Alignement de transcriptions imparfaites sur un flux de parole

Benjamin Lecouteux

▶ **To cite this version:**

Benjamin Lecouteux. Alignement de transcriptions imparfaites sur un flux de parole. [Rapport de recherche] LIA. 2005. hal-02094749

HAL Id: hal-02094749
https://hal.science/hal-02094749

Submitted on 9 Apr 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Alignement de transcriptions imparfaites sur un flux de parole

LECOUTEUX Benjamin

Laboratoire d'Informatique d'Avignon (CNRS FRE2487)
339, chemin des Meinajaries
Agroparc – B.P. 1228
F-84911 Avignon Cedex 9

benjamin.lecouteux@iup.univ-avignon.fr

RÉSUMÉ. *Dans de nombreuses situations, un texte descriptif peut être associé à un flux audio parlé : sous-titres de films, scénario/texte et théâtre, résumés et émissions radiophoniques, transcription réarrangée pour les débats politiques. Le texte correspond rarement à la transcription exacte de la parole : une pièce de théâtre est jouée différemment à chaque représentation et un présentateur s'éloigne parfois de son prompteur. Le but de ce travail est d'aligner un texte descriptif sur le flux parlé lorsqu'il lui correspond, et de laisser la main au système de reconnaissance lorsque la transcription s'en éloigne. Les applications visées sont multiples : permettre à des malentendants de suivre une pièce de théâtre en affichant le texte de la pièce aligné avec la parole correspondante en respectant les variations, suivre un film dans une langue en alignant au plus près les sous-titres avec une voix audio, suivre des débats, des réunions.*

ABSTRACT. *In many cases, a descriptive transcript can be associated to speech signal : movies subtitles, scenario and theatre, summaries and radio broadcast, rearranged transcription for political debates. Transcripts correspond rarely to the exact word utterance. An actor plays differently each play and a speaker is sometimes far of the promptor. The goal of this work is whether to align the given transcript when it matches the speech signal or to fall back on an automatic speech recognition (ASR) adapted with the transcript language. There are multiple applications : to help deaf people following a play with closed caption aligned to the voice signal (with respect to performers variations), to watch a movie in another language using aligned closed caption, to transcript in real time debates or meetings.*

MOTS-CLÉS : *reconnaissance de la parole, alignement automatique, sous-titrage automatique*

KEYWORDS: *speech recognition, automatic alignment, closed captions, subtitle*

Table des matières

Table des figures

Liste des tableaux

1. Introduction

Les systèmes de reconnaissance automatique de la parole continue grand vocabulaires sont capables d'obtenir des taux de reconnaissance corrects quand les conditions d'apprentissage et de test sont proches. Les résultats de la campagne d'évaluation ESTER [GEO 05], qui met en concurrence différents systèmes sur la transcription d'émissions radiophoniques francophones, montrent que dans ces conditions définies, les systèmes actuels obtiennent des taux d'erreurs mots compris entre 12% et 40%.

Cependant, ces systèmes sont généralement peu robustes dans des conditions éloignées de l'apprentissage : des variations entre les contextes de test et d'apprentissage provoquent fréquemment une dégradation significative du taux de reconnaissance.

Par ailleurs, dans certaines situations, il est possible d'utiliser des sources d'informations susceptibles de faciliter le décodage. C'est le cas pour des émissions comportant des sous-titres ou pour lesquelles un résumé peut être mis à disposition : scénario pour le cinéma, scripts et théâtre etc. Cette source d'information approchée peut être utilisée pour améliorer les performances. Ce problème est abordé dans la littérature dans le domaine de l'alignement automatique de textes sur des flux audio. Dans le cadre de cette étude, nous travaillons sur des transcriptions qui peuvent s'éloigner du signal audio. Ces divergences éventuelles changent profondément la nature du problème : il ne s'agit plus de faire un alignement forcé mais de retrouver, lorsque c'est nécessaire, le contenu réel du message dont la transcription s'est éloignée. Dans ce contexte, le problème peut être reformulé comme un problème de reconnaissance de la parole utilisant la transcription imparfaite comme une source d'information supplémentaire.

L'objectif de notre recherche est donc d'utiliser l'information contenue dans une transcription approchée afin d'aider le système de reconnaissance. Nous souhaitons aligner les transcriptions mises à disposition lorsqu'elles correspondent au flux audio et laisser le système de reconnaissance reprendre la main lorsque la transcription s'en éloigne. Nous allons proposer différentes méthodes pour modéliser les informations issues de la transcription approchée et différentes stratégies pour leur intégration dans le processus de décodage.

Dans une première partie, nous ferons une présentation succincte de l'architecture des systèmes de reconnaissance vocale grand vocabulaire et en particulier du système SPEERAL développé au LIA ; puis, dans une seconde partie, nous expliquerons quels sont les problèmes liés à l'alignement d'un texte approché sur un flux audio et quelles solutions existent actuellement. Nous décrirons ensuite les travaux que nous avons effectués sur le système SPEERAL pour l'intégration des informations issues des transcriptions approchées. En dernière partie, nous détaillerons les expériences menées et nous analyserons les résultats obtenus.

2. Outils du LIA utilisés pour la reconnaissance vocale

2.1. *Le ToolKit parole développé par le LIA*

Le LIA a développé un ensemble d'outils logiciels pour le traitement de la parole. Cette boite à outils permet de réaliser la plupart des traitements nécessaires à la réalisation d'un système de transcription automatique complet. Elle contient trois groupes de composants relativement indépendants : les outils de segmentation, les outils de modélisation acoustique et le moteur de reconnaissance lui-même :

– les outils de segmentation : il s'agit de composants destinés à préparer un flux audio au décodage de la parole et à en extraire des informations acoustiques de haut niveau (changement, identification ou suivi de locuteur, extraction des zones de parole, identification des conditions acoustiques, etc.) Cette partie est basée sur le ToolKit Alizé [ALI].

– la modélisation acoustique par Modèles de Markov Cachés (MMC) : ce module offre un ensemble de fonctionnalités pour la paramétrisation et la modélisation acoustique par MMC : estimation et adaptation des MMC, décodage acoustico-phonétique, alignements contraints ou partiellement contraints, génération des treillis de phonèmes, etc.

– le moteur de reconnaissance (SPEERAL) [NOC 02] : il s'agit d'un système de reconnaissance de la parole continue, grand vocabulaire, basé sur un dérivé de l'algorithme A^*. Le travail présenté ici concerne essentiellement cette partie du ToolKit du LIA.

2.2. *Le décodeur SPEERAL*

Le LIA a développé un système de reconnaissance de la parole grand vocabulaire et parole continue nommé SPEERAL [NOC 02]. Le moteur SPEERAL (figure 1) est un décodeur à pile asynchrone 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 fonction estimant le coût minimal en fin de chemin : la sonde acoustico-phonétique h est calculée par l'algorithme de Viterbi arrière sur le treillis de phonèmes associé à un décodage acoustico-phonétique (DAP). L'algorithme A^* utilise alors une fonction d'évaluation $F(n)$ pour chaque noeud exploré qui représente l'estimation du chemin de meilleur coût passant par ce chemin. La modélisation acoustique est basée sur des modèles de Markov cachés (HMM) et les modèles linguistiques sont des n-grammes [BAR 96]. SPEERAL est capable d'interpoler plusieurs modèles de langage. Avec l'utilisation de l'algorithme A^*, la progression des hypothèses se fait mot par mot puis la liste des meilleurs mots est prolongée.

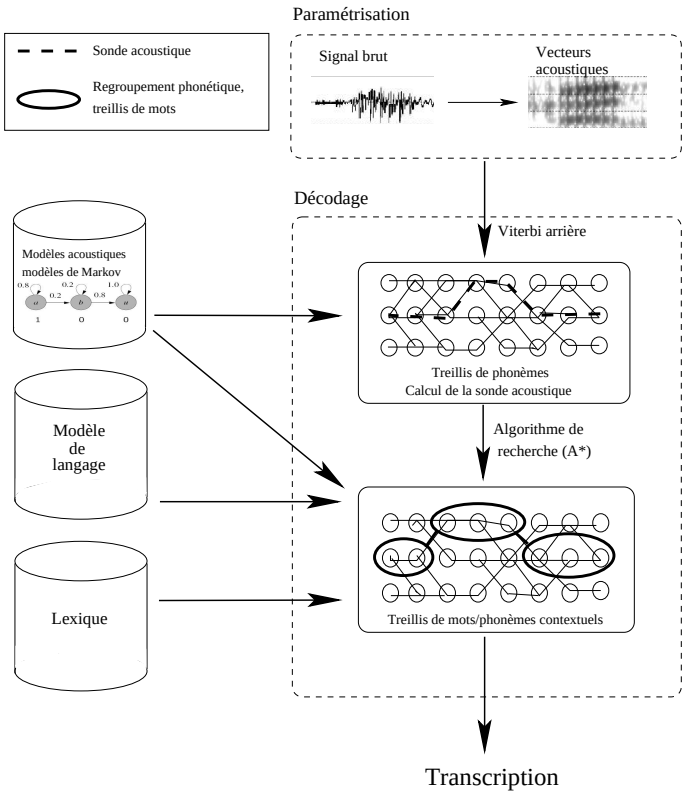

Figure 1. *Le système de reconnaissance vocale SPEERAL*

3. Alignement de la parole sur des textes approchés

3.1. *Présentation du problème*

Dans certaines conditions d'utilisation, il est possible d'avoir une connaissance à priori -et plus ou moins fiable- du texte qui va être prononcé. Un système de traitement automatique de la parole devrait pouvoir tirer parti de cette source d'information supplémentaire sur le contenu du message. Cette idée se retrouve dans la littérature pour l'indexation audio/vidéo automatisée ([JON 00]) ou la réestimation aveugle de modèles acoustiques ([JAN 99]) à partir de données non-transcrites. Il est formulé soit comme un problème de synchronisation de texte sur le flux audio, soit comme un problème de correction de transcriptions imparfaites. En effet, si des sous-titres ou une transcription approchée sont disponibles, il est possible de les exploiter au travers d'un système de reconnaissance automatique de la parole. Le système peut utiliser l'information supplémentaire pour décoder plus facilement l'audio. Malheureusement, ces

informations sont la plupart du temps approximatives. [PLA 96] mesure un taux de différence entre les sous-titres d'un film et la transcription exacte compris entre 10% et 20%. Ces divergences entre la transcription et le message tel qu'il a été dit augmentent considérablement la difficulté de l'alignement. Une approche différente d'un alignement forcé doit être utilisée. Cette étude aborde ces deux problèmes qui sont étroitement liés :

– aligner au mieux la transcription approchée sur le flux audio quand celui-ci lui correspond. Ce sujet a été traité dans quelques études qui montrent qu'en se basant sur un décodage de la parole contraint par la transcription, on obtient de bonnes performances en alignement.

– corriger les alignements pour qu'ils correspondent au flux audio quand la transcription s'en éloigne : un acteur s'éloigne de son texte, un journaliste ne suit pas toujours son prompteur. Par ailleurs, ces différences avec la transcription entraînent une perte d'information tout en augmentant la complexité de l'alignement.

Nous présenterons les techniques existantes actuellement pour faire de l'alignement sur un flux audio à partir d'une transcription exacte. Puis nous aborderons l'état de l'art de l'alignement sur des transcriptions approximatives.

3.2. *Alignement forcé avec transcription exacte*

Le sujet de l'alignement sur transcription exacte est abordé dans la littérature par [MOR 98] pour aligner de longs documents audio avec leur transcription dans le cadre d'une indexation automatique de documents multimédias. Ils observent notamment que la longueur des documents augmente leur variabilité acoustique et donc la difficulté d'alignement. De plus, l'alignement devient difficile, par manque de points de synchronisation. Sans ces derniers, le coût algorithmique devient trop important. [MOR 98] proposent une méthode basée sur la recherche de zones bien synchronisées, appelées *îlots de confi ance* : dans un premier temps, un modèle de langage est estimé sur la transcription exacte. Une première passe isole des zones où transcription à 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, dont l'application est restreinte aux transcriptions exactes, obtient d'excellents résultats : 99% des mots sont correctement alignés, même dans de mauvaises conditions acoustiques.

3.3. *Alignement de transcriptions approchées*

Le problème du traitement automatique de transcriptions approchées a été abordé par Paul Placeway et John Lafferty [PLA 96] qui ont expérimenté l'exploitation de sous-titres avec le système de reconnaissance SPHINX-3. Leurs expériences portaient sur une base de données de journaux diffusées en anglais par CNN datant d'août 1995.

Ils proposent d'utiliser les sous-titres S mis à disposition en estimant un modèle de langage sur ces derniers puis en alignant le flux audio A sur ces sous-titres. Pour déterminer la séquence de mots W la plus probable, ils cherchent la séquence de mots de vraisemblance maximale :

$$\widehat{W} = \underset{w}{\text{ArgMax}}\, P(W|A,S)$$
$$\widehat{W} = \underset{w}{\text{ArgMax}}\, P(A|S,W)P(W)P(S|W) \quad [1]$$

La probabilité $P(W)$ correspond à la probabilité linguistique de la séquence de mots W connaissant le modèle de langage. Le terme $P(S|W)$ correspond à la probabilité d'alignement des sous-titres sur la séquence de mots W. Par ailleurs, [PLA 96] émet l'hypothèse que le signal audio A et l'alignement sur les sous-titres S sont statistiquement indépendants. La tâche du décodeur est donc de trouver la séquence de mots W qui maximise le produit des vraisemblances linguistique ($P(W)$), d'alignement ($P(S|W)$) et acoustique ($P(A|W)$) :

$$\widehat{W} = \underset{w}{\text{ArgMax}}\, P(A|W) * P(S|W) * P(W) \quad [2]$$

On se retrouve alors dans le cas d'un système de reconnaissance standard où l'on rajoute la probabilité des sous-titres sachant une suite de mots.

Pour combiner l'information issue des sous-titres avec les modèles usuels du décodeur, Paul Placey et John Lafferty ont interpolé un modèle de langage générique avec un modèle estimé sur les sous-titres. Ces modèles combinés sont ensuite utilisés de façon classique par le système de reconnaissance de la parole. Leurs expérimentations sont faites avec des sous-titres comportant 9.7% d'erreurs par rapport à la transcription exacte sur des journaux radiophoniques de CNN 1995. Cette technique obtient une amélioration relative du taux d'erreurs mots de 15.4% (de 55.8% d'erreurs à 47.2% d'erreurs). Par ailleur, ils proposent d'intégrer un mécanisme d'alignement forcé sur les sous-titres, en plus de l'interpolation des modèles. Cette méthode apporte un gain relatif de 25.8% (de 47.2% d'erreurs à 35.0% d'erreurs).

3.4. *Alignement de sous-titres et indexation de documents vidéo*

[CHI 03] et [JON 00] utilisent les sous-titres présents dans les vidéos pour aider leur système de reconnaissance vocale afin de faire respectivement de l'alignement automatique de sous-titres et de la segmentation automatique. Le flux audio est synchronisé sur la vidéo, et apporte une information précise pour sa segmentation et son

indexation. Par ailleurs, les sous-titres apportent une information sur l'audio mais ne sont que partiellement synchronisés sur le début des phrases. La méthode de [CHI 03] consiste à séparer les sous-titres de la vidéo et à traiter d'abord le flux audio : un îlot de confiance est constitué avec les mots issus d'un sous-titre (dont on connaît le temps de départ) en effectuant un alignement sur une matrice D où i est l'index des mots issus du flux audio et j l'index des mots du sous-titre. La distance optimale est définie par :

$$D(i,j) = min \begin{cases} D(i-1,j) + ins(i,j) \\ D(i,j-1) + del(i,j) \\ D(i-1,j-1) + ms(i,j) \end{cases} \quad [3]$$

où $ins(i,j)$ (insertion), $del(i,j)$ (suppression) et $ms(i,j)$ (correct ou substitution) sont les trois types de transitions possibles. Par ailleurs [CHI 03] applique une pénalité d'insertion considérant le nombre moins important de mots contenus dans les sous-titres par rapport à la transcription du moteur de reconnaissance :

$$ins(i,j) = a + \frac{1-a}{e^{b*(t_{j+1}-t_j)}} \quad [4]$$

où $0 \leq a \leq 1$, b est une constante. t_{j+1} et t_j sont les temps d'apparition en secondes du $j+1^{ime}$ et du j^{ime} mots du sous-titre. Quand les mots sont peu éloignés dans le temps, l'insertion tend vers 1 : l'insertion sera difficile. Lorsque les mots s'éloignent dans le temps, l'exponentielle fait tendre l'insertion vers a, la rendant plus facile. Partant de ces considérations, [CHI 03] effectue un alignement automatique des sous-titres sur le flux audio en respectant les éloignements des deux flux.

L'approche de [JON 00] est différente : à partir des sous-titres, ils estiment un réseau de mots constitués d'unités phonétiques formées de HMM. Ils extraient la partie audio correspondant aux sous-titres sachant que le décalage entre l'apparition du sous-titre et l'audio est compris entre 2 et 7 secondes. Ils confrontent les segments de signal au réseau de mot jusqu'à obtention d'une corrélation suffisante (avec un seuil fixé au préalable). A chaque segment audio est ainsi associé au fur et à mesure un mot clef : ils réalisent ainsi un alignement entre sous-titres et flux audio. Avec cette méthode, ils arrivent à segmenter correctement 98% du document audiovisuel.

4. Modifications dans le moteur SPEERAL

Nous avons expérimenté deux méthodes exploitant la transcription approchée. 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 moteur de reconnaissance.

4.1. *Adaptation des modèles de langage de SPEERAL*

SPEERAL utilise des modèles de langage tri-grammes. Généralement, ces modèles sont estimés sur de grands corpus de diverses origines ; un algorithme de lissage (de type backoff) permet d'avoir une mesure plus fiable des fréquences.

SPEERAL peut interpoler linéairement ([MAS 00]) et dynamiquement plusieurs modèles de langages. Dans nos différentes expériences, nous avons utilisé des modèles spécifiques estimés sur les transcriptions exactes et approchées ([CHE 04]). L'information linguistique est extraite puis introduite dans le modèle de langage du moteur de reconnaissance. Le modèle trigramme obtenu est combiné à un modèle générique.

La probabilité linguistique de la séquence de mots W sachant le modèle interpolé est obtenue par combinaison linéaire des probabilités partielles sachant chacun des modèles initiaux :

$$P'(w|h) = \sum_{i=1}^{k} \alpha_i P_i(w|h) \qquad [5]$$

avec $0 < \alpha_i \leq 1$ et $\sum_i \alpha_i = 1$

où α_i représente le poids de chaque modèle.

4.2. *Moteur de reconnaissance et algorithme d'alignement*

La seconde méthode évaluée consiste à rajouter de l'information issue du flux transcrit via un alignement. Nous avons rajouté un module qui influence le score de l'hypothèse courante au sein de l'algorithme A^*. Notre méthode se fait en deux parties : la synchronisation entre la transcription et le flux audio puis la modification du score de l'hypothèse en fonction de l'alignement.

4.2.1. *Synchronisation entre le flux audio et la transcription imparfaite*

Le moteur de reconnaissance construit des hypothèses au fur et à mesure qu'il avance dans le treillis de phonèmes. A chaque nouveau mot rajouté dans une hypothèse, il aligne cette hypothèse sur la transcription approchée avec un algorithme d'alignement temporel (Dynamic Time Warping) introduit par [BER 94] pour la fouille de données. Cet algorithme consiste à effectuer une comparaison dynamique entre une matrice de référence et une matrice de test [KEO 00]. Avec deux vecteurs définis par

$$H = h_1, h_2, ..., h_i, ..., h_n \text{ et } T = t_1, t_2, ..., t_j, ..., t_n \qquad [6]$$

On construit une matrice de dimension $(n\text{x}m)$ dont (i, j) est la distance euclidienne entre les points H_i et T_j :

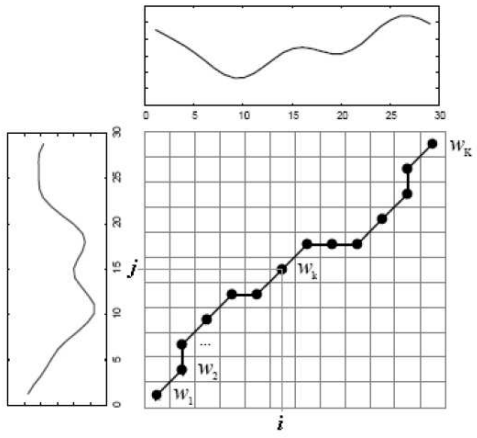

Figure 2. *Recherche d'un chemin DTW ([KEO 00])*

$$d(h_i, t_j) = (h_i - t_j)^2 \quad [7]$$

Un alignement entre T et H est le chemin W suivant les éléments contigus de cette matrice tel que

$$W = w_1, w_2, ..., w_k \text{ où } max(m, n) \leq m + n - 1 \quad [8]$$

Le chemin W doit commencer et finir aux extrémités d'une diagonale de la matrice et l'évolution est restreinte vers les éléments adjacents de la matrice. Les étapes successives dans le chemin sont réparties de manière monotone dans le temps. L'algorithme avance alors sur le chemin qui minimise la distance (figure 2) :

$$\gamma(i, j) = d(h_i, t_j) + min\ \gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1) \quad [9]$$

Le résultat est une mesure de la distance entre le test et la référence. Les scores de transition utilisés pour notre DTW sont les suivants : insertion = +4, suppression = +6, substitution = +12, 0 sinon. L'algorithme DTW sert essentiellement à synchroniser les mots proposés avec la transcription : le score le plus faible correspondant à l'alignement courant le plus probable.

4.2.2. *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. Dans SPEERAL, elle est composée d'une partie acoustico-phonétique et d'un terme d'anticipation linguistique. L'anticipation linguistique est basée sur une estimation des meilleurs trigrammes prolongeant l'hypothèse explorée. Dans le cadre de SPEERAL, l'algorithme A^* explore non pas un treillis de mots, mais un treillis de phonèmes. L'anticipation linguistique se fait donc à deux niveaux : au niveau des phonèmes et au niveau des mots. Le score linguistique est calculé par deux fonctions selon le type du dernier mot de l'hypothèse :

– la fonction Proba_Mot est appliquée si la nouvelle hypothèse est terminée par un mot : elle possède la liste des mots la constituant et détermine son score linguistique.

– la fonction Proba_Bout est appliquée si la nouvelle hypothèse se termine par un noeud de l'arbre correspondant à un début de mot. Alors, la fonction retourne un score linguistique basé sur l'anticipation des mots pouvant être générés à partir de ce noeud.

$$Proba\_Bout(m_1..m_k, n) = \underset{m_n}{\text{ArgMax}}(Proba\_Mot(m_1...m_k m_n)) \qquad [10]$$

Où m_n est une feuille (mot) du sous-arbre débutant au noeud n.

Cette anticipation permet de pénaliser rapidement les branches aboutissant à des mots improbables.

Notre méthode d'alignement a été effectuée au niveau de la fonction Proba_Mot. Ainsi, nous ne pondérons que le score linguistique des mots complets. Pour que notre alignement oriente le moteur de reconnaissance, il faut que ce dernier présente le mot complet. Le score de l'hypothèse courante sera alors modifié en conséquence : nous ne modifions pas l'exploration partielle (l'anticipation linguistique).

Le score de la probabilité d'un mot est $P(m|m_2m_1)$, m_2 et m_1 étant les deux mots précédants de l'hypothèse courante. Lorsque nous avons aligné l'hypothèse sur la transcription imparfaite, l'algorithme lui donne un score de confiance en fonction des mots précédants dans la transcription et dans l'hypothèse. Si les 2 mots précédants de l'hypothèse correspondent aux deux précédants dans la transcription, la confiance est maximale. Celle-ci décroît jusqu'à ce que le mot soit isolé.

Soient m_1 et m_2 les deux mots précédant le mot m dans l'hypothèse du moteur de reconnaissance. Et ma_1 et ma_2 les deux mots précédant le mot m dans la transcription et S le score retourné par notre fonction d'alignement :

$$S = \begin{cases} 0.6 & \text{si } m_1 = ma_1 \text{ et } m_2 = ma_2 \\ 0.8 & \text{si } m_1 = ma_1 \text{ et } m_2 \neq ma_2 \\ 0.8 & \text{si } m_1 \neq ma_1 \text{ et } m_2 = ma_2 \\ 0.95 & \text{si } m_1 \neq ma_1 \text{ et } m_2 \neq ma_2 \\ 1 & \text{si } m \text{ non trouvé} \end{cases} \qquad [11]$$

La probabilité $P(m|m_2m_1)$ retournée de l'hypothèse devient alors :

$$P'(m|m_2m_1) = S \,.\, P(m|m_2m_1) \quad [12]$$

Les scores de l'hypothèse sont ainsi pondérés en fonction des mots alignés avec la transcription.

5. Expérimentations

5.1. *Système de base pour les expériences*

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 **??**. Dans un premier temps nous présenterons les spécificités de ce système puis nous verrons sur quelles bases se sont appuyées nos expériences. Enfin nous présenterons les différentes expériences qui ont été effectuées.

5.1.1. *La segmentation*

Le but de la segmentation est de découper le signal audio en plusieurs classes comme la musique, la parole et les silences. Lors de la première étape, le signal est découpé en classes parole/non-parole. Ensuite une seconde segmentation est effectuée pour différencier les paroles de type "téléphone" et "radio".

5.1.2. *La paramétrisation et les modèles acoustiques*

La paramétrisation utilise un algorithme PLP qui génère des vecteurs de 12 coefficients plus l'énergie et les dérivées premières et secondes de ces 13 coefficients. Ensuite, une normalisation cepstrale est effectuée sur une fenêtre glissante de 500ms.

Les modèles acoustiques utilisés sont contextuels et appris sur 90 heures de données transcrites d'Ester. Chaque jeu de modèle acoustique contient 10000 HMM pour 3600 états émetteurs et 230000 gaussiennes.

5.1.3. *Les lexiques et modèles de langage*

Les ressources linguistiques sont extraites de deux corpus :

– le journal "Le Monde" de 1987 à 2003, représentant 330 millions de mots.

– ESTER représentant 960000 mots.

Le lexique contient 65000 mots, composé de tous les mots du corpus ESTER et des mots les plus fréquents présents dans "Le Monde". La phonétisation est effectuée à partir de deux sources :

– le lexique phonétique ILPHON.

– le logiciel de phonétisation LIA_PHON.

La phonétisation générée par LIA_PHON a été vérifiée et modifiée manuellement si nécessaire.

Les modèles de langage sont appris sur "Le Monde" et ESTER avec le SRI-LM Toolkit. Ces modèles sont ensuite combinés avec des modèles adaptés à l'utilisation. Ils sont constitués de trigrammes qui ont été lissés avec l'algorithme de backoff Kneser-Ney.

5.2. *Expérimentations*

5.2.1. *Corpus utilisé et transcription approchée*

Le corpus de test est constitué d'un extrait de 15 minutes d'émission radio issu de la base ESTER. Il s'agit d'une émission d'information de la Radio Télévision Marocaine (RTM). L'avantage de cette base est d'avoir à disposition sa segmentation de référence ainsi que sa transcription exacte. Par ailleurs, l'ensemble des outils du LIA sont développés pour ce type de données : les modèles de langage sont appris sur des corpus journalistiques et les modèles acoustiques sont appris sur des émissions radiophoniques. Nous avons également travaillé sur la transcription afin d'y simuler des erreurs et d'utiliser les informations qu'elle contient.

5.2.2. *Expériences de référence*

Pour les expériences de référence, nous avons cherché à identifier, aussi clairement que possible, les différents types d'erreurs commises par le système. Notre objectif est d'être capable de mesurer l'effet réel des techniques proposées sur les performances du décodeur. Nous avons effectué un premier décodage avec le modèle de langage générique : le taux d'erreurs mots est de 19.5%. Il y a quatre principales sources d'erreurs :

– les mots hors vocabulaire (MHV) : ces mots n'existent pas dans le lexique du moteur de reconnaissance. Lors du décodage, ils ne peuvent pas être reconnus. Ceci entraîne des effets de bord : avec les modèles probabilistes trigrammes, la non-reconnaissance d'un mot peut provoquer des erreurs sur les mots adjacents. De proche en proche, ces erreurs peuvent théoriquement se propager. Pratiquement, il est assez difficile des mesurer précisement l'impact des MHV sur les performances, sauf en les intégrant au lexique et en comparant les résultats obtenus. Nous avons analysé la transcription exacte pour en extraire les mots absents du lexique initial. Ils représentent 2% des mots. Ces mots ont ensuite été phonétisés et ajoutés au lexique. Le modèle de langage a été réestimé avec ce lexique enrichi.

– la mauvaise segmentation : le système "broadcast news" réalise automatiquement l'ensemble des traitements qui permettent de passer du flux audio brut à la transcription synchronisée. Une émission radiophonique comporte, en plus des segments de parole, des parties musicales, des événements acoustiques divers qui ne sont pas de la parole (bip, jingle, etc.). La première étape du processus global de décodage consiste à extraire les zones de parole du flux audio. Cette segmentation automatique

n'est pas exacte et les segments de parole sont parfois mal isolés. Le moteur cherche alors à décoder une zone qui ne contient pas de parole ou inversement, ne décode pas des zones de parole qui ont été supprimées par erreur par le segmenteur. Quelquesoit la stratégie de décodage utilisée en aval, les erreurs liées à la segmentation restent irrécupérables. De façon à clarifier l'influence des méthodes que nous proposons sur les performances du système, nous avons éliminé cette source d'erreurs incompressible en utilisant la segmentation de référence. Les zones de parole ainsi délimitées sont exactes ; elle correspondent à l'étiquetage réalisé par le concepteur de la base.

En résolvant ces deux problèmes, le taux d'erreurs mot passe de 19.5% à 14.2% sur notre émission de test.

– la linguistique : la qualité du modèle de langage dépend de l'adéquation du corpus d'apprentissage et des conditions d'utilisation du système. Un moteur de reconnaissance susceptible de décoder des messages linguistiquement variés devra utiliser un modèle de langage codant cette variabilité, ce qui nécessite des corpus représentatifs généralement très volumineux. Bien entendu, en augmentant le champ des hypothèses linguistiquement acceptables, on augmente aussi les risques de confusion. Ici, le domaine linguistique peut être réduit puisqu'on dispose d'une transcription exacte ou approchée du discours. Le gain maximal qu'on peut obtenir en réduisant globalement l'espace linguistique peut être estimé en apprenant un modèle de langage sur la transcription exacte elle même. Nous avons donc cherché à évaluer les performances d'un système qui disposerait d'un modèle de langage trigramme parfaitement adapté aux données à traiter. Avec ce modèle estimé sur la transcription exacte du corpus de test, le taux d'erreurs mot passe de 14.2% à 1.9%. Bien entendu, cette expérience ne permet pas d'imputer 12.3% des erreurs au seul modèle de langage, l'exploration du graphe d'hypothèses combinant simultanément les scores acoustiques et linguistiques (la bonne qualité linguistique d'une hypothèse peut compenser sa mauvaise qualité acoustique). De plus, ce taux très bas est obtenu en utilisant la transcription sans erreurs, ce qui ne correspond pas à un contexte d'utilisation réaliste.

– l'acoustique et les heuristiques de décodage : l'algorithme de recherche ne fait pas une exploration exhaustive du graphe d'hypothèses. La complexité d'un tel parcours serait bien trop importante pour des systèmes grand vocabulaire, et un certain nombre d'heuristiques accélèrent l'exploration en écartant des hypothèses jugées très improbables. Dans SPEERAL, les critères permettant de réduire l'espace de recherche sont à la fois d'ordre acoustique et linguistique. Théoriquement, ces coupures doivent introduire très peu d'erreurs de décodage ; cependant, lorsque le contexte acoustique est très mauvais, les meilleures hypothèses (en terme de taux d'erreur) peuvent se trouver exclues du faisceau de recherche. 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 (avec des seuils de coupure standards) pour faire un alignement forcé de la transcription exacte sur le signal. Nous obtenons dans ce cas un taux d'erreur mots ramené à 1.8%. On peut considérer ce niveau d'erreur comme minimal pour toute méthode ré-estimant les hy-

pothèses concurrentes dans le faisceau d'hypothèses sans remettre en cause le contenu même de ce faisceau.

Ces expériences permettent de mettre en place notre cadre expérimental. Le tableau 1 présente l'ensemble de ces expériences qui nous ont servi de référence.

	Taux d'erreur
ML-G sans mots hors-vocabulaire + seg auto	19.5%
ML-G	14.2%
ML-TrEx	1.9%
ML-TrEx + alignement TrEx	1.8%

Tableau 1. *Résultats des expériences de référence, avec ML-G : modèle de langage générique enrichi avec tous les mots hors-vocabulaire et TrEx : alignement sur la transcription exacte*

5.2.3. *Expériences avec modèle de langage appris sur la transcription approchée*

L'objectif de notre recherche s'appuie sur l'alignement de transcriptions approchées. Nous avons récupéré la transcription exacte d'un extrait d'Ester de 15 minutes. Nous y avons introduit manuellement 10% d'erreurs dans la transcription. Notamment, en supprimant toutes les répétitions, hésitations et en changeant la forme de certaines phrases. Nous avons pris soin de garder une forme journalistique correcte pour respecter les conditions classiques d'une émission radiophonique. A partir de cette transcription approchée, nous avons généré un modèle de langage. Un décodage de l'émission avec ce seul modèle de langage donne un taux d'erreurs de 9.6%. Ce taux est proche du taux d'erreurs introduits dans le texte. Le flux est correctement décodé lorsque le modèle de langage reste adapté. Cependant, lorsque les informations sont manquantes, le système ne peut les corriger : la majorité des erreurs de décodage correspondent à celles introduites dans la transcription. Le modèle adapté seul permet de réduire le taux d'erreurs relatif de 32% (cf. tableau 2). Cependant, les erreurs contenues dans la transcription approchée ne sont pas corrigées.

5.2.4. *Expériences avec interpolation de modèles de langage*

Les expériences suivantes ont combiné le modèle de langage appris sur la transcription erronée avec le modèle de langage générique. Il en ressort une légère amélioration des résultats. Le modèle de langage générique permet de corriger certaines erreurs. Cependant, l'amélioration relative du taux d'erreurs n'est que de 7.2% (de 9.6% d'erreurs mots à 8.9%). Si l'on privilégie le modèle de langage générique, le résultat a tendance à se dégrader (10.5% d'erreurs mots). L'interpolation permet d'amener des informations complémentaires, mais encore restreintes (cf. tableau 2).

Nous avons également testé différentes interpolations entre le modèle de langage générique et le modèle appris sur la transcription exacte. Le meilleur résultat est obtenu en donnant un poids plus élevé au modèle de langage appris sur la transcription (2% d'erreurs mots).

	Taux d'erreur
ML-TrEx	1.9%
ML-TrErr	9.6%
ML-G 70% + ML-TrEx 30%	3.6%
ML-G 50% + ML-TrEx 50%	2.4%
ML-G 30% + ML-TrEx 70%	2.0%
ML-G 70% + ML-TrErr 30%	10.5%
ML-G 50% + ML-TrErr 50%	9.2%
ML-G 30% + ML-TrErr 70%	8.9%

Tableau 2. *Résultats des expériences d'interpolation de modèles de langage avec ML-TrEx : modèle de langage appris sur la transcription exacte, ML-TrErr : modèle de langage appris sur la transcription approchée et ML-G : modèle de langage générique*

5.2.5. *Expériences avec interpolation de modèles et alignement*

Après avoir expérimenté des modèles interpolés, nous avons rajouté l'information temporelle sur l'apparition des mots. Ceci permet de palier aux problèmes de perplexité des petits modèles de langage. Pour rajouter cette information temporelle, nous avons intégré une DTW dans l'algorithme de recherche. En plus des modèles de langage interpolés, SPEERAL oriente le décodage en s'alignant si possible sur la transcription approchée (cf. tableau 3).

Les expériences de référence combinent l'interpolation des modèles de langage avec un alignement sur la transcription exacte. Les résultats sont très proches de ceux obtenus avec une simple interpolation de modèles de langages : l'information apportée par l'alignement est redondante avec celle du modèle de langage appris sur la transcription exacte.

Le meilleur résultat est obtenu en combinant le modèle de langage général à 30% avec le modèle appris sur la transcription approchée à 70%. L'alignement améliore le taux d'erreurs mots jusqu'à 8.7%.

	Taux d'erreur
ML-G 70% + ML-TrEx 30% + alignement TrEx	2.5%
ML-G 50% + ML-TrEx 50% + alignement TrEx	2.1%
ML-G 30% + ML-TrEx 70% + alignement TrEx	2.0%
ML-G 70% + ML-TrErr 30% + alignement TrErr	9.4%
ML-G 50% + ML-TrErr 50% + alignement TrErr	8.6%
ML-G 30% + ML-TrErr 70% + alignement TrErr	8.7%

Tableau 3. *Résultats des expériences avec interpolation et alignement forcé avec ML-TrEx : modèle de langage appris sur la transcription exacte, ML-TrErr : modèle de langage appris sur la transcription approchée, ML-G : modèle de langage générique, TrErr : transcription approchée et TrEx : transcription exacte*

5.2.6. *Expériences avec modèle générique et alignement*

Nous avons également fait une expérience sans interpoler les deux modèles de langage, en combinant le modèle générique avec un alignement sur la transcription approchée (cf. tableau 4). Cette expérience montre une amélioration relative beaucoup plus significative de 30% (8.6% d'erreurs à 6.0% d'erreurs) par rapport à l'utilisation d'un modèle de langage appris sur la transcription approchée. L'expérience précédente nous montre donc que l'information amenée par l'interpolation des modèles en plus de l'alignement est redondante. Le système est plus performant s'il utilise un modèle de langage générique accompagné d'un alignement sur la transcription imparfaite. Le modèle de langage générique est ainsi aidé par l'alignement au fur et à mesure. L'alignement permet d'éloigner certaines hypothèses du modèle de langage, et de confirmer les meilleures. Par ailleurs, le modèle de langage permet de corriger les erreurs présentes dans la transcription approchée. L'expérience combinant le modèle de langage appris sur la transcription approchée avec l'alignement sur cette dernière montre que le modèle de langage générique apporte une information nécessaire et supplémentaire pour améliorer le décodage.

La combinaison du modèle de langage générique avec un alignement sur la transcription exacte nous donne un taux d'erreurs mots de 5.5%. Résultat proche de celui avec la transcription approchée. Le modèle de langage générique corrige donc la majorité des erreurs de la transcription approchée, tout en corrigeant les erreurs qu'il aurait fait sans cette dernière.

	Taux d'erreur
ML-G + alignement TrEx	5.5%
ML-G + alignement TrErr	6.0%
ML-TrErr + alignement TrErr	9.3%

Tableau 4. *Résultats des expériences combinant le modèle de langage générique avec alignement forcé avec ML-TrErr : modèle de langage appris sur la transcription approchée, ML-G : modèle de langage générique, TrErr : transcription approchée et TrEx : transcription exacte*

6. Analyse des résultats

6.1. *Modèles de langage appris sur une transcription approchée*

L'ensemble des expériences montre 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 continuera 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 qu'on peut atteindre un taux d'erreur presque incompressible.

6.2. *Interpolation de modèles de langages*

Pour palier aux manques d'informations du modèle de langage appris sur la transcription exacte, l'interpolation avec un modèle de langage générique permet de supprimer une partie des erreurs. Cependant le gain reste faible : les erreurs introduites restent prépondérantes.

6.3. *Alignement et interpolation*

L'alignement permet d'apporter une information non fournie par le modèle de langage : l'information temporelle. L'utilisation d'un alignement DTW associé à l'interpolation des modèles montre à nouveau un léger gain. Cependant, la redondance des erreurs comprises dans l'alignement et le modèle de langage adapté limite l'amélioration du décodage.

6.4. *Modèle de langage générique et interpolation*

Cette expérience a montré 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 et en alignant sur la transcription approchée. 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. Dans ces conditions, le système tire avantageusement profit de la transcription approchée, tout en en respectant les écarts.

7. Conclusion

Nous avons proposé et évalué deux méthodes exploitant l'information contenue dans une transcription imparfaite. 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. Cependant l'information apportée par un modèle de langage adapté sur la transcription imparfaite ne permet pas un décodage correct dans les zones mal transcrites.

Notre seconde approche est d'utiliser l'information temporelle de la transcription en utilisant un algorithme DTW alignant les hypothèses du décodeur sur la transcription approchée. Cette méthode permet de combiner efficacement les scores linguistiques avec les scores d'alignement. Nous avons montré que l'utilisation d'un algorithme DTW couplé à l'algorithme de recherche A^* sur le moteur de reconnaissance SPEERAL permet d'atteindre notre objectif : aligner la transcription quand celle-ci

correspond au flux audio et laisser le modèle de langage reprendre la main lorsqu'on s'en éloigne. Cette méthode permet de corriger une partie des erreurs contenues dans la transcription approchée, et améliore sensiblement les résultats du système de reconnaissance.

Par ailleurs, l'algorithme de décodage A^* est particulièrement bien adapté à ce type d'alignement dynamique, où de très nombreuses hypothèses sont explorées de manière asynchrone. Ainsi, les hypothèses pondérées par l'alignement font converger l'algorithme plus rapidement vers la solution : l'espace de recherche est réduit. Partant d'un taux d'erreurs mots de 14.2%, notre méthode exploitant une transcription imparfaite permet d'améliorer ce taux jusqu'à 6%.

8. Perspectives

Ces premiers résultats montrent l'intérêt d'un alignement sur une transcription approchée. L'ensemble de nos premières expériences a été effectué dans des conditions contrôlées : faible bruit, segmentation de référence et paramétrisation pré-calculée. Nous envisageons d'utiliser nos méthodes dans des conditions plus difficiles : l'apport d'une transcription approchée pourrait alors ammener une plus grande robustesse au système de reconnaissance.

Par ailleurs, nous envisageons de développer des méthodes d'adaptation en ligne, au locuteur, à l'environnement, au thème et au contexte sémantique. Les modèles de langage peuvent être adaptés en fonction de leur synchronisation avec la transcription. L'information de la transcription approchée permet également d'envisager l'adaptation dynamique des modèles acoustiques au fur et à mesure du décodage. L'alignement DTW pourrait être adapté pour fonctionner au niveau acoustique. Actuellement, notre alignement ne s'effectue que sur des mots complets, alors qu'il pourrait agir au niveau des phonèmes proposant en priorité ceux qui correspondent à la transcription.

Nous envisageons également d'adapter SPEERAL pour qu'il fonctionne en flux, et puisse ainsi décoder en temps réel des flux audio pour lesquels il possèdera une transcription approximative. Les premiers résultats obtenus en utilisant l'information de transcriptions approchées montrent que ces objectifs sont envisageables. Nous préparons actuellement une maquette de ce type de système, dans le cadre d'une aide aux personnes malentendantes, afin d'automatiser le sous-titrage d'évènements culturels (théâtre, cinéma).

9. Bibliographie

[ALI] Alize toolkit, http ://www.lia.univ-avignon.fr/heberges/ALIZE/.

[BAR 96] BARRAS C., « Reconnaissance de la parole continue : adaptation au locuteur et contrôle temporel dans les modèles de Markov cachés », Thèse de doctorat, Université de Paris VI, 1996.

[BER 94] BERNDT D., CLIFFORD J., « Using Dynamic Time Warping to find Patterns in Time Series », , 1994, in Proc. of AAAI, Workshop on Knowledge Discovery in Databases, Seattle, Washington.

[CHE 04] CHEN L., GAUVAIN J.-L., LAMEL L., ADDA G., « Dynamic Language Modeling for Broadcast News », , 2004, CNRS-LIMSI.

[CHI 03] CHIH-WEI H., « Automatic closed caption alignment based on speech recognition transcripts », , 2003.

[GEO 05] GEOFFROIS E., « Bilan d'ESTER 2005 », Bilan de la campagne d'évaluation, 2005, LIA.

[JAN 99] JANG P. J., G.HAUPTMANN A., « Improving acoustic models with captioned multimedia speech », *IEEE International Conference on Multimedia Computing and Systems, Florence, Italy*, , 1999, Carnegie Mellon University.

[JON 00] JONGMOK S., JINWOONG K., KYUNGOK K., KEUNSUNG B., « Application of Speech Recognition with Closed Caption for Content-Based Video Segmentation », *IEEE DSPWorkshop*, , 2000, Kyungpook National University, Taegu, Korea.

[KEO 00] KEOG E., M.PAZZANI, « Scaling up Dynamic Time Warping for Datamining Applications », , 2000, in Proc. of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data mining, Boston.

[MAS 00] MASSONIÉ D., « Adaptation du modèle de langage par recherche dynamique d'information », rapport de dea, 2000, LIA.

[MOR 98] MORENO P. J., JOERG C., THONG J.-M. V., GLICKMAN O., « A recursive algorithm for the forced alignment of very long audio segments », *International Conference on Spoken Language Processing*, , 1998, Cambridge Research Laboratory and Compaq Computer Corporation.

[NOC 02] NOCERA P., LINARES G., MASSONIÉ D., « Principes et performances du décodeur parole continue Speeral », *XXIVèmes journées d'étude sur la parole*, , 2002, Laboratoire Informatique d'Avignon.

[PLA 96] PLACEWAY P., LAFFERTY J., « Cheating with imperfect Transcripts », *Proceedings of ICSLP*, , 1996, School of Computer Science Carnegie Mellon University.