Détection de locuteurs dans les séries TV

Xavier Bost* — **Georges Linarès***

* *Laboratoire Informatique d'Avignon*

RÉSUMÉ. *La segmentation de flux audio en locuteurs apparaît particulièrement délicate lorsqu'elle est appliquée à des films de fiction, où de nombreux personnages parlent dans des conditions acoustiques variables (musique de fond, bruitages, fluctuations dans l'intonation...). Au-delà d'une telle variabilité acoustique, ce type de films exhibe cependant de la régularité sur le plan visuel, particulièrement dans les passages dialogués. Nous introduisons dans ce papier une méthode en deux temps pour procéder à la segmentation en locuteurs d'épisodes de séries* TV *: un premier regroupement en locuteurs est effectué localement, dans les limites de scènes visuellement identifiées comme des dialogues ; les locuteurs conjecturés sont ensuite comparés lors d'une deuxième phase de regroupement afin de détecter les locuteurs récurrents : cette deuxième étape de regroupement a lieu sous la contrainte que les différents locuteurs impliqués dans un même dialogue soient assignés à des groupes distincts. Les performances obtenues par notre approche sont comparées à celles qu'on obtient en appliquant aux mêmes données des outils standards de segmentation en locuteurs.*

ABSTRACT. *Speaker diarization of audio streams turns out to be particularly challenging when applied to fictional films, where many characters talk in various acoustic conditions (background music, sound effects, variations in intonation...). Despite this acoustic variability, such movies exhibit specific visual patterns, particularly within dialogue scenes. In this paper, we introduce a two-step method to achieve speaker diarization in* TV *series: speaker diarization is first performed locally within scenes visually identified as dialogues; then, the hypothesized local speakers are compared to each other during a second clustering process in order to detect recurring speakers: this second stage of clustering is subject to the constraint that the different speakers involved in the same dialogue have to be assigned to different clusters. The performances of our approach are compared to those obtained by standard speaker diarization tools applied to the same data.*

MOTS-CLÉS : *Segmentation en locuteurs$_1$, structuration de vidéos$_2$, regroupement non supervisé$_3$.*

KEYWORDS: *Speaker diarization$_1$, video structuration$_2$, unsupervised clustering$_3$.*

1. Introduction

La recherche d'information dans les flux audiovisuels rencontre une difficulté majeure liée à l'absence fréquente de métadonnées décrivant les contenus et la façon dont ils sont structurés (Quénot *et al.*, 2010), (Lew *et al.*, 2006). Extraire l'organisation sous-jacente des flux de contenus est donc une étape critique du processus d'indexation, qui comporte deux volets. D'une part, il faut segmenter le flux continu en un ensemble de segments homogènes. D'autre part, il faut caractériser les segments extraits par des descripteurs qui peuvent être de natures très diverses (Adams *et al.*, 2003). Il peut s'agir de descripteurs visuels pour la segmentation en plans ou en scènes, de descripteurs des contenus sémantiques pour une segmentation thématique, des descripteurs caractérisant le style éditorial d'un segment dans la segmentation en genres vidéo... De nombreuses applications considèrent les contenus parlés et l'identité des locuteurs comme un élement essentiel à la structuration. Dans cet article, nous nous intéressons à la segmentation en locuteurs dans le cadre, assez général, de l'indexation de grandes bases audiovisuelles.

La détection des locuteurs impliqués dans des contenus audio est usuellement conçue comme une tâche de segmentation du flux, qui consiste à assigner les différents segments parlés à leurs locuteurs respectifs. Cette tâche est donc habituellement accomplie en deux temps, simultanés ou successifs : détection des points de changement entre locuteurs ; regroupement des segments résultants en classes de locuteurs. Ce regroupement est souvent mené en appliquant un algorithme de regroupement hiérarchique, soit ascendant, soit descendant (Evans *et al.*, 2012).

Le regroupement des segments en classes de locuteurs est mené de manière non supervisée : le nombre total de locuteurs en particulier n'est pas *a priori* fixé et le partitionnement optimal des segments parlés en classes de locuteurs doit être automatiquement déterminé.

Les systèmes de segmentation en locuteurs ont d'abord été conçus pour traiter des flux audio aux conditions acoustiques défavorables, mais bien balisées : conversations téléphoniques, journaux d'information, réunions de travail. Des travaux récents ont appliqué ces systèmes aux contenus vidéos, dont le contexte de production, moins bien défini, induit une plus forte variabilité.

Dans (Clément *et al.*, 2011), les auteurs appliquent des outils standards de segmentation en locuteurs à la source audio de documents vidéos de différents genres collectés sur le web. Les performances mesurées se dégradent sensiblement pour les dessins animés et les bandes-annonces de films, avec des taux d'erreurs de segmentation en locuteurs (*Diarization Error Rate*, DER) élevés : parmi les raison incriminées, les auteurs mentionnent le nombre élevé de locuteurs impliqués dans ces flux, ainsi que de fortes variabilités dans l'environnement acoustique (superposition de la parole et de la musique de fond, bruitages...).

De même, les taux d'erreurs reportés dans (Ercolessi, 2013) après application d'un outil standard de segmentation en locuteurs au canal audio de quatre épisodes d'une série TV apparaissent très élevés (DER $\simeq 70\%$).

Récemment, plutôt que d'aborder les flux vidéo par le seul canal audio, certains travaux se sont concentrés sur des approches multimodales pour procéder à la segmentation en locuteurs : dans (Friedland *et al.*, 2009), les auteurs évaluent une méthode fondée sur une fusion précoce de mélanges de gaussiennes audio et vidéo suivie de l'application d'un algorithme agglomératif au flux bicanal qui en résulte. Cette technique est évaluée sur le corpus AMI (Carletta *et al.*, 2006), constitué d'enregistrements audiovisuels de quatre participants jouant des rôles dans un scénario de réunion.

Appliquer un système de segmentation en locuteurs à des épisodes de séries télévisées, où le nombre de locuteurs est globalement plus élevé que dans des longs métrages, peut donc sembler particulièrement délicat. Toutefois, les films de fiction exhibent sur le plan visuel de nombreuses régularités formelles. Le montage des scènes dialoguées en particulier exige que la règle des « 180 degrés » soit respectée afin que deux interlocuteurs filmés alternativement donnent l'impression de se regarder en se parlant : pour suggérer le croisement des regards, l'un doit regarder vers la droite de l'écran et l'autre vers la gauche. Les deux caméras qui les filment doivent donc se situer du même côté d'une ligne imaginaire qui les relierait l'un à l'autre. L'observation d'une telle règle induit un motif d'alternance entre deux plans récurrents, caractéristique des scènes dialoguées.

Sur la base de ces régularités formelles, nous proposons ici de décomposer le processus de segmentation en locuteurs en deux étapes lorsqu'il est appliqué aux films de fiction : une première passe de regroupement des segments en classes de locuteurs est effectuée localement, dans les limites de courtes scènes visuellement détectées comme des dialogues ; une seconde étape vise à détecter les locuteurs récurrents d'une scène à l'autre en procédant à un second regroupement des classes de locuteurs conjecturées localement ; cette seconde phase de regroupement est opérée sous la contrainte que les locuteurs impliqués dans une même scène ne puissent pas être regroupés dans la même classe.

Un tel regroupement en deux temps est étroitement apparenté à ce qui est désigné dans (Tran *et al.*, 2011) sous le terme d'« architecture hybride » dans un contexte de segmentation en locuteurs d'émissions successives : les locuteurs des différentes émissions sont d'abord détectés indépendamment, avant que les locuteurs récurrents d'une émission à l'autre ne soient regroupés. Dans (Bendris *et al.*, 2013), les auteurs s'efforcent de combiner segmentation en locuteurs et regroupement des visages pour procéder à l'identification des personnes impliquées dans un débat filmé : la meilleure des deux modalités pour identifier une personne est retenue, avant que l'information acquise ne soit propagée aux autres éléments de la classe associée. Enfin, la segmentation en locuteurs a été appliquée aux séries TV, mais plutôt comme une méthode, parmi d'autres sources, de segmentation du flux en scènes présentant une unité narrative. Dans (Bredin, 2012), les performances obtenues par des approches mono-modales et multi-modales pour procéder à la segmentation en scènes sont évaluées et comparées.

Dans ce papier, plutôt que d'utiliser la segmentation en locuteurs pour structurer le film, on propose de s'appuyer sur sa structure narrative, telle qu'on peut l'inférer à partir d'indices visuels, pour améliorer la segmentation en locuteurs de tels contenus. La façon dont les scènes dialoguées sont visuellement détectées est décrite dans la section 2 ; les deux étapes de regroupement en locuteurs sont introduites dans les sections 3 et 4 ; les résultats expérimentaux obtenus sont donnés dans la section 5.

2. Détection visuelle de scènes dialoguées

Le flux vidéo peut être globalement considéré comme une suite finie d'images fixes projetées sur l'écran à un rythme constant propre à simuler pour l'œil humain la continuité du mouvement. Par ailleurs, selon (Koprinska et Carrato, 2001), un plan constitue une unité définie comme une « suite ininterrompue d'images prises par une caméra unique ».

Comme on l'a relevé dans la section 1, les scènes dialoguées sont caractérisées par des motifs spécifiques où des plans récurrents alternent.

Détecter ces motifs caractéristiques des dialogues exige donc d'une part de segmenter le flux vidéo en plans et d'autre part de déterminer les plans semblables.

2.1. *Segmentation en plans et détection des plans semblables*

Défini par la continuité des images qu'il contient, un plan peut aussi être défini par opposition aux images des plans qui lui sont temporellement contigus. La segmentation en plans est donc usuellement conçue comme une tâche de détection des ruptures, soit brutales (coupes), soit graduelles (fondus), dans la continuité des images constitutives du flux vidéo. Marginales dans les trois séries TV incluses dans notre corpus, les transitions graduelles sont ici ignorées. Une coupe entre plans est donc détectée dès que deux images consécutives diffèrent sensiblement. Inversement, deux plans sont considérés comme semblables si la dernière image du premier et la première image du second sont suffisamment proches.

Détecter les coupes entre plans et repérer les plans similaires suppose donc que la similarité entre deux images puisse être évaluée. A cette fin, chaque image est décrite par la distribution statistique des pixels qu'elle contient dans l'espace colorimétrique HSV (teinte, saturation, luminance) et la distance entre deux images est mesurée par la corrélation de leurs histogrammes colorimétriques respectifs. Toutefois, il n'est pas exclu que deux images différentes partagent le même histogramme colorimétrique : afin de prévenir de tels rapprochements erronés, l'information de localisation des couleurs sur l'image est réintroduite en divisant chaque image en 30 blocs, chacun étant associé à son propre histogramme de couleurs ; les images sont alors comparées bloc par bloc en mesurant la corrélation des histogrammes associés selon la méthode décrite dans (Koprinska et Carrato, 2001).

Les deux seuils de corrélation utilisés pour détecter les coupes et les plans semblables ont été estimés à partir d'expériences sur un sous-ensemble de développement.

2.2. *Détection des passages dialogués*

Soit $\Sigma = \{l_1, ..., l_m\}$ un ensemble de m étiquettes de plans, deux plans partageant la même étiquette s'ils sont conjecturés comme semblables au sens de la sous-section précédente.

En assimilant un plan à son étiquette, un film peut alors être décrit comme une chaîne finie de plans $\mathbf{s} = s_1 s_2 ... s_k$ ($s_i \in \Sigma$).

Pour tout couple de plans $(l_1, l_2) \in \Sigma^2$, l'expression régulière $r(l_1, l_2)$ qui suit désigne une partie de l'ensemble de toutes les suites possibles de plans $\Sigma^* = \bigcup_{n \geqslant 0} \Sigma^n$:

$$r(l_1, l_2) = \Sigma^* l_1 (l_2 l_1)^+ \Sigma^* \quad [1]$$

L'ensemble $\mathcal{L}(r(l_1, l_2))$ des chaînes désignées par l'expression régulière 1 correspond alors à toutes les suites de plans où le plan l_2 est inséré entre deux occurrences du plan l_1, avec une éventuelle répétition de la séquence $(l_2 l_1)$, quels que soient les plans qui précèdent et suivent cette séquence. Cette expression régulière vise à capturer les suites de plans alternants caractéristiques des passages dialogués.

La Figure 1 présente le type de séquence de plans capturée par l'expression régulière $r(l_1, l_2)$ et illustre la règle des « 180 degrés » évoquée dans la section 1 : les deux personnages regardent dans des directions opposées.

Figure 1. *Exemple de séquence de plans ...$l_1 l_2 l_1 l_2 l_1$... capturée par l'expression régulière 1 pour les deux étiquettes de plans l_1 et l_2.*

Pour un film caractérisé par la séquence de plans $\mathbf{s} = s_1 s_2 ... s_k$ ($s_i \in \Sigma$), on définit l'ensemble $\mathcal{P}(\mathbf{s}) \subseteq \Sigma^2$ de tous les couples de plans impliqués dans les séquences de plans alternants dénotées par l'expression 1 :

$$\mathcal{P}(\mathbf{s}) = \{(l_1, l_2) \mid \mathbf{s} \in \mathcal{L}(r(l_1, l_2))\} \quad [2]$$

Pour $(l_1, l_2) \in \mathcal{P}(\mathbf{s})$, l'ensemble $\mathbf{u}(l_1, l_2)$ regroupe alors tous les segments parlés couverts par les séquences où les plans l_1 et l_2 alternent selon les modalités définies par la règle 1.

Afin d'accroître la couverture des motifs de $\mathcal{P}(\mathbf{s})$ tout en réduisant leur dispersion, deux extensions de la règle 1 sont introduites :

1) Les expressions isolées du couple de plans alternants, de la forme $(l_1 l_2 | l_2 l_1)$, sont également prises en considération.

2) Par ailleurs, dans les cas où deux motifs (l_1, l_2) et (l_1, l_3) partagent un même plan l_1, les plans l_2 et l_3 sont considérés comme équivalents : en effet, de telles situations se produisent lors des scènes dialoguées quand un personnage est filmé de deux points de vue différents alors que son interlocuteur n'est filmé que d'un seul point de vue. La Figure 2 illustre une telle situation.

Figure 2. *Séquence de plans ...$l_1 l_2 l_1 l_3 l_1$... à la frontière de deux motifs adjacents (l_1, l_2) et (l_1, l_3) avec un plan en commun.*

Le taux de parole couverte par les motifs de plans alternants dans notre corpus (décrit dans la sous-section 5.1) est donné dans le tableau 1. Est également reporté le temps de parole moyen couvert par une séquence de plans alternants, ainsi que le nombre moyen de locuteurs impliqués dans une séquence. Ces données sont indiquées à la fois pour les séquences capturées par l'expression régulière r et pour les séquences capturées en utilisant sa version étendue.

Tableau 1. *Motifs de plans et parole : données statistiques*

	couverture (%)	**temps parole/motif (s.)**	**nb. locuteurs/motif**
r	49.51	11.07	1.77
ext. r	**51.99**	**20.90**	**1.86**

Comme on peut le noter, les séquences de plans alternants couvrent un peu plus de la moitié (51.99%) du temps de parole total contenu dans les épisodes de notre corpus. D'autre part, 69.85% des séquences contiennent deux locuteurs, 8.09% trois et 22.06% seulement un : les cas de séquences impliquant uniquement un locuteur se produisent essentiellement pour les courtes scènes dialoguées, où il peut arriver qu'un des deux personnages reste silencieux. Enfin, 97.96% des personnages parlant au moins 5% du temps total sont impliqués dans de telles séquences.

3. Détection locale des locuteurs

La segmentation du flux en locuteurs est accomplie en deux temps : d'abord localement en regroupant en classes de locuteurs l'ensemble des segments parlés inclus dans

une séquence de plans alternants ; dans un deuxième temps, les locuteurs conjecturés dans chaque scène dialoguée sont comparés afin de regrouper les locuteurs récurrents d'une scène à l'autre.

3.1. *Descripteurs acoustiques*

Facilement disponibles, les plages des sous-titres originaux sont ici utilisées pour estimer les limites temporelles des segments parlés. Malgré un léger décalage temporel, d'amplitude variable, le sous-titre original transcrit fidèlement le segment parlé et coïncide avec ses limites temporelles. Lorsque le temps de latence apparaissait trop important, les limites temporelles des segments parlés concernés ont été manuellement ajustées.

Par ailleurs, une même plage de sous-titre transcrit les propos d'un seul locuteur. Dans les rares cas où les propos de deux locuteurs distincts sont regroupés sur un seul sous-titre, les limites des tours de parole sont explicitement indiquées et le sous-titre peut-être subdivisé.

Chacun des segments parlés délimités d'après les plages de sous-titres ne peut donc être affecté qu'à un seul locuteur. Ainsi, le prérequis de la plupart des systèmes de segmentation en locuteurs, à savoir la détection des points de changement entre locuteurs dans le flux de parole, peut ici être contourné pour mieux se concentrer sur la phase de regroupement des segments en classes de locuteurs.

Pour caractériser les segments parlés, on extrait du signal audio les 19 premiers coefficients cepstraux, l'énergie ainsi que leurs dérivées première et seconde, pour un total de 60 composantes.

Afin d'extraire l'information acoustique pertinente pour caractériser le locuteur correspondant, on associe alors à chaque segment parlé un i-vecteur (Dehak *et al.*, 2011) de 60 composantes, après entraînement d'un modèle du monde de 512 composantes et apprentissage d'une matrice de variabilité totale sur un sous-ensemble du corpus.

3.2. *Regroupement hiérarchique ascendant*

Un première étape de regroupement hiérarchique ascendant (regroupement agglomératif) est accomplie localement, dans les limites de chacune des scènes dialoguées définies selon la méthode décrite dans la sous-section 2.2.

L'ensemble des segments parlés $\mathbf{u}(l_1, l_2)$ couverts par le motif de plans alternants (l_1, l_2) est alors partitionné en classes de locuteurs selon les modalité suivantes :

– La distance de Mahalanobis est choisie comme mesure de similarité entre les i-vecteurs associés aux segments parlés.

La matrice de covariance utilisée pour évaluer la distance de Mahalanobis entre i-vecteurs est la matrice de covariance intra-classe mentionnée dans (Bousquet *et al.*, 2011), apprise sur un sous-ensemble du corpus et dont la formule est donnée par :

$$W = \frac{1}{n} \sum_{s=1}^{S} \sum_{i=1}^{n_s} (\mathbf{u_i^s} - \overline{\mathbf{u_s}})(\mathbf{u_i^s} - \overline{\mathbf{u_s}})^T \quad [3]$$

où n désigne le nombre de segments parlés du corpus d'apprentissage, S le nombre de locuteurs et n_s le nombre de segments proférés par le locuteur s ; $\overline{\mathbf{u_s}}$ est la moyenne des i-vecteurs associés aux segments parlés du locuteur s et $\mathbf{u_i^s}$ désigne le i-vecteur associé au i-ème segment parlé du locuteur s.

– Lors du processus agglomératif, le critère d'agrégation de Ward est utilisé pour évaluer la distance $\Delta I(c, c')$ entre les classes c et c', selon la formule suivante :

$$\Delta I(c, c') = \frac{m_c m_{c'}}{m_c + m_{c'}} d^2(g_c, g_{c'}) \quad [4]$$

où m_c et $m_{c'}$ désignent les masses des deux classes c et c', g_c et $g_{c'}$ leurs centres de gravité respectifs et $d(g_c, g_{c'})$ la distance entre les deux centres de gravité.

– Enfin, la méthode Silhouette est utilisée pour couper le dendogramme issu de l'algorithme agglomératif à un niveau optimal et obtenir la partition finale des segments en classes de locuteurs. Décrite dans (Rousseeuw, 1987), la méthode Silhouette permet d'associer à chaque partition possible du jeu de données un score normalisé entre -1 et 1 : pour une partition donnée, si certaines instances apparaissent plus proches du centre d'une autre classe que du centre de leur propre classe, le score tend à décroître, et à croître si elles sont plus proches du centre de leur propre classe que du centre d'une autre classe.

4. Regroupement global sous contrainte

Une fois les segments parlés regroupés en classes de locuteurs dans chaque scène dialoguée, une seconde phase de regroupement a lieu pour regrouper au sein d'une même classe les locuteurs récurrents d'une scène à l'autre.

L'ensemble des segments parlés assignés lors de la première étape de regroupement à un même locuteur sont alors concaténés en un seul segment audio, qu'on associe après paramétrisation acoustique à un unique i-vecteur de dimension 60, caractéristique du locuteur conjecturé localement.

L'ensemble de i-vecteurs qui en résulte est alors partitionné après application d'un algorithme agglomératif selon les mêmes modalités que celles décrites dans la sous-section 3.2 : distance de Mahalanobis fondée sur une matrice de covariance intra-classe, critère d'agrégation de Ward, méthode de partionnement optimal Silhouette.

Toutefois, l'information de structure acquise après la première phase de partionnement est propagée à chaque étape de cette seconde étape de regroupement : dans la

recherche des locuteurs récurrents, on doit en effet empêcher que les locuteurs impliqués dans un même dialogue ne soient regroupés dans la même classe de locuteurs.

On peut donc contraindre cette seconde phase de regroupement en propageant à chaque étape du processus agglomératif ce que (Davidson et Ravi, 2009) décrivent comme une contrainte de type « impossible-à-relier » (*cannot-link*) qui interdit aux locuteurs simultanément impliqués dans un même dialogue d'être regroupés.

L'intégration d'une telle contrainte au processus de regroupement est réalisée selon les modalités suivantes :

– Dans la matrice de distance entre les i-vecteurs associés aux classes locales de locuteurs, la distance entre deux vecteurs et définie comme infinie si les deux locuteurs correspondants apparaissent conjointement dans le même dialogue :

$$d(\mathbf{s}, \mathbf{s}') = +\infty \Leftrightarrow \exists (l_1, l_2),\ \mathbf{u}(\mathbf{s}) \cup \mathbf{u}(\mathbf{s}') \subseteq \mathbf{u}(l_1, l_2) \quad [5]$$

où (l_1, l_2) désigne un motif dialogué, $\mathbf{u}(l_1, l_2)$, l'ensemble de segments parlés couverts par le motif (l_1, l_2) et $\mathbf{u}(\mathbf{s})$ l'ensemble des segments attribués au locuteur $\mathbf{s}$ après la phase de regroupement local .

– Lors de la réévaluation des distances entre classes après chaque itération du processus agglomératif, la contrainte *cannot-link* est propagée en situant à une distance infinie deux classes c et c' qui contiendraient respectivement des i-vecteurs incompatibles.

$$\Delta I(c, c') = +\infty \Leftrightarrow \exists (\mathbf{s}, \mathbf{s}') \in c \times c',\ d(\mathbf{s}, \mathbf{s}') = +\infty \quad [6]$$

où $\mathbf{s}$ et $\mathbf{s}'$ désignent les i-vecteurs correspondant à des classes de locuteurs conjecturées localement.

L'observation des règles 5 et 6 empêche deux locuteurs impliqués dans le même dialogue d'être regroupés lorsque les deux classes les plus proches sont fusionnées à chaque étape du processus agglomératif.

La Figure 3 illustre la première itération de l'algorithme agglomératif contraint ainsi que la manière dont la contrainte *cannot-link* se propage.

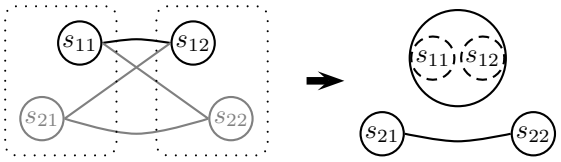

Figure 3. *Première itération de l'algorithme agglomératif contraint*

On se place dans un cadre simplifié où seules deux scènes dialoguées sont prises en considération : chacune d'entre elles est délimitée par des traits pointillés. On suppose

d'autre part avoir conjecturé deux locuteurs dans chacune d'entre elles : s_{ij} représente le i-ème locuteur de la j-ème scène. L'éventuelle arête reliant deux locuteurs figure la distance entre les deux i-vecteurs associés et l'absence de liens correspond à une distance infinie. L'algorithme va commencer par regrouper les deux locuteurs les plus proches, soit s_{11} et s_{12}, le groupe résultant pouvant correspondre à un même locuteur récurrent d'une scène à l'autre. L'absence de liens entre le locuteur récurrent et ses interlocuteurs respectifs dans les deux scènes montre alors comme la contrainte *cannot-link* est héritée par la classe issue de chaque itération de l'algorithme agglomératif : le seul regroupement possible à la seconde itération de l'algorithme (non représentée sur la Figure 3) serait de fusionner s_{21} et s_{22}. La propagation de cette contrainte vise à prévenir les regroupements précoces d'interlocuteurs dans une même classe : la musique de fond peut en effet, par exemple, recouvrir la variabilité entre locuteurs et provoquer prématurément un tel regroupement si aucune contrainte n'est posée.

D'autre part, on voit comment le processus agglomératif est bloqué par la propagation de ce type de contrainte : dans le petit exemple représenté sur la Figure 3, seules deux itérations de l'algorithme sont possibles, avec au minimum deux classes de locuteurs au terme du processus au lieu d'une seule en l'absence de contraintes.

La Figure 4 met en regard les dendogrammes issus de deux algorithmes agglomératifs, non-contraint et contraint, appliqués aux mêmes données. La partie haute de la figure représente le dendogramme issu du regroupement ascendant des classes de locuteurs : le processus peut être poursuivi jusqu'à ce qu'il ne reste plus qu'une classe unique rassemblant toutes les instances ; en revanche, dans le cas du regroupement effectué sous la contrainte de dissociation des interlocuteurs, le processus agglomératif est prématurément bloqué et débouche sur cinq dendogrammes disjoints, correspondant à cinq classes incompatibles : chaque classe contient au moins un locuteur en dialogue direct avec au moins un locuteur de chacune des quatre autres classes. On est donc en présence d'au moins cinq locuteurs récurrents.

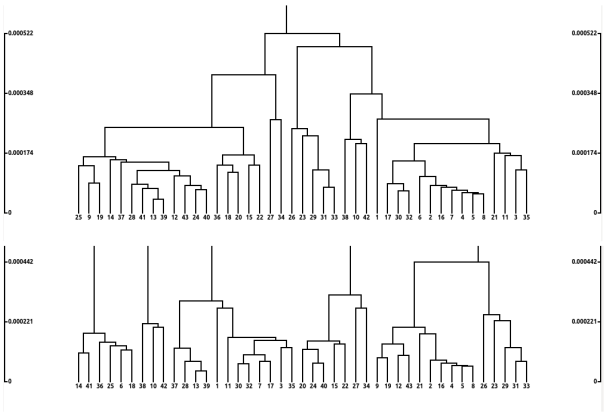

Figure 4. *Dendogrammes obtenus par regroupement agglomératif de locuteurs locaux, non-contraint (haut) ; contraint (bas)*

Le partitionnement final en classes de locuteurs est alors obtenu en partitionnant chacun des dendogrammes selon la méthode Silhouette décrite dans la sous-section 3.2.

Cette propriété d'irréductibilité (Davidson et Ravi, 2009) d'une partition, où les classes obtenues ne peuvent plus être réunies sans violer une contrainte de dissociation des instances, permet ainsi de contourner en partie le problème critique de la coupe optimale du dendogramme en bloquant prématurément le processus d'agglomération.

Toutefois, cette seconde phase de regroupement reste dépendante de la première phase de partitionnement des segments parlés en classes locales de locuteurs. Si les propos d'un même locuteur sont à tort distribués sur deux classes de locuteurs disjointes pendant la première phase, ils ne pourront plus être regroupés lors de la deuxième étape, les deux classes étant désormais considérées comme incompatibles. Cependant, même lors d'un regroupement non contraint, ces segments, dissociés au niveau local, ne seraient regroupés que tardivement au niveau global, éventuellement après le partitionnement jugé optimal.

D'autre part, une erreur peut être commise lors de la phase de regroupement global : pour des locuteurs a, b, et c, et en posant, en reprenant l'exemple de la Figure 3, $s_{11} := a$, $s_{21} := c$, $s_{12} := b$, $s_{22} := a$, la première itération de l'algorithme sous contrainte, en regroupant a et b, aura pour effet de placer le locuteur a dans deux dendogrammes disjoints, empêchant toute fusion ultérieure des deux occurrences de a, et dégradera la couverture des classes de locuteurs. Toutefois, dans le cas d'un partitionnement non contraint, les deux occurrences du locuteur a, non regroupées après la première itération, ne le seraient probablement que tardivement, éventuellement après que le partitionnement optimal aura été atteint.

Enfin, pour un ensemble donné d'instances entre lesquelles sont posées des contraintes de séparation à respecter lors du regroupement, le nombre final de classes de la partition irréductible semble dépendant de l'ordre dans lequel les regroupements deux à deux ont lieu (Davidson et Ravi, 2009). Toutefois, comme on le verra dans la section 5 qui suit, le nombre final de classes obtenues en partitionnant les dendogrammes issus du regroupement contraint est apparu expérimentalement offrir une approximation raisonnable du nombre réel de locuteurs impliqués.

5. Expériences et résultats

5.1. *Corpus*

Notre corpus est constitué d'épisodes des premières saisons de trois séries : *Breaking Bad* (abrégé *bb*), *Game of Thrones* (*got*) et *House of Cards* (*hoc*). Nous avons manuellement annoté trois épisodes de chaque série en indiquant les coupes de plans, les plans similaires, les segments parlés ainsi que les locuteurs associés.

Le temps de parole total contenu dans ces neuf épisodes représente un peu plus de trois heures (3 :12).

Un sous-ensemble de six épisodes (désigné DEV) a été utilisé à des fins de développement (seuils de coupes et de similarité entre plans) ou d'apprentissage (modèle GMM/UBM, matrice de variabilité totale, matrice de covariance intra-classe). Les trois épisodes restants (TEST) ont été utilisés à des fins de test.

5.2. *Détection des coupes et des similitudes entre plans*

L'évaluation de la méthode de détection des coupes repose sur une F1-mesure (Boreczky et Rowe, 1996) fondée sur le rappel (taux de coupes retrouvées parmi les coupes pertinentes) et la précision (taux de coupes pertinentes parmi celles qui ont été conjecturées). Pour la tâche de similarité entre plans, un F1-score analogue est utilisé : pour chaque plan, on compare la liste des plans conjecturés comme lui étant similaires à celle des plans annotés comme semblables : si l'intersection des deux listes est non vide, le plan est considéré comme correctement apparié. Les résultats sur les ensembles de DEV et de TEST sont présentés dans le Tableau 2.

Tableau 2. *Résultats obtenus pour la détection de coupes et de similitudes entre plans*

	coupes	**similitudes**		
	F1-score	précision	rappel	F1-score
bb-1	0.93	0.88	0.81	0.84
bb-2	0.99	0.90	0.83	0.86
got-1	0.97	0.88	0.84	0.86
got-2	0.98	0.89	0.90	0.90
hoc-1	0.99	0.91	0.92	0.92
hoc-2	0.98	0.93	0.97	0.95
moy. DEV	**0.97**	0.90	0.88	**0.89**
bb-3	0.98	0.83	0.84	0.83
got-3	0.99	0.92	0.89	0.91
hoc-3	0.99	0.98	0.96	0.97
moy. TEST	**0.99**	0.91	0.90	**0.90**

Le résultats obtenus pour les deux tâches relevant du traitement de l'image apparaissent élevés : le F1-score obtenu lors de la détection des images similaires en particulier (0.90) permet de déterminer avec confiance les limites des passages dialogués au sein desquels la segmentation en locuteurs sera effectuée.

5.3. *Segmentation locale en locuteurs*

Le taux d'erreur utilisé pour évaluer la segmentation locale en locuteurs est déterminé indépendamment pour chaque scène dialoguée, avant d'être estimé globalement par la somme des taux d'erreurs locaux pondérés par la durée de chaque scène

(*single-show Diarization Error Rate*, défini dans (Rouvier *et al.*, 2013)). Les résultats sont reportés dans le Tableau 3, à la fois en utilisant les similitudes de plans de référence (*entrée réf.*) et les similitudes de plans déterminées automatiquement (*entrée auto.*). D'autre part, à des fins de comparaison, le regroupement agglomératif (désigné *ra*) est comparée à un principe de regroupement « naïf » fondé sur la seule alternance des plans : chaque segment parlé est étiqueté selon le plan qui apparaît à l'écran, en supposant que l'alternance des plans coïncide avec l'alternance des locuteurs.

Tableau 3. *Segmentation locale en locuteurs :* DER

	entrée auto.		**entrée réf.**	
	naïf	RA	*naïf*	RA
bb-1	30.26	**19.11**	22.81	21.00
bb-2	22.06	22.51	19.78	**19.14**
got-1	22.16	23.70	19.46	**15.78**
got-2	26.19	18.78	22.80	**16.61**
hoc-1	17.23	13.36	16.31	**11.84**
hoc-2	30.66	**18.18**	31.87	19.12
moy. DEV	24.76	19.27	22.17	**17.25**
bb-3	40.45	21.15	24.31	**12.15**
got-3	33.45	17.43	35.43	**12.80**
hoc-3	24.44	12.83	22.95	**12.82**
moy. TEST	32.78	17.14	27.56	**12.59**

Les résultats obtenus par le regroupement des segments parlés sur un critère purement acoustique apparaissent supérieurs à ceux obtenus par la méthode naïve fondée sur la seule alternance des plans.

Par ailleurs, l'automatisation de l'étape précédente de détection des plans similaires n'affecte qu'en partie les résultats de la segmentation des locuteurs (augmentation du taux d'erreur de près de 5% sur les trois épisodes du TEST, mais seulement de 2% sur les six épisodes du DEV).

5.4. *Segmentation globale sous contrainte*

La tableau 4 regroupe les résultats obtenus lors de la seconde étape de regroupement global des locuteurs récurrents. Les résultats sont donnés à la fois en prenant en entrées les locuteurs conjecturés lors de la phase de segmentation locale (*entrée auto.*) et les locuteurs de référence (*entrée réf.*). Dans les deux cas, les résultats sont donnés pour la version non contrainte du regroupement (désignée *2S*) et pour sa version contrainte (*cnt. 2S*). Les résultats obtenus sont comparés à ceux obtenus par deux outils standards de segmentation en locuteurs fondés sur des algorithmes hiérarchiques (dénotés LIA et LIUM), auxquels on fournit en entrée les segments parlés de référence couverts par les dialogues capturés selon les modalités décrites dans la section 2.

Tableau 4. DER *Segmentation globale en locuteurs :* DER

	entrée auto.		**entrée réf.**		**segments réf.**	
	2S	*cnt. 2S*	*2S*	*cnt. 2S*	LIA	LIUM
bb-1	51.36	56.00	52.66	**48.10**	72.06	67.21
bb-2	**41.83**	65.07	58.76	49.49	77.03	76.79
got-1	70.13	**52.79**	70.67	53.87	65.57	58.49
got-2	67.28	**38.85**	70.32	41.24	65.29	60.80
hoc-1	**50.04**	55.61	52.70	52.15	60.26	62.37
hoc-2	64.91	56.40	63.65	**37.09**	67.05	59.00
moy.	57.59	54.11	61.46	**46.99**	67.88	64.11
bb-3	60.41	**33.94**	59.22	42.64	60.61	55.56
got-3	74.71	**49.31**	70.34	63.17	61.33	52.89
hoc-3	**57.68**	59.87	67.52	67.41	70.55	67.05
moy.	64.13	**47.71**	65.69	57.74	64.16	58.50

Bien qu'encore élevé, le taux d'erreur est en général réduit en propageant dans le processus de regroupement la contrainte de dissociation des interlocuteurs. En bloquant le processus de regroupement au moment où il devient irréductible, l'information de structure ainsi propagée permet de contourner en partie le problème délicat du niveau optimal où couper le dendogramme issu du processus agglomératif.

Le Tableau 5 donne, pour chacune des méthodes de segmentation en locuteurs présentée dans le Tableau 4, le nombre moyen de locuteurs par épisode, en regard du nombre réel, pour chacune des trois séries de notre corpus.

Tableau 5. *Nombre moyen de locuteurs par épisode*

	réf.	*2S*	*cnt. 2S*	LIA	LIUM
bb	**10.3**	7.3	**11**	6	25.7
got	**25.3**	4.7	15.7	9.3	**24**
hoc	**20.7**	3.7	**24**	6	27

Comme on peut le constater, deux des systèmes mentionnés, notre système non contraint (*2S*) et LIA, ont tendance à sous-estimer le nombre total de locuteurs par épisode en coupant le dendogramme à un niveau élevé. Inversement, le système désigné par LIUM a tendance à surestimer le nombre de locuteurs impliqués en opérant une coupe basse de l'arbre de regroupement. L'approche fondée sur la contrainte de dissociation des interlocuteurs (*cnt. 2S*) débouche en revanche sur une partition irréductible qui permet d'estimer raisonnablement le nombre de locuteurs.

6. Conclusion

Dans ce papier, nous avons proposé de procéder à la segmentation de séries télévisées en locuteurs en nous appuyant sur la structure narrative qui les sous-tend. En détectant les plans similaires, des séquences d'alternance de plans typiques des scènes dialoguées peuvent être délimitées et une première phase de regroupement des segments parlés en classes de locuteurs peut avoir lieu dans les limites de chacune des scènes isolées. Une seconde étape de regroupement, visant à détecter les locuteurs récurrents, est alors appliquée aux classes de locuteurs localement conjecturées : à chaque itération du processus d'agglomération, une contrainte de séparation des interlocuteurs impliqués dans une même scène est propagée. Le processus de regroupement se trouve alors bloqué avant que toutes les instances ne puissent être groupées. La partition irréductible qui en résulte permet d'estimer plus facilement le nombre total de locuteurs impliqués.

Il resterait à mener une étude plus systématique sur la manière dont l'ordre dans lequel les instances sont groupées sous une contrainte de séparation affecte le nombre final de sous-ensembles de la partition irréductible.

D'autre part, c'est essentiellement lors de la seconde phase de regroupement que le taux d'erreur se dégrade sensiblement : même en prenant en entrée les locuteurs de référence impliqués dans les différentes scènes dialoguées, certains des regroupements effectués apparaissent lourdement erronés. La variabilité dans l'environnement acoustique d'une scène à l'autre semble induire nombre de ces erreurs. Il faudrait donc étudier plus systématiquement, et éventuellement isoler, les sources de variabilité que peuvent constituer la musique de fond ou les bruitages.

Remerciements

Ce travail a été en partie soutenu par le projet GAFES (ANR-14-CE24-0022) de l'Agence Nationale de la Recherche, ainsi que par la Fédération de Recherche *Agorantic*, Université d'Avignon et des Pays de Vaucluse.

7. Bibliographie

Adams W., Iyengar G., Lin C., M. Naphade C. N., Nock H., Smith J., « Semantic Indexing of Multimedia Content using Visual, Audio and Text Cues », p. 1-16, 2003.

Bendris M., Favre B., Charlet D., Damnati G., Senay G., Auguste R., Martinet J., « Unsupervised face identification in tv content using audio-visual sources », *Content-Based Multimedia Indexing (CBMI), 2013 11th International Workshop on*, IEEE, p. 243-249, 2013.

Boreczky J. S., Rowe L. A., « Comparison of video shot boundary detection techniques », *Journal of Electronic Imaging*, vol. 5, n° 2, p. 122-128, 1996.

Bousquet P.-M., Matrouf D., Bonastre J.-F., « Intersession Compensation and Scoring Methods in the i-vectors Space for Speaker Recognition. », *INTERSPEECH*, Florence, Italy, p. 485-488, 2011.

Bredin H., « Segmentation of tv shows into scenes using speaker diarization and speech recognition », *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, IEEE, p. 2377-2380, 2012.

Carletta J., Ashby S., Bourban S., Flynn M., Guillemot M., Hain T., Kadlec J., Karaiskos V., Kraaij W., Kronenthal M., Lathoud G., Lincoln M., Lisowska A., McCowan I., Post W., Reidsma D., Wellner P., « The AMI Meeting Corpus : A Pre-announcement », *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, MLMI'05, Springer-Verlag, Berlin, Heidelberg, p. 28-39, 2006.

Clément P., Bazillon T., Fredouille C., « Speaker diarization of heterogeneous web video files : A preliminary study », *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, IEEE, Florence, Italy, p. 4432-4435, 2011.

Davidson I., Ravi S., « Using instance-level constraints in agglomerative hierarchical clustering : theoretical and empirical results », *Data Mining and Knowledge Discovery*, vol. 18, n° 2, p. 257-282, 2009.

Dehak N., Kenny P., Dehak R., Dumouchel P., Ouellet P., « Front-end factor analysis for speaker verification », *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, n° 4, p. 788-798, 2011.

Ercolessi P., Extraction multimodale de la structure narrative des épisodes de séries télévisées, PhD thesis, Université de Toulouse, Université Toulouse III-Paul Sabatier, 2013.

Evans N., Bozonnet S., Wang D., Fredouille C., Troncy R., « A comparative study of bottom-up and top-down approaches to speaker diarization », *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, n° 2, p. 382-392, 2012.

Friedland G., Hung H., Yeo C., « Multi-modal speaker diarization of real-world meetings using compressed-domain video features », *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, p. 4069-4072, April, 2009.

Koprinska I., Carrato S., « Temporal video segmentation : A survey », *Signal processing : Image communication*, vol. 16, n° 5, p. 477-500, 2001.

Lew M. S., Sebe N., Djeraba C., Jain R., « Content-based Multimedia Information Retrieval : State of the Art and Challenges », *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 2, n° 1, p. 1-19, February, 2006.

Quénot G., Tan T. P., Le V. B., Ayache S., Besacier L., Mulhem P., « Recherche par le contenu dans des documents audiovisuels multilingues », *Document Numérique*, vol. 13, n° 1, p. 229-246, 2010.

Rousseeuw P. J., « Silhouettes : a graphical aid to the interpretation and validation of cluster analysis », *Journal of computational and applied mathematics*, vol. 20, p. 53-65, 1987.

Rouvier M., Dupuy G., Gay P., Khoury E., Merlin T., Meignier S., « An open-source state-of-the-art toolbox for broadcast news diarization », *INTERSPEECH*, Lyon, France, 2013.

Tran V.-A., Le V. B., Barras C., Lamel L., « Comparing Multi-Stage Approaches for Cross-Show Speaker Diarization. », *INTERSPEECH*, Florence, Italy, p. 1053-1056, 2011.