Reconnaissance de la parole guidée par des transcriptions approchées

Benjamin Lecouteux, Georges Linarès, Pascal Nocera, Jean-François Bonastre

Laboratoire Informatique d'Avignon
339 chemin des Meinajaries, B.P 1228, 84911 Avignon, France
{benjamin.lecouteux, georges.linares, pascal.nocera, jean-francois.bonastre}@univ-avignon.fr

ABSTRACT

In many cases, an approximated transcript can be associated to speech signal : movies subtitles, scenario and theatre, summaries and radio broadcast. These transcripts correspond rarely to the exact word utterances. The goal of this work is to use these information to improve the performance of an automatic speech recognition (ASR) system with integration of information resulting from the transcripts. In this paper we use the partial transcript in order both to adapt the language model and to rescore the ASR word hypothesis when the partial transcript matches the input signal. Multiple applications are possible : to help deaf people to follow a play with closed caption aligned to the voice signal (with respect to performer variations), to watch a movie in another language using aligned closed captions, to transcript in real time debates or meetings.

1. INTRODUCTION

Dans certaines situations, des sources d'informations externes au signal de parole à transcrire peuvent être disponibles : émissions comportant des sous-titres ou pour lesquelles un résumé peut être mis à disposition, scénario pour le cinéma, scripts de pièces de théâtre, prompteur d'un journaliste... Ces informations peuvent être exploitées pour améliorer les performances d'un système de reconnaissance automatique de la parole (SRAP).

Ce problème est abordé dans le domaine de l'alignement automatique de textes sur des flux audio. Cependant, dès que les informations disponibles s'éloignent de ce qui est réellement dit, un simple alignement forcé entre le signal audio et le texte devient insuffisant : Il est alors nécessaire de retrouver le contenu du message par une transcription automatique de celui-ci, tout en tirant profit des informations évoquées précédemment.

L'utilisation de textes approchés pour le décodage audio a déjà été étudiée dans le cadre de l'indexation audio/vidéo automatisée ([2]) ou de la ré-estimation de modèles acoustiques ([4],[5]) à partir de données non transcrites. Cette tâche est présentée dans la littérature comme un problème de synchronisation de texte sur le flux audio, ou, plus rarement, comme un problème de correction de transcriptions imparfaites. Dans les deux cas, la principale difficulté de la tâche tient à la qualité, souvent approximative, des transcriptions : Placeway et Lafferty mesurent un taux de différence entre les sous-titres d'un film et la transcription exacte compris entre 10% et 20% [9]. Ces divergences entre la transcription et le message réel augmentent considérablement la difficulté de l'alignement.

Dans ce papier nous présentons tout d'abord les problèmes liés à l'alignement d'un texte approché sur un flux de parole. Puis nous décrivons notre solution pour la reconnaissance de la parole guidée par une transcription approchée. Cette solution est basée sur un décodeur à pile asynchrone permettant l'intégration des informations issues des transcriptions approchées. Enfin, les expériences menées et les résultats obtenus sont détaillés avant de présenter quelques conclusions et perspectives.

2. ÉTAT DE L'ART

2.1. Alignement forcé avec transcription exacte

Le sujet de l'alignement sur transcription exacte est abordé par Moreno et Joerg pour aligner de longs documents audio avec leur transcription dans le cadre d'une indexation automatique de documents multimédias [7]. Moreno et Joerg proposent une méthode basée sur la recherche de zones bien synchronisées, appelées *îlots de confiance* [7]. Dans un premier temps, un modèle de langage est estimé sur la transcription exacte. Une première passe isole des zones avec une forte correspondance entre transcription *a priori* et transcription automatique correspondent. Le document est alors segmenté par ces îlots de confiance ; sur chaque segment, un modèle de langage spécifique est estimé. L'algorithme est lancé récursivement sur chaque partie non alignée jusqu'à convergence. Cette méthode, restreinte aux transcriptions exactes, obtient d'excellents résultats : 99% des mots sont correctement alignés.

2.2. Alignement de transcriptions approchées

Le problème du traitement automatique de transcriptions approchées a été abordé par Placeway et Lafferty qui ont expérimenté l'exploitation de sous-titres avec un décodeur synchrone (SPHINX-3) [9]. Leurs expériences portent sur une base de données de journaux diffusés en anglais. Ils proposent d'utiliser les sous-titres en estimant un modèle de langage sur ces derniers puis en alignant le flux audio sur ces sous-titres.

Pour combiner l'information des sous-titres avec les modèles du décodeur, Placeway et Lafferty ont interpolé un modèle de langage générique avec un modèle estimé sur les sous-titres [9]. Ce modèle interpolé est ensuite utilisé par le SRAP. Leurs expérimentations sont menées avec des sous-titres comportant 9.7% d'erreurs par rapport à la transcription exacte. Cette technique améliore les performances du décodage de 15% relatifs de WER (Word Error Rate) par rapport au décodage initial (de 55.8% à 47.2%).

Par ailleurs, ils proposent d'intégrer un mécanisme d'alignement sur les sous-titres, en plus de l'interpolation des modèles : au fur et à mesure que le décodeur avance et propose sa liste de mots candidats, les mots correspondants à la transcription approchée sont favorisés dans le faisceau d'hypothèses. Cette méthode apporte un gain relatif de 37% WER comparé au décodage initial en ramenant le WER à 35%. Le résultat final reste cependant nettement inférieur à la qualité de la transcription approchée fournie au système (9.7% de WER).

3. RECONNAISSANCE DE LA PAROLE PAR DES TRANSCRIPTIONS APPROCHÉES

Notre objectif est d'exploiter des transcriptions approchées avec un décodeur asynchrone basé sur l'algorithme A*, dans le cadre d'un système de broadcast news en langue française. Nous présentons les particularités de ce type de décodeur et la solution pour intégrer l'information contenue dans les transcriptions imparfaites.

Deux méthodes exploitant la transcription approchée ont été expérimentées. La première consiste à combiner un modèle de langage générique et un modèle de langage estimé sur la transcription approchée. La seconde propose d'intégrer un algorithme d'alignement temporel au sein de l'algorithme A*, en influençant directement la fonction d'estimation de l'agorithme de recherche.

3.1. Adaptation des modèles de langage

Dans notre approche, la variabilité linguistique peut être réduite grâce à l'apport d'une transcription exacte ou approchée du discours. Un gain peut être obtenu en réduisant globalement l'espace linguistique, par estimation d'un modèle de langage sur la transcription elle même. Cependant, ce modèle de langage ne suffit pas lorsque le locuteur s'éloigne de la transcription. Un modèle de langage générique est donc interpolé au modèle de langage réduit issu de la transcription approchée.

3.2. Décodeur asynchrone et algorithme d'alignement

Le LIA a développé un système de reconnaissance de la parole grand vocabulaire et parole continue nommé SPEERAL [8]. Le décodeur de SPEERAL, à pile asynchrone, est basé sur l'algorithme de recherche A^*. L'exploration du graphe (le treillis de phonèmes) est dirigée par une fonction d'estimation basée sur deux informations : le score de l'hypothèse courante et une sonde estimant le coût minimal en fin de chemin. Cette sonde h combine un score acoustique et un score d'anticipation linguistique ([3]). Le terme acoustique est calculé par un décodage acoustico-phonétique réalisé par l'algorithme de Viterbi arrière sur le treillis de phonèmes.

L'algorithme A^* repose complètement sur la fonction d'estimation $F(n)$ qui évalue, pour chaque noeud exploré, le coût minimal des chemins passant par ce point. Cette fonction guide l'exploration du graphe d'hypothèses en orientant le décodage vers les chemins dont l'estimation partielle est bonne. Par ailleurs, la progression dans le graphe affine l'évaluation des chemins explorés, ce qui peut conduire à l'abandon de certaines voies. Dans ce cas, l'algorithme revient en arrière et explore d'autres branches du graphe, ce qui le désynchronise du flux audio.

Afin de pouvoir prendre en compte les informations issues de la transcription approchée, un module a été rajouté pour influencer le score de l'hypothèse courante au sein de l'algorithme de recherche. Ce mécanisme oriente le décodage en influençant dynamiquement le score de l'hypothèse évaluée. L'algorithme proposé se décompose en deux parties : la synchronisation entre la transcription et le flux de parole puis l'intégration de l'information issue de l'alignement synchrone à la fonction d'évaluation $F(n)$.

Synchronisation du flux audio et de la transcription imparfaite

Le moteur de reconnaissance construit des hypothèses au fur et à mesure qu'il avance dans le treillis de phonèmes. Les meilleures hypothèses à un instant t sont prolongées. Les modifications apportées au décodeur permettent d'aligner à la transcription approchée chaque nouveau mot et son historique. Ceci est réalisé par un algorithme d'alignement temporel (Dynamic Time Warping [1]). Ainsi, tous les mots de l'hypothèse courante du décodeur sont alignés sur un passage de la transcription. Chaque mot rajouté dans l'hypothèse est alors rescoré en fonction d'un indice de confiance issu de l'alignement.

La figure 1 présente l'évolution des hypothèses du décodeur à pile A* influencées par un alignement DTW sur une transcription approchée.

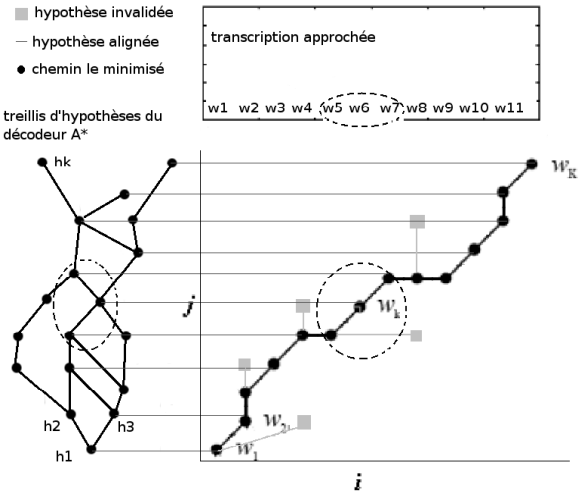

FIG. 1: Synchronisation du faisceau de recherche avec la transcription imparfaite par algorithme DTW

Pondération de l'hypothèse courante en fonction de l'alignement

La fonction d'estimation calcule, pour chaque noeud du graphe, les coûts du chemin exploré ainsi qu'une sonde minimisant le coût des chemins finaux. La qualité de cette sonde influence directement les performances de l'algorithme de recherche. La solution proposée réhausse le score des mots présents dans le faisceau de recherche lorsqu'ils sont alignés avec la transcription approchée ; s'ils ne sont pas présents dans le faisceau, l'algorithme d'alignement n'intervient pas. Pour que notre alignement oriente le moteur de reconnaissance, il faut que l'algorithme A* pré-

sente le mot : le score de l'hypothèse courante sera alors modifié en conséquence. Nous ne modifions pas les scores d'anticipation linguistique.

Une fois l'hypothèse synchronisée avec la transcription, l'algorithme estime un score de synchronie locale, calculé à partir du nombre de mots de l'historique correctement alignés à la transcription. Maximal lorsque le trigramme complet est aligné, ce score décroît en fonction des défauts d'alignement de l'historique.

4. EXPÉRIMENTATIONS

4.1. Cadre expérimental

L'ensemble des expériences a été effectué avec le système de "Broadcast news" développé au LIA qui a été engagé dans la campagne d'évaluation ESTER ([6]).

Corpus utilisé et transcription approchée

Le système est évalué sur une heure d'émission radio issu du corpus de développement d'ESTER (France Inter du 08-04-2003, de 7h à 8h). Ensuite, 10% d'erreurs ont été introduites manuellement dans la transcription, tout en prenant soin de garder une forme journalistique correcte pour respecter le style classique d'une émission radiophonique. Nous simulons ainsi une transcription imparfaite proche de ce que serait le script d'une émission de ce type.

Interpolation de modèles de langages

Des expériences préliminaires ont été menées afin d'identifier les gains potentiels de nos méthodes. Un modèle de langage a été appris sur la transcription exacte, puis combiné avec un modèle de langage générique (65000 mots appris sur Le Monde). L'objectif est d'être capable de mesurer l'effet réel des techniques proposées sur les performances du décodeur. Les mots hors vocabulaire ont été extraits de la transcription pour être phonétisés et ajoutés au modèle de langage. Le tableau 1 présente les résultats d'interpolation d'un modèle de langage appris sur la transcription exacte avec le modèle de langage générique. Par ailleurs, un décodage normal avec modèle générique présente un WER de 22.7%.

TAB. 1: Résultats des expériences de référence avec interpolation du modèle de langage générique (ML-G) et du modèle appris sur la transcription exacte (ML-TrExact)

	Taux d'erreur
ML-G seul	22.7%
ML-TrEx seul	5.2%
ML-G 70% + ML-TrExact 30%	13.0%
ML-G 50% + ML-TrExact 50%	11.5%
ML-G 30% + ML-TrExact 70%	10.8%

Ensuite, à partir de la transcription approchée, un modèle de langage a également été généré. Les expériences utilisant ce modèle de langage combiné au modèle de langage générique sont présentées dans le tableau 2.

Ces expériences montrent qu'un décodage restreint avec un modèle de langage appris sur une transcription approchée améliore sensiblement le taux d'erreurs. Cependant, sans autre source d'information, le moteur de reconnaissance continue à faire des erreurs sur les parties qui ont été mal apprises (les erreurs dans la transcription). Par ailleurs, un décodage avec un modèle de langage appris sur la transcription exacte montre que cette technique ne permet pas de descendre en deçà de 15% de WER.

TAB. 2: Résultats des expériences d'interpolation de modèles de langage Générique (ML-G) et appris sur la transcription approchée (ML-TrErr)

	Taux d'erreur
ML-TrErr seul	16.3%
ML-G 70% + ML-TrErr 30%	16.2%
ML-G 50% + ML-TrErr 50%	15.4%
ML-G 30% + ML-TrErr 70%	15.2%

Expériences avec interpolation de modèles et alignement

Après avoir expérimenté des modèles interpolés, nous avons évalué la méthode de décodage guidée par la transcription qui a été décrite précédemment. Bien que cette approche puisse permettre de lever certaines des limites observées dans la combinaison de modèles, des sources potentielles d'erreur subsistent. En particulier, des heuristiques sont utilisées dans le décodeur pour réduire l'espace de recherche et accélérer le décodage. En conditions normales, ces coupures ne doivent introduire que peu d'erreurs ; cependant, lorsque le contexte acoustique est très mauvais, les meilleures hypothèses peuvent se trouver exclues du faisceau de recherche. Ceci peut se produire plus fréquemment dans une configuration temps réel du système, pour laquelle les coupures sont plus strictes. Dans ce cas, une stratégie basée sur la "promotion" des hypothèses du faisceau coïncidant avec la transcription ne permet pas de récupérer ces erreurs. On peut quantifier de façon approximative la perte correspondant à cette situation en utilisant le moteur de reconnaissance pour faire un alignement forcé de la transcription exacte sur le signal.

Les expériences combinant l'interpolation des modèles de langage avec un alignement sur la transcription exacte sont présentées dans le tableau 3 .

TAB. 3: Résultats des expériences interpolant modèle de langage générique (ML-G) avec le modèle de langage appris sur la transcription exacte (ML-TrEx) et s'alignant sur la transcription exacte (alTrEx)

	Taux d'erreur
ML-G seul + alignement TrEx	6.1%
ML-TrEx seul + alTrEx	4.1%
ML-G70%+ML-TrEx30%+alTrEx	3.7%
ML-G50%+ML-TrEx50%+alTrEx	3.5%
ML-G30%+ML-TrEx70%+alTrEx	3.7%

Nous obtenons dans ce cas un taux d'erreur mots de 3.5%. Ce niveau d'erreur peut être considéré comme minimal pour une méthode ré-estimant les hypothèses concurrentes dans le faisceau d'hypothèses sans remettre en cause le contenu même de ce faisceau.

Le tableau 4 reprend les expériences précédentes mais en remplaçant la transcription exacte par la transcription approchée.

Le meilleur résultat est obtenu en combinant le modèle de

TAB. 4: Résultats des expériences avec interpolant le modèle de langage générique (ML-G) avec le modèle appris sur la transcription approchée (ML-TrEr), et s'alignant sur la transcription approchée (alTrEr)

	Taux d'erreur
ML-TrErr + alignement TrEr	9.9%
ML-G + alignement TrEr	7.7%
ML-G70%+ML-TrEr30%+alTrEr	7.2%
ML-G50%+ML-TrEr50%+alTrEr	7.4%
ML-G30%+ML-TrEr70%+alTrEr	8.6%

langage générique avec un poids de 70% avec le modèle appris sur la transcription approchée et en réalisant un alignement sur cette dernière. L'alignement fait descendre le taux d'erreurs mots jusqu'à 7.2%. Il permet d'apporter une information temporelle qui est mal prise en compte par le modèle de langage. L'utilisation d'un alignement DTW associé à l'interpolation des modèles montre à nouveau un gain. Cette expérience montre qu'un équilibre peut être atteint pour exploiter l'information approchée sans pour autant reproduire la majorité des erreurs qu'elle comporte. Les meilleurs résultats sont obtenus en utilisant le modèle de langage générique fortement pondéré par rapport au modèle de langage appris sur la transcription approchée et en alignant sur cette dernière. L'information erronée ne se trouve que dans la transcription sur laquelle le moteur essaye de s'aligner. Quand il ne trouve aucun alignement, il se replie exclusivement sur l'utilisation du modèle de langage générique. Par ailleurs, la légère interpolation avec le modèle de langage appris sur la transcription approchée permet de corriger certaines erreurs inhérentes au modèle de langage générique. Dans ces conditions, le système tire avantageusement parti de la transcription approchée lorsqu'elle est correcte et bascule en mode de reconnaissance automatique lorsque les observations acoustiques ne correspondent pas à la transcription proposée.

5. CONCLUSION

Nous avons proposé et évalué deux méthodes exploitant l'information contenue dans une transcription imparfaite pour améliorer les performances d'un SRAP. La première consiste à extraire du script l'information linguistique sous forme d'un modèle de langage trigramme appris sur la transcription approchée. Nos expérimentations montrent que l'interpolation de ce modèle avec le modèle de langage générique permet d'améliorer significativement le décodage. Il ne permet cependant pas de dépasser la qualité de la transcription approchée fournie, ce qui limite son intérêt. La seconde approche présentée consiste à orienter l'algorithme de recherche vers la transcription en synchronisant à la volée les hypothèses en cours d'évaluation et la transcription dont on dispose. Cette méthode permet de combiner efficacement les scores linguistiques avec les scores d'alignement. Partant d'un taux d'erreurs mots de 22.7%, notre méthode exploitant une transcription imparfaite permet de ramener ce taux jusqu'à 7.2%, montrant ainsi l'intérêt d'un alignement sur une transcription approchée : le décodage guidé permet de descendre le WER au dessous du WER de la transcription approchée (de 10.1% à 7.2%, soit 28% en relatif).

Un des intérêts du décodage basé sur A^* est la facilité avec laquelle des sources d'informations supplémentairespeuvent être intégrées au coeur même de l'algorithme de recherche. Ici, l'évaluation des hypothèses guidée par la transcription permet d'atteindre les objectifs fixés tout en accélérant le décodage. Ce gain au temps d'exécution est dû à la réduction de l'espace de recherche ainsi qu'à une meilleure anticipation des chemins optimaux, qui correspondent souvent aux hypothèses alignées. Cependant, le gain obtenu en terme de vitesse de décodage peut probablement être augmenté en introduisant plus tôt des heuristiques basées sur la transcription approchée, notamment au niveau de la sonde elle-même.

Bien que ces premiers résultats montrent l'intérêt d'un alignement sur une transcription approchée, ces expériences ont été effectuées dans des conditions contrôlées : niveau de bruit relativement réduit, transcription relativement proche de la transcription exacte, etc. Nous envisageons d'utiliser nos méthodes dans des conditions plus difficiles, par exemple sur le sous-titrage de films, ou l'alignement en temps réel de sous-titrages pour les pièces de théâtre.

RÉFÉRENCES

[1] D. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. 1994.

[2] Huang Chih-wei. Automatic closed caption alignment based on speech recognition transcripts. 2003.

[3] G. Linarès D. Massonié, P. Nocéra. Scalable language model look-ahead for lvcsr. *InterSpeech'05, Lisboa, Portugal*, 2005.

[4] Photina Jaeyung Jang and Alexander G.Hauptmann. Improving acoustic models with captioned multimedia speech. *IEEE International Conference on Multimedia Computing and Systems, Florence, Italy*, 1999.

[5] L. Lamel, J.L. Gauvain, and G. Adda. Lightly supervised and unsupervised acoustic models training. *Computer Speech and Language*.

[6] G. Linarès, P. Nocéra, D. Matrouf, F. Béchet D. Massonié, and C. Fredouille. Le système de transcription du lia pour ester-2005. 2005.

[7] Pedro J. Moreno, Chris Joerg, Jean-Manuel Van Thong, and Oren Glickman. A recursive algorithm for the forced alignment of very long audio segments. *International Conference on Spoken Language Processing*, 1998.

[8] Pascal Nocera, Georges Linares, and Dominique Massonié. Principes et performances du décodeur parole continue speeral. *XXIVées journées d'étude sur la parole*, 2002.

[9] Paul Placeway and John Lafferty. Cheating with imperfect transcripts. *Proceedings of ICSLP*, 1996.