Avancées dans le domaine de la transcription automatique par décodage guidé

Fethi Bougares[1] Yannick Estève[1] Paul Deléglise[1]
Mickael Rouvier [1] George Linarès[2]
(1) LIUM, Laboratoire d'Informatique de l'Université du Maine
(2) LIA, Laboratoire d'Informatique d'Avignon
[1]prenom.nom@lium.univ-lemans.fr, [2]prenom.nom@univ-avignon.fr

RÉSUMÉ

Dans cet article, nous présentons une méthode de combinaison de systèmes de reconnaissance automatique de la parole (SRAP) inspirée d'un algorithme de décodage guidé (DDA). La combinaison par décodage guidé est basée sur un alignement entre une transcription auxiliaire et l'hypothèse développée par un système primaire, suivi d'une ré-évaluation du score linguistique de cette dernière. Nous proposons une nouvelle méthode qui facilite la gestion des transcriptions auxiliaires pour effectuer cette ré-évaluation sans alignement. Les hypothèses auxiliaires sont groupées par segment sous forme de sac de n-grammes (BONG : Bag Of NGrams) et la ré-évaluation est réalisée en fonction du résultat de recherche dans le sac de trigrammes correspondant. Cette méthode permet de réduire le taux d'erreur mots du système primaire en utilisant des systèmes auxiliaires moins performants.

ABSTRACT

Improvements on driven decoding system combination

This paper proposes an improved driven decoding method for speech recognition system combination. The combination method involves the use of auxiliary transcription as external information source included on primary system decoding process. Auxiliary transcriptions are used to modify search space exploration via linguistic score reevaluation. it was shown that DDA outperforms ROVER when the primary system is guided by a more accurate system. In this paper we propose a new method to manage auxiliary transcriptions which are presented as a bag-of-n-grams (BONG) without temporal matching. These modifications allow to make easier the combination of several hypotheses given by different auxiliary systems and improves primary system WER even with less accurate auxiliary systems.

MOTS-CLÉS : Reconnaissance de la parole, combinaison de systèmes, décodage guidé.

KEYWORDS: Speech recognition, systems combination, driven decoding.

1 Introduction

Bien que la majorité des systèmes de reconnaissance de la parole (SRAP) soient, à l'heure actuelle, basés sur des méthodes statistiques, ils peuvent différer sur plusieurs points (méthodes de paramétrisation du signal, modélisation acoustique et linguistique, algorithmes de décodage ...).

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 795–802,
Grenoble, 4 au 8 juin 2012.

La combinaison de SRAP a pour objectif l'exploitation de ces différences pour construire une transcription finale améliorée. Le résultat de la combinaison de ces systèmes est directement lié à leur degré de complémentarité. En effet, la combinaison de deux systèmes qui font les mêmes types d'erreurs n'améliore pas la qualité de sortie finale.

Plusieurs méthodes de combinaison des SRAP ont été réalisées et testées à différents niveaux : dans le but d'exploiter les points forts de chaque méthode de paramétrisation, différents jeux de paramètres ont été combinés dans (Plahl *et al.*, 2011). La combinaison au niveau acoustique a été aussi testée via une adaptation croisée (cross-adaptation) dans (Stuker *et al.*, 2006). Les sorties de différents SRAP ont été aussi combinées dans un schéma de combinaison *a posteriori* (Fiscus, 1997).

La combinaison par décodage guidé a l'avantage d'être intégrée dans le processus de décodage. Contrairement aux méthodes de combinaison *a posteriori*, il n'est pas nécessaire d'attendre la fin du décodage de tous les systèmes utilisés pour pouvoir combiner leurs sorties.

Dans cet article, nous présentons une méthode de combinaison basée sur l'utilisation de différents systèmes auxiliaires pour la ré-évaluation des scores linguistiques d'un système primaire durant son processus de décodage. Cette méthode de combinaison est adaptée et améliorée dans l'optique de proposer un cadre de combinaison à la volée de systèmes de reconnaissance temps réel.

Cet article est organisé en quatre parties, la première partie présente le principe de décodage guidé, la deuxième détaille la méthode proposée et la troisième expose le cadre expérimental. Avant de conclure, la quatrième partie présente les résultats obtenus.

2 Principe de décodage guidé

Le décodage guidé modifie dynamiquement l'exploration de l'espace de recherche(Lecouteux *et al.*, 2007), il procède par la recherche de points de synchronisation entre les hypothèses d'un système primaire et celles d'un système auxiliaire. Cette recherche est réalisée en utilisant un alignement dynamique (DTW) entre les sorties du système auxiliaire et les résultats partiels de décodage du système primaire. Ensuite un score de correspondance est calculé selon le nombre de mots correctement alignés. Le score de correspondance est utilisé pour modifier la probabilité linguistique des hypothèses du système primaire en utilisant la formule suivante :

$$L(w_i|w_{i-2},w_{i-1}) = P(w_i|w_{i-2},w_{i-1})^{1-\alpha(w_i)}$$

Avec $P(w_i|w_{i-2},w_{i-1})$ la probabilité initiale du trigramme (w_i,w_{i-2},w_{i-1}) et $\alpha(w_i)$ le score de correspondance calculé via une mesure de similarité entre hypothèses du système primaires hw_i et celles du système auxiliaire w_i. Ce score de correspondance est donné par :

$$\alpha(w_i) = \begin{cases} \frac{\phi(w_i)+\phi(w_{i-1})+\phi(w_{i-2})}{3} & \text{if } (hw_i,hw_{i-1},hw_{i-2}) = (w_i,w_{i-1},w_{i-2}) \\ \frac{\phi(w_i)+\phi(w_{i-1})}{2} & \text{if } (hw_i,hw_{i-1}) = (w_i,w_{i-1}) \\ \phi(w_i)-\gamma & \text{if } (hw_i) = (w_i) \text{ and } \phi(w_1) \geq \gamma \\ 0 & \text{if } \phi(w_1) < \gamma \end{cases}$$

Avec $\phi(w_i)$ la mesure de confiance du mot w_i et γ un seuil fixé empiriquement.

Dans (Lecouteux *et al.*, 2008), l'auteur propose une généralisation de la combinaison DDA. La généralisation consiste à guider le système primaire par un réseau de confusion de mots (WCN : Word Confusion Network) construit à partir des hypothèses de plusieurs systèmes auxiliaires. La généralisation de la combinaison par WCN n'apporte pas d'amélioration par rapport à l'utilisation de la meilleure hypothèse d'un seul système auxiliaire.

3 Décodage guidé par sac de trigrammes

Dans la formulation initiale de DDA, l'hypothèse auxiliaire est considérée comme une séquence de mots. Notre proposition est de relâcher partiellement cette contrainte de séquentialité et de représenter chaque segment de l'hypothèse auxiliaire comme un sac de trigrammes. Cette simplification est raisonnable, car les segments ont une durée de 10 secondes ce qui fait une moyenne de vingtaine de mots par segment. Ces modifications permettent une accélération du processus de combinaison et rendent l'intégration de plusieurs systèmes auxiliaires simple et efficace (Bougares *et al.*, 2011). L'architecture du décodage guidé par sac de n grammes (Bag Of NGram (BONG) driven decoding) est présentée dans la figure 1.

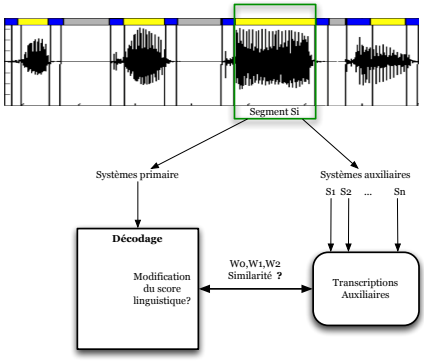

FIGURE 1 – *Décodage guidé par sac de trigrammes (BONG)*

4 Cadre expérimental

Aujourd'hui, la majorité des SRAP utilisent souvent une stratégie multi-passes avec différentes méthodes d'adaptation de modèles acoustiques et l'utilisation d'un modèle de langage plus performant lorsque l'espace de recherche est figé. L'architecture multi-passes permet la réduction de l'espace de recherche et le raffinement des modèles utilisés dans chaque passe en exploitant les sorties de la passe précédente. Cela permet aussi l'amélioration de la qualité de la transcription en utilisant des informations supplémentaires et des modèles plus complexes à chaque itération. En revanche, une telle architecture nécessite un décodage supplémentaire par passe et un temps de traitement plus important.

Contrairement à l'architecture multi-passes et à la combinaison *a posteriori*, L'objectif de la combinaison intégrée n'est pas limité à l'amélioration de la qualité de transcription, mais cherche aussi à accélérer le processus de combinaison. En effet, nous proposons une méthode permettant l'exploitation de l'ensemble de l'information qu'un système de transcription est capable de fournir avant de terminer son décodage. Bien que ces informations partielles sont incomplètes et moins précises (par exemple absence de mesure de confiance sur les mots), elles représentent un compromis entre la qualité des transcriptions et la rapidité du système.

Nos expériences seront donc limité à une seule passe de décodage, sans utilisation de mesures de confiance qui sont généralement calculées *a posteriori*. De plus, nous utilisons des sacs de n-grammes d'ordre 3 ($n = 3$).

4.1 Systèmes de reconnaissance

Trois différents systèmes de reconnaissance ont été utilisés pour tester la méthode de combinaison : le système du LIUM (Sphinx), le système du LIA (Speeral) et le système du RWTH (RASR). Le but étant d'améliorer la qualité de transcription finale, le système le plus performant a été choisi comme système primaire qui va être guidé par les autres systèmes dits auxiliaires.

Comme l'unité d'échange d'information inter-systèmes est le segment (voir section 3), et puisque l'échange doit concerner le même signal décodé par l'ensemble de systèmes, la segmentation est identique pour les trois systèmes.

4.1.1 Le système du LIUM (Sphinx)

Le système de transcription du LIUM est basé sur le système libre CMU-SPHINX amélioré et adapté à la langue française par le LIUM (Deléglise *et al.*, 2009). Durant le processus de segmentation (Meignier et Merlin, 2010), chaque segment est caractérisé par des conditions acoustiques spécifiques (parole téléphonique ou en studio, présence de parole, présence de musique, genre du locuteur, identité du locuteur...). Ces indications sont utilisées par la suite pour choisir les modèles acoustiques les plus appropriés pour décoder le segment considéré. Le processus de décodage comporte 5 passes ; les deux premières passes utilisent le décodeur Sphinx 3, tandis que Sphinx 4 est utilisé pour la suite.
Dans ce travail, seule la première passe sera utilisée, elle consiste à un décodage (beam search) avec un modèle de langage trigramme appris sur les données fournies pendant la compagne d'évaluation ESTER 2 et augmentées par le corpus Giga Word Corpus (environ 1 milliard de mots) et par des données provenant du web (80 millions de mots) pour un total d'environ 1,1 milliard de mots. Le modèle acoustique est spécialisé par bande (Large/Étroite) et par genre (Homme/Femme) pour modéliser un jeu de 35 phonèmes en contexte (triphone) avec une paramétrisation de 39 coefficients : 12 descripteurs PLP (Hermansky, 1990) plus l'énergie ainsi que leurs dérivées et dérivées secondes. En fin le lexique contient environ 122000 mots.

4.1.2 Le système du LIA (SPEERAL)

SPEERAL est un système de reconnaissance grand vocabulaire pour la parole continue (Nocera *et al.*, 2004). Le processus de décodage est basé sur un algorithme A^* avec un lexique d'environ

85.000 mots, une modélisation linguistique type n-gramme et des modèles acoustiques basés sur des Modèles de Markov Cachés (MMC) contextuels à états partagés.

Nous utilisons un modèle de langage quadri-gramme estimé sur environ 1 milliard et 200 millions de mots du journal Le Monde, sur environ 1 million de mots du corpus d'entrainement de la campagne d'évaluation ESTER-1 et ESTER-2 et sur 600 millions de mots extraits du de brèves d'informations de l'AFP (Gigaword). Les modèles acoustiques sont dépendants du genre et de la bande, ils sont entrainés sur les corpus ESTER-1 et 2 (environ 190 heures d'émission journalistique). Les paramètres acoustiques utilisés sont composés de 12 coefficients PLP plus l'énergie et leurs dérivées première et seconde, soit 39 coefficients.

4.1.3 Le système du RWTH (RASR)

Le système de transcription automatique de la parole RASR (RWTH ASR) a été développés par le groupe RWTH à l'université de Aachen (Allemagne). RASR est gratuitement téléchargeable [1] sous une licence dérivée de la Licence Publique Q (QPL).

Le système est basé sur un décodeur *Beam search*, une modélisation n-gramme du langage et des modèles de Markov cachés contextuels (triphone inter et intra-mots) à états partagés. La matrice de covariance est commune à l'ensemble des états. Une description plus détaillée du système est présente dans (Lööf *et al.*, 2007).

Contrairement aux autres systèmes, le décodage est réalisé avec un seul modèle acoustique (indépendant du genre et de la bande) appris sur les données d'ESTER-1 et d'ESTER-2 avec une paramétrisation $MFCC$ à 15 coefficients plus l'énergie et leurs dérivées première. Le lexique et le modèle de langage sont ceux utilisés dans le système Sphinx, avec une différence puisque les variantes de prononciation d'un mot dans le lexique sont équiprobables dans RASR.

Nous utilisons deux variantes du système RASR en modifiant la paramétrisation acoustique : d'abord, nous utilisons directement les 15 coefficients $MFCC$, ensuite nous concaténons les coefficients de 9 trames consécutives pour capturer l'information sur une fenêtre temporelle à plus long terme et nous appliquons une LDA (Haeb-Umbach et Ney, 1992) dessus pour obtenir un vecteur acoustique de 45 coefficients.

4.2 Corpus d'évaluation

Il est bien connu que la combinaison de systèmes donne des meilleurs résultats lorsque les systèmes utilisés ont des performances comparables, de ce fait nous avons choisi, dans un premier temps, d'évaluer notre méthode de combinaison uniquement sur la partie où les systèmes utilisés ont des performances proches.

Étant donné que RASR, contrairement aux autres systèmes, utilise un seul modèle acoustique, l'évaluation est faite sur la partie $STUDIO$ du corpus de développement de la compagne d'évaluation ESTER 2. Cette partie est plus proche du corpus d'apprentissage et son utilisation pour l'évaluation réduit l'écart entre le système RASR et les autres systèmes. Le corpus de développement contient initialement 6 heures d'émission radiophonique et la partie $STUDIO$ utilisée représente 5 heures.

1. `http://www-i6.informatik.rwth-aachen.de/rwth-asr/`

4.3 Performance du systèmes

Les données expérimentales sont initialement transcrites par les trois systèmes. Le taux d'erreur mots (WER : Word Error Rate) de chaque système est reporté dans le tableau 1. En se basant sur ces résultats, nous avons assigné à chaque système son rôle, ainsi le système Sphinx sera le système primaire. Dans la suite, chaque système utilisé sera identifié par son rôle.

Système	Rôle	WER
Sphinx	prim	**32,3** %
SPEERAL	aux_1	32,8 %
RASR-LDA	aux_2	34,1 %
RASR	aux_3	34,4 %

TABLE 1 – Taux d'erreur mots du système primaire (Sphinx) et des systèmes auxiliaires (SPEERAL, RASR et RASR-LDA)

5 Résultats

Les résultats sont présentés séparément selon le nombre de systèmes auxiliaires utilisés. Pour plus d'un système auxiliaires, l'amélioration obtenue est comparée à la combinaison ROVER.

5.1 Combinaison avec un seul système auxiliaire

Contrairement à ROVER, la combinaison par décodage guidé reste applicable même si on dispose uniquement de deux systèmes de transcription avec des sorties sans mesure de confiance. Dans un premier temps, nous utilisons séparément les sorties des systèmes auxiliaires dont on dispose pour guider le meilleur système.

Système	WER
prim	32,3 %
BONG-aux_1	**29,2** %
BONG-aux_2	29,9 %
BONG-aux_3	30,1 %

TABLE 2 – Taux d'erreur mots de système primaire et de sa combinaison avec les systèmes auxiliaires

Le meilleur gain est obtenu lorsque le meilleur système auxiliaire (aux_1) est utilisé dans la combinaison. Dans le tableau 2 l'utilisation de système aux_1 permet un gain absolu de 3,1 points de taux d'erreur. Il est intéressant de noter aussi les gains obtenus avec des systèmes auxiliaires ayant, initialement, de taux d'erreurs plus élevés de presque deux points. En effet, lorsque le système auxiliaire aux_3 (34,4 de WER) est utilisé on obtient un gain absolu de 2,2 points de WER.

5.2 Combinaison avec plusieurs systèmes auxiliaires

La généralisation de la combinaison par décodage guidé est directe ; en cas de présence de plusieurs systèmes, les hypothèses auxiliaires issues de ces systèmes sont groupées dans le même

sac de n-grammes à utiliser pendant la ré-évaluation linguistique.

En premier lieu, les systèmes auxiliaires sont utilisés par groupe de deux avec le système primaire. Les résultats de combinaison sont comparés à un ROVER entre les sorties de trois systèmes auxiliaires. Les résultats obtenus sont rapportés dans le tableau 3 :

Système	WER
Rover-prim-aux_1-aux_2	30,1 %
Rover-prim-aux_1-aux_3	**29,9** %
Rover-prim-aux_2-aux_3	30,7 %
Rover-aux_1-aux_2-aux_3	31,3 %
BONG-aux_1-aux_2	**28,7** %
BONG-aux_1-aux_3	**28,7** %
BONG-aux_2-aux_3	29,5 %

TABLE 3 – Comparaison de taux d'erreur mots de la combinaison ROVER et la combinaison BONG avec deux systèmes auxiliaires.

L'utilisation de deux systèmes auxiliaires réduit le taux d'erreur mot de 1,2 point par rapport à la meilleure combinaison ROVER. L'intégration de couple de systèmes aux_1-aux_2 et aux_1-aux_3 rapporte plus d'information par rapport à l'utilisation de couple aux_2-aux_3. Cette différence est liée au degré de complémentarité entre les systèmes, en effet les systèmes aux_2 et aux_3 utilisent le même modèle de langage et les mêmes données d'apprentissage des modèles acoustiques (voir section 4.1.3).

Nous effectuons ensuite un ROVER sur l'ensemble de systèmes en remplaçant le système primaire par la sortie de la combinaison BONG. Les résultats sont présentés dans le tableau 4.

Système	WER
ROVER-prim-aux_1-aux_2-aux_3	28,3 %
Bong_aux_1-aux_2-aux_3	28,6 %
ROVER-BONG$_{ALL}$-aux_1-aux_2-aux_3	**27,4** %

TABLE 4 – Taux d'erreur mots de la combinaison BONG et ROVER en utilisant tous les systèmes.

Bien que le passage d'une combinaison BONG avec un seul système vers une combinaison avec deux systèmes auxiliaires améliore le WER de 0,5 point (de 29,2 à 28,7), l'ajout d'un troisième système n'apporte pas un gain significatif. Si l'on dispose de quatre systèmes le ROVER donne un meilleur résultat que le BONG, ce dernier en fournissant une sortie qui permet d'améliorer le ROVER final. La combinaison BONG modifie le processus d'exploration de l'espace de recherche, ainsi le décodeur garde de chemins qui auraient été élagués sans l'intégration des hypothèses auxiliaires. De ce fait, l'intégration de la sortie de combinaison BONG dans le schéma de ROVER réduit encore le WER de 0,9 point absolu par rapport au ROVER initial.

6 Conclusion

Dans cet article, nous avons présenté une méthode de combinaison par décodage guidé en utilisant des sacs des trigrammes (BONG) issus de systèmes auxiliaires. La combinaison BONG permet une réduction du WER avec une intégration simple et efficace des transcriptions auxiliaires.

Ces sources d'informations supplémentaires modifient les décisions prises par le décodeur et donnent plus de chance aux hypothèses proposées à la fois par le système primaire et les systèmes auxiliaires. La combinaison BONG a l'avantage d'être applicable même avec un seul système auxiliaire, et en l'absence de mesure de confiance. Elle offre aussi un cadre simple pour l'intégration de nouveaux systèmes auxiliaires. La sortie de la combinaison BONG peut être utilisée pour enrichir la combinaison ROVER et réduit de 15% relatif ainsi significativement le taux d'erreur mots du meilleur système individuel avec **4,9** points absolus (15% relatifs). Actuellement, nous avons testé et comparé la combinaison BONG avec ROVER. Dans la suite, nous envisageons d'étendre la combinaison BONG vers une combinaison à la volée où tous les systèmes décodent en parallèle. Le système primaire utilise les hypothèses partielles du systèmes auxiliaires pour guider son décodage et augmenter sa performance.

Références

Bougares, F., Estève, Y., Déléglise, P. et Linarès, G. (2011). Bag Of N-Gram driven decoding for LVCSR system harnessing. *In ASRU*.

Deléglise, P., Estève, Y., Meignier, S. et Merlin, T. (2009). Improvements to the LIUM French ASR system based on CMU Sphinx : what helps to significantly reduce the word error rate ? *In Interspeech*, Brighton, UK.

Fiscus, J. (1997). A post-prcessing system to yield reduced word error rates : recogniser output voting error reduction (ROVER). *In ASRU*, pages 347–354.

Haeb-Umbach, R. et Ney, H. (1992). Linear discriminant analysis for improved large vocabulary continuous speech recognition. *In Proceedings of the 1992 IEEE international conference on Acoustics, speech and signal processing - Volume 1*, pages 13–16.

Hermansky, H. (1990). Perceptual linear predictive (plp) analysis for speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752.

Lecouteux, B., Linarès, G., Estève, Y. et Gravier, G. (2008). Generalized driven decoding for speech recognition system combination. *In ICASSP*, Las Vegas, Nevada, USA.

Lecouteux, B., Linarès, G., Estève, Y. et Mauclair, J. (2007). System combination by driven decoding. *In ICASSP*.

Lööf, J., Gollan, C., Hahn, S., Heigold, G., Hoffmeister, B., Plahl, C., Rybach, D., Schlüter, R. et Ney, H. (2007). The rwth 2007 tc-star evaluation system for european english and spanish. *In Interspeech*, Antwerp, Belgium.

Meignier, S. et Merlin, T. (2010). LIUM SpkDiarization : an open source toolkit for diarization. *In CMU SPUD Workshop*, Dallas, Texas, USA.

Nocera, P., Fredouille, C., Linares, G., Matrouf, D., Meignier, S., Bonastre, J., Massonié, D. et Béchet, F. (2004). The LIA's french broadcast news transcription system. *In SWIM : Lectures by Masters in Speech Processing*, Maui, Hawaii.

Plahl, C., Schlïter, R. et Ney, H. (2011). Improved acoustic feature combination for lvcsr by neural networks. *In Interspeech 2011*, Florence ,Italie.

Stuker, S., Fugen, C., Burger, S. et Wofel, M. (2006). Cross-system adaptation and combination for continuous speech recognition : The influence of phoneme set and acoustic front-end. *In Interspeech 2006*, Pittsburgh, USA.