Participation du CLIPS aux Evaluations ESTER : Tâches TRS et SRL

Laurent Besacier, Brigitte Bigi, Daniel Moraru, Richard Lamy

Laboratoire CLIPS / IMAG
Université Joseph Fourier, BP 53 - 38041 GRENOBLE Cedex 9, France
Tél.: ++33 (0)4 76 63 56 95- Fax: ++33 (0)4 76 63 55 52
Mél: laurent.besacier@imag.fr

Avertissement : Ceci est un document de travail très descriptif qui résume les systèmes présentés par le CLIPS à la campagne d'évaluation ESTER pour les tâches TRS et SRL et présente les performances associés.

1. INTRODUCTION

La campagne d'évaluation ESTER[1] vise à l'évaluation des performances des systèmes de transcription d'émissions radiophoniques. Les transcriptions seront enrichies par un ensemble d'informations annexes, comme le découpage automatique en tours de paroles, le marquage des entités nommées, etc. Cette campagne est organisée dans le cadre du projet EVALDA sous l'égide scientifique de l'Association Francophone de la Communication Parlée avec le concours de la Délégation Générale de l'Armement et de ELDA.
Le laboratoire CLIPS s'est engagé à participer à certaines tâches de cette campagne d'évaluation. Cet article décrit les travaux du laboratoire sur les données radiophoniques, dans les tâches de transcription automatique (TRS, section 2), et de segmentation et regroupement de locuteurs (SRL, section 3). Nous présentons également les résultats obtenus sur ces tâches lors de l'évaluation finale qui a eu lieu en Janvier 2005.
La section 2 de cet article présente le système complet de transcription automatique (TRS), en décrivant tout d'abord le segmenteur, puis le vocabulaire, le modèle de langage, le modèle acoustique, et enfin les résultats. La section 3 s'attachera à décrire la segmentation en locuteurs (SRL) et les résultats associés.

2. TRANSCRIPTION AUTOMATIQUE (TRS)

Le signal à transcrire est d'abord segmenté automatiquement en morceaux de petite taille (découpage du signal). Nous catégorisons ensuite chaque morceau de signal avec une étiquette de qualité : PA (parole) ou MU (musique seule). L'étiquette MU est utilisée pour retirer les zones de musique de l'ensemble des signaux à décoder. Le décodeur, qui utilise la boîte à outils Janus 3.2 [1], est ensuite appliqué sur chaque morceau de signal, avec les modèles acoustiques et de langage décrits plus loin dans les sections 2.3 et 2.4. Avec la taille du treillis choisie finalement pour l'évaluation, le système final est environ 40 fois temps-réel.

2.1. Segmenteur

Découpage automatique du signal

Nous utilisons tout d'abord un détecteur de silence pour trouver, sur le signal à traiter, toutes les zones contenant au moins 0,3s de silence. Pour cela, nous utilisons l'outil « audioseg » mis à disposition par l'IRISA. Ensuite, le signal est découpé en morceaux (ou utterances) en utilisant ces zones de silence comme séparateurs. Si il existe des zones plus longues que 30s, nous réalisons à nouveau un découpage de ces zones en détectant les silences d'au moins 0.15s. Ainsi, sur des signaux d'un heure, nous obtenons en général un peu moins de 1000 utterances, ce qui semble cohérent avec le nombre d'utterances des transcriptions manuelles.

Segmentation en qualité

Nous utilisons une méthode proche de la quantification vectorielle (VQ) [2]. Cette segmentation s'applique sur le signal complet non découpé en utterances. Cependant, des étiquettes de qualité sont ensuite attribuées, en fonction de cette segmentation, à chaque utterance issue de l'étape de découpage du signal. Cette segmentation étiquette le signal suivant deux qualités : PA (parole) ou MU (musique seule). Des détails sur cette segmentation se trouvent dans [3].

2.2. Vocabulaire et dictionnaire phonétique

La couverture lexicale du vocabulaire doit être la plus élevée possible afin de limiter le nombre de mots hors vocabulaire (OOV). Il faut donc définir un vocabulaire de taille importante. Cependant, une taille trop grande de vocabulaire peut amener différents problèmes :
-le corpus d'apprentissage du modèle de langage doit être d'autant plus grand. En effet, le problème de la couverture lexicale est un sous-problème de la couverture en n-grams du modèle de langage,
-des mots trop proches acoustiquement vont certainement provoquer des erreurs de substitution qui risquent également d'engendrer des erreurs sur les mots qui les suivent.

Pour l'évaluation finale, notre vocabulaire est composé de 21298 mots, extraits de la façon suivante :

- 16800 mots les plus fréquents du corpus LE_MONDE des années 2001 à 2003.

- Rajout des mots de ESTER TRAIN (> 2 occurences)
- Rajout des mots de ESTER DEV (> 1 occurence)

Précisons que ceci est le vocabulaire final utilisé pour l'évaluation sur les données de TEST ESTER. Lorsque dans cet article, nous décrirons des résultats obtenus sur

[1] http://www.afcp-parole.org/ester/index.html

les données de DEV ESTER, le vocabulaire sera légèrement différent puisqu'il ne contiendra pas alors les mots issus du DEV.
Au final, le nombre de mots hors vocabulaire sur les données de test, avec notre vocabulaire de 21298 mots est de 3.61%. Pour plus de précisions sur les critères d'extraction du vocabulaire, voir [7].

Les mots ont ensuite été phonétisés. La phonétisation est issue de BDLex pour les mots connus. Pour les mots inconnus de BDLex, le phonétiseur du LIA (liaphon) est utilisé et chaque phonétisation est ensuite vérifiée et/ou complétée manuellement. Chaque mot est décrit comme une suite d'unités phonétiques, choisies parmi 43 classes pré-définies. Au final, on obtient 38000 entrées dans le dictionnaire de prononciation, représentant les 21298 mots du vocabulaire.

2.3. Modèles de langage

Nettoyage des corpus

Dans un premier temps, nous débruitons le corpus et ajoutons des marqueurs spécifiques, tels que les débuts et fins d'articles, de paragraphes, de thèmes et le titre. Puis, nous normalisons la notation des caractères (l'encodage choisi est iso-8859-1), et ajoutons les marques de début et de fin de phrases, respectivement notées <s> et </s>. Le corpus est alors segmenté en "mots", selon des règles d'usage et selon un lexique le plus complet possible. Ensuite, certaines notations, notamment les notations chiffrées, sont converties en leur forme textuelle. Les expressions du langage ainsi que les noms de personnes sont alors regroupées. Finalement, le corpus est réduit en minuscules et nous avons choisi de supprimer la ponctuation.

Apprentissage

En plus des transcriptions radio, nous utilisons les années 1992 à 2003 (environ 1.8 Go de données nettoyées) du journal « Le Monde ». Ce corpus est filtré afin de sélectionner les phrases qui contiennent uniquement des mots du vocabulaire (fixé dans la section 2.2).

Au final, notre modèle de langage utilisé pour l'évaluation finale résulte d'une interpolation entre un modèle appris sur les données LE_MONDE (1992-2003), un modèle appris sur les données ESTER (TRAIN+DEV) et un modèle appris sur le Web (données collectées sur des sites de journaux et radios entre juin 2003 et mars 2004) :

ML-final= 0.52*LE_MONDE92_03 +
0.4*TRANS_ESTER +
0.08*WEB_Juin03_Mars04

L'apprentissage des modèles de langage a été réalisé avec la boîte à outils SRILM [4].
Dans le cas du système testé sur les données de DEV, les données ESTER (TRAIN+DEV) sont remplacées par les données ESTER (TRAIN) uniquement.

2.4. Modèles acoustiques

Les vecteurs de paramètres sont de dimension 24 et résultent d'une transformation LDA réalisée à partir d'un espace de 42 paramètres acoustiques (13 MFCC, 13 ΔMFCC, 13 ΔΔMFCC, ΔE, ΔΔE, zero-crossing) extraits toutes les 10ms. Une soustraction de moyenne cepstrale (CMS) est appliquée au préalable sur les coefficients MFCC et le paramètre statique de l'énergie n'est pas utilisé. Les modèles acoustiques dépendants du contexte sont appris uniquement sur le corpus d'apprentissage de la phase 2 d'ESTER (162 fichiers représentant 82h de signal). Nous apprenons un seul modèle « multi-condition » indépendant de la qualité du signal et du sexe du locuteur. Chaque modèle est un HMM de 3 états (sauf pour le silence qui a 4 états) avec 16 gaussiennes par état. Les distributions multigaussiennes sont partagées entre différents états de différents HMM (*tying*) et le nombre total de distributions multigaussiennes (i.e. d'états) réellement utilisées est environ 1500 (soit 24000 gaussiennes au total environ).

2.5 Résultats sur la tâche TRS

Progression depuis le test à blanc

Le tableau 1 illustre la progression de notre système depuis le test à blanc [3], les tests étant effectués sur la partir DEV de la phase 2 d'ESTER (8h).

Système	Test à blanc (1)	Nouveau ML (2)	Nouveau MA (3)	Final Eval (4)
%WER	45.2	42.4	38.2	36.7

Tableau 1 : Progression de notre système depuis le test à blanc (tests effectués sur DEV ESTER2, 8h de parole)

Le système (1) correspond à celui développé pour le test à blanc (décrit dans [1]) et testé sur les données de DEV ESTER Phase 2. Le système (2) est le même mais avec un nouveau modèle de langage et un nouveau vocabulaire, adaptés avec les transcriptions de TRAIN ESTER Phase2. Le système (3) correspond au système (2) mais avec de nouveaux modèles acoustiques, réappris sur les 82h du TRAIN ESTER Phase2. La progression observée à ce niveau là peut paraître un peu décevante. Finalement, le système (4) correspond à une optimisation du système (3) en terme de taille de treillis et pondérations ML/MA. Ce système (4) évalué sur le DEV est celui présenté officiellement pour l'évaluation de Janvier 2005 à une petite différence près : l'ajout d'informations issues des transcriptions de DEV au vocabulaire et au modèle de langage, comme cela est décrit dans la section 2.3.

Résultats TRS du CLIPS sur le test final (10h test)

Le tableau 2 présente les résultats officiels du CLIPS (système primaire) pour l'évaluation finale de la tâche

TRS. Les performances obtenues (40.9%WER) sont légèrement moins bonnes que celles obtenues sur les données de DEV (36.7%WER). Il semblerait que les données issues de la RTM notamment, soient beaucoup plus difficiles que dans le cas du DEV.
De façon assez surprenante, nous avons par ailleurs observé que notre système secondaire, qui n'utilise aucune détection de zones de musique pure, obtient exactement les mêmes performances que le système primaire. Ceci peut s'expliquer par le faible nombre de zones de musique pure prises en compte pour l'évaluation, et par la qualité assez moyenne de notre détecteur qui n'a pas été retouché depuis le test à blanc.

Signal	%WER
20041008_1800_1830_INFO_DGA	38,1
20041012_1800_1830_INFO_DGA	37,6
20041013_1700_1800_INFO_DGA	36,6
20041007_0800_0900_INTER_DGA	35,2
20041011_1300_1400_INTER_DGA	46,1
20041006_0700_0800_CLASSIQUE	34,7
20041006_0800_0900_CULTURE	45,9
20041025_1930_2000_RFI_ELDA	40,6
20041026_1930_2000_RFI_ELDA	48,6
20041027_1230_1300_RFI_ELDA	36,6
20041124_1230_1300_RFI_ELDA	34,6
20041217_1300_1322_RTM_ELDA	44,1
20041218_1300_1314_RTM_ELDA	40,9
20041219_1300_1314_RTM_ELDA	44,9
20041220_1300_1314_RTM_ELDA	47,7
20041221_1300_1321_RTM_ELDA	45,8
20041222_1300_1320_RTM_ELDA	50,5
20041223_1300_1318_RTM_ELDA	51,7
moyenne	**40,9**

Tableau 2 : résultat officiel du système primaire du CLIPS sur la tâche TRS (10h de test)

3. SEGMENTATION EN LOCUTEURS (SRL)

3.1. Système

Le système présenté par le CLIPS [5,6] repose sur une détection de changement de locuteurs, suivie d'un procédé de regroupement (clustering) hiérarchique. La détection des changement de locuteurs est effectuée par utilisation du critère BIC (Bayesian Information Criterion), à l'aide de fenêtres glissantes adjacentes (de 1,75 s.). Les fenêtres sont modélisées par des gaussiennes à matrices diagonales. Un procédé de seuillage permet de sélectionner les points de changement les plus vraisemblables. L'étape de clustering commence par l'apprentissage d'un modèle du monde GMM à 32 composantes diagonales, en utilisant le fichier complet et en maximisant le critère ML (Maximum Likelihood). Les modèles de chaque segment sont alors adaptés, à partir du modèle du monde, par MAP (les moyennes seules sont adaptées). Ensuite, des distances BIC sont calculées entre les segments pour fusionner les deux plus proches, jusqu'à obtenir N modèles (i.e. N locuteurs).
Le nombre de locuteurs présents dans la conversation (NSp) est estimé automatiquement à l'aide d'un score BIC pénalisé. Le nombre de locuteurs est contraint entre 1 et 40. La limite supérieure est ajustée en fonction de la durée du document. Le nombre de locuteurs (NSp) maximise :

$$BIC(M)=\log L(X;M)-\lambda\frac{m}{2}NSp\log NX$$

où M est le modèle composé des NSp modèles de locuteur détectés, NX est le nombre total de trames (de parole) du document, m est un paramètre dépendant de la complexité des modèles et λ un paramètre de réglage expérimentalement fixé à 0.6. La légère différence entre notre système primaire et notre système secondaire réside dans cette estimation du nombre de locuteurs. Pour le système primaire, nous appliquons une pénalisation du nombre de locuteurs estimé dans le cas des fichiers courts de la RTM : dans ce cas, le nombre de locuteurs finalement estimé correspond au nombre de locuteurs estimés par l'équation ci-dessus (BIC) moins trois (ce détail nous fait gagner environ 3% d'erreur sur les données DEV). Pour le système secondaire : le nombre de locuteurs est estimé uniquement avec l'équation donnée plus haut (BIC).

Les silences et les zones de musiques ne sont pas retirés dans les résultats du système soumis pour l'évaluation, puisque des expériences préliminaires conduites sur les données de DEV (avec les anciens fichiers .uem) avaient montré que ceci dégradait légèrement les résultats. Le système n'a pas du tout été optimisé et tourne en 30 fois le temps réel environ pour des fichiers d'une heure (plus vite pour les fichiers plus courts).

3.2. Résultats sur la tâche SRL

Le tableau 3 présente un bilan des performances de notre système de segmentation (toujours le même), testé sur les données du test à blanc, de DEV, et de TEST d'ESTER.

Données de test	Test à blanc (4h40)	DEV ESTER2 (8h)	TST ESTER2 (10h)
%err. seg.	17.7	19.8	24.3

Tableau 3 : évolution des performances du système TRS CLIPS sur ESTER.

Nous observons une dégradation des résultats sur les données bien que notre système demeure inchangé. Nous n'avons pas encore analysé en détail les données de TST pour comprendre les raisons de cette dégradation.

Le tableau 4 présente quand à lui les résultats officiels du CLIPS (système primaire) pour l'évaluation finale de la tâche SRL.

Signal	%err seg
20041008_1800_1830_INFO_DGA	11.98
20041012_1800_1830_INFO_DGA	23.77
20041013_1700_1800_INFO_DGA	27.63

20041007_0800_0900_INTER_DGA	37.85
20041011_1300_1400_INTER_DGA	24.29
20041006_0700_0800_CLASSIQUE	22.14
20041006_0800_0900_CULTURE	36.46
20041025_1930_2000_RFI_ELDA	25.50
20041026_1930_2000_RFI_ELDA	34.29
20041027_1230_1300_RFI_ELDA	9.77
20041124_1230_1300_RFI_ELDA	26.93
20041217_1300_1322_RTM_ELDA	21.14
20041218_1300_1314_RTM_ELDA	1.76
20041219_1300_1314_RTM_ELDA	21.09
20041220_1300_1314_RTM_ELDA	1.54
20041221_1300_1321_RTM_ELDA	5.09
20041222_1300_1320_RTM_ELDA	24.79
20041223_1300_1318_RTM_ELDA	20.40
moyenne	**24.33**

Tableau 4 : résultat officiel du système primaire du CLIPS sur la tâche SRL (10h de test, segments sans locuteurs ignorés)

Concernant l'évaluation il est important de noter que lorsque seulement les segments *nontrans* sont ignorés, notre système est pénalisé et l'erreur de segmentation passe de 24.33% à 27.38%.

4. CONCLUSION

Ces travaux du laboratoire CLIPS pour la campagne d'évaluation ESTER montrent l'état actuel de nos systèmes de reconnaissance et de segmentation et les performances associées.

BIBLIOGRAPHIE

[1] T. Zeppenfeld, M. Finke, K. Ries, M. Westphal, A. Waibel "Recognition of conversational telephone speech using the Janus speech engine" *IEEE International Conference on Acoustics, Speech and Signal Processing*, Munich, Germany, 1997.

[2] R. Lamy, L. Besacier, "Non-linear acoustical pre-processing for multiple sampling rates ASR and ASR in noisy condition", In *NOLISP Worshop*, Le Croisic, France, 2003.

[3] "Premiers pas du CLIPS sur les données d'évaluation ESTER", R. Lamy, D. Moraru, B. Bigi, L. Besacier, *JEP 2004*, Fès, Maroc, Avril 2004

[4] A. Stolcke "SRILM -- An Extensible Language Modeling Toolkit". *Proc. Intl. Conf. on Spoken Language Processing*, vol. 2, pp. 901-904, Denver, USA, 2002.

[5] D. Moraru, S. Meignier, L. Besacier, J.-F. Bonastre, and I. Magrin-Chagnolleau, "The ELISA consortium approaches in speaker segmentation during the NIST 2002 speaker recognition evaluation". *ICASSP'03*, Hong Kong.

[6] D. Moraru, S. Meignier, C. Fredouille, L. Besacier, J.-F. Bonastre, "The Elisa Consortium Approaches in Broadcast News Speaker Segmentation During The Nist 2003 Rich Transcription Evaluation", *Proc of ICASSP 2004,* Montreal, Canada, May 2004

[7] B. Bigi "Automatic Vocabulary Evaluation and Selection", soumis à *Eurospeech 2005*. Lisbonne, Portugal, Septembre 2005.