PRES LUNAM
École Doctorale STIM
Sciences et Technologies de
l'Information et Mathématiques

Spécialité : Informatique
Laboratoire : LIUM
Equipe : LST

Segmentation et Regroupement en Locuteurs adapté au traitement de collection d'émissions audio

Dupuy Grégor
Mél : gregor.dupuy@lium.univ-lemans.fr

Résumé : Dans cet article nous présentons un système de segmentation et regroupement en locuteur adapté au traitement des collections de documents audio volumineuses. L'objectif est de détecter les locuteurs intervenant dans plusieurs émissions afin de les identifier par une même étiquette dans toutes les émissions de la collection. Dans notre approche, chaque émission est d'abord traitée individuellement avant de considérer la collection dans son intégralité. Le processus de regroupement global est réalisé par la résolution d'un problème ILP dans lequel chacun des locuteurs issus du traitement individuel est représenté par une classe. Le problème ILP cherche à minimiser le nombre de classes ainsi que la dispersion intra-classe. Ce système a été évalué sur un ensemble de collections constituées à partir des données de la campagne d'évaluation ESTER2. Notre processus de regroupement global par ILP permet d'obtenir des taux d'erreur similaires à ceux obtenu avec les méthodes état-de-l'art tout en réduisant considérablement le temps de calcul. Cette méthode est, par conséquent, adaptée au traitement de collections volumineuses.

Mots clés : *Segmentation et regroupement en locuteur, architecture hybride, regroupement ILP, i-vectors.*

1 Introduction

La tâche de segmentation et de regroupement en locuteurs (SRL) "Qui parle, quand ?", ou *speaker diarization*, a été définie par le NIST [1] lors des campagnes d'évaluation *Rich Transcription* comme le découpage d'un flux audio en tours de parole et le regroupement des plages associées à un même locuteur. Le procédé de SRL est appliqué individuellement sur chacun des enregistrements audio d'un corpus, sans utiliser de connaissances *a priori* sur les locuteurs. Dans ce contexte, les locuteurs détectés par les systèmes sont identifiés par des étiquettes anonymes propres à chaque enregistrement : un même locuteur intervenant dans deux enregistrements différents est identifié par deux étiquettes différentes (SRL *single-show*).

En considérant la quantité toujours croissante de ressources multimédia, la non prise en compte des interventions récurrentes de certains locuteurs dans plusieurs émissions se révèle être un obstacle à de nombreuses applications de traitement automatique de la parole. Cette situation de récurrence est pourtant très fréquente dans les émissions journalistiques où, généralement, les présentateurs, journalistes et autres invités qui les animent apparaissent régulièrement. La SRL sur une collection d'émissions est une approche récemment introduite visant à détecter et regrouper globalement les locuteurs sur l'ensemble des émissions d'une collection. Ainsi, un locuteur intervenant dans plusieurs émissions est identifié par la même étiquette dans chacune de ces émissions (SRL *cross-show*). Dans cet article, nous présentons un système de SRL adapté au traitement de collections volumineuses. Ce système implémente une architecture *hybride* [1][2] dans laquelle les émissions de la collection sont d'abord traitées individuellement, avant d'être considérées globalement par la résolution d'un problème de Programmation Linéaire en Nombres Entiers (ILP *Integer Linear Programming*) [3].

Dans les paragraphes suivants, nous décrivons l'architecture implémentée en présentant d'abord le système de SRL *single-show* du LIUM [2], utilisé pour traiter les émissions individuellement, et ensuite, le problème ILP permettant de traiter la collection dans son ensemble (SRL *cross-show*). Nous présentons, avant de conclure, les corpus utilisés pour évaluer notre approche ainsi que nos résultats expérimentaux.

2 Architecture du système

Avec l'architecture *hybride* [1][2], schématisée en figure 2, les différentes émissions de la collection sont d'abord considérées individuellement en réalisant une SRL propre à chacune (SRL *single-show*). Cette

1. National Institute of Standards and Technology
2. Laboratoire d'Informatique de l'Université du Maine

première étape est accomplie avec le système développé par le LIUM, présenté en section 2.1. À l'issue de ce processus, les locuteurs d'une émission sont identifiés par des étiquettes anonymes et associés aux régions temporelles de l'enregistrement audio.

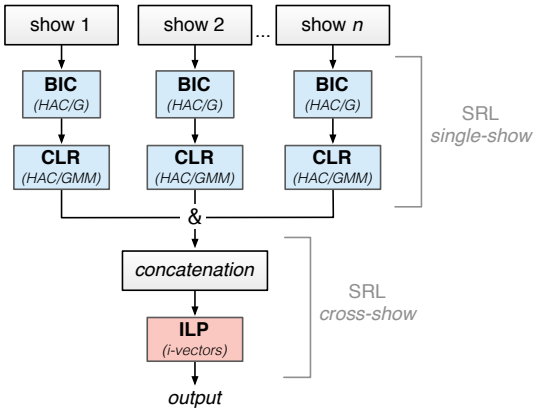

FIGURE 1 – Architecture du système de SRL pour une collection d'émissions, présentant les traitements individuel *(single-show)* et global *(cross-show)*.

Les sorties de cette première étape sont ensuite *concaténées* afin d'être traitées de manière globale par ILP (SRL *cross-show*). Dans ce processus global, chacun des locuteurs issus de la première étape est modélisé par une classes représentant les caractéristiques acoustiques de sa voix. Le but de cette deuxième étape est de regrouper les classes correspondant à un même locuteur, en fonction de leur similarité, afin de lui attribuer une étiquette unique au sein des différentes émissions. Ce regroupement global est détaillé en section 2.2.

2.1 SRL *single-show*

Le système utilisé lors de nos expériences, le *LIUM_SpkDiarization toolkit*[3] [4], a été développé pour la campagne d'évaluation française ESTER2 [5], où il a obtenu les meilleurs résultats dans la tâche de SRL sur des émissions journalistiques.

Le *LIUM_SpkDiarization* est composé d'une segmentation acoustique et d'une classification hiérarchique utilisant BIC (Bayesian Information Criterion) comme mesure de similarité entre les locuteurs et comme critère d'arrêt. Chaque locuteur est modélisé par une gaussienne à matrice de covariance pleine. Une segmentation en zones de parole/non-parole est également réalisée afin de retirer les zones de non-parole des segments. Segmentation, classification et décodage sont réalisés à partir de 12 paramètres MFCC (Mel-Frequency Cepstral Coefficients), complétés de l'énergie.

À ce stade, chaque locuteur n'est pas forcément représenté par une seule classe. Le système réalise alors une classification hiérarchique utilisant un rapport de vraisemblance croisé (CLR) [6] comme mesure de similarité entre les classes ainsi que comme critère d'arrêt. Contrairement aux étapes précédentes, les paramètres acoustiques sont normalisés (centrés/réduits + feature warping calculé sur chaque segment). L'objectif de la normalisation des paramètres est de minimiser la contribution du canal. Les modèles de locuteur sont obtenus par une adaptation MAP (Maximum A Posteriori) des moyennes d'un modèle du monde (UBM - Universal Background Model) sur les données de chaque classe.

2.2 SRL *cross-show*

Les i-vecteurs, utilisés principalement dans le domaine de la vérification du locuteur [7], permettent de réduire de grandes quantités de données acoustiques en vecteurs de dimensions réduites, en ne conservant que les informations pertinentes des locuteurs. Cette approche a été adaptée à la SLR en utilisant l'algorithme k-means, appliqué à la distance entre les i-vecteurs, pour détecter les interventions des locuteurs au sein de corpus où le nombre de locuteurs est *a priori* connu [8].

Ici, le nombre de locuteurs est inconnu. Un i-vecteur j est extrait à partir de chacune des classes j issues de la classification *single-show*. Le problème de classification consiste à minimiser, d'une part, le

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

nombre K de classes centrales choisies parmi les N i-vecteurs et, d'autre part, la dispersion des i-vecteurs au sein de ces classes (la valeur $K \in \{1, \ldots, N\}$ devant être déterminée automatiquement).

Nous proposons d'exprimer ce problème de classification à l'aide d'un Programme Linéaire en Nombre Entier, où la fonction objective de résolution (eq. 1) est minimisée en vérifiant les contraintes :

Minimize

$$\sum_{k=1}^{N} x_{k,k} + \frac{1}{D} \sum_{k=1}^{N} \sum_{j=1}^{N} d(k,j) x_{k,j} \qquad (1)$$

Subject to

$$x_{k,j} \in \{0,1\} \qquad \forall k, \forall j \qquad (1.2)$$

$$\sum_{k=1}^{N} x_{k,j} = 1 \qquad \forall j \qquad (1.3)$$

$$d(k,j) x_{k,j} \leq \delta \qquad \forall k, \forall j \qquad (1.4)$$

$$x_{k,j} - x_{k,k} \leq 0 \qquad \forall j \qquad (1.5)$$

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 K est implicitement inclus dans l'équation 1 ($K = \sum_{k=1}^{N} x_{k,k}$). La distance $d(k,j)$ est calculée en utilisant la distance de *Mahanalobis* entre les i-vecteurs k et j [9]. D est un facteur de normalisation égal à la plus grande distance $d(k,j)$ pour chaque k et j. 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 et unique centre k (eq. 1.3). Le i-vecteur j associé au centre k (*i.e.* $x_{k,j} = 1$) doit avoir une distance $d(k,j)$ inférieure à un seuil δ déterminé expérimentalement (eq. 1.4). L'équation 1.5 assure qu'un seul i-vecteur k est sélectionné pour être un centre.

3 Expériences

3.1 Données d'évaluation

Les données sélectionnées pour réaliser nos expériences représentent l'intégralité du corpus d'apprentissage de la campagne d'évaluation française ESTER2 [5]. Ces données représentent exactement 100 heures d'émission radiophoniques, enregistrées entre 1999 et 2003 et manuellement annotées en locuteur. Ce corpus d'apprentissage a été divisé en treize sous-ensembles sur lesquels nous avons réalisé des expériences indépendantes. La distribution des données au sein des treize collections a été effectuée en fonction de l'année et de la plage horaire de chaque émission. Nous avons utilisé le premier de ces treize corpus comme corpus de développement, afin de configurer au mieux notre système, et nous avons considéré les douze autres corpus comme des corpus de test, sur lesquels le système configuré a été appliqué tel quel.

3.2 Métriques d'évaluation

La métrique d'évaluation choisie pour mesurer les performances est le DER (Diarization Error Rate), introduit par le NIST comme la fraction de temps de parole qui n'est pas attribuée au bon locuteur, en utilisant une correspondance optimale entre l'étiquetage des locuteurs des références et des hypothèses. On distingue deux taux d'erreur différents : le DER *single-show*, lorsque l'évaluation est réalisée en considérant les émissions indépendamment les unes des autres, et le DER *cross-show*, lorsque l'évaluation est réalisée simultanément sur toutes les émissions de la collection. Le DER *single-show* correspond à la moyenne des DER mesurés sur chaque émission, pondérés par leurs durées. Le DER *cross-show* tient compte de la réapparition des locuteurs dans plusieurs émissions.

3.3 Résultats

Afin de mesurer l'impact de notre méthode de regroupement global par ILP, nous avons réalisé les mêmes expériences avec un système contrastif implémentant un regroupement agglomératif hiérarchique global, utilisant la mesure CLR à la fois comme critère de similarité entre les classes et comme critère d'arrêt du processus de regroupement.

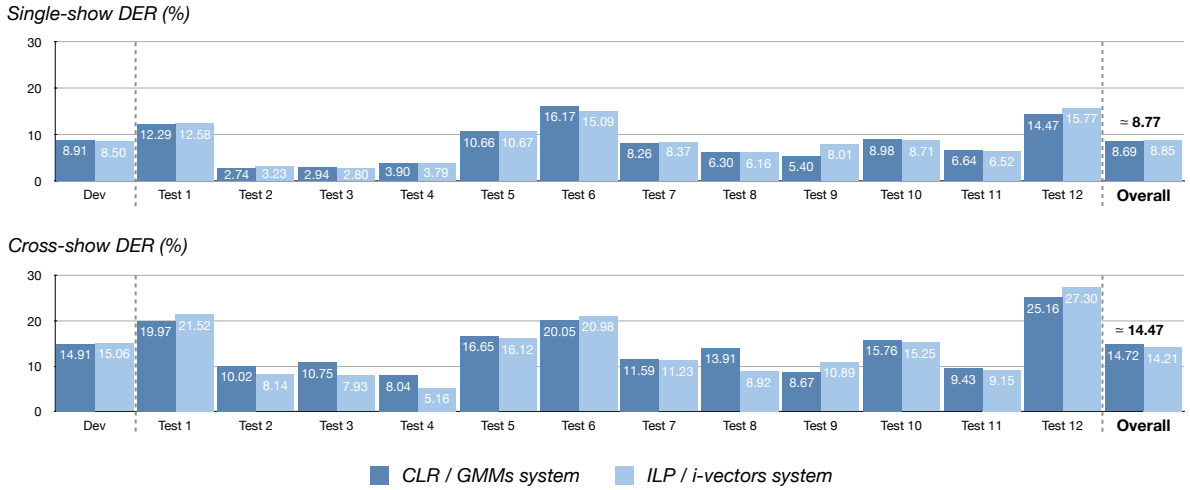

FIGURE 2 – Architecture du système de SRL pour une collection d'émissions, présentant les traitements individuel *(single-show)* et global *(cross-show)*.

Ce système contrastif, bien que couteux en ressource, permet d'obtenir les meilleurs résultats en termes de taux d'erreur [1] [2]. Les expériences ont montré, d'une part, que les taux d'erreur *single-show* et *cross-show* sont similaires entre les deux systèmes et, d'autre part, que le regroupement global ILP est beaucoup rapide. Nous avons mesuré le temps de calcul nécessaire au regroupement global des deux systèmes, sur chaque collection-test. À titre de comparaison, le temps moyen de calcul pour les regroupement globaux par ILP et CLR sont les suivant :

Collections de 5 heures : 18 minutes vs 1 :50 heures,
Collections de 10 heures : 1 :58 heures vs 10 :57 heures,
Collections de 15 heures : 3 :55 heures vs 60 :24 heures.

4 Conclusion

Nous avons proposé une nouvelle approche adaptée à la tâche segmentation et de regroupement en locuteurs pour une collection de documents. Dans cette approche, les locuteurs sont modélisés par des i-vecteurs et la classification en elle-même est exprimée sous la forme d'un problème ILP sur la distance entre les i-vecteurs. Les performances du système implémentant cette approche sont comparables, en termes de DER, à celles du système contrastif implémentant la classification globale par CLR. Néanmoins, le regroupement global par ILP est beaucoup plus rapide et consomme moins de ressources.

Références

[1] Viet-Anh Tran, Viet Bac Le, Claude Barras, and Lori Lamel. Comparing multi-stage approaches for cross-show speaker diarization. In *Proceedings of Interspeech*, Florence, Italie, 2011.

[2] Qian Yang, Qin Jin, and Tanja Schultz. Investigation of cross-show speaker diarization. In *Proceedings of Interspeech*, Florence, Italie, 2011.

[3] Mickael Rouvier and Sylvain Meignier. A global optimization framework for speaker diarization. In *Odyssey Workshop*, Singapore, 2012.

[4] Sylvain Meignier and Teva Merlin. LIUM SpkDiarization : an open-source toolkit for diarization. In *CMU SPUD Workshop*, Dallas, Texas (USA), 2009.

[5] Sylvain Galliano, Guillaume Gravier, and Laura Chaubard. The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. In *Proceedings of Interspeech*, Brighton, UK, 2009.

[6] Viet Bac Le, Odile Mella, and Dominique Fohr. Speaker diarization using normalized cross-likelihood ratio. In *Proceedings of Interspeech*, Antwerp, Belgique, 2007.

[7] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. In *Proceedings of IEEE TASLP*, volume 19, pages 788–798, 2011.

[8] Stephen Shum, Najim Dehak, Ekapol Chuangsuwanich, Douglas Reynolds, and Jim Glass. Exploiting intra-conversation variability for speaker diarization. In *Proceedings of Interspeech*, Florence, Italie, 2011.

[9] Pierre-Michel Bousquet, Driss Matrouf, and Jean-François Bonastre. Intersession compensation and scoring methods in the i-vectors space for speaker recognition. In *Proceedings of Interspeech*, Florence, Italie, 2011.