Segmentation et regroupement en locuteurs pour la parole conversationnelle

E. El Khoury [(1)], S.Meigner[(2)], C.Sénac[(1)]

IRIT[(1)], Université Paul Sabatier, Toulouse, France
LIUM[(2)], Université du Maine Le Mans, France
{elie.el-khoury,christine.senac}@irit.fr
sylvain.meignier@lium.univ-lemans.fr
http://www.irit.fr/-Equipe-SAMoVA-/ http://www-lium.univ-lemans.fr/

ABSTRACT

In the context of conversational speech, we present a hybrid speaker diarization system based on the combination of the LIUM and IRIT systems. It contains a first step of speech detection followed by speaker segmentation. Then we apply a speaker clustering in two steps: a first clustering based on Bayesian Information Criterion (BIC) is followed by a clustering based on the Cross Likelihood Ratio (CLR). Moreover, we make some improvements by optimizing clustering thresholds and by purifying the BIC clustering using feature F0. Results show this hybrid system is suitable on one hand for traditional corpus as ESTER and on the other hand for conversational data as used in EPAC project.

Keywords: speaker diarization, segmentation, clustering, Bayesian Information Criterion (BIC), Generalized Likelihood Ratio (GLR), Cross Likelihood Ratio (CLR).

1. INTRODUCTION

De nombreuses méthodes de segmentation et de regroupement en locuteurs (SRL) ont été proposées. Pour la segmentation en locuteurs, on peut citer les approches basées sur une métrique comme la distance de Kullback-Leibler [1] (ce type d'approche nous a posé problème dans le cas de données bruitées) ou bien les approches basées sur la sélection d'un modèle comme les méthodes GLR [2] ou BIC [3]. La plupart des regroupement en locuteurs s'effectuent de manière hiérarchique ascendante : les clusters les plus proches (au sens d'une distance ou d'une similarité) sont regroupés de façon itérative. Les méthodes diffèrent principalement dans le choix de la distance (essentiellement distance BIC ou Kullback-Leibler) et dans le critère d'arrêt (le plus souvent un seuil).

Récemment, le LIUM et l'IRIT ont développé leurs propres systèmes de SRL. Cependant, à travers le projet EPAC[1], qui vise à traiter de la parole conversationnelle enregistrée dans différents environnements sonores, nous avons mis en évidence des lacunes dans nos systèmes traditionnels de SRL. Nous avons donc construit un système hybride qui combine les points forts de nos deux systèmes et qui au final est plus performant à la fois sur des corpus traditionnels de type ESTER et sur des données conversationnelles.

Dans les paragraphes 2 et 3 nous décrivons succinctement nos deux systèmes de base et le système hybride est ensuite détaillé dans le paragraphe 4. Ensuite nous donnons les conditions d'expériences et les résultats associés dans le paragraphe 5.

Le prétraitement acoustique est similaire pour les trois systèmes : 12 MFCC sont extraits toutes les 10ms (fenêtre de 20ms). Pour les étapes de séparation parole/non parole et de regroupement CLR, nous avons également utilisé les dérivées de l'énergie et des MFCC (fenêtre de 5 trames) et nous avons normalisé les MFCC (centrés et réduits).

2. LE SYSTÈME DE L'IRIT

Ce système [4] ne nécessite aucune connaissance a priori et opère sans segmentation préliminaire en parole et non parole. Le signal acoustique est segmenté avec un double critère basé sur le GLR et sur le BIC (c.f. 4.2). Le regroupement repose sur la méthode EVSM (Eigen Vector Space Models) introduite par Tsai et al. [5]. Une matrice de similarité est exprimée sur l'ensemble des segments (représentés sous forme de vecteurs) de façon à regrouper dans un même cluster les segments les 'plus proches' au sens d'une métrique. Cette dernière repose à la fois sur l'expression du cosinus de l'angle formé par 2 vecteurs à comparer et sur l'utilisation du paramètre F0, ce qui permet ainsi d'éliminer des erreurs de regroupements abusifs (c.f. 4.3).

La segmentation GLR-BIC de ce système permet de détecter des segments courts : en effet, les points de rupture sont détectés très finement. Par contre, le critère d'arrêt du regroupement est basé sur un seuil fixe qui n'est pas optimal en présence d'une variation importante du nombre de locuteurs.

3. LE SYTÈME DU LIUM

Ce système [6] a été développé dans le cadre de la tâche de transcription de la campagne d'évaluation ESTER [7]. Il repose sur une segmentation acoustique de type BIC suivie d'un regroupement hiérarchique également de type BIC. Ensuite un décodage Viterbi permet d'ajuster les bornes des segments. Un nouveau regroupement de type CLR est alors appliqué sur les clusters générés par le décodage Viterbi. Finalement les régions correspondant à de la musique ou des *jingles* sont éliminées par un

[1] http://epac.univ-lemans.fr

décodage Viterbi (8 GMMs appris sur le corpus d'apprentissage d'ESTER sont nécessaires).

Le regroupement de ce système est meilleur que celui du système de l'IRIT : cependant il introduit des erreurs sur les petits segments. De plus, on retrouve un problème général : l'utilisation de seuils fixes que ce soit pour le regroupement BIC ou GLR, n'est pas adapté quand la structure des données est très variable.

4. LE SYSTÈME HYBRIDE

La mise en évidence des avantages et des faiblesses de chacun des systèmes précédents nous a permis de construire un système hybride, décrit dans la figure 1, plus robuste aux variations de données. Il se présente sous la forme de trois modules principaux décrits ci-dessous : séparation de parole/non parole, segmentation GLR/BIC et regroupement.

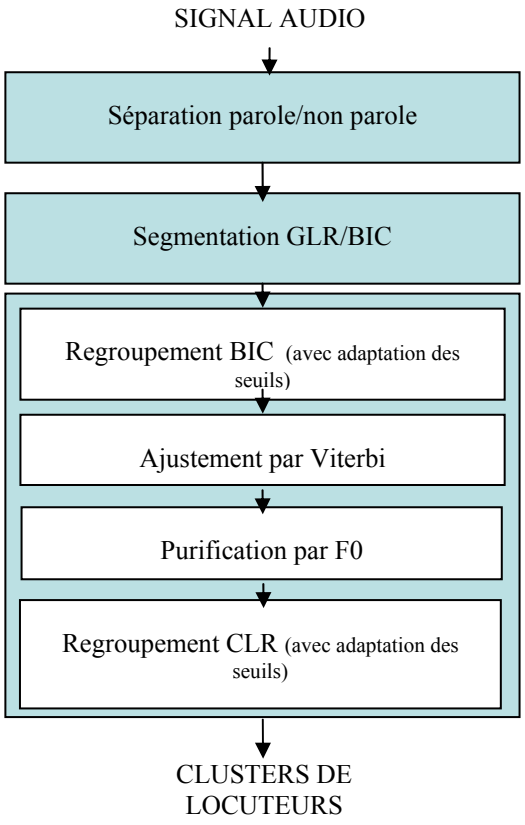

Figure 1 : Le système hybride.

4.1. Séparation Parole / Non Parole

Cette séparation est basée sur l'utilisation conjointe de modèles GMM (matrice de covariance diagonale et 512 gaussiennes par modèle) et de la modulation de l'énergie à 4Hz. Ce système [8] a obtenu les meilleurs résultats lors de la campagne ESTER.

4.2. La segmentation GLR/BIC

Cette méthode proposée dans nos précédents travaux [4], consiste à appliquer l'algorithme GLR jusqu'à convergence vers la meilleure répartition gaussienne et ensuite à appliquer le BIC pour choisir les véritables points de rupture. L'intérêt de cette méthode réside dans sa précision et dans la simplification que nous avons apportée dans l'écriture du coefficient de pénalité de l'expression BIC puisque celui-ci apparaît sous forme de constante contrairement à son expression dans les méthodes traditionnelles utilisant le BIC.

4.3. Le regroupement

Le regroupement BIC : Il s'agit d'un regroupement hiérarchique ascendant dans lequel chaque cluster est modélisé par une gaussienne de covariance pleine. Les deux plus proches clusters sont regroupés à chaque itération jusqu'à convergence sur le critère d'arrêt. La métrique BIC est utilisée à la fois comme critère d'arrêt et pour sélectionner les clusters c_i et c_j à regrouper :

$$BIC_{i,j} = (n_i+n_j) \log |\Sigma| - n_i \log |\Sigma_i| - n_j \log |\Sigma_j| - \lambda P \qquad (1)$$

Avec λ le poids de pénalité; $|\Sigma|$, $|\Sigma_i|$ et $|\Sigma_j|$ les déterminants des covariances des modèles des clusters ($c_i + c_j$), c_i et c_j.

Le facteur de pénalité P de la métrique BIC est exprimé en fonction de n_i et n_j qui sont respectivement les longueurs totales (en terme de paramètres) de c_i et c_j, et de d la dimension des paramètres :

$$P = \tfrac{1}{2} (d + d(d+1)/2) \log(n_i+n_j) \qquad (2)$$

Barras et al. ont montré que ce facteur de pénalité, parce qu'il s'appuie uniquement sur la taille des deux clusters et non sur la taille complète des données, donne les meilleurs résultats [9]. Cependant, nous avons remarqué :

- qu'en fonction de la durée du fichier et du nombre de locuteurs, le regroupement optimal peut s'effectuer plus ou moins rapidement ;
- que la vitesse de regroupement optimal est proportionnelle à la différence du nombre de clusters obtenus pour λ=1.5 et λ=3.5 : plus cette différence est élevée, plus vite le regroupement optimal est atteint et donc plus vite le regroupement doit être stoppé.

Nous avons donc appris le poids de pénalité λ pour qu'il s'adapte automatiquement en fonction de la structure des données du fichier (le système du LIUM utilisait pour λ un seuil fixe de 4.5). Le corpus d'apprentissage est composé de 30 fichiers de tailles différentes. Pour chaque fichier d'apprentissage, le regroupement hiérarchique par BIC a été effectué pour 4 valeurs différentes de λ : pour λ=1.5 et λ=3.5 le nombre de clusters a été mémorisé, pour λ=4 et λ=4.5 les DER (NIST speaker Diarization Error Rate) ont été calculés. La valeur optimale de λ pour chaque fichier d'apprentissage est la valeur de l'ensemble {4, 4.5} donnant le plus faible DER. Les

tables 1 et 2 montrent le nombre de clusters obtenus en fonction de λ pour deux fichiers d'une heure chacun du corpus d'apprentissage. Cet apprentissage a permis de déduire que lors des tests, le choix de λ est ramené à la décision binaire suivante :

si $NC(1.5) - NC(3.5) > k.S$ alors $\lambda = 4$ sinon $\lambda = 4.5$

avec *NC(i)* le nombre de clusters du fichier test pour $\lambda=i$, *k* une constante apprise (ici $k=1.1$) et *S* la durée du fichier de test en minutes.

Table 1 : Exemple sur un fichier d'apprentissage de 1h. La différence du nombre de clusters est élevée (110 clusters pour λ=1.5 et 41 clusters pour λ=3.5). Le regroupement est stoppé pour λ = 4 (ce qui correspond bien au DER le plus faible)

λ	1.5	3.5	**4**	4.5
Nombre de clusters (NC)	110	41	**39**	36
DER	-	-	**10.85**	11.55

Table 2 : La différence du nombre de clusters est faible (106 clusters pour λ=1.5 et 46 clusters pour λ=3.5). Le regroupement est stoppé pour λ = 4.5

λ	1.5	3.5	4	**4.5**
Nombre de clusters (NC)	106	47	43	**38**
DER	-	-	9.47	**8.99**

L'ajustement par Viterbi : Cet ajustement permet de remettre en cause le regroupement précédent. Chaque locuteur est modélisé par un HMM à un état avec une matrice de covariance diagonale (8 GMM) appris par l'algorithme EM-ML sur l'ensemble de ses segments. Une pénalité permettant de basculer sur un autre locuteur a été fixée expérimentalement.

La purification par F0 : Il s'agit encore ici de remettre en cause le regroupement : en effet, nous avons observé que l'utilisation du F0 permet d'éliminer des erreurs de regroupement. Pour cela, nous avons extrait le F0 toutes les 10ms sur les zones voisées avec le logiciel ESPS. Ensuite, nous avons calculé la moyenne des valeurs de F0 de chaque segment d'un cluster. Une trop grande différence ΔF_0 (i.e. si ΔF_0 >40Hz, ce seuil a été expérimentalement fixé) entre la moyenne de F0 du segment et la moyenne de F0 du cluster conduit à l'exclusion du segment de ce cluster.

Le regroupement CLR : Lors de la segmentation acoustique, l'environnement sonore aide le système à détecter les changements de locuteurs : les paramètres ne sont donc pas normalisés. Les résultats de la segmentation mettent parfois en évidence l'existence de plusieurs clusters pour le même locuteur en fonction de l'environnement sonore (bruit, musique, calme,…). Par exemple, dans les journaux d'information, l'annonce des titres débute toujours sur un fond musical qui s'atténue progressivement jusqu'à disparaître. L'annonceur risque donc, à tord, d'être modélisé par deux classes différentes. La contribution de l'environnement sonore doit alors être réduite et normalisée dans les modèles des clusters.

Un regroupement hiérarchique ascendant est donc effectué sur les cluster issus de la purification : les paramètres de chaque segment sont normalisés et un modèle du monde (appris sur 4h du corpus d'apprentissage d'ESTER) est adapté (MAP) pour chaque cluster. A chaque itération, sont regroupés les clusters qui maximisent le critère CLR :

$$\mathrm{CLR}_{i,j} = \frac{\mathrm{L}(c_i/M_j)}{\mathrm{L}(c_i/\mathrm{UBM})} * \frac{\mathrm{L}(c_j/M_i)}{\mathrm{L}(c_j/\mathrm{UBM})} \qquad (3)$$

avec M_i, M_j les modèles des clusters c_i et c_j ; *UBM* le modèle du monde universel et *L(.)* la vraisemblance. Le regroupement s'arrête lorsque $CLR_{i,j}$ dépasse un seuil *clr_thr*. Après observation des résultats, nous avons conclu que le seuil optimal dépend de la durée du fichier traité : plus le fichier est court plus son seuil optimal sera élevé. En s'appuyant sur les 30 fichiers d'apprentissage déjà utilisés pour le regroupement BIC, nous avons modélisé cette dépendance par :

$$Clr\_thr = a\,L + b \qquad (4)$$

avec *L* la longueur du fichier en minutes et *a* et *b* déterminés par apprentissage (ici : a = 0.013 et b = 2.22). De plus, ici comme pour le regroupement BIC, la différence du nombre de clusters affecte le seuil : en suivant le même schéma que précédemment, le seuil est donc multiplié par un coefficient appris pouvant prendre deux valeurs (0.95 ou 1).

5. EXPERIENCES ET ANALYSE DES RESULTATS

Le corpus d'apprentissage (20 heures d'informations françaises, échantillonnées à 16KHz) est issu des deux phases de la campagne ESTER. Les trois systèmes décrits précédemment ont tout d'abord été évalués sur les fichiers test de la phase II d'ESTER (10 heures). La seconde évaluation s'est faite sur de la parole conversationnelle qui apporte davantage de spontanéité et d'interaction entre les locuteurs. Nous avons pour cela utilisé le pré corpus EPAC composé de 10 émissions issues de trois radios différentes et d'une durée totale de 7 heures. Les zones de parole multi locuteurs représentent 5.4% du temps de ce corpus (contre 0.3% dans le corpus ESTER).

Les résultats obtenus pour chacun de ces corpus sont reportés dans les tables 3 à 5 ci-dessous. Il est à noter que les étapes de segmentation GLR/BIC, de regroupement BIC avec adaptation des seuils, de purification par F0 et de regroupement CLR avec adaptation des seuils du système hybride apportent chacune un gain minimum de 0.3 par rapport au système de base du LIUM. Les

résultats des tables 3 et 4 montrent la difficulté de traiter de la parole conversationnelle : en effet pour le corpus traditionnel ESTER le DER est de 10.4% mais il monte à 23.8% pour le corpus conversationnel d'EPAC. La table 5 montre le même type de résultats sur le même corpus mais en excluant cette fois-ci les zones multi locuteurs de l'évaluation: les résultats sont meilleurs que pour la table 4, cependant on n'atteint pas les taux affichés dans la table 3. Ceci est sûrement dû aux segments très courts de parole qui, même s'ils sont bien détectés, sont parfois regroupés de façon erronée par manque d'information.

Table 3 : Résultats sur le corpus ESTER pour les trois systèmes.

	IRIT	LIUM	**IRIT+LIUM**
Détection manquée de parole	1.7%	1.1%	**0.8**%
Fausse détection de parole	1.3%	0.8%	**0.8**%
Substitution de locuteurs	11.6%	9.7%	**8.8**%
DER (NIST speaker Diarization Error Rate)	14.6%	11.5%	**10.4**%

Table 4: Résultats sur le corpus conversationnel EPAC pour les trois systèmes.

	IRIT	LIUM	**IRIT+LIUM**
Détection manquée de parole	10.2%	8.5%	**10.7%**
Fausse détection de parole	3.1%	0.6%	**0.3%**
Substitution de locuteurs	22.2%	16.4%	**12.8%**
DER	35.4%	25.5%	**23.8**%

Table 5: Résultats sur le corpus conversationnel EPAC pour les trois systèmes : les zones multi locuteurs sont exclues de l'évaluation.

	IRIT	LIUM	**IRIT+LIUM**
Détection manquée de parole	5.3%	3.4%	**5.9%**
Fausse détection de parole	3.4%	0.7%	**0.3%**
Substitution de locuteurs	21.4%	17.1%	**13.2%**
DER	30.2%	21.2%	**19.4**%

6. CONCLUSION

Nous avons proposé un système hybride de speaker diarization qui conjugue les meilleurs aspects des systèmes du LIUM et de l'IRIT.

Le système du LIUM est un état de l'art basé sur une segmentation BIC suivie d'un regroupement CLR.

Le système de l'IRIT propose une séparation parole / non parole ainsi qu'une segmentation originale plus efficace que la méthode traditionnelle BIC.

Nous avons mis en évidence certains problèmes de seuillages liés à la nature du corpus : la parade a été d'utiliser des seuils adaptables au corpus de test.

D'autre part, outre les traditionnels types de corpus, nous avons évalué nos systèmes sur le pré corpus conversationnel du projet ANR EPAC pour lequel les résultats, bien qu'inférieurs que pour un corpus traditionnel, sont encourageants.

BIBLIOGRAPHIE

[1] M.A. Seigler, U. Jain, B. Raj and R.M Stern. Automatic Segmentation, Classification and Clustering of Broadcast News Audio. In *DARPA Speech Recognition Workshop*, 1997.

[2] M. Siu, H. Gish, and R. Rohlicek. Segregation of speaker for speech recognition and speaker identification. In *ICASSP*, pages 873-876, 1991.

[3] S.S. Chen and P.S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion. In *DARPA Speech Recognition Worshop* 1998.

[4] E. El Khoury, C. Sénac and R. André-Obrecht. Speaker Diarization: Towards a more Robust and Portable System. In *ICASSP,* pages 489-492, 2007.

[5] W.H. Tsai, S.S. Cheng, Y.H. Chao and H.M. Wang. Clustering speech utterances by speaker using Eigenvoice-Motivated vector space models. In *ICASSP*, pages 725-728, 2005.

[6] P. Deléglise, Y. Estève, S. Meignier, T. Merlin. The LIUM speech transcription system: a CMU Sphinx III-based system for french broadcast news. In *Interspeech*, pages 1653-1656, 2005.

[7] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J. Bonastre, and G. Gravier. The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In *Eurospeech*, pages 1149-1152, 2005.

[8] J. Pinquier, J-L Rouas, R. André-Obrecht. A Fusion Study in Speech/Music Classification. In *ICASSP*, pages 17-20, 2003.

[9] C. Barras, X. Zhu, S. Meignier, J.-L. Gauvain. Multi-stage speaker diarization of broadcast news. In *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pages 1505-1512, 2006.