Décodage interactif de la parole

Grégory Senay, Georges Linarès, Benjamin Lecouteux, Stanislas Oger, Thierry Michel

Décodage interactif de la parole

Grégory Senay, Georges Linarès, Benjamin Lecouteux, Stanislas Oger, Thierry Michel

Laboratoire Informatique d'Avignon,
CERI, Université d'Avignon et des Pays de Vaucluse,
Xtensive Technologies
{gregory.senay, georges.linares, benjamin.lecouteux, stanislas.oger}@univ-avignon.fr
thierry.michel@xtensive.com

ABSTRACT

Speech recognition technology suffers from a lack of robustness which limits its usability for fully automated speech-to-text transcription, and manual correction is generally required to obtain perfect transcripts. In this paper, we propose a general scheme for semi-automatic transcription, in which the system and the transcriptionist contribute jointly to the speech transcription. In order to reduce the correction time, we evaluate various strategies aiming to guide the transcriptionist towards the critical areas of transcripts. These strategies are based on graph density-based criterion and two semantic consistency criterion ; using a corpus-based method and a web-search engine. Results show semantic information must be integrated into the interactive decoding process.

Keywords: semi-automatic transcription, speech recognition, transcription aids, speech understanding

1. Introduction

Malgré les efforts réalisés par la communauté scientifique ces vingts dernières années, les performances des systèmes de reconnaissance automatique de la parole (RAP) sont étroitement liées à différents paramètres, notamment les conditions acoustiques, la couverture lexicale et, plus généralement, l'adéquation entre le corpus d'apprentissage et les conditions d'utilisation. Dans un contexte réel d'utilisation, le manque de robustesse des systèmes conduit à réaliser une correction manuelle des transcriptions. Le coût de la transcription doit donc intégrer cette étape de post-traitement manuel.

Dans cet article nous proposons un processus de RAP dans lequel système et opérateur collaborent à l'élaboration d'une transcription conforme aux besoins de l'application.

Plusieurs études ont évalué l'intérêt d'utiliser un système de RAP pour améliorer la productivité des transcripteurs [1]. La plupart d'entre elles propose un processus séquentiel se décomposant en deux étapes, la première consiste à utiliser un système de RAP pour produire une hypothèse de transcription (éventuellement un réseau de confusion), et la seconde consiste à corriger manuellement ces hypothèses. Les résultats montrent un gain assez variable : le temps moyen de correction est réduit de 20 à 80%, en fonction des performances du système de RAP initial.

Ici, nous proposons d'utiliser un processus interactif dans lequel le système et l'opérateur collaborent à l'écriture des transcriptions. Cette interactivité repose sur l'exploitation par le système de RAP des actions correctives de l'opérateur, ce qui lui permet de produire automatiquement une meilleure hypothèse, qui sera à son tour corrigée. Dans cette boucle du processus de correction, la position dans la phrase où les corrections sont appliquées peut être crutiale pour l'efficacité du redécodage ; nous proposons différentes stratégies de guidage du transcripteur basées sur des mesures de confiance et des critères de consistance sémantique.

La section suivante décrit l'algorithme de décodage intéractif que nous proposons. Les differéntes stratégies de guidage du correcteur sont ensuite détaillées, puis évaluées dans le section 4. La dernière section propose un bilan général de l'approche proposée et présente quelques perspectives.

2. Décodage intéractif

2.1. Principe

La meilleure hypothèse de reconnaissance ne constitue qu'une partie de l'information que le système de RAP possède sur les données à transcrire. En effet, le système évalue généralement un grand nombre d'hypothèses concurrentes. Proposer des alternatives au correcteur pourrait améliorer son efficacité [9], par exemple en lui évitant la saisie de certains mots. Cette approche présente néanmoins un certain nombre de difficultés, en particulier la grande diversité de variantes possibles qui de plus ne diffèrent souvent que de quelques mots [4]. Comme proposé dans [2], nous utilisons une représentation basée sur des réseaux de confusion (RC), qui sont plus compacts que les treillis de mots, et de fait plus *lisibles*. Avec une telle représentation, les actions correctives sont réduites à de simples actions d'édition du réseau : sélection d'un mot, suppression ou ajout d'une alternative manquante.

Chaque action corrective effectuée sur le réseau de confusion est suivie d'un redécodage de celui-ci, guidé par l'historique des corrections de l'utilisateur. L'objectif de ce redécodage contraint par les corrections est d'améliorer la transcription, en particulier au voisinage des corrections dont le contexte linguistique se trouve changé. Cette modification locale peut de plus changer globalement le segment de transcription, car

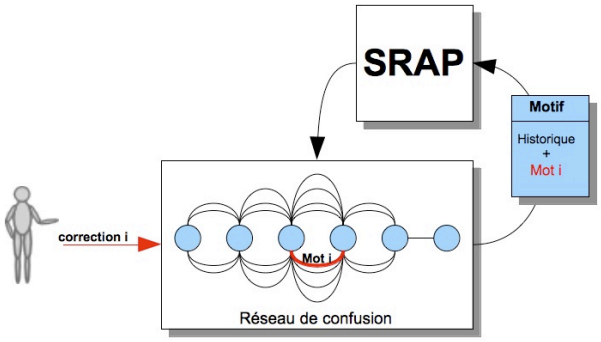

Fig. 1: Decodage interactif

la contrainte locale peut induire des modifications qui se propagent.

2.2. Décodage contraint par un motif

Le principe général du décodage contraint est de fournir au décodeur un motif de phrase, dans lequel les corrections effectuées par l'utilisateur (cf Figure 1) apparaissent comme des mots figés et les zones à redécoder comme des jokers que le système doit trouver. Techniquement, le décodage contraint est implémenté par l'algorithme de décodage guidé que nous avons présenté dans des travaux précédents, en particulier pour l'alignement de transcriptions imparfaites [6] ou la combinaison de systèmes [7]. Nous proposons donc un décodage guidé par le motif issu des corrections appliquées par le correcteur au RC.

Considérant le scénario de correction incrémentale des transcriptions semi-automatiques, on peut penser que l'ordre et les zones dans lesquels les corrections sont appliquées sont suceptibles de modifier sensiblement l'efficacité du décodage. Nous allons évaluer différentes méthodes visant à guider le correcteur vers les zones qui sont supposées être les plus profitables au système en terme de réduction du taux d'erreur mot (TEM) pour chaque acte correctif.

3. Stratégies d'orientation du correcteur

Nous proposons d'exploiter les capacités d'autodiagnostic d'un système pour orienter le correcteur vers les zones qui sont susceptibles d'être utiles au redécodage. La stratégie la plus naturelle est une progression de gauche à droite ; suivant le sens de lecture. On peut aussi penser que les corrections des zones massivement erronées seront particulièrement utiles au système. Une dernière stratégie consiste à estimer les zones sémantiquement incohérentes, dont la correction permet d'améliorer l'intelligibilité de la transcription. De plus, cette approche dans le contexte d'une application de compréhension ou d'indexation, privilégie la qualité sémantique à la classique mesure du TEM.

3.1. Utilisation de la densité du graphe de mots

Les mesures de confiance ont pour but d'estimer la probabilité qu'un mot de la transcription soit correct. Notre objectif est d'utiliser ces mesures pour orienter la correction, afin de maximiser le gain en TEM pour chaque action corrective. La largeur du graphe à un instant t a souvent été utilisée comme un indicateur pour l'estimation des scores de confiances [5], une "explosion" de la largeur étant caractéristique d'une situation d'incertitude de l'algorithme.

En s'inspirant de cette mesure, nous évaluons une stratégie dans laquelle le correcteur est orienté prioritairement vers les zones les plus développées du graphe de mots. Le système sélectionne automatiquement la zone la plus dense du graphe, et propose au transcripteur de la corriger. Chaque action corrective produit un motif de la phrase qui est réinjecté dans le système de reconnaissance via un décodage guidé. L'efficacité de cette méthode est estimée en calculant le gain en TEM issu de cette dernière phase de décodage.

3.2. Utilisation de la consistance sémantique sur un corpus fermé

Cette méthode a pour but d'estimer la consistance sémantique de la transcription par l'utilisation d'un très grand corpus de nouvelles journalistiques. Le principe général est de considérer qu'un segment est inconsistant s'il est significativement distant de la dépêche du corpus la plus similaire. Une fois la plus proche dépêche trouvée, la transcription est découpée en fenêtres plus courtes (10 mots pertinents en moyenne) ; pour chacunes d'entres elles, leurs consistances sémantiques sont évaluées et le correcteur est dirigé en premier vers la fenêtre de moindre consistance.

La similarité entre le segment de transcription et les dépêches repose sur une mesure *Cosinus* [10].Afin de prendre en compte uniquement des mots significatifs, la transcription et le corpus sont lemmatisés et filtrés via une stop-liste contenant les mots les plus fréquents du corpus. Les scores obtenus sont utilisés pour guider le correcteur vers la fenêtre du segment ayant le score le plus faible.

Les données utilisées pour nos expériences sont issues du corpus Français Gigaword qui est une archive de données journalistiques acquise pendant plusieurs années par le Linguistic Data Consortium (LDC). Le corpus a été collecté à partir de deux sources internationales distinctes. La première est l' *Agence France-Presse*, qui fournit des données journalistiques entre 1994 et 2006. Cette première partie contient plus de 480000 mots différents. La seconde vient de l'*Associated Press French Service*, couvrant la même période et contenant environ 180000 mots différents. Tout ce contenu est écrit dans un style journalistique : les dépêches sont relativement courtes ; avec en moyenne une quinzaine de mots par phrase. Les documents sont structurés en paragraphes. Chacun d'entre eux correspond à une dépêche, se focalisant sur un sujet précis. Le corpus contient environ 2 millions de phrases et 250 millions de mots.

3.3. Utilisation de la consistance sémantique à l'aide du Web

L'idée de cette approche est de détecter des ruptures sémantiques dans une phrase en estimant à quel point les mots porteurs de sens sont cohérents entre eux dans leur contexte. Nous considérons que les mots porteurs de sens sont ceux différents des mots-outils de la langue. Nous proposons d'utiliser le Web car il offre une couverture très large de la langue. Le Web est alors utilisé comme une très grande base de documents dans laquelle chacun est vu comme un sac de mots proches d'un point de vue sémantique.

Pour mesurer à quel point un mot porteur de sens est cohérent avec son contexte, et donc à quel point il apparaît dans ce contexte sur le Web, nous proposons d'utiliser la probabilité d'apparition d'un mot dans un document Web, sachant que les mots de son contexte y apparaissent. Cette probabilité est formalisée dans l'équation 1, pour un mot w_i et une probabilité d'ordre n, et donc avec un contexte gauche de taille $n-1$, noté $\psi_i^n = w_{i-n-1}, \ldots, w_{i-1}$:

$$P_s(w_i|\psi_i^n) = \frac{WC(\psi_i^n, w_i)}{WC(\psi_i^n)} \quad (1)$$

Avec $WC(\psi_i^n, w_i)$ le nombre de documents dans lesquels les mots ψ_i^n et w_i apparaissent quelque soit leur position dans le document. Comme avec un modèle n-gramme classique, lorsqu'un mot n'apparaît dans aucun document web où apparaissent les mots de son contexte, on se replie sur la probabilité obtenue avec un contexte plus court, pondérée par un coefficient de repli déterminé empiriquement :

$$\hat{P}_s(w_i|\psi_i^n) = \begin{cases} P_s(w_i|\psi_i^n), \text{ si } WC(\psi_i^n, w_i) > 0 \\ \alpha \cdot P_s(w_i|\psi_i^{n-1}), \text{ sinon} \end{cases} \quad (2)$$

La mesure de cohésion sémantique d'ordre n de la phrase est donc :

$$SC(w_1 \ldots w_i) = P_s(w_2|w_1) \times P_s(w_3|w_1, w_2) \times P_s(w_i|\psi_i^n) \quad (3)$$

Les résultats expérimentaux obtenus avec cette approche sont présentés dans la section 4.2.

4. Expériences

Le système de RAP utilisé est SPEERAL [8], développé au *Laboratoire Informatique d'Avignon*, et la manipulation des RC et des treillis de mots est réalisée avec la boite à outils SRILM [11] développée par l'entreprise *SRI International*. Les expériences sont conduites sur les données de développement de la campagne d'évaluation ESTER 2005 [3], qui est composée de 8 heures de journaux d'informations français provenant de 4 radios différentes. Le décodage initial des données est obtenu avec le système rapide en deux fois le temps réel, sans seconde passe comprenant l'adaptation des modèles acoustiques au locuteur. Dans ces conditions de décodage, le TEM est de 32,6% sur l'ensemble du corpus.

4.1. Protocole

La correction Gauche-Droite, sans décodage réactif, est simulée en utilisant un alignement fourni par l'outil de mesure de la campagne ESTER : *Sclite*[1]. Chaque erreur rencontrée est marquée par *Sclite* comme insertion, suppression ou substitution. L'action corrective est appliquée sur l'hypothèse en fonction de l'indication fournie par cet outil. Les corrections sont effectuées suivant les différentes stratégies de guidage de la correction : Gauche-Droite (GD-ID), densité de graphe (DG-ID), correction sémantique par l'approche basée sur le corpus (Corp-ID) et par l'approche basée sur le Web (Web-ID).

Nous évaluons, dans la partie 4.2, les performances en mesurant la réduction du TEM après chaque action corrective faite dans un segment. Le nombre de corrections effectuées dans un segment est limité à 20. Afin de mettre en valeur les performances de notre approche dans différentes situations, les segments sont séparés en deux classes : la première contient ceux dont le TEM du décodage initial est inférieur ou égal à 40% et la seconde ceux dont le TEM du décodage initial est supérieur à 40%.

4.2. Résultats

Tab. 1: *TEM selon le nombre d'actions correctives, pour les segments dont la transcription initiale est ≤ à 40% de TEM*

#c	1	3	10	20
Manuelle	25.22	22.98	17.23	9.44
GD-ID	24.28	**20.82**	**11.88**	**5.26**
DG-ID	26.58	25.38	16.62	11.76
Corp-ID	**23.90**	21.15	13.93	8.51
Web-ID	24.33	21.10	12.21	7.40

Tab. 2: *TEM selon le nombre d'actions correctives, pour les segments dont la transcription initiale est > à 40% de TEM*

#c	1	3	10	20
Manuelle	55.91	54.05	47.81	40.14
GD-ID	54.95	49.77	37.71	**25.36**
DG-ID	57.51	53.52	44.05	36.99
Corp-ID	54.19	49.37	39.06	29.54
Web-ID	**51.88**	**48.32**	**37.49**	29.49

Dans les tableaux 1 et 2, nous présentons les résultats obtenus en terme de TEM, pour deux classes de segments dont chacun a subi un, trois, dix et vingt actes correctifs (#c). Ces deux classes représentent respectivement 46% et 54% du corpus. Nous constatons que le décodage réactif améliore le TEM dans toutes les configurations par rapport à la correction manuelle (qui pour rappel est sans redécodage).

Pour la première classe (TEM ≤ 40%), la comparai-

[1] Vous pouvez trouver la dernière version de l'outil à cette adresse : http ://www.itl.nist.gov/iad/mig/tools/

son entre les différentes stratégies guidées montre que la correction par densité de graphe (DG-ID) est plutôt inefficace, son rendement étant nettement moins bon que la méthode classique Gauche-Droite (Manuelle - sans décodage interactif). Les stratégies basées sur la sémantique obtiennent de meilleurs résultats, plus particulierement l'approche Web (Web-ID). Néanmoins, elles restent légèrement moins performantes que le décodage interactif utilisant une méthode Gauche-Droite (GD-ID), qui serait probablement plus confortable pour le correcteur.

Le tableau 2 présente les résultats obtenus sur les segments massivement erronés (TEM $>$ 40%). Les résultats obtenus par le guidage sémantique du correcteur sont, dans ce cas, sensiblement meilleurs que toutes les autres stratégies guidées lors des 10 premières corrections. En particulier, la méthode basée sur la consistence sémantique à l'aide du Web (Web-ID) qui procure un gain très significatif dès la première correction (-3,07% absolus). On peut noter que cette méthode est la plus efficace pour 10 corrections, le gain est alors de 10,32% absolus comparativement à la méthode manuelle.

Globalement, l'approche basée sur le Web obtient de meilleurs résultats que la méthode basée sur le corpus, en dépit du fait que les données d'ESTER (journaux radiophoniques) sont étroitement liées au corpus que nous utilisons, composé de dépêches d'informations. Les bénéfices de l'approche Web pourraient être encore plus importants sur les tâches qui se sont pas couvertes par le corpus.

5. Conclusion et perspectives

Nous avons présenté et évalué une approche intéractive du décodage de la parole qui vise à minimiser le coût global de la transcription. L'idée principale est d'alterner des phases de correction et de transcription automatique qui prennent en compte les corrections de l'utilisateur. Considérant que la correction d'une zone bien précise de la transcription pouvait être particulièrement rentable durant un décodage réactif, nous avons proposé différentes stratégies afin d'orienter le correcteur vers les zones "critiques" de la transcription.

Les résultats montrent que le décodage interactif apporte une amélioration notable dans l'efficacité de la correction, comparé à une correction uniquement manuelle. De plus, la comparaison entre les différentes stratégies d'orientation du correcteur permet de tirer plusieurs conclusions. Focaliser les actions correctives sur les zones denses du graphe de recherche est peu efficace. Une des raisons de ce résultat décevant est qu'un système en situation d'échec majeur ne peut pas se raccrocher à un faible nombre de corrections pour se remettre sur le chemin d'un décodage réussi. L'intégration d'un mot correct dans une zone dans laquelle le système hésite n'a pas l'efficacité d'une correction par bloc telle qu'elle est réalisée lors d'une correction gauche-droite. Les stratégies de guidage sémantique n'améliorent pas clairement les parties du corpus où le taux d'erreur mot est bas, mais elles deviennent très efficaces lorsque la transcription est de mauvaise qualité.

Ces résultats ouvrent des perspectives intéressantes dans des contextes d'indexation par le contenu ou plus généralement, d'interprétation ou d'analyse de la parole. Nous envisageons une stratégie permettant d'identifier et de corriger uniquement les parties de la transcription mal décodées afin de les rendre correctement indexables ou interprétables. C'est dans cette perspective que nous développerons ce travail.

Références

[1] Thierry Bazillon, Yannick Estève, and Daniel Luzzati. Manual vs assisted transcription of prepared and spontaneous speech. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA).

[2] G. Riccardi D. Falavigna, R. Gretter. Acoustic and word lattice based algorithms for confidence scores. pages 1621–1624, 2002.

[3] 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 *European Conference on Speech Communication and Technology, Interspeech*, Lisbon, Portugal, 2005.

[4] J.Ogata and M.Goto. Speech repair : Quick error correction just by using selection operation for speech input interfaces. In *International Conference on Speech Communication and Technology, Interspeech*, pages 133–136, Lisboa, Portugal, 2005.

[5] Thomas Kemp and Thomas Schaaf. Estimating confidence using word lattices. In *Proc. Eurospeech '97*, pages 827–830, Rhodes, Greece, 1997.

[6] Benjamin Lecouteux, Georges Linarès, J.F. Bonastre, and Pascal Nocera. Imperfect transcript driven speech recognition. In *Interspeech'06-ICSLP*, Pittburgh, Pensylvania, USA, 2006.

[7] Benjamin Lecouteux, Georges Linarès, Yannick Estève, and Guillaume Gravier. Generalized driven decoding for speech recognition system combination. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Las Vegas, USA, 2008.

[8] Georges Linarès, Dominique Massonié, Pascal Nocera, and Christophe Lévy. The lia speech recognition system : from 10xrt to 1xrt, 2007.

[9] Hiroaki Nanjo, Yuya Akita, and Tatsuya Kawahara. Computer assisted speech transcription system for efficient speech archive. In *Western Pacific Acoustic conference*, Seoul, Korea, 2006.

[10] C. Van Rijsbergen. Information retrieval. Newton, MA, USA, 1979. Butterworth-Heinemann.

[11] A. Stolcke. SRILM-an extensible language modeling toolkit. In *Proceedings of International Conference on Spoken Language Processing*, volume 2, pages 901–904, Denver, Colorado, USA, 2002.