Speaker diarization de fichiers vidéos hétérogènes issus du web

Segmentation et du regroupement en locuteurs de fichiers vidéos hétérogènes issus du web : une étude préliminaire

Pierre CLÉMENT

Laboratoire Informatique d'Avignon
339, chemin des Meinajaries - Agroparc BP 91228
84911 AVIGNON Cedex 9
pierre.clement@univ-avignon.fr

RÉSUMÉ. *Ces dix dernières années, internet a significativement changé. Le principal changement est certainement le contenu proposé, que ce soit dans sa quantité, sa diversité ou encore le média utilisé pour le présenter. Concernant le média audio/video, l'évolution la plus impressionnante est le succès continuellement grandissant des sites de partage de vidéos. Mais ce succès entraîne des diffcultés à indexer efficacement le contenu de ces documents. La segmentation et le regroupement en locuteurs (speaker diarization) est une tâche importante s'inscrivant dans ce processus. Ce papier décrit la base de données audio/video, construite spécialement pour la speaker diarization, basée sur différents genres vidéo. Au travers d'expériences préliminaires, ce papier met en évidence les difficultés rencontrées pour cette tâche dans ce contexte, difficultés principalement dûes à l'hétérogénité de la base de données.*

ABSTRACT. *In the last ten years, Internet changed significantly. The main change is certainly the content of the Internet, in its quantity, its variety or the media used to show it. Regarding multimedia, the most impressive evolution is the continuous growing success of the video sharing websites. But, with this success come the difficulties to efficiently search, index and access relevant information about these documents. Speaker diarization is an important task in the overall information retrieval process. This paper describes the audio/video database, especially built for the speaker diarization task, based on different video genres. Through some preliminary experiments, it highlights the difficulties encountered in this context, mainly linked to the database heterogeneity.*

MOTS-CLÉS : *segmentation-regroupement en locuteurs, vidéos hétérogènes, erreur de diarization*

KEYWORDS: *speaker diarization, heterogeneous web videos, diarization error rate*

1. Introduction

Beaucoup de fichiers multimédias sont envoyés chaque jour sur différents sites internets, dont les plus connus sont probablement YouTube ou DailyMotion. Cette quantité de données ne cessant pas de s'agrandir, il est devenu très important de rechercher, indexer et accéder à certaines informations à propos de ces vidéos et ce, de manière automatique. Ces informations peuvent être du contenu linguistique, des informations concernant le contexte de la vidéo etc. La segmentation et le regroupement en locuteurs (Speaker Diarization) est donc une tâche importante dans la recherche d'informations dans les fichiers vidéos. En effet, le but de la speaker diarization est de fournir de manière automatique des régions où une même personne parle dans un flux média, sans informations à priori sur le nombre total de locuteurs ni leur identité. Depuis une dizaine d'années, les systèmes de speaker diarization ont été évalués sur différents domaines d'application, comme les conversations téléphoniques (deux locuteurs dans un document audio), des journaux d'informations (Broadcast news, BN) ou encore de la téléconférence (meeting). Chaque type de données ayant ses propres spécificités, chaque domaine d'application implique une configuration différente du système. Il est donc intéressant d'étudier le comportement de ce type de système lorsqu'on l'applique à un autre domaine particulier qu'est la vidéo issue du web, où il peut y avoir une grande variabilité du type de contenu audio. Dans ce papier, le comportement du système de speaker diarization du Laboratoire Informatique d'Avignon (LIA) va être comparé au système du Laboratoire Informatique de l'Université du Maine (LIUM) qui est lui un système bottom-up. La section 2 décrira la base de données de fichiers vidéos hétérogènes issus du web. La section 3 détaillera les deux systèmes de speaker diarization dont les résultats expérimentaux sont fournis en section 4. Enfin, des conclusions et perspectives sont données en section 5.

2. Base de données web

2.1. *Présentation générale*

La base de données LIA GERARD (multi**GE**nre and multispeake**R** **A**udio/video F**R**ench **D**atabase) est constituée de 856 vidéos téléchargées depuis différents sites internet. Ces vidéos peuvent être classées suivant sept catégories, comme indiqué dans (Rouvier *et al.,* 2010). Dans cette étude, seuls cinq types de vidéos seront utilisés : documentaries, movie trailers, cartoons, commercials and news, laissant de côté le sport (car seule la bande audio était disponible) et la musique (car la voix est utilisée comme un instrument de musique). Cette base de données ne contient pas de vidéos amateurs, pour éviter une variabilité supplémentaire liée au matériel utilisé pour l'enregistrement (comme par exemple un téléphone portable). La langue principale de LIA GERARD est le français, sauf pour certains documentaires qui sont en anglais ou avec la traduction par une voix-off superposée.

2.2. *Annotation manuelle*

Une petite partie de la base de données à été annotée manuellement afin de mesurer à la fois les performances du système de détection de parole mais aussi le système

de speaker diarization. Les fichiers ont été segmentées, les tours de parole et de locuteur ont été relevés en prenant en compte la parole superposée, les bruits de fond et la musique ont également été relevés. Ainsi, environ 130 vidéos des cinq catégories sélectionnées ont été annotées et seront utilisées lors des expériences. Une description de ce corpus par catégorie est donnée dans le tableau 1. Les vidéos choisies durent entre 1 et 10 minutes. La catégorie contenant la plus grande quantité de données est les cartoons (environ 4h12min), suivi par les documentaries (3h20min). Les commercials ne contiennent que 16min de vidéos, car seules 10 vidéos de LIA GERARD respectaient la contrainte de durée minimum (1min) fixée pour cette étude. On peut remarquer que la durée moyenne d'un tour de parole est malheureusement basse (6,31s pour le plus long).

Catégorie	**Nb de fich.**	**Durée totale**	**Durée moy. par fich**	**Part de parole (%)**	**Nb moy. de loc par fich**	**Tours de parole par fich (moy)**	**Durée moy d'un tour de parole (s)**
Documentary	29	3 :19 :06	0 :06 :51	71.78	8.2	84	3.51
Movie trailer	30	1 :07 :05	0 :02 :14	53.73	9.4	35	2.06
Cartoon	30	4 :11 :40	0 :08 :23	64.41	10.9	113	2.87
Commercial	10	0 :15 :40	0 :01 :34	68.09	5.2	25	2.56
News	30	1 :32 :40	0 :03 :05	88.65	4.5	26	6.31

Tableau 1. *Informations issues de l'annotation manuelle de la partie sélectionnée de LIA GERARD, par catégorie : le nombre de fichiers, la durée totale des fichiers choisis, le % de parole. Par fichier et catégorie, la durée moyenne, le nombre moyen de locuteur par fichier, de tours de parole et leur durée moyenne.*

2.3. *Caracteristiques par genre vidéo*

Chaque vidéo traitée dans ce papier a son propre niveau de difficulté. Ces difficultés, comme montrées plus bas, peuvent être liées à l'audio, la vidéo (non traitée ici) ou les deux.

– **Documentaries :** peuvent contenir beaucoup de parole superposée, lié au doublage. Une voix-off peut être présente et peut être source d'erreur en cas de croisement entre du traitement audio et vidéo.

– **Movie trailers :** sont très intéractifs. Un grand nombre de locuteurs peuvent parler tour à tour, comme montré dans le tableau 1, impliquant un grand nombre de très courts tours de parole. Il y a également beaucoup d'effets sonores et de musique (46% de non-parole), de parole sur de la musique ou des bruits forts, beaucoup de changement de plans et de voix-off. Cette catégorie fût la plus complexe à annoter.

– **Cartoons :** un doubleur peut doubler plusieurs personnages, donc la voix de différents personnages peut être similaire. Selon la scène, un personnage peut prendre une voix différente. Comme les movie trailers, les cartoons sont très interactifs, avec beaucoup de musique et d'effets sonores. Comparé aux autres catégories, le nombre de locuteurs est plutôt haut, alors que la durée d'un tour de parole est plutôt courte (environ 2,9s).

– **Commercial :** le style des commercials varie d'une vidéo à l'autre (certaines peuvent être plus musicales par exemple). Ces vidéos sont très courtes, et la part de

non-parole est plutôt élevée (32%), alors que le nombre de tours de parole est plutôt haut.

– **News :** il y a deux types de news : celles enregistrées en studio, et celles dans un environnement non-contrôlé. Cette catégorie contient plus de parole (90%) avec les plus long tours de parole (6,3s).

3. Présentation des systèmes de speaker diarization

Traditionnellement, les systèmes de speaker diarization impliquent deux grandes étapes. La première étape est la segmentation, qui a pour but de détecter les changements de locuteurs, fournissant ainsi les tours de paroles de locuteurs (speaker turn). La seconde est le regroupement des segments par similarité des locuteurs (clustering). Concernant le clustering, il ressort deux approches : le bottom-up et le top-down. Le **système de speaker diarization top-down du LIA**, expliqué en détails dans (Fredouille *et al.,* 2009), est basé sur un modèle de Markov caché évolutif (Evolutive Hidden Markov Model, E-HMM), utilisant la boîte à outils libre ALIZE (Bonastre *et al.,* 2005). Le système est composé de trois étapes principales :

– (1) la Speech Activity Detection (SAD), fournit une segmentation en parole/non-parole sur laquelle s'appuient les étapes de segmentation et regroupement. L'algorithme de la SAD est similaire à celui utilisé lors de la campagne d'évaluation RT'09 (Fredouille *et al.,* 2009). Elle est basée sur un HMM sur lequel un décodage Viterbi et une adaptation de GMM sont appliqués pour fournir la segmentation

– (2) la segmentation et le regroupement en locuteurs, basée sur un E-HMM, dans lequel chacun des états est un locuteur et chacune des transitions est un changement de locuteur. Dans cette étape, le signal est caractérisé par 21 coefficients, 20 LFCC plus l'énergie. Le processus commence par initialiser l'E-HMM avec seulement un seul état pour tout le flux, représentant ainsi un seul locuteur. Un processus itératif ajoute alors un locuteur via un décodage Viterbi et une boucle d'apprentissage d'un modèle pour ce locuteur. Ensuite, les segments de parole du fichier sont comparés au modèle et affectés ou non au locuteur courant. Enfin, cette boucle s'arrête lorsque l'on ne peut plus ajouter de locuteur.

– (3) une étape de resegmentation, qui a pour but de corriger la sortie de le segmentation et supprimer les locuteurs peu pertinents (trop peu de parole). Un HMM est généré à partir de la segmentation courante et l'adaptation Maximum A Posteriori (MAP, impliquant un Universal Background Model - UBM) est utilisée au lieu de EM/ML pour calculer les GMM de locuteurs.

Le second système de speaker diarization utilisé dans cette étude est **développé par le LIUM**. Distribué librement, ce système bottom-up exécute différentes étapes, pleinement détaillées dans (Meignier *et al.,* 2010). La grande différence entre les deux systèmes (bottom-up et top-down) est que le bottom-up calcule une première segmentation avec beaucoup de locuteurs (généralement plus que le nombre réel de locuteurs), et ensuite réduit ce nombre de locuteurs en cherchant les locuteurs similaires, contrairement aux systèmes top-down qui eux partent d'un seul locuteur et cherchent

à en ajouter. Quelle que soit la stratégie, le but est d'atteindre un nombre de locuteurs optimum.

4. Résultats et comparaison

Cette section montre les résultats obtenus par les systèmes de speaker diarization du LIA et du LIUM sur différents jeux de données : RT'09(NIST, 2009), ESTER'08(Galliano *et al.,* 2009) (données de développement et d'éval), LIA GERARD décrit précédement. Ces tests ont été effectués afin de comparer l'effet des données multigenres par rapport aux données monogenres sur ces systèmes, mais aussi comparer les stratégies bottom-up et top-down. Les systèmes sont évalués en calculant de Diarization Error Rate (DER), pleinement décrit dans (NIST, 2009). Le tableau 2

	RT'09 eval		**Ester'08 dev**		**Ester'08 eval**		**LIA GERARD (SAD auto.)**		**LIA GERARD (SAD man.)**	
Type	Meeting		BN				Heterogeneous			
Nb fich.	7		18		26		129			
Durée tot.	3 :00 :58		9 :27 :50		7 :10 :22		9 :52 :47			
Système	LIA	LIUM	LIA	LIUM	LIA	LIUM	LIA	LIUM	LIA	LIUM
DER	18.9%	NA	14.6%	8.8%	15.5%	8.2%	73.2%	55.6%	38.7%	34.3%
E_{missed}	0.5%	NA	1.8%	0.5%	1.3%	0.2%	9.5%	13.9%	0%	0%
E_{fa}	2.9%	NA	0%	2.5%	1.7%	1.6%	27.2%	13%	0%	0%
E_{spkr}	15.5%	NA	12.8%	5.8%	12.5%	6.4%	36.5%	28.7%	38.7%	34.3%

Tableau 2. *Résultats préliminaires des systèmes de speaker diarization du LIA et du LIUM sur différents jeux de données (RT'09 non test'e avec le système du LIUM).*

montre les résultats obtenus avec les systèmes du LIA et LIUM. Il peut être observé que le système du LIUM (bottom-up) obtient de meilleurs résultats par rapport au système du LIA (top-down) sur les domaines d'application classiques (BN et meeting). On observe également une chute des performances des deux systèmes avec les données web. Cette baisse est en partie dûe au processus de SAD, qui semble être inefficace sur ce type de données. Pour cette raison, les expériences ont été ensuite menée avec une segmentation parole/non-parole manuelle (5ième colonne du tableau 2), dont les résultats ont sont analysés dans la suite du papier. Bien qu'utilisant tous deux une segmentation manuelle, on peut constater que les deux systèmes obtiennent quand même de mauvais résultats (DER de 34% pour le LIUM et 38% pour le LIA). Afin de mesurer l'influence du genre vidéo sur les scores sur les stratégies de speaker diarization, les performances globales par genre des systèmes sont montrées dans le tableau 3. Les documentaries et les news obtiennent les meilleurs résultats, un DER

Category	**Nb moy. de locs. trouvés**		**DER (in %)**		**DER min**		**DER max**	
Système	LIA	LIUM	LIA	LIUM	LIA	LIUM	LIA	LIUM
News	2.9	4.6	25.7	12.8	0.1	0	66.4	61.6
Documentary	4	7	26.4	22	4.2	0	65.6	73.8
Movie trailer	1.4	2.8	54.3	51.1	18.9	17.7	79.2	79.2
Cartoon	4.3	10.2	49.7	53.1	22.1	31.6	72.2	73.4
Commercial	1.5	3.9	33.6	27.7	0	0	62	45.9

Tableau 3. *Résultats par genre donnés pour les systèmes du LIUM et du LIA.*

de 26% pour le LIA et respectivement 22 et 12% pour le LIUM. Les performances du système du LIA sont donc stables sur ces 2 genres, alors que les résultats sont plsu

contrastés pour le LIUM. Concernant les trois autres catégories, le DER augmente beaucoup (environ 45% de moyenne sur les trois catégories pour les deux systèmes). Cette dramatique baisse semble être dûe à la très courte durée des tours de paroles et le nombre plus importants de locuteurs. Ces spécifications compliquent la détection des locuteurs car la quantité de donnée disponible pour un locuteurs est alors réduite. Ceci est appuyé par le fait le système du LIA utilise le E-HMM dont la faiblesse reconnue pour cette tâche est de mieux détecter les locuteurs principaux que les "petits" locuteurs. L'autre raison de cette énorme dégradation des performances peut être mise en relation avec l'environnement d'enregistrement de ces données par rapport aux données classiques (paroles sur de la musique, effets sonores ou musique). Il est intéressant de noter que le nombre moyen de locuteurs trouvé par le système bottom-up du LIUM est assez proche du nombre effectif de locuteurs (tableaux 1 et 3). Néanmoins, les relativement faibles performances de ce système montre que les segments associés aux locuteurs ne sont pas ceux attendus. Enfin, on peut aussi relever les mêmes tendances pour les deux systèmes, la dégradation des résultats relative à la catégorie est sensiblement équivalente pour les deux systèmes.

5. Conclusion et perspectives

Cette étude préliminaire montre les difficultées rencontrées par les systèmes de speaker diarization, plus particulièrement pour les systèmes top-down, sur les données audio/vidéo hétérogènes. De nouvelles techniques doivent être imaginées pour traiter ces données particulières, notament pour dépasser l'obstacle de l'environnement. L'utilisation du Factor Analysis va être étudié dans ce contexte particulier. Deuxièmement, le flux vidéo contient des informations à propos du locuteur qui devraient être utilisées pour aider le système à prendre sa décision lorsqu'il y a une ambiguité sur le locuteur détécté via le flux audio.

6. Remerciements

Nous sommes reconnaissants à Sylvain MEIGNIER et Teva MERLIN de nous avoir fournis leur système et leur aide.

7. Bibliographie

Bonastre J.-F., Wils F., Meignier S., « ALIZE, a free toolkit for speaker recognition », *Proc. ICASSP'05*, vol. 1, Philadelphia, USA, p. 737-740, March, 2005.

Fredouille C., Bozonnet S., Evans N. W. D., « The LIA-EURECOM RT'09 Speaker Diarization System », *RT'09, NIST Rich Transcription Workshop*, Florida, USA, 2009.

Galliano S., Gravier G., Chaubard L., « The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts », *Interspeech'09*, Brighton, UK, 2009.

Meignier S., Merlin T., « LIUM_SPKDIARIZATION : An open source toolkit for diarization », *CMU SPUD Workshop*, 2010.

NIST, « The NIST Rich Transcription (RT'09) Evaluation », rt09-meeting-eval-plan-v2.pdf, 2009.

Rouvier M., Linares G., Matrouf D., « On-the-fly Video genre classification by combination of audio features », *ICASSP 2010*, Dallas, US, 2010.