Segmentation selon le locuteur : les activités du Consortium ELISA dans le cadre de Nist RT03

Daniel Moraru [(1)], Sylvain Meignier [(2)],
Corinne Fredouille [(2)], Laurent Besacier [(1)], Jean-François Bonastre [(2)]

[1] CLIPS-IMAG (UJF & CNRS) - BP 53 - 38041 Grenoble Cedex 9 - France
[2] LIA-Avignon - BP1228 - 84911 Avignon Cedex 9 – France

(daniel.moraru,laurent.besacier)@imag.fr
(sylvain.meignier,corinne.fredouille,jean-francois.bonastre)@lia.univ-avignon.fr

ABSTRACT

This paper presents the ELISA consortium activities in automatic speaker diarization (also known as speaker segmentation) during the NIST Rich Transcription (RT) 2003 evaluation. The experiments were achieved on real broadcast news data (HUB4), in the framework of the ELISA consortium. The paper firstly shows the interest of segmentation in acoustic macro classes (like gender or bandwidth) as a front-end processing for segmentation/diarization task. The impact of this prior acoustic segmentation is evaluated in terms of speaker diarization performance. Secondly, two different approaches from CLIPS and LIA laboratories are presented and different possibilities of combining them are investigated. The system submitted as ELISA primary obtained the second lower diarization error rate compared to the other RT03-participant primary systems. Another ELISA system submitted as secondary outperformed the best primary system (i.e. it obtained the lowest speaker diarization error rate).

1. INTRODUCTION

La segmentation selon le locuteur est une tâche relevant du domaine du traitement automatique de la parole. Cette tâche est née relativement récemment pour répondre au besoin créé par le nombre toujours croissant de documents multimédia devant être archivés et accédés. Les tours de parole et l'identité des locuteurs constituent une intéressante clé d'accès à ces documents. Le but de la segmentation selon le locuteur est donc de segmenter en tours de parole (un tour de parole est un segment contenant une intervention d'un locuteur) un document audio contenant N locuteurs et d'associer chaque tour de parole au locuteur l'ayant prononcé. En général, aucune information *a priori* n'est disponible, sur le nombre de locuteurs ou leurs identités.

Ce papier résume les activités en segmentation automatique selon le locuteur réalisées par le Consortium ELISA [1] dans le cadre de la campagne d'évaluation NIST/USA RT03[1] (Rich Transcription). La *partie 2* présente le système de segmentation en macro classes acoustiques réalisé au LIA. La segmentation en macro classes est souvent présente dans la littérature [2][3][4], en particulier comme une aide aux systèmes de transcription d'émissions radiophoniques. Cependant, son apport dans le cadre d'un système de segmentation selon le locuteur reste encore à évaluer. La *Partie 3* présente deux systèmes – provenant de deux laboratoires différents, le CLIPS et le LIA – basés sur des stratégies de segmentation très différentes. La *Partie 4* propose différentes méthodes pour combiner des experts en segmentation sur le locuteur et les expérimente à partir des différents systèmes proposés précédemment. La *Partie 5* de ce papier présente les résultats expérimentaux liés à cet article. Les données proviennent de la campagne RT03 ainsi que la majorité des protocoles. Enfin, la dernière partie de l'article est consacrée à une conclusion, assortie de quelques perspectives.

2. SEGMENTATION EN MACRO CLASSES ACOUSTIQUES

La segmentation en macro classes acoustiques est nécessaire pour supprimer les parties du document ne contenant pas de parole (comme la musique, les silences…) ou pour réaliser des traitements spécifiques à des conditions acoustiques données (genre des locuteurs, parole téléphonique, parole au dessus de la musique…). Le processus de segmentation acoustique proposé dans ce papier réalise une segmentation en trois niveaux : parole/non parole, parole propre/parole avec musique/parole téléphonique et homme/femme. La classification est réalisée suivant un procédé hiérarchique en trois étapes :

- Le premier niveau de segmentation correspond à une séparation "parole/non parole". Le procédé est basé sur une modélisation statistique des deux classes. Il consiste en une discrimination trame à trame suivie d'un ensemble de règles morphologiques. Ces dernières permettent de définir des contraintes sur les segments, comme leur durée minimale.
- La deuxième étape de segmentation consiste à répartir les zones étiquetées "parole" en trois classes : "parole propre", "parole et musique" et "parole téléphonique". Cette étape repose sur un décodage de type Viterbi associé à un HMM ergodique.
- La dernière étape est dédiée à la séparation "homme/femme". Un procédé de même type que pour l'étape précédente est employé, avec des états dépendant de la classe acoustique et du genre (une classe "parole dégradée" est ajoutée, pour augmenter la robustesse du procédé).

[1] Voir http://www.nist.gov/speech/tests/rt/rt2003/index.htm pour plus de détails

Les modèles acoustiques décrivant les classes sont des modèles de mélange de lois Gaussiennes (GMM pour Gaussian Mixture Models) diagonaux [5] contenant de 512 à 1024 composantes, à l'exception de la classe "non parole" représentée par un modèle monogaussien. L'ensemble des paramètres (GMM et transitions) ont été appris à partir de la base HUB4 (1996).

3. LES SYSTEMES DE SEGMENTATION SELON LE LOCUTEUR

Les différents systèmes présentés ont été développés dans le cadre du Consortium ELISA en utilisant AMIRAL, la plate-forme de reconnaissance du locuteur du LIA [6]. Ils utilisent la macro segmentation décrite dans la *Partie 2* comme pré-traitement et sont appliqués séparément sur les macro classes acoustiques détectées par ce procédé.

3.1. Le système LIA

Le système proposé par le LIA est basé sur une modélisation de la conversation par un HMM [8][9]. Chaque état du modèle caractérise un locuteur et les transitions modélisent les changements de locuteur. Les différentes segmentations obtenues pour chaque macro classe acoustique sont fusionnées et un procédé de resegmentation affine le résultat. Le modèle HMM de la conversation est généré par un procédé itératif qui détecte les locuteurs un à un et les ajoute au modèle, i.e. qui ajoute l'état correspondant au locuteur. Ce procédé comporte quatre étapes :

- *1-Initialisation*. Un modèle initial de "locuteur" est entraîné à partir de l'ensemble du document (il s'agit plus d'un modèle générique que du modèle d'un locuteur donné). La conversation est décrite par un HMM à un seul état ; l'ensemble du document est attribué à ce locuteur/état.
- *2-Détection et ajout d'un locuteur*. Un nouveau modèle de locuteur est entraîné à partir de 3 secondes de signal, sélectionnées pour maximiser le rapport de vraisemblance entre le modèle initial (étape 1) et un modèle du monde (appris à partir de données externes). La segmentation courante et le HMM sont modifiés pour intégrer le nouveau locuteur (i.e., le nouvel état).
- *3-Optimisation des modèles de locuteur*. Les modèles associés aux états (modèles de locuteur) sont optimisés en utilisant la segmentation courante, en optimisant un critère de type MAP. Un décodage de type Viterbi est alors effectué pour produire une nouvelle segmentation. Cette étape d'optimisation est répétée tant que la segmentation évolue entre deux itérations.
- *4-Critères d'arrêt*. La meilleure segmentation actuelle, à N locuteurs, est comparée à la meilleure segmentation de l'itération précédente, à N-1 locuteurs, en utilisant la version courante du HMM. Le critère d'arrêt est estimé atteint si la dernière solution n'obtient pas une meilleure vraisemblance[2] [8]. Une heuristique est également vérifiée : si le dernier locuteur ajouté est seulement associé à un segment (<4sec.), l'ajout de locuteur (étape 2) est relancé sur un segment de parole différent.

Après fusion des sous segmentations, une resegmentation similaire à l'étape 3 est exécutée. La différence majeure consiste à utiliser des variantes différentes de MAP pour l'adaptation des modèles de locuteur (des GMM). L'algorithme défini dans [5] est utilisé pour la resegmentation alors qu'une variante du LIA [6], adaptée à des segments très courts, est employée dans l'étape 3[3]. Le LIA a également présenté une variante[4] utilisant une troisième version de MAP, basée sur une combinaison linéaire [10] entre connaissances a priori et données d'adaptation.

Les modèles acoustiques sont des GMM à 128 composantes (à matrices diagonales). Le modèle du monde est appris à partir d'un sous ensemble de HUB4 (96).

3.2. Le système CLIPS

Le système présenté par le CLIPS [10] repose sur une détection de changement de locuteurs, suivie d'un procédé de clustering hiérarchique.

La détection des changement de locuteurs est effectuée par BIC [7] (Bayesian Information Criterion), à l'aide 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 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 :

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

où M est le modèle composé des NSp modèles de locuteur détectés, NX est le 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.

4. STRATEGIES DE COMBINAISON DES SYSTEMES

Nous proposons deux stratégies pour combiner les systèmes, par hybridation ou par fusion.

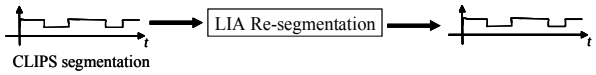

Figure 1 : Système "piped"

4.1 Hybridation (système "piped")

L'idée de l'hybridation consiste à initialiser le système du LIA (étape de resegmentation) par une segmentation issue du système du CLIPS (figure 1). Cette solution associe les avantages des longs (et relativement purs) segments (issus du CLIPS), utiles pour initialiser les modèles du HMM, avec le pouvoir de modélisation et de décodage des HMM (système LIA).

[2] ou si il ne reste plus de parole pour initialiser un nouveau locuteur.

[3] Dans les deux cas, seules les moyennes des gaussiennes sont adaptées.

[4] Cette variante a été utilisée durant la campagne NIST 2002 et est employée dans le 4.2.

4.2 Fusion (système "fusion")

Le principe de la fusion est d'utiliser des segmentations provenant d'autant d'experts que possible, quatre dans ce papier (figure 2) : le système du CLIPS, le système principal du LIA, la variante du LIA et le système "piped".

La stratégie de fusion proposée, originale, consiste en deux étapes : une première phase de génération exhaustive d'étiquettes et de pseudo locuteurs communs aux experts (regroupant toutes les informations provenant des divers experts) suivie d'une étape de resegmentation, chargée de produire la segmentation optimale.

La première étape repose sur une décision trame à trame qui consiste à grouper les étiquettes proposées par chacun des experts. Par exemple :

- Trame *i*: Sys1="S1", Sys2="T4", Sys3="S1", Sys4="F1" → Etiquette produite "S1T4S1F1",
- Trame *i+1* : Sys1="S2", Sys2="T4", Sys3="S1", Sys4="F1" → Etiquette produite "S2T4S1F1".

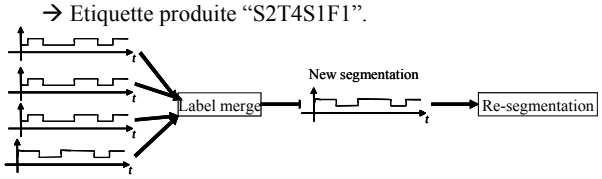

Figure2 : système "fusion"

Cette méthode génère un grand nombre de pseudo-locuteurs regroupés en :

- Des pseudo-locuteurs associés à de nombreuses données acoustiques. Ces locuteurs étant de fait proposés conjointement par une majorité d'experts, ils correspondent certainement à une bonne hypothèse.
- Des pseudo-locuteurs proposés par peu de systèmes (étiquettes "courtes"), en général associés à un (ou quelques) court segment (~3s à 10s). Ces hypothèses peuvent être rejetées, leur poids sur la qualité du résultat étant marginal.
- Des pseudo-locuteurs associés à une faible quantité de données acoustiques séparée en de multiples (courts) segments. Ces hypothèses correspondent à des zones d'indécision et peuvent être supprimées.

La re-segmentation du LIA est appliquée sur la segmentation obtenue. En fonction des remarques précédentes, avant chaque étape d'adaptation et de décodage, les pseudo-locuteurs associés à moins de 3s de signal sont supprimés (les données correspondantes seront attribuées aux autres locuteurs durant l'itération suivante). Ce procédé permet de réduire automatiquement le nombre de locuteurs (d'environ 150 à 50 locuteurs dès la première itération).

Cependant, cette stratégie de fusion ne permet pas de corriger certaines erreurs des experts, comme un locuteur divisé en deux pseudo-locuteurs chacun associé à un long segment de parole. La resegmentation ne remettant pas en cause le clustering (si les pseudo-locuteurs sont associés à des segments de longueur suffisante), cette erreur a peu de chance d'être corrigée.

5. EXPERIENCES

Les expériences décrites dans ce papier ont été réalisées dans le cadre de NIST/RT'03. Deux corpus d'émission radiophoniques étaient disponibles :

- *Dev,* composé de 6 enregistrements de 10mn chacun.
- *Eva* composé de 3 émissions de 30 minutes et contenant de 10 à 27 locuteurs (distribué pour la phase de test de RT).

Pour la segmentation en locuteurs, dans RT, la mesure des performances est basée sur la meilleure correspondance entre les locuteurs de la référence et les locuteurs de la segmentation à évaluer. L'erreur globale de segmentation est définie par le pourcentage de temps attribué à tord au mauvais locuteur. Cette mesure prend en compte les erreurs de locuteurs mais aussi les erreurs liées à la détection parole/non parole.

5.1. Performance de la macro segmentation, isolément.

La Table 1 donne les performances de la pré segmentation acoustique en macro classes. Durant NIST RT03, le système a obtenu une erreur de détection « parole/non parole » de 4,5% (en % de temps) à comparer à 4,4% pour le meilleur système. Les erreurs de détection de genre oscillent entre 1,5% (Dev) et 5,5% (Eva). La détection de la parole téléphonique a une précision estimée à moins de 0,1% d'erreur sur Dev et 3 % sur Eva.

Corpus	Missed Speech Error	False Alarm Speech Error	Gender Error	Telephone / Non telephone Speech error
Dev	**2.3%**	2.2%	1.5%	0.09 %
Eva	1.8%	2.7%	5.5%	3 %

Table 1 : Taux d'erreur de la segmentation en macro classes

5.2 Influence de la macro segmentation sur la segmentation en locuteur.

Pour évaluer l'influence de la macro segmentation sur les performances de la segmentation en locuteurs, différents niveaux de macro segmentation ont été testés :

- parole / non parole (S/NS),
- S/NS + le genre (S/NS-Gender),
- S/NS-Gender + telephone / non telephone (S/NS-Gender-T/NT),
- S/NS-Gender + téléphone/parole propre/parole et musique/parole dégradée (S/NS-Gender-T/S/MS/DS).
- Pour aider à l'évaluation, des résultats obtenus avec une segmentation manuelle en "parole / non parole + genre + téléphone / non téléphone" sont proposés (Hand S/NS-Gender-T/NT).

NB : *les publicités ont été supprimées manuellement, ce qui induit une légère différence avec les résultats officiels de l'évaluation RT (proposés dans la section suivante).*

	LIA +resegmentation		CLIPS +resegmentation	
Acoustic segmentation	***Dev***	***Eva***	***Dev***	***Eva***
Hand S/NS-Gender-T/NT	*10.8%*	*13.2%*	*15.7%*	*9.4%*
S/NS	15.5%	26.6%	19.3%	15.7%
S/NS-Gender	13%	24.9%	18.2%	15.3%
S/NS-Gender-T/NT	12.8%	14.1%	19.0%	13.9%
S/NS-Gender-T/S/MS/DS	15.6%	14.3%	18.7%	15.1%

Table 2 : Erreur de segmentation en locuteur pour différents niveaux de macro segmentation (avec resegmentation).

La Table 2 présente l'erreur de segmentation en locuteurs, en fonction de la macro segmentation choisie, du corpus et du système concerné, LIA ou CLIPS. Il est à noter que, dans cette expérience, une étape de resegmentation LIA est systématiquement utilisée après regroupement des segmentations réalisées sur les différentes classes acoustiques.

Les résultats montrent :

- un gain de performance quand une macro segmentation est utilisée,
- l'influence de la qualité de la segmentation en macro classes : les meilleurs résultats sont obtenus avec la pré segmentation manuelle,
- un gain de performance lorsque la granularité de la pré segmentation augmente (de S/NS à S/NS-Gender-T/NT),
- cependant, la segmentation la plus fine n'amène pas d'amélioration (S/NS-Gender-T/S/MS/DS),

5.3 Performance des systèmes ELISA durant RT03

	Miss Speech	FA Speech	SPK ERR	**ERR**
CLIPS primary	2.0%	2.9%	14.3%	**19.25%**
LIA primary	1.1%	3.8%	12.0%	**16.90%**
LIA second	1.1%	3.8%	19.8%	**24.71%**
ELISA "merged"	1.1%	3.8%	9.3%	**14.24%**
ELISA "piped"	1.1%	3.8%	8.0%	**12.88%**

Table 3 : Résultats officiels d'ELISA durant RT03

La Table 3 résume les résultants des différents systèmes proposés dans ce papier, durant l'évaluation RT03. Quelques commentaires permettent de synthétiser cette table :

- Les systèmes principaux du LIA, du CLIPS et d'ELISA (système "merged") ont obtenu des performances très honorables durant NIST RT03. Le meilleur système principal (ELISA) a obtenu la deuxième place avec 14,24% d'erreurs.
- Le deuxième système ELISA, le système « piped » (présenté comme variante), a obtenu le plus faible taux d'erreur de la campagne RT03 avec 12,88% d'erreur.
- La stratégie de fusion obtient des résultats mitigés (le système "piped" est l'un des experts). Cependant, une analyse plus fine montre que la majeure partie des erreurs vient du dernier enregistrement. Ces résultats semblent liés à la faiblesse de la méthode identifiée en 4.2 et particulièrement marquée par la méthode de mesure de performance de RT.

Nb : même si l'ensemble des systèmes utilisent la même segmentation "parole / non parole", les erreurs correspondantes (*Miss Speech* et *False Alarm Speech)* sont différentes. Ces différences proviennent d'un alignement des frontières de segments toutes les 0,2 s. pour les systèmes LIA et ELISA.

Une expérience complémentaire menée avec le système du CLIPS montre que, pour ce système, l'estimation du nombre de locuteurs pèse pour environ 3% dans le taux d'erreur global.

6. CONCLUSIONS

Ce papier résume les activités du consortium ELISA pour la tâche de segmentation en locuteurs de la campagne NIST RT03.

Ce papier a d'abord montré qu'une étape de segmentation en macro classes acoustiques était utile pour la segmentation en locuteur.

Deux approches différentes de la segmentation en locuteur ont ensuite été décrites. L'une (système du LIA) est basée sur une modélisation de la conversation par un HMM, associée à un algorithme itératif de détection des locuteurs et de construction du HMM. La deuxième (système du CLIPS) met en œuvre une approche plus classique, basée sur BIC, enchaînant une détection des changements de locuteurs et une étape de clustering. Les deux systèmes proposés ont obtenu des performances intéressantes durant la campagne NIST RT03 (19,25% d'erreur pour le système CLIPS et 16,90% d'erreur pour le système LIA).

Deux techniques de combinaison des systèmes de segmentation ont également été proposées. La première utilise le système par détection de changement de locuteur (CLIPS) pour initialiser le système HMM (LIA). Cette association améliore nettement les performances, en obtenant avec 12,88% d'erreurs le plus faible taux d'erreur de la campagne NIST RT03 (ce système était présenté comme système secondaire par le Consortium ELISA). Ce résultat montre qu'initialiser la méthode HMM avec de bonnes hypothèses de locuteurs (nombre et segments associés) permet à cette modélisation d'exprimer son potentiel. La deuxième technique de combinaison repose sur une approche originale de fusion d'experts, qui génère de façon exhaustive des pseudo-locuteurs (à partir des segmentations fournies par les experts) avant de réévaluer les hypothèses par une resegmentation HMM (système du LIA). Cette technique de fusion permet, sans effort particulier, d'utiliser des informations venant d'un nombre quelconque d'experts. Elle ne peut cependant corriger certaines erreurs pouvant provenir d'un seul expert (un locuteur divisé en deux pseudo-locuteurs chacun associé à un long segment de parole). Cependant, le potentiel de cette méthode reste important et l'ajout d'une étape de clustering est une solution à envisager.

BIBLIOGRAPHIE

[1] I. Magrin-Chagnolleau, G. Gravier, and R. Blouet for the ELISA consortium, "Overview of the 2000-2001 ELISA consortium research activities," *A Speaker Odyssey*, pp.67–72, Chania, Crete, June 2001.

[2] P.C. Woodland, "The development of the HTK Broadcast News transcription system: An overview", *Speech Communication*, Vol. 37, pp. 291-299, 2002.

[3] T. Hain, and P.C. Woodland, "Segmentation and Classification of Broadcast News audio", *ICSLP'98*, Sydney, Australia.

[4] J.L. Gauvain, L. Lamel, and G. Adda. "The LIMSI Broadcast News Transcription System". *Speech Communication*, 37(1-2):89-108, 2002.

[5] D.A. Reynolds, T.F. Quatieri and R.B. Dunn, "Speaker Verification Using Adaptation Mixture Models". *Digital Signal Processing*, Vol. 10, No. 1-3, January/April/July 2000.

[6] C. Fredouille, J.-F. Bonastre, and T. Merlin, "AMIRAL: a block-segmental multi-recognizer architecture for automatic speaker recognition," Digital Signal Processing, Vol. 10, No. 1-3, January/April/July 2000.

[7] P. Delacourt and C. Wellekens, "DISTBIC: a speaker-based segmentation for audio data indexing," *Speech Communication,* Vol. 32, No. 1-2, September 2000.

[8] S. Meignier, J.-F. Bonastre, and S. Igounet, "E-HMM approach for learning and adapting sound models for speaker indexing," *A Speaker Odyssey*, pp.175-180, Chania, Crete, June 2001.

[9] S. Meignier, J.-F. Bonastre, C. Fredouille, and T. Merlin, "Evolutive HMM for Multi-Speaker Tracking System". *ICASSP'00*, 5-9 June 2000, Istanbul, Turkey.

[10] 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.