Ancres macrophonétiques pour la transcription automatique

Daniel Moraru, Guillaume Gravier

IRISA
Campus Universitaire de Beaulieu, 35042 Rennes Cedex, France
daniel.moraru@irisa.fr, guillaume.gravier@irisa.fr

ABSTRACT

Automatic speech recognition mainly rely on hidden Markov models (HMM) which make little use of phonetic knowledge. As an alternative, landmark based recognizers rely mainly on precise phonetic knowledge and exploit distinctive features. We propose a theoretical framework to combine both approaches by introducing prior (phonetic) knowledge in a non stationary HMM decoder. To demonstrate the potential of the method, we investigate how broad phonetic landmarks could be used to improve a HMM decoder by focusing the best path search. We show that, assuming error free landmark detection, every broad phonetic class brings a small improvement. The use of all the classes reduces the error rate from 22% to 14% on a broadcast news transcription task. We also experimentally validate that landmarks boundaries does not need to be detected precisely and that the algorithm is robust to non detection errors.

1. INTRODUCTION

Dans le cadre de la reconnaissance automatique de la parole par modèle de Markov caché, le décodage consiste à construire un graphe incluant l'ensemble des sources d'information disponibles (modèle de langage, lexique de prononciations, modèles acoustiques) et à effectuer une recherche du meilleur chemin dans ce graphe pour trouver la séquence de mots

$$\hat{w} = \arg\max_{w} p(y|w)p(w) \qquad (1)$$

à l'aide de l'algorithme de Viterbi. L'approche par modèle de Markov caché se base au niveau acoustique sur l'apprentissage et ne permet pas de prendre en compte des connaissances phonétiques précises. A l'inverse de cette approche, de nombreux travaux ont porté sur l'utilisation explicite de traits phonétiques pour représenter le signal de parole à des fins de reconnaissance automatique [7, 5, 4]. Les approches basées sur des traits phonétiques reposent sur une détection de traits fins comme les *onsets* et *offsets* des fricatives et des consonnes voisées [5] ou encore des traits distinctifs [7, 4]. Cependant, en pratique, la détection automatique de tels traits est délicate, notamment sur des signaux bruités ou sur de la parole spontanée.

Nous proposons une approche permettant d'introduire des connaissances (phonétiques) a priori dans le formalisme des modèles de Markov cachés et appliquons cette approche à la connaissance des macroclasses phonétiques composant le signal. Alors que les systèmes à base de traits phonétiques utilisent ces derniers comme descripteurs du signal, nous proposons de les utiliser pour guider la recherche du meilleur chemin dans le graphe de décodage. En ce sens, les connaissances a priori sont utilisées comme *points d'ancrage* de l'algorithme de décodage. Dans le cadre de cette étude, nous cherchons à valider le modèle proposé et à évaluer l'impact sur le système de reconnaissance de la parole. Dans ce but, nous utilisons des points d'ancrage déterminés manuellement. Cependant, le cadre proposé permet de prendre en compte les erreurs de détection des points d'ancrage.

Après une description du principe du décodage avec ancrage, nous étudions l'apport des différentes macroclasses phonétiques au décodage. Nous étudions ensuite l'impact de la précision de la détection des points d'ancrage macrophonétiques sur les performances avant de conclure sur des perspectives.

2. DÉCODAGE AVEC CONTRAINTES

La plupart des systèmes de transcription utilisent plusieurs passes de transcription. Chaque passe permet de générer la meilleure phrase selon l'Eq. (1) ainsi qu'un graphe de mots, ce dernier représentant de manière compact un ensemble de phrases. Chaque passe est basée sur le même principe de recherche du meilleur chemin dans un treillis en utilisant des modèles et des connaissances de plus en plus fines. En raison de la taille des treillis, une recherche exhaustive du meilleur chemin est impossible en pratique et l'on a recours à une recherche approximative dite en faisceau. Nous décrivons succinctement cet algorithme de recherche avant de l'étendre pour introduire des connaissances a priori.

2.1. Décodage en faisceau

Le décodage en faisceau se base sur l'algorithme de Viterbi pour la recherche du meilleur chemin dans un treillis. Cet algorithme procède de manière incrémentale en recherchant l'hypothèse partielle optimale menant à un instant donné dans un état (j,t) du treillis, selon

$$S(j,t) = \max_{i} S(i,t-1) + \ln(a_{ij}) + \ln(p(y_t|j)) \ , \qquad (2)$$

où $\ln(a_{ij})$ est le poids d'une transition de l'état i vers l'état j et $p(y_t|j)$ la vraisemblance d'un descripteur y_t conditionnellement à l'état j. $S(i,t)$ représente le score du meilleur chemin partiel menant à l'état i à l'instant t.

Dans la recherche en faisceau, à chaque instant, seules les hypothèses les plus prometteuses, vérifiant

$$S(i,t) \geq \max_{j} S(j,t) - \alpha \ , \qquad (3)$$

sont conservées, α étant un facteur d'élagage contrôlant

la taille du faisceau. En général, on fixe également un nombre maximum d'hypothèses à chaque instant.

2.2. Introduction de points d'ancrage

Nous étendons le formalisme de l'algorithme de Viterbi pour y introduire des contraintes liées à une connaissance a priori sur le meilleur chemin. Par exemple, si l'on sait a priori qu'une portion donnée d'un segment correspond à une voyelle, il est possible de pénaliser, voire de supprimer, les chemins dans le treillis qui ne concordent pas avec cette connaissance.

D'un point de vue formel, une approche possible consiste à considérer le graphe de décodage comme non stationnaire. L'idée est que, si une transition amène dans un état du treillis (i, t) incompatible avec les connaissances a priori, alors le coût de cette transition augmente. Pour refléter la présence d'une connaissance a priori, vue comme une contrainte sur les chemins dans le procédure de décodage, nous remplaçons dans (2), les poids de transition a_{ij} par

$$\ln(a_{ij}(t)) = \ln(a_{ij}) - \lambda(t) I_j(t) \quad (4)$$

où $I_j(t)$ est une indicatrice valant 0 si le noeud (j, t) du treillis est en accord avec la connaissance a priori dont on dispose. La pénalité $\lambda(t) \geq 0$ infligée à une transition incompatible avec nos connaissances a priori dépend de la confiance accordée à ces connaissances. Par exemple, si $\lambda(t) = \infty$, on accorde alors une confiance totale aux connaissances a priori disponibles et la recherche du meilleur chemin se limitent aux seuls chemins valides par rapport aux a priori. Si $\lambda(t) = 0$, on retrouve alors un décodage sans a priori. Dans la suite de ce travail, nous considérons des a priori fiables car extraits manuellement et posons donc $\lambda(t) = \infty, \forall t$.

Dans cette approche, les connaissance a priori sont utilisées comme comme *points d'ancrage* ou *ancres* pour guider la recherche vers les solutions plausibles par rapport aux connaissances disponibles. Dans le cas extrême où $\lambda = \infty$, on peut également parler de *contraintes* sur l'espace des solutions. Soulignons que, à la différence des décodeurs à base de traits phonétiques, la non détection de points d'ancrage n'a que peu d'impact sur le décodage.

3. ANCRES MACROPHONÉTIQUES

Dans un premier temps, nous exploitons le formalisme défini précédemment pour étudier l'apport d'une connaissance sur le contenu macrophonétique du signal à reconnaître. Nous considérons les cinq macroclasses phonétiques définis par SAMPA : voyelles, occlusives, fricatives, nasales et *glides*, cette dernière classe regroupant semi-voyelles, latérales et vibrantes. L'idée est de segmenter le signal selon ces cinq classes et d'utiliser le résultat de cette segmentation pour obtenir des points d'ancrage afin de ne garder que les hypothèses valides par rapport à ces points d'ancrage. Dans la mesure où il est illusoire de détecter précisément les frontières dans la segmentation, nous n'utilisons que la partie centrale d'un segment comme ancre. Cette étude ayant pour but principal de déterminer si l'information macrophonétique est pertinente, nous utilisons des points d'ancrage déterminés manuellement.

Plusieurs considérations justifient le choix d'une information macrophonétique. D'une part, nous avons remarqué que, malgré l'élagage, la liste des hypothèses concurrentes à un instant donnée contient des hypothèses correspondant à différentes classes phonétiques. D'autre part, si la détection automatique de traits phonétiques fins est délicate, la détection/segmentation de macroclasses phonétiques donnent de bonnes performances [1, 3, 6].

3.1. Corpus

Les expériences sont réalisées sur un corpus de quatre heures d'émissions radiophoniques, extrait du corpus de développement de la campagne ESTER [2]. Le corpus est essentiellement composé de parole planifiée de haute qualité avec quelques entrevues comportant de la parole plus spontanée et/ou des conditions acoustiques dégradées. Afin de dériver des points d'ancrage, nous avons réalisé un alignement phonétique forcé du corpus. Pour une macroclasse donnée, les points d'ancrage sont obtenus à partir de l'alignement phonétique de référence. Pour chaque phone appartenant à la classe considérée, on génère une ancre centrée sur le phone et dont la durée est proportionnelle à celle du phone. Dans cette première expérience, la durée d'une ancre est fixée à 50% de la durée du phone correspondant.

3.2. Systèmes de transcription

Nous utilisons les points d'ancrage dans deux systèmes de transcriptions. Pour les deux systèmes, une première passe permet de générer un graphe de mots à l'aide d'un modèle de langage trigrame, le graphe de mots étant ensuite réévalué avec des modèles de phones contextuels inter-mots. La différence entre les deux systèmes résident dans la finesse des modèles acoustiques utilisés pour la première passe : des modèles hors contexte dans un cas et des modèles contextuels intra-mots dans l'autre. Les modèles hors-contexte possèdent 114 états à 128 gaussiennes tandis que les modèles contextuels possèdent 4 019 états à 32 gaussiennes.

La taille du treillis de la deuxième passe est relativement petite puisque limitée par le graphe de mots. En revanche, le treillis exploré lors de la première passe est constitué par la composition des graphes représentant le modèle de langage, les prononciations et les modèles acoustiques. Nous considérons donc l'introduction de points d'ancrage au niveau de la première passe afin de faciliter la recherche du meilleur chemin dans le treillis le plus complexe. L'utilisation de points d'ancrage lors de la première passe vise également à limiter la taille du graphe de mots généré.

3.3. Résultats

Le tableau 1 regroupe les taux d'erreurs obtenus pour les deux systèmes avec chacune des classes phonétiques étudiées. Nous reportons également la proportion de signal correspondant à un point d'ancrage (signal/ancre).

L'utilisation de l'ensemble des points d'ancrage permet une amélioration très nette du taux d'erreur après chacune des passes, aussi bien pour les monophones que pour les triphones. Le gain de performance est évidemment plus conséquent lorsque l'on utilise des monophones, ces derniers donnant lieu à plus de confusions phonétiques lors d'un décodage acoustico-phonétique. L'introduction d'une connaissance supplémentaire permet alors de lever plus d'ambiguïtés que pour les triphones. En revanche,

TAB. 1: Taux d'erreur de transcription (en %) après chaque passe pour les différentes classes d'ancrage.

ancres		aucune	toutes	voy.	occ.	fri.	nas.	gli.
signal/ancre		0,0	43,6	18,3	9,0	7,3	2,8	6,2
monophones	passe 1	29,2	15,3	26,6	26,5	27,5	27,8	25,1
	passe 2	22,3	13,9	21,2	20,7	21,0	21,5	20,1
triphones	passe 1	27,3	19,6	27,0	26,3	26,0	26,4	24,9
	passe 2	21,3	15,0	20,7	20,4	20,3	20,7	19,6

il est intéressant de constater que les deux systèmes obtiennent des taux d'erreur comparables, voire légèrement meilleur pour le système monophone, après la deuxième passe. Soulignons également le faible taux d'erreur obtenu avec les monophones après la première passe, comparable au taux d'erreur avec les triphones après les deux passes. Ce résultat suggère que l'utilisation de points d'ancrage en conjonction avec des modèles phonétiques simples permet une nette amélioration des performances pour des temps de décodage faible, le nombre moyen d'hypothèses actives par trame étant divisé par 4 avec les points d'ancrage.

Si l'on regarde l'influence des points d'ancrage de chacune des classes phonétiques, il apparaît clairement que chaque classe apporte une information supplémentaire. En effet, les classes considérées séparément apportent chacune un léger gain de performances, le plus grand gain étant obtenu avec les *glides*. Cette dernière classe correspond à des phonèmes transitoires, fortement affectés par la coarticulation, pour lesquels les modèles acoustiques font de nombreuses confusions. Le gain obtenu avec des points d'ancrage sur les voyelles est plus surprenant dans la mesure où les modèles des deux systèmes de reconnaissance font peu de confusions entre les voyelles et les autres classes. Ceci s'explique par le fait que la liste des nœuds du treillis actifs à un moment donné dépend non seulement du score $p(y_t|i)$ de la trame courante mais également de tout le chemin parcouru jusqu'à ce nœud. Il se peut donc que même si, localement, les probabilités $p(y_t|i)$ sont plus élevés pour les nœuds i correspondant à une voyelle, certains chemins incompatible avec le point d'ancrage ait un bon score global. Il est probable qu'une technique d'anticipation phonétique (phone look-ahead) permettrait de limiter cet effet.

L'utilisation de points d'ancrage lors de la première passe du système de transcription permet, outre une amélioration du taux d'erreur, de limiter la taille du graphe de mots produit. Le tableau 2 indique la taille moyenne des graphes de mots obtenus pour chaque classe de points d'ancrage avec les monophones et les triphones. La taille des graphes est donnée en nombre d'arcs et de nœuds par trames, avec un débit de 100 trames par seconde. Les nœuds du graphe correspondent aux frontières de mots tandis que les arcs correspondent aux mots. Nous donnons également le taux d'erreur oracle (GER) après alignement du graphe avec la transcription de référence. Les conclusions sur la taille des graphes sont similaires en tout point à celles énoncées précédemment sur les taux d'erreur. L'ensemble des points d'ancrage permettent une forte réduction de la taille des graphes de mot, réduction plus marquée pour les monophones qui produisent des graphes de mots plus grands en l'absence de contraintes. Chaque classe d'ancrage apporte une réduction du graphe de mot, les *glides* offrant la plus grande réduction.

Ces résultats montrent que l'utilisation de l'ensemble des ancres macrophonétiques permettrait d'une part de faire grandement diminuer le taux d'erreur des systèmes de transcription, alors que l'apport de chacune de ces classes d'ancrage est minime, et, d'autre part, d'accélérer le décodage. Dans la mesure où les points d'ancrage limitent le faisceau de décodage, il s'avère avantageux de les utiliser avec des modèles acoustiques peu sélectifs qui présentent a priori moins de conflits avec les points d'ancrage.

4. PRÉCISION DES POINTS D'ANCRAGE

Dans les expériences précédentes, nous utilisons des points d'ancrage représentant 50% de la durée des phones correspondant, c'est-à-dire qu'un [a] d'une durée de τ secondes donnera lieu à une ancre de type voyelle d'une durée de $\tau/2$ secondes, centrée sur le milieu du phone. Nous étudions dans cette partie l'importance de la précision temporelle des points d'ancrage en variant la durée des points d'ancrage de 10% à 70% de la durée du phone. Le tableau 3 regroupe les résultats obtenus pour les triphones avec l'ensemble des points d'ancrages, en terme de taux d'erreur après chacune des passes (WER 1 et WER 2), de taille de l'espace de recherche (# hyps, le nombre moyen d'hypothèses actives par trame) et de qualité du graphe de mot généré par la première passe. Ces résultats montrent que, bien évidemment, lorsque la taille des points d'ancrages augmente, le taux d'erreur diminue ainsi que le nombre d'hypothèses propagées et la taille du graphe de mot. De manière plus surprenante, on note que des points d'ancrages d'une durée égale à 10% de la durée des phones permettent déjà une très nette amélioration des performances ainsi qu'une réduction sensible du nombre moyen d'hypothèses actives et de la taille des graphes de mots. Notons que dans cette configuration, les points d'ancrage correspondent seulement à 8,7% de la durée totale du signal. La même expérience en n'utilisant que les voyelles comme points d'ancrage montre une tendance similaire avec une baisse du taux d'erreur de 20,7% à 19,5% pour une durée des points d'ancrage de 10% (soit 3,6% de la durée totale du signal correspondant à des points d'ancrage).

Ces résultats établissent clairement que la détection des points d'ancrage n'a pas besoin d'être précise puisque, avec seulement 10% de la durée réelle, le taux d'erreur baisse de manière significative. Une plus grande précision temporelle dans la détection des ancres permet une légère amélioration des performances et surtout une accélération du décodage en limitant le nombre de chemins explorés dans le treillis de décodage.

Nous avons également pu montré la robustesse de l'approche proposée face aux erreurs de non détection des ancres, en simulant ce type d'erreurs pour différents taux de faux rejet. Sur une heure d'émission, le taux d'erreur

TAB. 2: Taille et taux d'erreur des graphes de mots générés avec les différentes classes d'ancrage.

	monophones			triphones		
contraintes	# arcs	# nœuds	GER (%)	# arcs	# nœuds	GER (%)
aucune	54,8	8,2	7,9	37,0	6,9	7,5
toutes	23,4	4,0	3,0	24,3	4,8	5,0
voyelles	43,6	7,0	7,3	36,4	6,8	7,7
occlusives	45,9	7,1	7,0	36,5	6,8	7,0
fricatives	47,5	7,3	7,2	35,5	6,7	6,9
nasales	49,7	7,6	7,3	35,7	6,7	7,2
glides	39,1	6,3	6,6	31,8	6,1	6,9

TAB. 3: Taux d'erreur, tailles des graphes de mots et de l'espace de recherche en fonction de l'étendue des points d'ancrage. Les résultats sont données sur une heure d'émission (France-Info), pour le système triphones avec l'ensemble des points d'ancrage.

étendue (en %)	0	10	20	30	40	50	60	70
WER 1 (%)	26,3	19,9	19,4	18,8	18,8	18,5	18,8	20,0
WER 2 (%)	20,7	15,2	15,1	15,0	14,8	14,4	14,3	14,7
# hyps	51 730	31 346	29 444	27 240	24 813	22 421	20 091	18 311
# arcs	38,1	30,5	29,8	28,7	26,8	24,9	22,7	20.96
# nœuds	7,0	5,9	5,8	6,0	5,2	4,9	4,5	4,2
GER (%)	6,5	4,5	4,5	4,5	4,6	4,5	4,7	5,1

pour le système monophone est de 13.9% avec l'ensemble des ancres et de 22.3% sans points d'ancrage. Ce taux d'erreur est respectivement de 15,8% et 17,9% pour des taux de non détection des ancres de 25% et de 50%.

5. DISCUSSION

Ce travail propose un cadre théorique pour l'introduction de connaissances a priori dans un système de décodage basé sur la recherche d'un chemin optimal dans un treillis. Nous appliquons ce cadre à la reconnaissance de la parole en introduisant dans le décodage une connaissance oracle des macroclasses phonétiques présentes dans le signal. Les résultats montrent que cette simple information permet une très nette amélioration des performances en transcription, même en conjonction avec des modèles de phones peu performants. De plus, nous mettons en évidence qu'il n'est pas nécessaire de détecter les points d'ancrage avec une bonne précision temporelle. Ces premiers résultats montrent donc l'intérêt de travailler sur un détecteur fiable d'ancres macrophonétiques.

La première perspective de ce travail est donc l'étude du comportement de cette approche avec une détection automatique des points d'ancrage, notamment en ce qui concerne la robustesse aux erreurs de la détection automatique. Une piste de travail consiste à exploiter des mesures de confiances sur la détection des points d'ancrage pour adapter la pénalité $\lambda(t)$ en conséquence. Nous pensons également que l'utilisation d'ancres macrophonétiques devrait permettre une robustesse accrue par rapport aux changement de condition acoustique ou du style de parole, dans la mesure où l'on est en droit d'attendre plus de robustesse d'une segmentation macrophonétique que de modèles de phones.

Une deuxième perspective réside dans l'étude d'autres points d'ancrage, comme les traits distinctifs. Enfin, dans le strict cadre de la reconnaissance de la parole, l'intégration de connaissance phonétique est possible à d'autres niveaux, par exemple à travers un choix dynamique du vocabulaire et/ou du modèle de langage ou encore un élagage du graphe de mots. Notre approche réalise d'une certaine manière une stratégie d'adaptation dynamique du modèle de langage en pénalisant les mots incompatibles avec les points d'ancrage.

Pour conclure cette discussion, notons que le cadre générique proposé s'applique à n'importe quel problème pouvant se formuler sous la forme d'un décodage par optimisation dans un graphe et possèdent donc des applications dans de nombreux domaines autre que la reconnaissance de la parole.

RÉFÉRENCES

[1] M. Chen. Nasal landmark detection. In *Intl. Conf. Speech and Language Processing*, pages 636–639, 2000.

[2] S. Galliano, E. Geoffrois, J.-F. Bonastre, G. Gravier, D. Mostefa, and K. Choukri. Corpus description of the ESTER evaluation campaign for the rich transcription of french broadcast news. In *Language Resources and Evaluation Conference*, 2006. to appear.

[3] A. Howitt. Vowel landmark detection. In *Intl. Conf. Speech and Language Processing*, 2000.

[4] John Hopkins University, Center for Language and Speech Processing. *Landmark-based speech recognition : report of the 2004 John Hopkins Summer Workshop*, 2005.

[5] A. Juneja. *Speech recognition based on phonetic features and acoustic landmarks*. PhD thesis, University of Maryland, 2004.

[6] J. Li and C.-H. Lee. On designing and evaluating speech event detectors. In *European Conf. on Speech Communication and Technology*, 2006.

[7] S. A. Liu. *Landmark detection for distinctive feature-based speech recognition*. PhD thesis, Massachusetts Institute of Technology, 1995.