Segmentation et Regroupement en Locuteur pour le traitement incrémental des collections volumineuses

Grégor Dupuy, Sylvain Meignier, Yannick Estève

Segmentation et Regroupement en Locuteur pour le traitement incrémental des collections volumineuses

Grégor Dupuy, Sylvain Meignier et Yannick Estève
LUNAM Université, LIUM, Le Mans, France
prenom.nom@lium.univ-lemans.fr

RÉSUMÉ

Les systèmes de Segmentation et Regroupement en Locuteurs *cross-show* actuels reposent principalement sur un processus de regroupement global qui traite collectivement chaque émission d'une collection. Cette approche a déjà été étudiée dans diverses situations et semble être le meilleur moyen à ce jour pour atteindre des taux d'erreur satisfaisants, dans une durée de traitement raisonnable. Néanmoins, ce processus montre ses limites dans un contexte applicatif réaliste où de grandes et dynamiques collections doivent être traitées. Dans cet article, nous étudions l'utilisation d'un regroupement cross-show incrémental pour traiter de manière itérative des émissions devant être insérées dans une collection existante. Les nouvelles émissions à insérer sont traitées les unes après les autres, selon l'ordre chronologique de diffusion. Les expériences ont été menées sur les enregistrements LCP et BFMTV distribués au cours des campagnes d'évaluation françaises ETAPE et REPERE. L'ensemble représente 67 heures de données annotées, réparties sur 310 enregistrements, couvrant une période d'environ deux ans (de septembre 2010 à octobre 2012).

ABSTRACT

Cross-show Speaker Diarization to Incrementally Process Large Volume of Data

Current cross-show diarization systems are mainly based on an overall clustering process that handles collectively each show of a collection. This approach has already been studied in various situations and seems to be the best way so far to achieve low error rates. Nevertheless, that process shows its limits in a realistic applicative context where large and dynamically increasing collections have to be processed. In this paper we investigate the use of an incremental clustering cross-show speaker diarization architecture to iteratively process new shows within an existing collection. The new shows to be inserted are processed one after another, according to the chronological order of broadcasting. Experiments were conducted on the LCP and the BFMTV show recordings distributed among the ETAPE and the REPERE French evaluation campaigns. It represents 67 hours of annotated data, distributed among 310 shows, and covering a 2-years period (from Sept. 2010 to Oct. 2012).

MOTS-CLÉS : SRL, architecture incrémentale, regroupement PLNE global, i-vecteurs.

KEYWORDS: speaker diarization, incremental architecture, cross-show, ILP clustering, i-vectors.

1 Introduction

La tâche de Segmentation et Regroupement en Locuteurs (SRL) joue un rôle essentiel dans de nombreuses applications de traitement automatique de la parole, comme la transcription automatique, la détection des entités nommées et du rôle des locuteurs. Cette tâche a été définie par le NIST dans le cadre des campagnes d'évaluation *Rich Transcription* comme le partitionnement d'un flux audio en segments, et le regroupement de ces segments en fonction de l'identité des locuteurs. Ce processus traite individuellement les enregistrements audio, et les locuteurs détectés sont identifiés par des étiquettes anonymes spécifiques à chaque enregistrement : un même locuteur impliqué dans deux émissions différentes est identifié par deux étiquettes différentes. Un inconvénient à cette approche est de ne pas prendre en compte les interventions de certains locuteurs récurrents dans plusieurs émissions, bien que cette situation se produise fréquemment dans les programmes audiovisuels d'information. En conséquence, la tâche de SRL a récemment été considérée dans un contexte plus large, dans lequel les émissions qui composent un corpus sont traitées collectivement pour regrouper les locuteurs impliqués dans plusieurs émissions. Ainsi, un locuteur cross-show est toujours identifié par la même étiquette dans chacune des émissions de la collection.

Cette extension de la tâche "Qui parle ? Quand ?" a conduit à l'émergence du concept de *collection*, qui se réfère à un ensemble de données avec une ou plusieurs caractéristiques communes entre les émissions. L'un des questionnements actuels concerne les collections dont le volume augmente dynamiquement au cours du temps. Des études antérieures ont montré que les architectures de SRL cross-show les mieux adaptées pour faire face à ce genre de situation sont de type *incrémentales* (Tran *et al.*, 2011)(Yang *et al.*, 2011), c'est à dire, traiter les émissions les unes après les autres en utilisant les modèles de locuteur des émissions déjà traitées dans le but d'aider la SRL de l'émission courante. L'objet de cet article est d'étudier l'utilisation d'un système de SRL dans un contexte applicatif réaliste : nous avons observé l'effet de l'insertion de nouveaux enregistrements dans une collection existante, en ajoutant et traitant les émissions les unes après les autres en fonction de leur date de diffusion.

Nous donnons un aperçu de l'état de l'art de la SRL incrémentale dans la partie 2, puis nous décrivons l'architecture incrémentale que nous avons expérimentée dans la section 3. Nous présentons ensuite les données, les métriques d'évaluation et les résultats expérimentaux dans la partie 4. Enfin, nous concluons dans la partie 5.

2 Travaux connexes

La tâche de SRL cross-show vise à étendre la tâche de SRL à un contexte plus large, où les locuteurs qui apparaissent dans plusieurs enregistrements d'une même collection (les *locuteurs cross-show*) sont toujours identifiés par la même étiquette. Les systèmes de SRL qui travaillent au niveau "enregistrement", et ne se préoccupent pas des relations entre les locuteurs impliqués dans plusieurs émissions, sont nommés, de manière contrastive, systèmes *single-show*. La tâche de SRL cross-show partage le même objectif que la tâche de "liaison des locuteurs" (speaker linking) (Ferràs et Bourlard, 2013)(Bourlard *et al.*, 2013)(Ghaemmaghami *et al.*, 2013). Cependant, dans ce procédé l'appariement global des étiquettes de locuteur est indépendant de la tâche SRL.

Les expériences menées précédemment sur le regroupement incrémental ont été réalisées sur

des collections de petite taille (23 enregistrements pour une durée totale d'environ 4h) (Tran *et al.*, 2011)(Yang *et al.*, 2011). L'architecture de SRL cross-show par regroupement incrémental proposé par les auteurs (Tran *et al.*, 2011)(Yang *et al.*, 2011), illustrée en Figure 1, fonctionne de la manière suivante : chaque enregistrement de la collection est d'abord traité individuellement jusqu'à un regroupement agglomératif hiérarchique BIC. Dans ce premier regroupement, les locuteurs sont modélisés par des Gaussiennes à matrice de covariance pleine, afin de maximiser la pureté des classes.

Puisqu'aucune information *a priori* n'est disponible pour traiter le premier enregistrement de la collection, la segmentation résultant du regroupement BIC est utilisée telle quelle pour procéder au regroupement agglomératif hiérarchique CLR. Dans ce second regroupement, les locuteurs sont désormais modélisés par des Modèles de Mixture Gaussiens (GMMs). Un module d'identification en locuteur (OSI module) a été inséré entre les regroupements BIC et CLR. Ce module mémorise les modèles de locuteur des enregistrements déjà traités. À partir du second enregistrement, les GMMs déjà extraits des enregistrements précédemment traités peuvent être utilisés dans le regroupement CLR courant. Après avoir traité intégralement une émission, le module d'identification en locuteur est mis à jour avec les nouveaux modèles de locuteurs, et les modèles de locuteur existant sont actualisés avec les données additionnelles que la dernière émission traitée peut fournir.

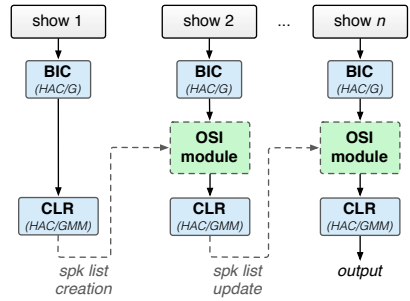

FIGURE 1 – Architecture incrémentale basée sur des regroupements agglomératifs hiérarchiques et des GMMs, incluant un module d'identification en locuteur (OSI module).

Cette approche cross-show de type incrémentale permet de traiter l'ensemble des enregistrements d'une collection bien plus rapidement que les architectures basées sur un regroupement de "global", qui traiteraient l'ensemble des enregistrements simultanément. Cette architecture est également adaptée aux situations où de nouveaux enregistrements viendraient régulièrement compléter une collection existante. Elle présente cependant deux inconvénients : le taux d'erreur DER (Diarization Error Rate) est plus élevé que celui obtenu par une approche cross-show basée sur un regroupement global, et l'ordre dans lequel les émissions sont traitées influence les résultats.

3 Regroupement incrémental global par PLNE

Motivés par l'approche incrémentale présentée dans la partie précédente, ainsi que par les récents résultats en SRL cross-show que nous obtenons avec le regroupement global par Programmation Linéaire en Nombres Entiers (PLNE), nous avons expérimenté un système de SRL regroupant ces deux aspects : architecture incrémentale et regroupement global par PLNE.

Nous avons considéré une collection initiale dans laquelle de nouveaux enregistrements ont été insérés et traités les uns après les autres (Figure 2). L'ordre dans lequel les émissions sont insérées dans la collection initiale suit l'ordre chronologique de diffusion. En considérant cet ordre chronologique, les n premières émissions ont été sélectionnées pour constituer la

collection initiale. Cette collection initiale est traitée avec une approche cross-show basée sur un regroupement global par PLNE (Dupuy *et al.*, 2012). La collection initiale a pour but d'amorcer le processus de SRL cross-show incrémental en considérant la segmentation cross-show et les modèles de locuteur de la collection initiale.

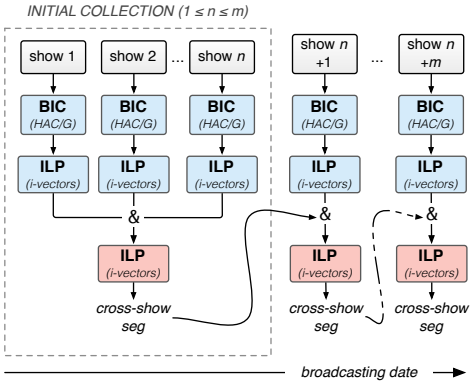

FIGURE 2 – Architecture cross-show incrémentale reposant sur un regroupement global par PLNE.

Le processus incrémental est réalisé comme suit : d'abord, chaque enregistrement de la collection est traité individuellement avec système de SRL single-show (*cf.* partie 3.1). Le traitement cross-show de la collection initiale consiste en un regroupement global par PLNE sur la concaténation des n segmentations single-show de la collection initiale (*cf.* section 3.2).

Pour chaque insertion d'une nouvelle émission, sa segmentation single-show est concaténée à la segmentation cross-show courante (qui résulte du traitement incrémental de l'émission précédemment insérée dans la collection). Un regroupement global par PLNE est alors appliqué, produisant une nouvelle segmentation cross-show dans laquelle la nouvelle émission a été prise compte. Le processus est ainsi répété jusqu'au traitement de la dernière émission de la collection.

3.1 SRL single-show

Le système de SRL single-show a été mis en place à l'aide de la boîte à outils *LIUM_SpkDiarization* [1] (Meignier et Merlin, 2009)(Rouvier *et al.*, 2013). Les systèmes conçus à partir de cette boîte à outils ont réalisé d'excellentes performances dans les tâches de SRL single- et cross-show lors des dernières campagnes d'évaluation francophones sur des émissions journalistiques, radio et TV, comme ETAPE (2011) (Gravier *et al.*, 2012) et REPERE (Jan. 2012, 2013 et 2014) (Galibert et Kahn, 2013).

Le système de SRL single-show est basé sur une segmentation acoustique et un regroupement agglomératif hiérarchique utilisant le critère BIC (Bayesian Information Criterion), avec des distributions Gaussiennes à matrices de covariance pleines. Les frontières des segments ont été ajustées par un décodage Viterbi utilisant des GMMs à 8 composantes appris sur les données de chaque locuteur via l'algorithme EM (Expectation-Maximization). Un autre décodage Viterbi, avec 8 HMMs (Hidden Markov Models) mono-état, représentés par des GMMs à 64 composantes entraînés par EM-ML (Expectation-Maximization Maximum Likelihood), est utilisé pour retirer les zones de non-parole. Segmentations, regroupements et décodages sont réalisés avec 12 paramètres MFCC (Mel-Frequency Cepstral Coefficients), complétés par l'énergie.

À cet instant, chaque locuteur n'est pas nécessairement représenté par une seule et unique classe. Un autre regroupement est donc effectué, exprimé sous la forme d'un problème de PLNE. Dans ce regroupement, les classes (les locuteurs) sont modélisées par des i-vecteurs et la similarité

1. `http://www-lium.univ-lemans.fr/en/content/liumspkdiarization`

entre les classes est estimée avec une distance de Mahalanobis (Rouvier et Meignier, 2012). Les i-vecteurs sont devenus état de l'art dans le domaine de la vérification du locuteur (Dehak *et al.*, 2011). Ils permettent de réduire les données acoustiques d'un locuteur en un vecteur de faible dimension, ne retenant que les informations dîtes "pertinentes" à son propos. L'approche à base de i-vecteurs a d'abord été adaptée à la SRL par l'intermédiaire de l'algorithme *K-means*, appliqué aux distances entre les i-vecteurs ,dans un corpus où le nombre de locuteurs est connu *a priori* (Shum *et al.*, 2011). Dans nos travaux, le nombre de locuteurs est inconnu.

Un i-vecteur de dimension 50 est extrait pour chacun des regroupements de la segmentation BIC, à l'aide d'un GMM-UBM de 256 composantes. Les i-vecteurs sont normalisés par leur longueur dans un processus itératif (Garcia-Romero et Espy-Wilson, 2011)(Bousquet *et al.*, 2011). Les paramètres acoustiques utilisés pour l'extraction des i-vecteurs correspondent à 12 MFCCs, complétés de l'énergie et des dérivées premières. Les paramètres acoustiques sont normalisés par la moyenne et la variance. Le problème de regroupement consiste, d'une part, à minimiser le nombre de classes centrales C choisi parmi les N i-vecteurs de la segmentation BIC, et d'autre part, à minimiser la dispersion des i-vecteurs au sein des classes. La valeur $C \in \{1, \ldots, N\}$ devant être déterminée automatiquement. Ce problème de classification est exprimé sous la forme d'un problème de PLNE (eq. 1), où la fonction objective de résolution est minimisée en vérifiant les contraintes suivantes :

$$\text{let } C \in \{1 \ldots N\}, \text{ let } K_{j \in C} = \{k / d(k,j) < \delta\}$$

$$\text{Minimize :} \qquad \sum_{k \in C} x_{k,k} + \frac{1}{N\delta} \sum_{j \in C} \sum_{k \in K_j} d(k,j) x_{k,j} \qquad (1)$$

$$\text{Subject to :} \qquad x_{k,j} \in \{0,1\} \qquad k \in K_j, j \in C \qquad (1.2)$$

$$\sum_{k \in K_j} x_{k,j} = 1 \qquad j \in C \qquad (1.3)$$

$$x_{k,j} - x_{k,k} < 0 \qquad k \in K_j, j \in C \qquad (1.4)$$

Où $x_{k,k}$ (eq. 1) est une variable binaire égale à 1 lorsque le i-vecteur k est un centre. Le nombre de centres C est implicitement inclus dans l'équation 1 ($C = \sum_{k=1}^{N} x_{k,k}$). La distance $d(k,j)$ est estimée par une distance de *Mahalanobis* entre les i-vecteurs k et j. $1/N\delta$ est un facteur de normalisation. La variable binaire $x_{k,j}$ est égale à 1 quand le i-vecteur j est assigné au centre k. Chaque i-vecteur j doit être associé à un seul centre k (eq. 1.3). L'équation 1.4 assure la sélection du centre k si un i-vecteur lui a été assigné. Les distances sont implicitement prises en compte dans les équations 1.2, 1.3 et 1.4, en utilisant l'ensemble K_j à la place de C. Pour une valeur j donnée, l'ensemble K_j représente l'ensemble des valeurs possibles de k (prise en 1 et N), pour lesquelles la distance entre les classes k et j sont plus petites que le seuil δ.

Le corpus de test fourni lors de la campagne d'évaluation française ESTER1(Galliano *et al.*, 2009) a été utilisé pour l'apprentissage du GMM-UBM, et le corpus d'apprentissage ESTER1 a été utilisé pour de l'apprentissage des i-vecteurs (étape de normalisation). Le problème de PLNE est résolu par l'algorithme *Branch and Bound* implémenté dans le solveur *glpsol* de la suite *GNU Linear Programming Toolkit* [2].

2. `http://www.gnu.org/software/glpk/`

3.2 SRL cross-show

La SRL cross-show consiste en un regroupement global par PLNE, dans lequel l'ensemble des émissions est traité simultanément. Ce regroupement global par PLNE est réalisé de la même manière que décrit dans la sous-partie précédente, cependant, la configuration à changé : le GMM-UBM de 256 composantes a été remplacé par un GMM-UBM à 1024 composantes, et les paramètres acoustiques correspondent désormais à 19 MFCCs, complétés par l'énergie ainsi que les dérivées premières et secondes. Ces paramètres restent normalisés par la moyenne et la variance.

L'aspect "cross-show" de ce regroupement PLNE repose sur la concaténation des segmentations single-show de plusieurs émissions. Le système de SRL cross-show présenté dans cet article diffère quelque peu de celui décrit dans (Dupuy *et al.*, 2012).

- Plutôt que de ré estimer les modèles de locuteur à partir de la concaténation des segmentations single-show, nous avons "recyclé" les modèles de locuteur (i-vecteurs) correspondant aux classes identifiées comme des centres, dans les segmentations single-show. Cette approche n'est possible que dans la mesure où les segmentations dont la concaténation est donnée en entrée ont été obtenues à partir de la même formulation du problème de PLNE.

- D'autre part, l'équation 4 du problème de PLNE, formulée dans la sous-partie précédente, est utilisée en amont : la matrice associant les distances aux couples de i-vecteurs peut être interprétée comme un graphe complètement connecté, dans lequel les nœdus représentent les locuteurs, et les arêtes représentent les distances entre les i-vecteurs. Il est possible d'identifier les sous-composantes connexes de ce graphe en éliminant les arêtes correspondant aux distances inférieures au seuil δ. Chaque sous-composante connexe représente un sous-ensemble indépendant du problème de regroupement. La recherche de sous-composantes connexes permet ainsi de découper le problème de PLNE, tel qu'originalement formulé, en un certain nombre de sous-problèmes élémentaires, qui ne consistent qu'en un faible nombre de variables et contraintes. Considérer le problème de regroupement en locuteur comme la recherche des sous-composantes connexes d'un graphe totalement connecté a permis de diviser par 5 le temps de calcul de ce regroupement global par PLNE sur les 142h de données que nous introduisons dans la partie 4.1 (2'47 minutes au lieu de 13'23 minutes).

Il est important de noter que nous avons exprimé notre problème de regroupement global par PLNE de manière à manipuler globalement la concaténation des segmentations single-show : le système de SRL cross-show est libre de regrouper les classes issues d'une même émission. En effet, nous avons constaté que le regroupement global par PLNE permet souvent d'améliorer, quelque peu, les segmentations single-show. Cette particularité explique pourquoi le DER single-show est amené à évoluer au cours des itérations de l'approche cross-show incrémentale.

4 Expériences

4.1 Données

Les données considérées pour réaliser les expériences consistent en un regroupement de toutes les émissions BFMTV et LCP distribuées dans les corpus d'apprentissage, de développement et

de test des campagnes françaises ETAPE et REPERE Phase 1. Nous avons également inclus les données d'apprentissage fournies pour la phase 2 du défi REPERE. L'ensemble représente un total de 142 heures de données audio, réparties sur 310 fichiers, enregistrés entre septembre 2010 et octobre 2012. La distribution des données en termes de nombre d'émissions et de durées est présentée dans le Tableau 1. Seule une partie de ces données a été annotée et évaluée : les durées présentées dans le Tableau 1 (exprimées en heures et minutes) ont été calculées d'après les fichiers UEM (Unpartitioned Evaluation Map) spécifiant les régions à évaluer.

Show name	Whole data		Initial collection	
	# Show	Duration	# Show	Duration
BFMStory	37	21 :21	7	05 :01
CultureEtVous	54	01 :42	0	-
PlaneteShowbiz	73	02 :24	1	02 :07
All BFMTV	*164*	*25 :27*	*8*	*07 :08*
CaVousRegarde	20	09 :02	9	04 :29
EntreLesLignes	24	08 :20	10	04 :05
LCPInfo13h30	31	07 :57	0	-
LCPInfo20h30	4	00 :40	1	00 :10
PileEtFace	32	08 :15	12	04 :07
TopQuestions	35	07 :41	9	02 :03
All LCP	*146*	*41 :55*	*41*	*15 :44*
BFMTV+LCP	*310*	*67 :22*	*49*	*22 :52*

TABLE 1 – Distribution des données, en nombre d'émissions et en durées (h :min), pour la totalité des données BFMTV+LCP, ainsi que pour la collection initiale.

Les 49 premières émissions, selon l'ordre chronologique de diffusion, ont été sélectionnées pour constituer la collection initiale. Les émissions restantes ont été itérativement ajoutées à la collection initiale. Cette collection initiale, qui représente environ la moitié des données en termes de durée, a été utilisée pour déterminer les seuils des systèmes de SRL single- et cross-show.

4.2 Résultats et discussion

Le DER (Diarization Error Rate) est la métrique utilisée pour mesurer les performances en SRL. Le DER fut introduit par le NIST comme la fraction du temps de parole qui n'est pas attribué au bon locuteur, en utilisant la meilleure correspondance entre les étiquettes des références et des hypothèses. L'outil d'évaluation a été développé par le LNE [3] dans le cadre des campagnes ETAPE et REPERE. Cet outil d'évaluation a été conçu de manière à calculer les DER single-show et cross-show.

Les DER single- et cross-show obtenus lors de chaque itération de l'expérience incrémentale sont présentés dans la Figure 3. Les droites horizontales représentent les DER single- et cross-show obtenus par regroupement global PLNE sur l'ensemble des 310 émissions de la collection. La collection est composée d'émissions enregistrées sur une période de deux ans, avec quelques ruptures temporelles notables entre deux enregistrements consécutifs. Les droites verticales interrompues, sur la Figure 3, représentent les interruptions temporelles de plus de 12 jours consécutifs. En moyenne, la durée écoulée entre deux enregistrements consécutifs est d'une journée.

3. Laboratoire National de métrologie et d'Éssais

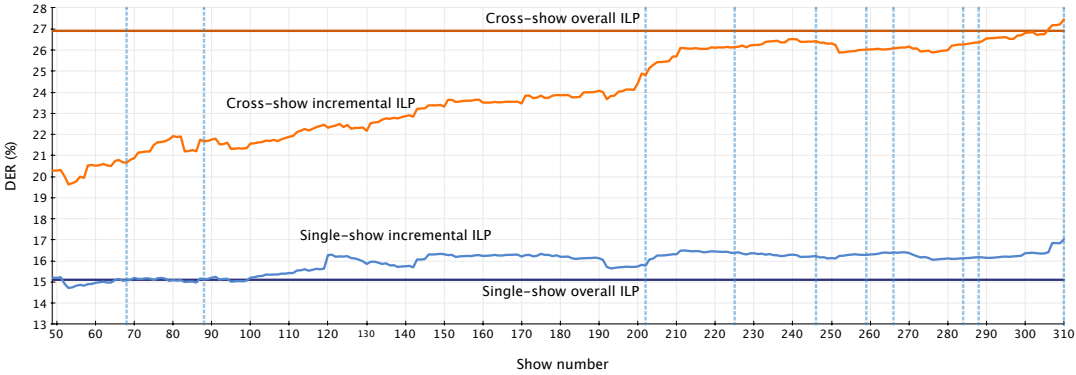

FIGURE 3 – DER single- et cross-show pour chaque itération de l'expérience incrémentale

Le DER cross-show obtenu avec le regroupement global par PLNE sur les 310 émissions de la collection est élevé (26,91%), comparé à celui obtenu sur les 49 enregistrements de la collection initiale (20,28%). Le DER obtenu avec l'approche incrémentale, après que la dernière émission ait été traitée, est de 27,45%. L'évaluation single-show des segmentations cross-show donne des DER plus élevés avec l'approche incrémentale (17,02%), comparé aux DER obtenus par regroupement global PLNE sur l'ensemble des données (15,11%). Une explication par rapport à la brutale augmentation des DER entre les émissions 200 et 210 pourrait être l'interruption temporelle entre les émissions 201 et 202, qui est de 24 jours. Cependant, ce comportement étrange ne s'observe pas au niveau des autres interruptions temporelles. Retirer ce bloc d'émissions de la collection permet d'obtenir des résultats globalement meilleurs, il convient donc de penser que certains locuteurs posent problème dans ces émissions. 55,8% des locuteurs de la collection initiale sont également présents dans le reste des données, qui sont traitées itérativement. De plus, 12 locuteurs sont particulièrement présents (la plupart d'entre eux correspondent à des présentateurs et des personnalités fortement médiatisées). Le temps de parole de ces 12 locuteurs représente environ 30,6% du temps de parole évalué sur les 310 émissions.

5 Conclusion

La complexité de la tâche de Segmentation et Regroupement en Locuteurs cross-show, en ce qui concerne les programmes audiovisuels d'information, augmente en fonction de la quantité de données à traiter. L'approche incrémentale peut convenir au traitement des collections volumineuses, ou susceptibles de prendre du volume au cours du temps. Les résultats sont très proches de ceux obtenus par regroupement global sur l'ensemble des données, mais ce qui rend l'approche incrémentale vraiment intéressante est le rapport résultats/temps de calcul. Considérer le problème de regroupement en locuteur comme la recherche des sous-composantes connexes d'un graphe totalement connecté a permis de diviser par 5 le temps de calcul. Le "recyclage" des modèles de locuteurs en fonction les regroupements effectués en amont permet d'économiser les ressources de manière non négligeable et d'effectuer les regroupements à la chaîne. Traiter une collection de 142 heures de données avec un regroupement global par PLNE reste une solution envisageable, cependant, cette approche globale risque de montrer ses limites avec des collections bien plus conséquentes en volume.

Références

Bourlard, H., Ferras, M., Pappas, N., Popescu-Belis, A., Renals, S., McInnes, F., Bell, P. et Guillemot, M. (2013). Processing and Linking Audio Events in Large Multimedia Archives : The EU inEvent Project. *In Proceedings of Interspeech satellite workshop on Speech, Language and Audio in Multimedia (SLAM)*, Marseille, France.

Bousquet, P.-M., Matrouf, D. et Bonastre, J.-F. (2011). Intersession Compensation and Scoring Methods in the I-vectors Space for Speaker Recognition. *In Proceedings of Interspeech*, Florence, Italie.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P. et Ouellet, P. (2011). Front-End Factor Analysis for Speaker Verification. *In Proceedings of IEEE TASLP*, volume 19, pages 788–798.

Dupuy, G., Rouvier, M., Meignier, S. et Estève, Y. (2012). I-vectors and ILP Clustering Adapted to Cross-Show Speaker Diarization. *In Proceedings of Interspeech*, Portland, Oregon (USA).

Ferràs, M. et Bourlard, H. (2013). Speaker Diarization and Linking of Large Corpora. *In Proceedings of IEEE Spoken Language Technology Workshop*, Miami, Floride (USA).

Galibert, O. et Kahn, J. (2013). The first official REPERE evaluation. *In Proceedings of Interspeech satellite workshop on Speech, Language and Audio in Multimedia (SLAM)*, Marseille, France.

Galliano, S., Gravier, G. et Chaubard, L. (2009). The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts. *In Proceedings of Interspeech*, Brighton, UK.

Garcia-Romero, D. et Espy-Wilson, C. Y. (2011). Analysis of I-vector Length Normalization in Speaker Recognition Systems. *In Proceedings of Interspeech*, Florence, Italie.

Ghaemmaghami, H., Dean, D. et Sridha, S. (2013). Speaker Attribution of Australian Broadcast News Data. *In Proceedings of Interspeech satellite workshop on Speech, Language and Audio in Multimedia (SLAM)*, Marseille, France.

Gravier, G., Adda, G., Paulsson, N., Carré, M., Giraudel, A. et Galibert, O. (2012). The ETAPE Corpus for the Evaluation of Speech-based TV Content Processing in the French Language. *In Proceedings of LREC*, Istanbul, Turkey.

Meignier, S. et Merlin, T. (2009). LIUM SpkDiarization : An Open-Source Toolkit for Diarization. *In CMU SPUD Workshop*, Dallas, Texas (USA).

Rouvier, M., Dupuy, G., Gay, P., Khoury, E., Merlin, T. et Meignier, S. (2013). An Open-source State-of-the-art Toolbox for Broadcast News Diarization. *In Proceedings of Interspeech*, Lyon, France.

Rouvier, M. et Meignier, S. (2012). A Global Optimization Framework For Speaker Diarization. *In Odyssey Workshop*, Singapore.

Shum, S., Dehak, N., Chuangsuwanich, E., Reynolds, D. et Glass, J. (2011). Exploiting Intra-Conversation Variability for Speaker Diarization. *In Proceedings of Interspeech*, Florence, Italie.

Tran, V.-A., Le, V. B., Barras, C. et Lamel, L. (2011). Comparing Multi-Stage Approaches for Cross-Show Speaker Diarization. *In Proceedings of Interspeech*, Florence, Italie.

Yang, Q., Jin, Q. et Schultz, T. (2011). Investigation of Cross-show Speaker Diarization. *In Proceedings of Interspeech*, Florence, Italie.