Le système de transcription du LIA pour ESTER-2005

G. Linares, P. Nocera, D. Matrouf, F. Béchet, D. Massonié, C. Fredouille

Laboratoire Informatique d'Avignon
Université d'Avignon et des Pays de Vaucluse
georges.linares@univ-avignon.fr

Abstract

Nous présentons le système de transcription automatique d'émissions radiophoniques du LIA. Il s'agit d'un système complet, intégrant une première phase de segmentation et d'extraction d'informations de haut niveau ; cette étape préliminaire est suivie du processus de décodage lui même, réalisé par le système de reconnaissance grand vocabulaire du LIA (Speeral).
Les configurations du système *non contrainte* (tâche *TRS*) puis *temps réel* (tâche *TTR*) qui ont été évaluées lors de la campagne ESTER sont décrites successivement. Les résultats obtenus sont satisfaisants ; ils valident les choix qui ont été faits et permettent d'identifier des axes de travail prioritaires.

1. Introduction

Outre les problèmes classiques de reconnaissance de la parole, la transcription d'émissions radiophonique pose des problèmes spécifiques liés notamment au format des émissions et à la diversité des conditions. Ces particularités nécessitent des traitements adaptés, destinés à la fois à extraire des informations utiles au décodage et à mettre le signal audio brut sous une forme compatible avec les capacités du système de reconnaissance.
Le système de transcription du LIA intègre une chaîne complète de traitements permettant de passer du signal audio brut à la transcription orthographique finale.
L'architecture générale du système est décrite dans une première partie de cet article, puis les phases de pré-traitement sont détaillées. Une seconde partie présente le système de reconnaissance de la parole (algorithmes et ressources) qui est au cœur du système de transcription. Enfin, les résultats obtenus dans le cadre de la campagne ESTER sont analysés et discutés.

2. Architecture générale de la machine de transcription

La chaîne de traitement comporte deux parties relativement indépendantes. La première réalise un ensemble de pré-traitements destinés à préparer le signal en vue du décodage ultérieur. Par ailleurs, cette phase de pré-traitement extrait des informations de haut niveau, à partir desquelles le système de reconnaissance sera configuré.
Les zones de parole sont d'abord isolées du flux audio ; un segmenteur bande-large/bande-étroite permet ensuite d'identifier les zones de parole téléphonique ou assimilées (enregistrement en bande étroite).
Les segments isolés par cette première segmentation en macro-classes acoustiques sont fréquemment d'une taille incompatible avec les capacités des algorithmes de reconnaissance ; par ailleurs, l'identification des locuteurs réguliers peut permettre d'appliquer des traitements spécifiques au locuteur. Nous re-segmentons donc les zones extraites par un algorithme de segmentation en locuteur.
Enfin, un système de suivi de locuteur reconnaît, parmi les groupes de segments issus de l'étape de segmentation en locuteurs, ceux qui sont prononcés par des locuteurs réguliers des stations de radios du corpus d'apprentissage. Cet étiquetage des segments affectés à des locuteurs référencés nous permettra d'utiliser des modèles acoustiques spécifiques pendant la transcription.
Le deuxième phase effectue la reconnaissance proprement dite. Une première passe génére une transcription intermédiaire, en utilisant les modèles dépendant du locuteur lorsqu'ils existent (ie lorsque le locuteur est référencé dans un dictionnaire constitué à priori).
Ce premier décodage permet d'appliquer une adaptation aveugle des modèles acoustiques ; une deuxième passe utilise ces modèles adaptés pour produire la transcription finale.

3. Segmentation

3.1. Segmentation en macro-classes acoustiques

Cette première passe de segmentation extrait la parole du flux audio, puis re-segmente les zones de parole en bande-large/bande-étroite. Ce système est évalué par ailleurs dans la campagne ESTER (tâche SES), et décrit dans [1]

3.2. Segmentation en locuteur

Il s'agit du système de segmentation du LIA participant à la tâche SRL. Il repose sur un mécanisme d'identification incrémentale de nouveaux locuteurs [2].
Par ailleurs, un système secondaire (*TRS-2)*, utilise un algorithme de segmentation rapide fourni par le LIUM. Il s'agit d'un segmenteur mono-gaussien basé sur un critère BIC [Meignier].

3.3. Suivi de locuteur

Un bonne partie des données d'apprentissage est produite par les intervenants « réguliers » des stations de radios du corpus. Cette observation nous à conduit à construire un dictionnaire des locuteurs les plus fréquents et à entraîner des modèles acoustiques spécifiques pour chacun d'eux.
La phase de suivi de locuteur doit permettre d'identifier les segments réalisés par ces locuteurs.
Le dictionnaire est constitué des 100 locuteurs les plus fréquents du corpus ; chacun d'eux est modélisé par un GMM estimé par adaptation MAP d'un modèle du monde appris sur l'ensemble des données d'apprentissage. Un segment est attribué au locuteur i si le rapport des vraisemblances R du segment S sachant le modèle G_i sur la vraisemblance sachant le modèle du monde G_w est supérieur à un seuil t fixé à priori

$$R = \frac{LogL(S|G_i)}{LogL(S|G_w)} \geq t$$

3.4. Recouvrement

Le traitement, par un système de reconnaissance, des segments issus des premières phases de segmentation du flux audio peut poser un certain nombre de problèmes techniques. D'une part, une séquence trop longue peut amener le décodeur à demander des ressources mémoire importantes; en limitant la taille des segments à une valeur compatible avec les capacités de la machine de transcription, on introduit des discontinuités linguistiques, voire des coupures intra mot, qui peuvent se révéler préjudiciables à la qualité du décodage. Pour limiter ces effets de bord, les segments adjacents issus d'un même locuteur sont augmentés de façon à se recouvrir de 3 secondes., la longueur maximale d'un segment étant limitée à 30 secondes. Un post traitement fusionnera les hypothèses concurrentes en fin de décodage.

4. Transcription non contrainte

4.1. Paramétrisation

Nous avons utilisé une paramétrisation *PLP* 12 coefficients, plus l'énergie, et les dérivées premières et secondes de ces 13 coefficients. Après la segmentation en locuteurs, les paramètres sont centrés et réduits sur une fenêtre glissante de 500 ms. Les outils utilisés lors du pré-traitement acoustique ont été développés au LIA.

4.2. Les modèles acoustiques

Le système utilise des modèles acoustiques contextuels appris sur l'ensemble des données transcrites du corpus ESTER, corpus de développement (DEV) compris. Il s'agit de triphones avec partage d'états. La classification des états est faite à priori, par des arbres de décisions dont les questions portent sur le contexte acoustique de l'état à modéliser. Tous les modèles utilisés sont des GMM d'au plus 64 gaussiennes. L'ensemble regroupe environ 3600 états émetteurs pour environ 10000 HMM et 230 000 gaussiennes.
Nous avons utilisé 2 modèles génériques : un modèle large-bande et un modèle bande-étroite, tous deux indépendants du genre.
Les modèles dépendants du locuteur sont estimés par une double transformation. La première repose sur une MLLR supervisée classique. Les modèles intermédiaires issus de la MLLR sont ensuite transformés selon une méthode d'adaptation structurelle développée au LIA (SGMAP, [3]).
Tous les outils utilisés pour l'apprentissage et l'adaptation des HMM ont été développés au LIA.

4.3. Lexique et modèles de langage

Les ressources linguistiques sont extraites de deux corpus :

- journal Le Monde de 1987 à 2003 (330 Millions de mots)
- corpus d'apprentissage ESTER (960K mots)

Le lexique utilisé contient 65K mots. Il est composé de tous les mots du corpus ESTER, ainsi que des mots les plus fréquents du corpus Le Monde 1987-2003. La phonétisation provient de 2 sources, le lexique phonétique ILPHO et le phonétiseur LIA_PHON pour les mots inconnus de ILPHO. Les formes générées par LIA_PHON ont été partiellement vérifiées et corrigées manuellement (en particulier pour les noms propres).
Le modèle de langage a été appris sur les corpus Le Monde et ESTER à l'aide de la boite à outils SRILM. Il est obtenu par uNe combinaison linéaire de trois modèles ; le premier a été appris sur les données du *Monde 1987-2002*, le second sur *Le Monde 2002-2003*, et le dernier sur le corpus *ESTER*. Tous ces modèles sont des modèles trigrammes avec méthode de backoff Kneser-Ney modifiée (vocabulaire ouvert). Les coupes sur les ngrammes ont été fixées à 0 pour les modèles appris sur ESTER, à 1 pour les bigrammes du Monde et à 2 pour les trigrammes du Monde.
Enfin, ces modèles sont mélangés au sein d'un même modèle avec des coefficients d'interpolation déterminés par l'entropie des modèles sur le corpus de développement d'ESTER (poids relatifs de *0.41, 0.24 et 0.35* pour respectivement *Le Monde 1987-2002, Le Monde 2002-2003 , ESTER*).

4.4. Moteur de reconnaissance

4.4.1. Principe

Il s'agit d'un décodeur à pile, asynchrone, dérivé d'un algorithme de recherche A*. Il utilise des modèles de Markov contextuels pour la modélisation acoustique et des modèles de langage trigrammes. Les contextes acoustiques inter-mots sont utilisés.
Dans le cadre de la tâche TRS de la campagne ESTER, deux passes sont effectuées; la première produit une transcription intermédiaire qui permet une adaptation aveugle des modèles acoustiques. La seconde passe génère, à partir de ces modèles adaptés, la transcription finale.

4.4.2. Algorithme de recherche

Le moteur est fondé sur un algorithme de recherche A* opérant sur un treillis de phonèmes. La fonction d'estimation somme, à chaque nœud du graphe, les coûts du chemin exploré et une sonde minimisant le coût des chemins finaux. La qualité de cette sonde est déterminante pour les performances de l'algorithme de recherche. Dans Speeral, elle est composée d'une partie purement acoustique et d'un terme d'anticipation linguistique. L'approximation des scores acoustiques se fait à partir d'un Viterbi arrière sur les modèles non-contextuels, puis par une re-estimation du score obtenu par des modèles pseudo-contextuels [5]. L'anticipation linguistique est basée sur une estimation des meilleurs trigrammes prolongeant l'hyptohèse explorée.
Par ailleurs, un certains nombre d'heuristiques permettent de réduire la dimension de l'espace de recherche, les retours arrières et les explorations répétées de parties du graphe.

4.5. Expériences

Nous avons d'abord évalué les performances du système (dans sa configuration TRS-2) sur le corpus de développement. Nous obtenons un taux d'erreur mot de 22.3% à l'issu de la première passe, qui descend à 21.2% lors de la seconde passe.

Table 1: Résultats des passes 1 et 2 du système secondaire (TRS-2) sur le corpus de développement

Système	WER - PASSE 1	WER –PASSE 2
TRS-2	22,30%	21,20%

Le système a ensuite été évalué sur le corpus de test. Pour ce test, nous avons intégré le corpus de développement dans l'apprentissage des modèles acoustiques et du modèle de langage, qui sont chronologiquement plus proches de la période du test que le corpus d'apprentissage. Cet apport de données permet une réduction de 0.7% et 1.1% de taux d'erreur mot sur les systèmes TRS-1 et TRS-2.

Table 2: Résultats des systèmes primaire (TRS-1) et secondaire (TRS-2) sur le corpus test sans données de développement dans l'apprentissage, puis avec ('TRS-1-DEV, TRS-2-DEV).

Système	WER – PASSE 1	WER – PASSE 2
TRS-1	29,40%	28,00%
TRS-2	29,60%	27,20%
TRS-1-DEV	/	26,90%
TRS-2-DEV	/	26,50%

4.6. Discussion

L'analyse des résultats montre d'abord un écart important entre les scores obtenus sur le DEV et sur le TEST; l'écart relatif est de l'ordre de 25% d'erreurs supplémentaires. Plusieurs points expliquent cette baisse des performances :

- le taux de mots hors vocabulaire est plus élévé sur le test (de l'ordre de 1.5% pour 0.5% sur le dev)
- la couverture par les modèles spécifiques est bien plus faible : 45 % des données de développement étaient traitées par des modèles dépendants du locuteur, contre seulement 8% dans le test;
- une quantité significative des donnés de test sont des enregistrements de mauvaise qualité, en bande étroite mais contenant un bruit parasite dans les parties hautes du spectre. Notre segmenteur bande-étroite n'a pas étiqueté correctement ces signaux parasités; nous avons menés un certain nombre d'expériences complémentaires sur cet aspect particulier. En améliorant l'apprentissage des GMM dédiés à l'identification des segments bande-étroite et en filtrant la partie haute du spectre de ces signaux, ont obtient un gain compris entre 3 et 7% (WER absolu) sur les émissions de la Radio Télévision Marocaine, qui sont les plus touchées par ce phénomène.

5. Système temps-réel

5.1. Généralités

Il s'agit du système primaire du LIA engagé dans la tâche TTR. Il dérive directement du système secondaire (TRS-2) décrit au chapitre précédent. Les différences majeures portent sur 6 points :

- les ressources : les corpus utilisés sont les mêmes, mais les modèles sont plus compacts : modèles acoustiques à 60K gaussiennes au lieu des 230k du système TRS-1; le lexique est réduit à 27K mots, au lieu des 65K du système TRS.
- la segmentation est fournie par le LIUM. Il s'agit d'une méthode très rapide, l'approche adoptée au LIA pour la tâche SRL n'autorisant pas le temps réel.
- nous utilisons une méthode de calcul et d'approximation rapide des vraisemblances; cette méthode, développée au LIA, permet de limiter le cacul des vraisemblance à environ 8% du nombre de gaussiennes initial sans perte significative des performances [6]. De plus, l'optimisation des fonctions de calcul des vraisemblances à permis de réduire encore le temps de calcul des vraisemblances d'un facteur 2 (assembleur, utilisation d'instructions SIMMD).
- la sonde linguistique intègre un certain nombre d'heuristiques visant à limiter le nombre de trigrammes évalués; ces techniques ont permis de réduire d'environ 20% le temps de décodage total du système de temps-réel.
- une coupure suplémentaire sur la taille du fasisceau d'hyptohèses à été introduite; elle est basé sur les scroes issus de la sonde acosutique,
- une seule passe est effectuée; il n'y a pas l'adaptation non supervisée réalisée par le système TRS

5.2. Expériences

Les paramètres du système ont été calibrés sur le corpus de développement de façon à obtenir un décodage temps réel du corpus de développement sur une machine ordinaire, en l'occurence un Pentium IV à 3Ghz équipé d'un Go de mémoire vive. Ce système a ensuite été évalué sur le corpus de test.
Il s'agit d'une première expérience dans le cadre du temps réel. Les résultats nous semblent très prometteurs : l'écart absolu avec le système TRS en première passe est de l'ordre de 6.1%, et passe à 7.3% après adaptation non-supervisée (effectuée uniquement sur le système TRS).
Ce résultat valide la stratégie de décodage globale, est celle du système standard, les modifications apportées pour le système temps réel visant à accélérer différentes parties du processus en préservant le modèle fonctionnel.
Par ailleur, les résultats comparés sur le corpus de développement et celui de test montrent un dégradation assez importante des taux d'erreur mots, acompagnée d'une durée de décodage augmentée d'environ 23%. Les causes de cette dégradation coorespondent à celles évoquées pour le système TRS. Le ralentissement du décodage est, quand à lui, lié aux stratégies de coupures dynamiques et à l'algorithme de calcul des vraisemblances :

- les conditions acoustiques moins favorables dégradent les performances de la sonde, et augmente la dimension de l'espace de recherche,
- l'algorithme de sélection de gaussienne réalise un bi-partitionnement dynamique de l'ensemble des gaussiennes sur un critère de minimisation de l'erreur d'approximation. La dispersion des formes acoustiques conduit à une sélection plus large des vraisemblances à calculer.

Ces approches dynamiques doivent permettre au système de s'adapter à des conditions adverses en limitant la baisse des performances, au prix d'une exploration plus profonde de l'espace de recherche.

Table 3: Résultats du système temps-réel sur les corpus de développement et de test.

	DEV	TEST
WER	28,40%	36,80%
RT Ratio	0.99%	1.23%

6. Conclusions

Les résultats obtenus valident l'ensemble de la plateforme de reconnaissance de la parole développée au LIA, à la fois en terme d'outils logiciels et d'approches algorithmiques : les principaux outils utilisés tout au long de la chaîne de traitement sont issus du LIA, de la segmentation au décodage lui-même. De plus, le coeur du système de transcription repose sur une stratégie de décodage spécifique et intégrant des techniques originales développées localement (adaptation structurelle, calcul rapide des vraisemblances, etc.).
Cependant, les résultats obtenus permettent d'identifier des directions à privilégier pour continuer à développer le système ; l'intégration de techniques d'apprentissage discriminant (type MMI ou MPE) permettrait sans doute d'améliorer la robustesse. D'autre part, nos expériences montrent un certaine sensibilité du système aux données d'apprentissage. Les performances gagneraient probablement à l'intégration des données plus récentes et plus variées; de ce point de vue, l'exploitation des données non-transcrites fournies à l'occasion de la campagne sont probablement une piste à explorer.
Enfin, les premiers résultats obtenus par la version temps réel de Speeral montrent le potentiel de la stratégie de décodage adoptée. Ce type d'approche est généralement réservé à des systèmes sans contraintes fortes en terme de ressources matérielles (mémoire et CPU). L'intégration d'heuristiques dans l'agorithme de recherche et l'utilisation d'une sonde performante à la fois en précision et en vitesse permet d'atteindre le temps réel en préservant l'intérêt de la transcription.

7. References

[1] **Scheffer N., Istrate D., Fredouille C., Bonastre J.F., "Les systèmes du LIA pour les tâches de segmentation et de suivi de locuteur : SES, SRL, SVL", Actes du Workshop Ester, Mars 2005**
.

[2] **M**eignier, S. and Moraru, D. and Fredouille, C. and Besacier, L. and Bonastre, J.-F, "Benefits of prior acoustic segmentation for automatic speaker segmentation", ICASSP-04, Mai 2004,Montreal, Canada,

[3] D. Matrouf, O. Bellot, P. Nocera, G. Linares, J.-F. Bonastre, "Structural Linear Model-Space Transformations for Speaker Adaptation", 2003 Eurospeech 2003, Genève

[4] Nocera P. Linares G., D. Massonié D., Lefort L."Phoneme lattice based A* search algorithm for speech recognition", Sept. 2002, Brno, TSD2002

[5] **Linares G., Nocera P., Matrouf D.**, "**Partitionnement dynamique des distributions pour le calcul des vraisemblances dans un DAP Markovien"**, JEP, Juin 2000 , Aussois