Combinaison de systèmes par décodage guidé

Benjamin Lecouteux [(1)], *Georges Linarès* [(1)], *Yannick Estève* [(2)], *Guillaume Gravier* [(3)]

LIA, Avignon (France) [(1)], LIUM, Le Mans (France) [(2)], IRISA, Rennes (France) [(3)]

ABSTRACT

In this paper, we propose an integrated approach for system combination named Driven Decoding Algorithm (DDA). It consists in guiding the search algorithm of a primary ASR system by the outputs of a auxiliary systems. We first evaluate this method in simple configuration in which the primary search is driven by the one-best hypothesis of a single auxiliary system. Then, we generalize DDA to confusion-network driven decoding and we propose a general combination schemes for multiple system combination. The proposed extended DDA is evaluated using 3 ASR systems from different labs. Results show that generalized-DDA outperforms significantly ROVER method : we obtain a 15.7% relative word error rate improvement with respect to the best single system, as opposed to 8.5% with the ROVER combination.
Keywords:speech recognition, system combination, decoding strategies

1. Introduction

La combinaison de systèmes de reconnaissance de la parole a suscité un intérêt croissant ces dernières années. Diverses approches ont été proposées, opérant à différents niveaux du processus de décodage. [1, 2, 3, 4] évaluent des combinaisons strictement acoustiques, au niveau des paramètres et/ou des modèles acoustiques. Cependant, la plupart des techniques de combinaison reposent sur une ré-estimation *a posteriori* des hypothèses générées par les différents systèmes. Ce type de combinaison peut être implémenté par un vote [5] ou par fusion de réseaux de confusion [6].

Bien que l'efficacité de ces approches ait été clairement mise en évidence dans la littérature, elles comportent un certain nombre d'inconvénients. D'une part, le fonctionnement parallèle des systèmes combinés conduit à une sélection prématurée d'hyptohèses jugées improbables par un ou plusieurs systèmes. D'autre part, en opérant sur les sorties des systèmes codées sous forme de réseau de confusion, des informations relatives au décodage sont perdues, notamment les frontières de mots. Nous pouvons espérer obtenir une plus grande précision en intégrant directement l'ensemble des informations disponibles dans l'algorithme de recherche [7]. Dans cette optique, nous proposons un modèle de combinaison dans lequel un système "primaire" intègre, dans sa propre fonction de coût, un ensemble d'informations issues des systèmes de reconnaissances auxiliaires. Dans cet article, nous décrivons le principe de cet algorithme de décodage guidé par une transcription auxiliaire puis nous généralisons cette méthode d'abord en injectant des réseaux de confusion, puis en intégrant plusieurs systèmes auxiliaires au sein de l'algorithme de recherche primaire.

La section suivante présente le principe général de l'algorithme de décodage guidé par une transcription auxiliaire. Dans la seconde partie, nous étendons cette méthode au guidage par des réseaux de confusion. Dans la troisième section, nous présentons l'intégration de plusieurs systèmes par décodage guidé et nous confrontons les résultats à ceux obtenus par ROVER. Enfin nous concluons et discutons sur les améliorations possibles.

2. Décodage guidé par une one-best

Le décodage guidé (Driven Decoding Algorithm, DDA) consiste à intégrer la sortie d'un système auxiliaire au sein de l'algorithme de recherche d'un système primaire. Cette intégration repose sur deux étapes. Premièrement, l'hypothèse courante et l'hypothèse auxiliaire sont alignées en minimisant leur distance d'édition. Une fois alignées, les probabilités linguistiques sont combinées en fonction d'un score d'alignement et des probabilités *a posteriori* . Les deux prochains paragraphes expliquent en détail le DDA.

2.1. L'algorithme A de Speeral*

Le système de reconnaissance automatique du LIA Speeral [8, 9] est utilisé comme système principal. Il est basé sur l'algorithme A* opérant sur un treillis de phonèmes. Le décodage se base sur une fonction d'estimation $F(h_n)$ évaluant la probabilité de l'hypothèse sur un nœud n

$$F(h_n) = g(h_n) + p(h_n), \qquad (1)$$

où $g(h_n)$ est la probabilité de l'hypothèse partielle arrivant au nœud n et $p(h_n)$ est une sonde estimant la probabilité du nœud n à la fin du graphe. Afin de prendre en compte l'information provenant du système auxiliaire, la partie linguistique g dans (1) est modifiée en fonction de l'hypothèse auxiliaire.

2.2. Algorithme de décodage guidé

Le système de reconnaissance Speeral développe des hypothèses en explorant le treillis de phonèmes. Les meilleures hypothèses à un temps t sont étendues, en fonction de la probabilité de l'hypothèse courante ainsi que du résultat de la sonde. Dans le but de combiner les informations fournies par la transcription auxiliaire H_{aux} et l'algorithme de recherche, un point de synchronisation doit être trouvé pour chaque nœud évalué. Ces points sont trouvés en alignant dynamiquement la transcription sur l'hypothèse courante. Ceci permet d'identifier dans la transcription

auxiliaire H_{aux}, la séquence correspondant à l'hypothèse partielle h_{cur}. Cette sous-séquence notée h_{aux} est utilisée pour une nouvelle estimation du score linguistique, en se basant à la fois sur un score d'alignement $\theta(w_i)$ et sur les probabilités *a posteriori* $\phi(w_i)$ de h_{aux}. $\theta(w_i)$ est le nombre de mots correspondants entre h_{cur} et h_{aux}. Ce score est combiné aux probabilités *a posteriori* pour modifier les probabilités linguistiques suivant l'équation

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

où $L(w_i|w_{i-2}, w_{i-1})$ est le score linguistique résultant, $P(w_i|w_{i-2}, w_{i-1})$ la probabilité initiale du trigramme, β un fudge calculé empiriquement et $\alpha(w_i)$ le score de confiance de w_i donné par :

$$\text{si } \theta(w_i) > 0 \text{ alors } \alpha(w_i) = \phi(w_i) \cdot \frac{\theta(w_i)}{\gamma} \text{ et } \beta = 0.6$$
$$\text{sinon } \beta = 0$$

où γ est la taille de la fenêtre d'analyse où est calculée la distance d'édition ($\gamma = 4$) et $\phi(w_i)$ est le postérior du mot w_i du système auxiliaire.

2.3. Cadre expérimental

Le corpus d'évaluation Les expériences se basent sur le corpus français ESTER [10]. Ce corpus est composé d'heures de radio journalistiques incluant des interviews, des locuteurs non-natifs et les transcriptions associées. Les résultats sont reportés sur un ensemble de trois heures de radios différentes (F.Inter, F.Info et RFI), extraites du corpus de développement d'ESTER.

Trois systèmes de reconnaissance ont été utilisés pour tester le DDA : le système du LIA *Speeral*, le système du LIUM et le système de l'IRISA *Irene*. Le système du LIA est utilisé comme système principal, tandis que les systèmes du LIUM et de l'IRISA sont utilisés comme systèmes auxiliaires. Ces trois systèmes sont brièvement décrits dans les prochaines sections. Tous les systèmes ont été entraînés sur les même ressources issues d'ESTER, pour leurs modèles acoustiques et linguistiques. Les données d'entraînement sont composées de 80 heures de données audio annotées manuellement (environ un million de mots), et environ 200 millions de mots extraits du journal "Le Monde".

Le système de transcription du LIA Le système de transcription du LIA est basé sur le décodeur Speeral et la boîte à outils Alizé pour la segmentation. Ce système utilise des HMM contextuels et des modèles de langage trigrammes, avec un vocabulaire d'environ 65000 mots. Le système fonctionne en deux passes, la première fournissant les transcriptions intermédiaires qui sont nécessaires à une adaptation au locuteur des modèles acoustiques.

Le système de transcription du LIUM Le système de transcription du LIUM est basé sur le moteur du CMU Sphinx 3.3 [11]. Ce décodeur utilise des modèles acoustiques continus avec trois ou cinq états HMM gauche-droite. Le LIUM a également ajouté un module d'adaptation au locuteur, une ré-estimation avec des quadrigrammes sur le treillis de mots, ainsi qu'un outil de segmentation automatique [12]. La première passe utilise des modèles acoustiques dépendants du sexe et de la largeur de bande (téléphone ou non). Le modèle de langage est trigramme avec un vocabulaire de 65000 mots. La seconde passe utilise les modèles acoustiques adaptés sur la première, ainsi que le treillis de mots ré-estimé avec un modèle de langage quadrigramme [12].

	F. Inter	F. Info	RFI
LIA	21.1	22.2	**24.6**
LIUM	**18.5**	**18.9**	25.6
IRISA	21.4	21.8	25.6
DDA-IRISA-P1	19.6	19.3	23.5
DDA-IRISA-P2	18.7	18.7	22.2
DDA-LIUM-P1	17.8	18.1	22.4
DDA-LIUM-P2	**17.2**	**17.8**	**21.5**

Tab. 1: WER pour la combinaison DDA de Speeral avec le système du LIUM (*DDA-LIUM*) et celui de l'IRISA (*DDA-IRISA*) avec (*P1*) et sans (*P2*) adaptation acoustique.

Le système de transcription de l'IRISA Irene utilise des modèles acoustiques de type HMM et un modèle de langage trigramme comprenant un vocabulaire de 64000 mots. Le système fonctionne en trois étapes auxquelles s'ajoute un processus de réestimation linguistique. La première étape utilise des modèles acoustiques non-contextuels avec un modèle de langage trigramme pour générer un treillis de mots. Ce dernier est ré-évalué avec un modèle de langage quadrigramme et des modèles acoustiques contextuels. Un treillis est généré dans une troisième passe après une adaptation MLLR des modèles acoustiques sur les différents locuteurs. Finalement, un nouveau décodage est appliqué sur les 1000 meilleures hypothèses en combinant les scores acoustiques, linguistiques et morpho-syntaxiques [13].

2.4. Résultats

Les résultats sont reportés dans la table 1 pour chaque système auxiliaire combiné avec Speeral, avant et après l'adaptation des modèles acoustiques. Une stratégie en deux passes est testée après l'adaptation, basée sur la transcription du premier décodage guidé. Nous présentons également les WER (Word Error Rate) pour chaque système.

Ces résultats montrent une amélioration significative avec la combinaison des systèmes par rapport aux systèmes seuls. La combinaison avec le système du LIUM est meilleure que celle obtenue avec Irene (environ 1% de WER en absolu). Cependant, la combinaison avec le système de l'IRISA montre une réelle amélioration par rapport à la performance initiale de Speeral.

3. Décodage guidé par des réseaux de confusion

L'information utilisée par le décodage guidé et basée sur les transcriptions de systèmes auxiliaires peut sembler restreinte. Nous abordons dans cette partie l'utilisation d'une information plus complète que la transcription des systèmes auxiliaires. Nous avons étendu l'idée en intégrant les réseaux de confusion générés par les systèmes auxiliaires.

3.1. Principe

Le principe est similaire à celui utilisé avec une transcription. La combinaison s'effectue au niveau de l'algorithme d'exploration du graphe en alignant dynamiquement l'hypothèse courante avec le réseau de confusion. Ceci est effectué en minimisant la distance d'édition entre l'hypothèse et le réseau de confusion. Cependant, cette opération demandant beaucoup de

	F.Inter	F.Info	RFI
LIUM	18.5	18.9	25.6
DDA-LIUM-P1	17.8	18.1	22.4
DDA-LIUM-P2	**17.2**	**17.8**	**21.5**
DDA-WCN-LIUM-P1	17.7	18.1	22.3
DDA-WCN-LIUM-P2	**17.2**	**17.8**	**21.5**

Tab. 2: WER pour le décodage guidé par les réseaux de confusion (*DDA-WCN*). Les résultats sont comparés aux systèmes seuls (LIUM) et au décodage guidé par une one-best (*DDA-LIUM*).

	F.Inter	F.Info	RFI
LIUM	18.5	18.9	25.6
ROVER-3	17.1	18.2	22.5
2-Level DDA-ROVER	16.8	17.3	21.3
DDA-3	16.7	17.0	**20.6**
DDA-3+ROVER	**16.0**	**16.4**	20.7

Tab. 3: WER en fonction de la combinaison utilisée : la référence ROVER avec les trois systèmes (*ROVER-3*), la méthode DDA-ROVER (*2-Level DDA-ROVER*), la combinaison des 3 systèmes par DDA (DDA-3), et le ROVER intégrant le DDA (DDA-3+ROVER). Ce dernier améliore de 15.7% relatifs le WER par rapport au meilleur des systèmes initiaux (LIUM).

calculs, les chemins partiels sont sauvés et restaurés à la demande en fonction de l'historique exploré. Le temps requis par cette étape devient alors négligeable en comparaison de l'ensemble du décodage. L'étape d'alignement permet d'extraire la meilleure projection de l'hypothèse sur le réseau de confusion. Une fois cet alignement effectué, le problème est similaire à un décodage guidé par une transcription : les probabilités linguistiques sont ré-estimées en fonction d'un score d'alignement avec le réseau de confusion, ainsi qu'avec les probabilités *a posteriori* du réseau de confusion (cf. equation 2).

3.2. Résultats

Nous avons utilisé le décodage guidé par les réseaux de confusion du LIUM. Les résultats sont reportés dans la table 2. Nous observons une amélioration significative en comparaison des deux systèmes seuls (-1.5% de WER en absolu). Cependant, le gain comparé à celui obtenu avec la one-best est négligeable (environ -0.15% de WER) pour la première passe et nul après l'adaptation locuteur. Deux raisons peuvent expliquer ces faibles améliorations :

– le décodage guidé par la one-best utilise à la fois les mesures de confiance et la décision prise par le système auxiliaire. Ce dernier guidant la recherche parmis les meilleures hypothèses, c'est probablement une stratégie qui écarte les mauvaises hypothèses et qui rend la combinaison plus robuste.
– les mesures de confiance utilisées dans la one-best sont plus fiables que les probabilités *a posteriori* utlisées dans le réseau de confusion. Ceci est en particulier dû au fait qu'ils ont été ré-estimés avec un modèle de language quadrigramme. Étant donné que le score de confiance est crucial pour ré-estimer la partie linguistique, ce point peut impacter significativement les résultats finaux.

4. Combinaison de plusieurs systèmes

Jusqu'ici, le DDA est utilisé avec un seul système auxiliaire. Dans cette section nous proposons une extension qui permet de généraliser le DDA à plusieurs systèmes auxiliaires.

4.1. Principe

En reprenant le principe général de la combinaison par décodage guidé, combiner plusieurs systèmes peut s'envisager de deux manières.

La première consiste à fusionner l'ensemble des hypothèses auxiliaires en utilisant un vote basé sur un ROVER. L'hypothèse résultante guide alors le décodage.

La seconde consiste à conserver indépendamment les hypothèses auxiliaires. Ces hypothèses sont alors intégrées, chacune séparément au sein du décodage guidé. Dans cette approche, chaque transcription auxiliaire est synchronisée avec l'hypothèse courante du décodeur. Les scores d'alignement sont indépendants, tout comme les probabilités *a posteriori* relatives aux transcriptions auxiliaires. Finalement, l'ensemble des scores est combiné pour ré-estimer la partie linguistique : aucune information n'est perdue.

Nous avons testé et comparé ces deux approches sur trois configurations différentes. Finalement nous testons une dernière approche, où tous les systèmes ainsi que la sortie du DDA sont combinés par un ROVER.

4.2. Combinaison ROVER-DDA

Dans cette approche nous fusionnons, dans une première étape, l'ensemble des transcriptions auxiliaires. Nous avons utilisé ROVER pour fusionner les systèmes du LIUM et de l'IRISA. Les scores de confiance finaux sont composés de la moyenne des scores de chacun des systèmes. La transcription obtenue est alors utilisée comme hypothèse auxiliaire, de la même façon qu'un simple décodage guidé.

4.3. Combinaison intégrée

Dans cette méthode toutes les transcriptions auxiliaires sont soumises indépendamment à l'algorithme de recherche. Un score d'alignement est calculé pour chacune et les scores linguistiques sont mergés via une combinaison log-linéaire étendue à n systèmes :

$$L(w_i|w_{i-2},w_{i-1}) = P(w_i|w_{i-2},w_{i-1})^{1-\beta}.\frac{1}{N}\sum_{k=0}^{N}\alpha_k(w_i)^{\beta_k}$$

où β est la moyenne des β_k comme défini dans l'équation 2, α_k sont les probabilités *a posteriori* fournies par les autres systèmes k et N est le nombre de systèmes auxiliaires.

4.4. Résultats

La table 3 compare les résultats obtenus par les différentes stratégies. Nous observons que l'addition d'un troisième système améliore systèmatiquement les performances. Cependant, le ROVER des 3 systèmes obtient un résultat identique à celui de la meilleure combinaison de deux systèmes (-0.2% de WER en absolu). La méthode en deux passes permet d'obtenir une baisse du WER plus significative (1.1% de mieux que le DDA-LIUM), mais cette approche reste cependant moins performante que la combinaison intégrant directement les trois systèmes (un gain de 0.4% de WER).

La dernière méthode de combinaison consiste à fusionner toutes les sorties disponibles (celle du DDA comprise). Cette méthode améliore encore légèrement

	F.Inter	F.Info	RFI
DDA-3	16.7	17.0	20.6
ORACLE-3	10.3	10.5	14.5
DDA-3+ROVER	16.0	16.4	20.7
ORACLE DDA+ROVER	9.8	10.0	13.6

Tab. 4: Comparaisons entre le DDA, l'Oracle et le Rover.

le système d'environ 0.3% de WER absolu.

Au final, notre meilleure configuration de combinaison permet d'améliorer le meilleur des systèmes d'environ 3.3% de WER en absolu, bien plus que la combinaison ROVER classique (-1.6% de WER absolu).

Les résultats obtenus confirment l'idée qu'une information auxiliaire peut être intégrée au sein de l'algorithme de recherche, permettant aussitôt que possible d'exploiter l'information disponible.

4.5. Analyse du décodage guidé

Afin de compléter notre analyse de la combinaison par décodage guidé nous avons réalisé quelques expériences supplémentaires.

Nous avons essayé de savoir si le DDA permet de trouver des hypothèses qui ne sont présentes dans aucune des hypothèses primaires. Ceci a été réalisé en comparant les résultats obtenus à ceux d'un Oracle (*ORACLE DDA+ROVER*) entre tous les systèmes (*ORACLE-3*). Les résultats reportés dans la table 4 montrent que ré-estimer les scores linguistiques permet de guider la recherche sur d'autres chemins. Ceci confirme que le DDA n'est pas seulement un vote en ligne, mais une approche intégrée apportant une information supplémentaire à la fonction de coût qui explore le graphe.

De plus, il est important de noter qu'une combinaison ROVER du DDA avec tous les autres systèmes améliore encore le résultat du DDA. Ceci montre que le DDA trouve de nouveaux chemins corrects, mais aussi en supprime certains qui étaient présents dans les systèmes initiaux. Cette constatation suggère qu'il est encore possible d'améliorer le DDA pour qu'il prenne systèmatiquement les bonnes hypothèses trouvées dans les systèmes auxiliaires.

5. Conclusion

Dans cet article, nous avons proposé une approche intégrée pour la combinaison de systèmes de reconnaissance de la parole. Le modèle de combinaison proposé est fondé sur l'intégration, dans le moteur de reconnaissance d'un système primaire, des sorties de systèmes auxiliaires. Différentes configurations ont été évaluées sur la base ESTER-2005. Les résultats montrent que cette approche permet une réduction très sensible du taux d'erreur mots. Par ailleur, nos expériences montrent que le décodage guidé par la meilleure hypothèse auxiliaire obtient de meilleurs résultats que le guidage par réseau de confusion. Enfin, l'intégration de plusieurs systèmes auxiliaires (au lieu d'un seul) apporte un gain additionnel très substantiel et dépasse significativement la combinaison ROVER des 3 systèmes. Finalement en utilisant le DDA avec une dernière passe en ROVER nous obtenons un gain global d'environ 3.3% de WER (15.7% relatifs) par rapport au meilleur des systèmes initiaux.

Références

[1] L. Barrault, C. Servan, D. Matrouf, G. Linarès, and R. De Mori, "Frame-based acoustic feature integration for speech understanding," in *International Conference on Acoustic, Speech and Signal Processing (ICASSP'08)*, 2008.

[2] B. Hoffmeister, T. Klein, R. Schluter, and H. Ney, "Frame based system combination and a comparison with weighted ROVER and CNC," in *International Conference on Spoken Language Processing, Interspeech*, 2006, pp. 537–540.

[3] O. Siohan, B. Ramabhadran, and B. Kingsbury, "Constructing ensembles of ASR systems using randomized decision trees," in *IEEE International Conference on Acoustics, Speech and Language Processing*, Philadelphia, PA, March 2005, vol. 1, pp. 197–200.

[4] R. Prasad, S. Matsoukas, C.-L. Kao, J.Z. Ma, D.-X. Xu, T. Colthurst, O. Kimball, R. Schwartz, J.L. Gauvain, L. Lamel, H. Schwenk, G. Adda, and F. Lefevre, "The 2004 BBN/LIMSI 20xRT English Conversational Telephone Speech Recognition System," in *InterSpeech 2005*, Lisbon, 2005.

[5] J.M Fiscus, "A post processing system to yield reduced word error rates : Recognizer output voting error reduction (rover)," in *IEEE ASRU Workshop*, 1997, pp. 347–352.

[6] G. Evermann and 2000. P. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *NIST Speech Transcription Workshop*, 2000.

[7] I-Fan Chen and Lin-Shan Lee, "A new framework for system combination based on integrated hypothesis space," in *Interspeech'06-ICSLP*, Pittburgh, Pensylvania, USA, 2006.

[8] Pascal Nocéra, Georges Linarès, and Dominique Massonié, "Phoneme lattice based a* search algorithm for speech recognition," *Text, Speech and Dialogue : 5th International Conference, TSD 2002, Brno, Czech Republic*, 2002.

[9] Pascal Nocéra, Georges Linarès, and Dominique Massonié, "Principes et performances du décodeur parole continue speeral," *XXIVées journées d'étude sur la parole*, 2002.

[10] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News," in *Interspeech'05-Eurospeech*, Lisbon, Portugal, 2005.

[11] K. Seymore, C. Stanley, S. Doh, M. Eskenazi, E. Gouvea, B. Raj, M. Ravishankar, R. Rosenfeld, M.A. Siegler, R. Stern, and E. Thayer, "The 1997 CMU Sphinx-3 english broadcast news transcription system," in *DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, USA, February 1998.

[12] P. Deléglise, Y. Estève, S. Meignier, and T. Merlin, "The LIUM speech transcription system : a CMU Sphinx III-based system for french broadcast news," in *Interspeech'05-Eurospeech*, Lisbon, Portugal, September 2005.

[13] G. Gravier S. Huet and P. Sébillot, "Morpho-syntactic processing of N-best lists for improved recognition and confidence measure computation," in *European Conf. on Speech Communication and Technology – Interspeech*, 2007.