). Crf-based combination of contextual features to improve a posteriori word-level confidence measures. *In Interspeech'10*, Makuhari, Japan.

Galliano, S., Gravier, G. et Chaubard, L. (2009). The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. *In Interspeech'09*, pages 2583–2586.

Hori, T., Hetherington, I. L., Hazen, T. J. et Glass, J. R. (2007). Open-vocabulary spoken utterance retrieval using confusion neworks. *In ICASSP'07*, pages 73–76.

Huet, S., Gravier, G. et Sébillot, P. (2010). Morpho-syntactic post-processing of n-best lists for improved french automatic speech recognition. *Computer Speech and Language*, (24):663–684.

Mohri, M. (2002). Edit-distance of weighted automata. *In CIAA'02*, pages 1–23. Springer Verlag.

Saraclar, M. et Sproat, R. (2004). Lattice-based search for spoken utterance retrieval. *In HLT-NAACL'04*, pages 129–136.

Yu, P. et Seide, F. (2004). A hybrid-word/phoneme-based approach for improved vocabulary-independent search in spontaneous speech. *In Interspeech'04, Korea*, page 293296.