Premiers pas du CLIPS sur les données d'évaluation ESTER

Richard Lamy, Daniel Moraru, Brigitte Bigi, Laurent Besacier

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: {richard.lamy ; daniel.moraru ; brigitte.bigi ; laurent.besacier}@imag.fr

ABSTRACT

This paper presents the first steps of the CLIPS laboratory in automatic speech transcription and speaker diarization on radio broadcast news data. Systems and results obtained on the first part (dry run) of the french ESTER evaluation are presented in this paper.

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 premiers 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 du test à blanc qui a eu lieu en Janvier 2004.

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 premiers résultats. La section 3 s'attachera à décrire la segmentation en locuteurs (SRL) et les résultats associés.

2. TRANSCRIPTION AUTOMATIQUE

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 : BL (parole bande large), BE (parole bande étroite) ou MU (musique seule). Pour l'instant, seule l'étiquette MU est utilisée pour retirer les zones de musique de l'ensemble des signaux à décoder. L'information BL/BE n'est pour l'instant pas exploitée puisque nous utilisons un seul modèle acoustique « multiconditions » . 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. Le système complet est environ 15 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. Ainsi, sur des signaux d'un heure, nous obtenons un peu moins de 1000 utterances, ce qui semble cohérent avec le nombre d'utterances des transcriptions manuelles. La taille des morceaux (ou utterances) qui seront ensuite décodés varie, par exemple de 0.44s à 45.6s pour le signal 19981217_0700_0800_inter, ce qui évite d'avoir de trop longs morceaux (> 1min) à décoder.

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 trois qualités : BL (parole bande large), BE (parole bande étroite) ou MU (musique seule). Des dictionnaires de vecteurs (ou codebooks) de taille 1024 sont appris sur les données d'apprentissage (20min / qualité) pour chaque qualité. Les vecteurs sont composés de 15 coefficients statiques (13MFCC + énergie + taux de passage par zéro). Quatre classes sont en fait utilisées pour la segmentation (musique seule MU, parole bande large SP-BL, parole bande étroite SP-BE, parole indéterminée SP), qui a lieu en deux passes :

- une première passe étiquette le signal en musique seule, parole indéterminée ou inconnu (UNK). Cette dernière classe est attribuée lorsque la décision d'une des deux classes musique ou parole est faite avec un seuil de confiance insuffisant,
- une deuxième passe étiquette en bande large ou bande étroite les zones parole issues de la première passe. Sur les zones inconnues (UNK), une décision est prise entre les trois classes parole bande large, parole bande étroite et musique seule.

La figure 1 résume l'arbre de décision utilisé pour cette segmentation en qualité. Une longueur minimale de 0,5s de zone étiquetée est imposée dans la première passe.

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

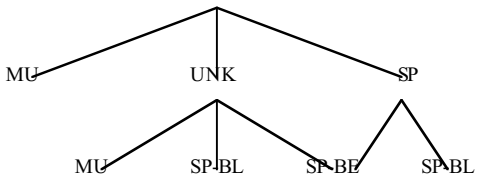

Figure 1 : Arbre de décision pour détecter la qualité

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.
Afin de définir le vocabulaire de manière optimale selon nos différents critères (compromis taux de mots hors vocabulaire / taille finale du vocabulaire), nous avons choisi de conserver pour le test à blanc tous les mots obtenus avec les transcriptions des données TRAIN et DEV d'ESTER (environ 24 000 mots). Ces 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.

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 disposons des années 1987 à 2002 (environ 2,3 Go de données nettoyées) du journal « Le Monde ». Ce corpus est filtré afin de ne sélectionner que les phrases qui contiennent uniquement des mots du vocabulaire (fixé dans la section 2.2). Par ailleurs, afin d'obtenir plus de corpus, nous appliquons la méthode des blocs-minimaux sur les phrases rejetées (qui contiennent au moins un mot inconnu). Par rapport à l'utilisation complète du corpus, ce filtrage permet d'éliminer les suites qui alternent mots inconnus et mots connus de faible taille (qui ont un effet négatif sur l'estimation des probabilités unigrammes et bigrammes essentiellement).
L'apprentissage des modèles de langage a été réalisé avec la boîte à outils SRILM [3]. Deux modèles de langage – issus des transcriptions ou du journal Le Monde - sont appris à vocabulaire fermé, c'est à dire qu'ils n'attribuent pas de probabilité au mot inconnu, ceci étant dû à la méthode de filtrage utilisée.
Nous interpolons ensuite les deux modèles obtenus en utilisant encore la boîte à outils SRILM. Elle propose une fonction afin de déterminer le meilleur coefficient d'interpolation, en l'estimant avec la valeur de perplexité sur le corpus de développement. Les meilleurs coefficients et les perplexités associées, suivant les conditions expérimentales sont reportés dans la section 2.5.

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 43 paramètres acoustiques (13 MFCC, 13 ΔMFCC, 13 ΔΔMFCC, E, ΔE, ΔΔE, zero-crossing) extraits toutes les 10ms. Les modèles acoustiques dépendants du contexte sont appris uniquement sur le corpus d'apprentissage de la phase 1 d'ESTER (30h40 de signal). Pour l'instant, nous n'exploitons pas les informations bande large / bande étroite issues du système de segmentation en qualité puisque nous apprenons un seul modèle « multi-condition » indépendant de la qualité du signal et du sexe du locuteur. Une évolution vers des modèles adaptés est évidemment la prochaine étape la plus importante afin d'améliorer les performances de notre système. Chaque modèle est un HMM de 3 états (sauf pour le silence qui a 4 états) avec 16 gaussiennes par état. Les distributions gaussiennes sont partagées entre différents états de différents HMM (*tying*) et le nombre total de distributions réellement utilisées est environ 750.

2.5 Premiers Résultats sur la tâche TRS

Les tests réalisés pour optimiser le système complet ont été réalisés sur les données de développement (DEV, 4h40) d'ESTER. Tous les résultats reportés dans cette section sont donc obtenus sur ces données, ou une partie de ces données. A la fin de la section, les résultats obtenus sur les données de test (TST, 4h40), soumis lors du test à blanc, sont également présentés.

Evaluation de la segmentation en qualité

Pour cette évaluation, nous avons utilisé comme références les étiquettes du LIA obtenues par segmentation automatique et vérifiées manuellement. Le critère d'évaluation est l'erreur de segmentation (*diarization error*) donnée par un script d'évaluation utilisé généralement pour évaluer un système de segmentation en

locuteurs[2]. Ce script compare et aligne l'hypothèse de segmentation avec la référence et calcule un score correspondant au pourcentage de la durée totale du signal contenant des erreurs de segmentation.

	% de musique détectée	% de musique réel	% erreur seg. BE/BL/MU
19981217_0700_0800_inter	1.01%	1.06%	1.29%
19981217_0800_0900_inter	0.57%	0.52%	0.78%
19990622_1900_1920_inter	0.96%	1.18%	0.57%
19990623_1900_1920_inter	1.12%	1.22%	1.95%
20000907_0930_1030_rfi	5.97%	5.12%	3.04%
20000907_1130_1230_rfi	8.68%	6.12%	4.84%
TOTAL	**3.67%**	**2.95%**	**2.20%**

Table 1 : Résultats de la segmentation en qualité (parole bande large / parole bande étroite / musique seule) et détails sur la quantité de musique seule (DEV ESTER)

La Table 1 présente le pourcentage d'erreur de segmentation en trois qualités. Nous constatons que notre système détecte plus de musique que les références. Ceci est dû d'une part aux erreurs de notre système bien sûr, et d'autre part aux parties de musique non référencées telles que les zones d'inspiration dans une zone avec musique de fond. Les résultats détaillées par classe montrent que les zones MU sont biens étiquetées à 94%, les zones BL à 98% et les zones BE à 97.8%. Il y a peu de zones mal étiquetées, cependant les début et fin de zones sont parfois imprécis et cela représente une part non négligeable dans le résultat final.

Influence du découpage automatique du signal sur les performances de transcription

Nous avons ici testé l'influence d'un découpage automatique du signal en utterances, par rapport à l'utilisation du découpage manuel fourni avec les transcriptions. La table 2 reporte les taux d'erreur de mots obtenus pour la transcription du premier fichier de DEV seulement, avec un découpage manuel (colonne 1), automatique sans retrait des zones de musique seule (colonne 2) et automatique avec retrait des zones de musique seule (colonne 3).
Ce résultat montre que le découpage automatique en utterances ne dégrade pas (et améliore même légèrement) les performances de transcription. En revanche, le retrait des zones contenant de la musique seule n'apporte pas grand chose sur les signaux France-Inter, car il y en a très peu, et parmi ces zones, certaines se trouvent en plus sur des parties non évaluées du signal. Nous verrons plus loin qu'il est quand même intéressant de retirer les zones de musique seule sur les signaux de type RFI.

Type de Découpage	**Manuel (Références)**	**Auto. (sans retrait mus.)**	**Auto. (avec retrait mus.)**
%Err	**33.9%**	**33.6%**	**33.5%**

Table 2 : Influence du découpage automatique sur les perf. de transcription (fic 19981217_0700_0800_inter seulement : 1h) ; %OOV=0, LZ/LP[3]=18/0

Performances de différents modèles de langage

Vocabulaire et ML	%Err	PPL(DEV)
Vocab. : TRAIN(ester), sans séquences %OOV=4.24 ML : TRAIN(ester)	49.2	245
ML : Le Monde filtré (87_02)	41.7	153
Interpolation (0.7/0.3)	39.6	109
Vocab. : TRAIN(ester), avec séquences %OOV=4.24 ML : TRAIN(ester)	X	250
ML : Le Monde filtré (87_02)	X	145
Interpolation (0.7/0.3)	39.5	111
Vocab. : TRAIN+DEV(ester), avec séquences ; %OOV=0 ML : TRAIN(ester)	X	250
ML : Le Monde filtré (87_02)	X	148
Interpolation (0.75/0.25)	34.7	135

Table 3 : Influence du vocabulaire et du ML sur les perf. de transcription (fic 19981217_0700_0800_inter) ; découpage manuel ; LZ/LP=25/6

Les résultats reportés dans la Table 3 montrent l'intérêt de l'interpolation entre un modèle appris sur une grande quantité de données (Le Monde) et un modèle plus représentatif de la tâche (Transcriptions Ester TRAIN). En revanche, l'intérêt d'ajouter des séquences n'apparaît pas évident sur ces résultats.

Résultats complets du « test à blanc »

	DEV Avec MU	DEV Sans MU	TST Avec MU	TST Sans MU
1217(18)_0700_0800_inter	33.6	**33.5**	33.1	**33.2**
1217(18)_0800_0900_inter	41.3	**41.2**	41.6	**41.4**
622(624)_1900_1920_inter	37.8	**37.7**	47.8	**48**
623(625)_1900_1920_inter	35.4	**35.1**	43	**42.9**
907(908)_0930_1030_rfi	48.7	**48.0**	55.2	**54.6**
907(908)_1130_1230_rfi	52.8	**52.0**	55.1	**54.4**

Table 4 : Résultats complets (%Err) de la tâche TRS du test à blanc ; Vocab. TRAIN +DEV ; ML : cf dernière ligne Tab. 3 ; découpage automatique ; LZ/LP=18/0.

Les résultats complets du test à blanc sont présentés dans la table 4. Ils montrent l'intérêt du retrait automatique des zones de musique sur les signaux RFI (gain de 0.6-0.7%) où celles-cis sont les plus nombreuses. Le taux d'erreur moyen sur toutes les données TST est de 45.1%. Ce chiffre constitue le résultat de référence que le CLIPS s'attachera à améliorer dans les mois à venir.

[2] http://www.nist.gov/speech/tools/seg_scr.v21.tarZ.htm

[3] pondération modèle de langage / modèle acoustique

3. SEGMENTATION EN LOCUTEURS

3.1. Système

Le système présenté par le CLIPS [4,5] 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 25. La limite supérieure est ajustée en fonction de la durée du document. Le nombre de locuteurs (NSp) maximise :

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

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.
Les silences et les zones de musiques ne sont pas retirés dans les résultats du système soumis pour le test à blanc, puisque des expériences préliminaires conduites sur les données DEV ont montré que ceci dégradait légèrement les résultats. Ceci peut s'expliquer par le fait que dans la procédure d'évaluation définie pour ESTER, de nombreuses zones de signal sont retirées avant le scoring, et ces zones contiennent déjà pour la plupart des silences et des portions de musique.

3.2. Résultats sur la tâche SRL

	% err. seg. DEV	% err. seg. TST
1217(18)_0700_0800_inter	12.0	16.9
1217(18)_0800_0900_inter	11.3	22.4
622(624)_1900_1920_inter	6.9	12.7
623(625)_1900_1920_inter	8.0	6.4
907(908)_0930_1030_rfi	13.9	13.1
907(908)_1130_1230_rfi	23.1	25.1
TOTAL	**13.8**	**17.7**

Table 5 : Résultats complets de la tâche SRL du test à blanc sur tout DEV et tout TST

Ces premiers résultats sur la tâche SRL d'ESTER obtenus par CLIPS sont cohérents avec ceux obtenus lors des évaluations NIST RT03 sur la même tâche, où 19.2% d'erreur de segmentation étaient obtenus sur des données de journaux télévisés. Cependant, dans l'évaluation RT03, les erreurs sur la détection d'activité vocale étaient également comptées dans le score de segmentation (elles étaient de 4.9% sur notre système, soit une erreur de segmentation en locuteurs pure de 19.2-4.9=14.3%). Dans notre test à blanc ESTER, les erreurs de détection d'activité vocale sont nulles ce qui peut s'expliquer par le fait que toutes les zones prises en compte pour le calcul du score de segmentation ne contiennent que de la parole. Ce détail mis à part, il ne semble pas y avoir de différence fondamentale de difficulté entre la segmentation de journaux télévisés et de journaux radiophoniques.

4. CONCLUSION

Ces premiers travaux du laboratoire CLIPS pour la campagne d'évaluation ESTER montrent l'état actuel de nos systèmes de reconnaissance et de segmentation. Nous travaillons actuellement sur plusieurs points : pre-segmentation en genre (réalisée a posteriori sur le résultat de la segmentation en locuteurs) ; apprentissage de modèles acoustiques spécifiques (bande étroite / bande large, homme / femme, etc ..).

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

[4] 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.

[5] 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", Accepted to *Proc of ICASSP 2004*, Montreal, Canada, May 2004