Autoapprentissage pour le regroupement en locuteurs : premières investigations

Gaël Le Lan, Sylvain Meignier, Delphine Charlet, Anthony Larcher

Autoapprentissage pour le regroupement en locuteurs : premières investigations

Gaël Le Lan[1,2] Sylvain Meignier[1] Delphine Charlet[2] Anthony Larcher[1]
(1) LIUM, Université du Maine, France
(2) Orange Labs, France
prenom.nom@lium.univ-lemans.fr, prenom.nom@orange.com

RÉSUMÉ

Cet article traite de l'autoapprentissage d'un système *i-vector/PLDA* pour le regroupement en locuteurs de collections d'archives audiovisuelles françaises. Les paramètres d'extraction des i-vectors et du calcul des scores PLDA sont appris de façon non supervisée sur les données de la collection elle-même. Différents mélanges de données cibles et de données externes sont comparés pour la phase d'apprentissage. Les résultats expérimentaux sur deux corpora cibles distincts montrent que l'utilisation des données des corpora en question pour l'apprentissage itératif non supervisé et l'adaptation des paramètres de la PLDA peut améliorer un système existant, appris sur des données annotées externes. De tels résultats indiquent que la structuration automatique en locuteurs de petites collections non annotées ne devrait reposer que sur l'existence d'un corpus externe annoté, qui peut être spécifiquement adapté à chaque collection cible. Nous montrons également qu'une collection suffisamment grande peut se passer de l'utilisation de ce corpus externe.

ABSTRACT

First investigations on self trained speaker diarization

This paper investigates self trained cross-show speaker diarization applied to collections of French TV archives, based on an *i-vector/PLDA* framework. The parameters used for i-vectors extraction and PLDA scoring are trained in a unsupervised way, using the data of the collection itself. Performances are compared, using combinations of target data and external data for training. The experimental results on two distinct target corpora show that using data from the corpora themselves to perform unsupervised iterative training and domain adaptation of PLDA parameters can improve an existing system, trained on external annotated data. Such results indicate that performing speaker indexation on small collections of unlabeled audio archives should only rely on the availability of a sufficient external corpus, which can be specifically adapted to every target collection. We show that a minimum collection size is required to exclude the use of such an external bootstrap.

1 Introduction

La tâche de Segmentation et Regroupement en Locuteurs (SRL) vise à étiqueter les locuteurs dans un ou plusieurs enregistrements audios, sans connaissance a priori des locuteurs. L'augmentation constante des volumes de données nécessite une indexation efficace. La SRL appliquée à des collections est une tâche globale qui consiste à traiter un ensemble d'enregistrements audios bruts (non segmentés en tours de parole) afin d'identifier de manière unique les tours de paroles de chaque locuteur. Cette tâche se décompose généralement en deux étapes : la SRL intra-enregistrement, où il

s'agit de segmenter et regrouper les occurrences des locuteurs au sein d'un même enregistrement, et le regroupement inter-enregistrements, qui vise à regrouper les locuteurs intra-enregistrements (Meignier *et al.*, 2002). D'autres implémentations sont possibles, où tous les enregistrements peuvent être concaténés en un super-enregistrement, pour ensuite être traités comme un problème artificiel de SRL intra-enregistrement (Tran *et al.*, 2011). L'approche en deux étapes, plus naturelle, est la plus couramment utilisée. Elle permet de traiter de grands volumes de données comme des archives audiovisuelles ou radiophoniques (Le Lan *et al.*, 2016; Dupuy *et al.*, 2014a; Yang *et al.*, 2011; Tran *et al.*, 2011; Van Leeuwen, Proc Odyssey 2010), des enregistrements téléphoniques (Shum *et al.*, Proc Odyssey 2014, 2013; Karam & Campbell, 2013; Ghaemmaghami *et al.*, 2012), ou encore d'enregistrements de réunions (Ferràs & Bourlard, 2012).

L'état de l'art en reconnaissance du locuteur repose sur le système *i-vector/PLDA*, qui requiert des données annotées en locuteurs, indiquant l'identité et les tours de parole de ceux-ci. La modélisation de la variabilité inter-locuteur est une étape clé. Pour la réaliser, le corpus d'apprentissage doit contenir plusieurs occurrences d'un même locuteur dans des conditions acoustiques variées. Les annotations manuelles en locuteurs sont coûteuses et rarement disponibles. Ainsi, deux approches peuvent être utilisées pour entraîner un système automatique : 1) un apprentissage non supervisé sur les données cibles, 2) une approche supervisée sur des données d'entraînement annotées en locuteurs. Dans cette deuxième approche, la différence de conditions acoustiques entre le corpus d'apprentissage et le corpus cible induit une dégradation des performances sur la collection cible. Des solutions de compensation ont déjà été proposées pour la tâche de vérification du locuteur, basées sur l'adaptation au domaine de manière non supervisée (Shum *et al.*, Proc Odyssey 2014). Dans le contexte de l'apprentissage non supervisé d'un modèle PLDA [1], la notion d'apprentissage itératif (Liu *et al.*, 2014) a été proposée pour améliorer la qualité des modèles. Dans les deux articles précités, l'apprentissage était fait sur des enregistrements mono locuteur dont l'identité est inconnue. Dans (Le Lan *et al.*, 2016), nous avons montré qu'un système de SRL pour une collection, entraîné de façon non supervisée sur des enregistrements multi locuteurs non segmentés en tour de parole, pouvait être aussi performant qu'un système appris de manière supervisée sur les mêmes données. Ces expériences nous ont montré que les annotations d'un corpus d'apprentissage ne sont pas obligatoires.

Cette étude porte sur l'autoapprentissage des modèles *i-vector/PLDA* utilisés lors des phases de regroupement intra-et inter-enregistrements à partir des enregistrements eux-mêmes, sans l'utilisation de données d'apprentissage ou d'adaptation externes. L'idée est de proposer un système de SRL utilisant le minimum de connaissances a priori. Dans ce contexte, nous étudions plusieurs variantes du système de regroupement *i-vector/PLDA*, appris avec ou sans données externes. Nous évaluons aussi des modèles appris à partir de la combinaison de données étiquetées et non étiquetées en locuteur. Nous commençons par décrire le système de SRL en précisant le périmètre de l'autoapprentissage pour la tâche visée. Ensuite, nous détaillons les données utilisées pour les expériences avant de conclure avec l'étude des performances du système proposé et des améliorations possibles.

2 Systèmes de SRL intra-et inter-enregistrements

La figure 1 décrit les traitements pour l'étape de SLR inter-enregistrements et l'étape intra-enregistrement résumés dans les deux paragraphes ci-dessous. Nous invitons le lecteur à se reporter à (Le Lan *et al.*, 2016) pour une description complète du processus de SRL.

1. PLDA : Probabilistic Linear Discriminant Analysis

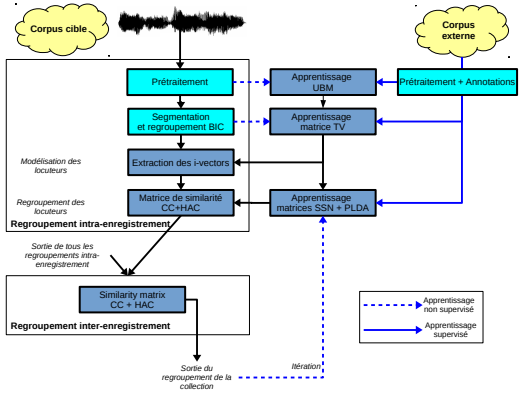

FIGURE 1 – Vue d'ensemble du système de SRL, entraîné de façon supervisé (traits bleus continus), non supervisé (traits pointillés bleus) et semi-supervisé (tous les traits bleus).

SRL intra-enregistrement : après l'extraction de 13 MFCC et une étape de détection de la parole, un système classique de segmentation et de classification BIC (Barras *et al.*, 2006) permet d'obtenir des classes très pures contenant les tours de parole d'un locuteur. À ce stade, nous disposons de suffisamment de données pour apprendre à partir de MFCC normalisés (moyenne et variance) et étendus avec leurs dérivés premières et secondes des modèles plus complexes comme les *i-vectors* (Kenny, 2010). Après leur apprentissage, ces modèles sont normalisés par la méthode décrite dans (Bousquet *et al.*, 2012).

Un rapport de vraisemblance entre chaque paire de *i-vectors* est calculé en utilisant un modèle PLDA (dimension 200 pour la matrice *inter-classes* et sans matrice *intra-classes*). Ces paires forment une matrice de scores sur laquelle nous appliquons l'algorithme de regroupement $CC + HAC$ à base de graphe et de regroupement hiérarchique décrit dans (Dupuy *et al.*, 2014b).

SRL inter-enregistrements : après avoir traité chaque émission séparément, la collection est considérée dans son ensemble et le même algorithme $CC + HAC$ est appliqué afin de regrouper les classes obtenues dans chaque émission, chacune étant représentée par un unique *i-vector*.

3 Autoapprentissage du système automatique

Après avoir montré dans (Le Lan *et al.*, 2016) que, pour un corpus de grande dimension, l'apprentissage non supervisé d'un modèle de SRL pouvait donner des performances comparables à un apprentissage supervisé, nous étudions ici le cas où aucun corpus externe (annoté ou non) n'est disponible pour l'apprentissage et où les seules données d'apprentissage sont issues du corpus cible lui-même, qui est non annoté.

La figure 1 représente les processus d'apprentissage d'un système de SRL supervisé (traits continus bleus) et non supervisé (traits pointillés bleus). Ce dernier est similaire à celui décrit dans (Le Lan *et al.*, 2016) mais ici, le corpus cible sert de corpus d'apprentissage. Nous comparons cette approche à un système de référence, appris de façon classique sur un corpus d'apprentissage annoté et à un système oracle appris sur le corpus cible avec des annotations manuelles. Enfin, nous adaptons

itérativement la PLDA du système de référence avec le corpus cible (traits bleus continus et pointillés).

Le système de SRL nécessite l'apprentissage d'un modèle du monde (UBM), d'un extracteur de *i-vectors* (TV) et d'un modèle PLDA. Leur apprentissage non supervisé nécessite d'extraire des informations sur les locuteurs de façon automatique comme suit. Un UBM à 256 gaussiennes est appris sur des segments de parole détectés grâce à un système parole/non-parole à base de GMMs. L'apprentissage de TV, de rang 200, nécessite des segments contenant un unique locuteur. Ceux-ci sont obtenus grâce à un système de segmentation BIC qui produit, d'après notre expérience, des classes pures. L'apprentissage des paramètres de la normalisation sphérique et du modèle PLDA nécessite des annotations en locuteurs et sessions afin de modéliser les variances inter-et intra-locuteurs. Nous ne considérons, pour apprendre ces paramètres, que les locuteurs apparaissant dans au moins trois émissions pour une durée minimum de 10 secondes par émission. Ces annotations en locuteur sont obtenues grâce à un système non supervisé de SRL similaire à celui décrit à la section 2 mais utilisant une distance cosinus.

4 Contexte expérimental

Les modèles contrastifs pour la SRL ont été appris sur trois corpus distincts, tirés d'environ 200 heures d'émissions audiovisuelles des campagnes d'évaluation REPERE (Galibert & Kahn, 2013), ETAPE (Galibert *et al.*, 2014) et ESTER(Galliano *et al.*, 2009). Ces corpus, où les locuteurs sont identifiés par leur nom et prénom, contiennent plusieurs sessions pour un grand nombre de locuteurs. Les locuteurs apparaissant dans plus d'un épisode sont appelés locuteurs récurrents (r.), par opposition aux locuteurs ponctuels (p.), qui ne sont présents que dans un épisode.

Nous définissons deux corpus $cible$, construits à partir des corpus officiels d'apprentissage et de test de REPERE (Galibert & Kahn, 2013). Le premier, qu'on appellera LCP_{cible}, est la collection de tous les épisodes disponibles de l'émission $LCPInfo$. Le second, qu'on appellera BFM_{cible}, contient tous les épisodes disponibles de l'émission $BFMStory$. Ces deux émissions ont été choisies parce qu'elles comptent un nombre suffisamment important d'épisodes (plus de quarante chacune), et contiennent de nombreux locuteurs récurrents, qui parlent pour plus de 50% du temps de parole total de la collection. Quelques chiffres à propos des deux corpus sont présentés dans le tableau 1. Les deux corpus étant partiellement annotés, seuls les chiffres des locuteurs annotés sont présentés.

Corpus	LCP_{cible}	BFM_{cible}
Nombre d'épisodes	45	42
Durée d'un épisode	25m	60m
Durée de parole annotée	10h08m	19h57m
Locuteurs ponctuels	127	345
Locuteurs récurrents (2 occurrences ou plus)	93	77
Locuteurs récurrents (3 occurrences ou plus)	48	35
Nombre total de locuteurs	220	422
Locuteurs p., part du temps de parole total	20.12%	44,84%
Locuteurs r. (2+ occ.), part du temps de parole total	79,88%	55,16%
Locuteurs r. (3+ occ.), part du temps de parole total	67,06%	45,94%
Durée de parole moyenne d'un locuteur, par épisode	1m08s	1m58s

TABLE 1 – Composition des deux corpus cibles. Seuls les locuteurs annotés sont présentés. La durée de parole annotée correspond à la durée de parole comptant pour l'évaluation.

Le corpus d'entraînement, ou $train$, est utilisé pour des expériences complémentaires. Il contient 344 enregistrements audios issus des corpus d'entraînement et de développement des campagnes citées précédemment, pour un total de 200 heures de parole annotée. Le corpus contient 3888 locuteurs uniques, dont 391 répondent aux critères minimaux choisis pour l'estimation des paramètres de la PLDA : ils apparaissent dans au moins trois enregistrements avec un temps de parole minimal de dix secondes par enregistrement. Par conséquent, ce corpus est bien adapté pour l'apprentissage d'un système *i-vector/PLDA*.

5 Expériences

La métrique utilisée pour mesurer la performance de SRL est le DER (pour Diarization Error Rate, taux d'erreur de SRL). Il a été introduit par le NIST comme la part de temps de parole qui n'est pas attribuée au bon locuteur, en utilisant la meilleure correspondance entre la référence et l'hypothèse. L'outil d'évaluation utilisé (Galibert, 2013) sert à mesurer la performance de la SRL intra-enregistrement et inter-enregistrements. Dans ce dernier cas, on s'attend à ce que le système identifie chaque locuteur récurrent par la même étiquette dans tous les enregistrements de la collection. Nous reportons les résultats intra-et inter-enregistrements, cependant nous commenterons principalement le DER inter-enregistrements. Les résultats intra-enregistrement montrent dans la plupart des cas un très faible impact du calcul du score entre les paires de locuteurs (tableau 2). Pour le calcul du DER, une erreur de 250ms aux frontières des segments est tolérée et la parole superposée est aussi évaluée.

5.1 $Référence$ et $Oracle$

Le système $référence$ est un système *i-vector/PLDA* supervisé, entraîné sur le corpus $train$ et appliqué sur les deux corpus $cible$. Ce système représente la stratégie usuelle lorsqu'on souhaite traiter un nouveau corpus $cible$ non annoté et qu'on dispose déjà d'un corpus d'entraînement annoté. Pour chaque corpus $cible$, un système $oracle$ est également appris de manière supervisée, à partir des annotations de référence des corpus $cible$.

Enfin, un système est appris de façon non supervisée, uniquement à partir d'annotations automatiques de chaque corpus $cible$. Il s'agit d'un système *i-vector utilisant une distance cosinus* qui ne nécessite pas d'information sur les locuteurs récurrents. Ce système est appelé $référence_{cible}$. Les résultats présentés dans le tableau 2 montrent que sans aucune information a priori sur le corpus $cible$, les performances pour le regroupement inter-enregistrements du système $référence_{cible}$ sont bien plus mauvaises que la $référence$, utilisant des données externes pour l'apprentissage (un DER inter-enregistrements de 29,68% ($référence_{cible}$) contre 17,72% ($référence$) et de 27,62% contre 13,22% respectivement). Ceci s'explique par le fait que le système $référence_{cible}$ ne modélise en aucune façon la variabilité inter-locuteurs.

Si l'on s'intéresse à l'expérience $oracle$, l'apprentissage d'une PLDA supervisée sur le corpus $cible$ n'est pas toujours possible : pour le corpus BFM_{cible}, l'apprentissage de la PLDA ne converge pas, probablement dû à un nombre de locuteurs récurrents trop faible (35). En revanche, lorsque le corpus $cible$ contient suffisamment de locuteurs récurrents, les résultats montrent que l'utilisation de l'information a priori donne les meilleurs résultats avec un DER inter-enregistrements de 10,87%.

5.2 Critères limites pour l'apprentissage de la PLDA

Les résultats $oracle$ montrent que pour un corpus $cible$, il pourrait ne pas y avoir assez de locuteurs récurrents pour apprendre la PLDA. La figure 2 montre que pour le corpus LCP_{cible}, un système PLDA efficace peut être entraîné à partir de 37 épisodes, comprenant alors 40 locuteurs récurrents. Chaque locuteur apparaît alors en moyenne dans 7 épisodes différents. Ces résultats montrent qu'un nombre minimal de locuteurs récurrents est nécessaire pour l'autoapprentissage de la PLDA sur un corpus $cible$. Cependant, la définition des critères minimaux pour l'estimation des paramètres de la PLDA n'est pas encore élucidée. Lorsqu'il y a trop peu de données pour l'apprentissage, l'algorithme EM d'estimation des paramètres ne converge pas.

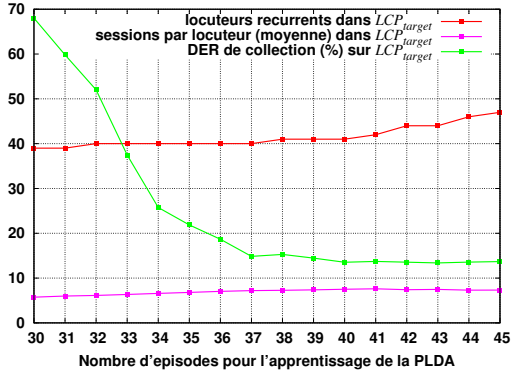

FIGURE 2 – Métriques du regroupement en locuteurs du corpus LCP_{cible}, dont la PLDA est auto apprise avec les annotations. Les valeurs sont fonction du nombre d'épisodes utilisés pour l'apprentissage. Les épisodes ont été sélectionnés dans l'ordre chronologique.

Expérience	UBM + TV appris sur	SNN + PLDA appris sur	LCP_{cible} DER I (%)	LCP_{cible} DER C (%)	BFM_{cible} DER I (%)	BFM_{cible} DER C (%)
$référence$	$train_{sup}$	$train_{sup}$	7,77	17,72	9,92	13,22
$oracle$	$cible_{sup}$	$cible_{sup}$	6,68	10,87	X	X
$référence_{cible}$	$cible_{nonsup}$	sans objet (cosinus)	7,04	29,68	12,46	27,62
$PLDA_{sup}$	$cible_{nonsup}$	$train_{sup}$	7,80	16,58	8,30	12,60
$PLDA_{adapt}$	$cible_{nonsup}$	$train_{sup} + cible_{nonsup}$	7,67	15,60	8,58	11,38

TABLE 2 – DER inter-enregistrement (I) et DER intra-enregistrements (C) des corpus $cible$, pour toutes les expériences.

5.3 Apprentissage non supervisé

Dans la suite de cet article, l'utilisation des annotations fournies avec les deux corpora $cible$ est interdite. Dans (Le Lan *et al.*, 2016), nous avons montré que l'utilisation de sessions issues d'un regroupement BIC pour l'apprentissage non supervisé d'un système *i-vector/PLDA* pouvait fonctionner aussi bien qu'un système supervisé appris sur des données manuellement annotées.

Nous avons donc entraîné un UBM et une matrice TV à partir des sessions BIC. Sur cette base, nous avons appris une PLDA supervisée sur les données du corpus $train$ complémentaire (expérience $PLDA_{sup}$). Les résultats, présentés dans le tableau 2, montrent que l'utilisation d'un UBM et d'une matrice TV appris sur les données $cible_{nonsup}$ à la place des données $train_{sup}$ améliore le DER inter-enregistrements d'environ 1% absolu (16,58% contre 17,72% et 12,60% contre 13,22%).

Dans un second temps, nous avons appris un modèle PLDA à partir des locuteurs issus de la sortie du système de SRL de l'expérience $PLDA_{sup}$. En effet, certaines classes sont étiquetées comme locuteurs récurrents par le système de SRL. L'apprentissage de la PLDA sur ces données n'a pas fonctionné, l'algorithme EM ne convergeant pas. Ceci s'explique probablement par le fait que ces classes ne sont pas suffisamment pures, ne permettant pas d'agréger des statistiques suffisantes à la convergence de l'algorithme EM.

La composition des données pour l'apprentissage supervisé et non supervisé des UBM, TV et PLDA pour les deux corpora est présentée dans le tableau 3. La quantité de données utilisées pour l'apprentissage supervisé est deux fois moins importante que pour l'apprentissage non supervisé. On rappelle que seulement 50% des corpus sont annotés, alors que le système de SRL automatiquement est appliqué sur l'intégralité de l'émission. Cette différence se traduit aussi par un nombre plus important de locuteurs pour l'apprentissage non supervisé.

	LCP_{cible}		BFM_{cible}	
	sup	$nonsup$	sup	$nonsup$
Données utilisées pour l'apprentissage de UBM/TV	9h56m	19h17m	19h50m	39h09m
Durée moyenne d'une session pour l'apprentissage de UBM/TV	1m08s	4m23s	1m58s	4m10s
Nombre de classes-locuteurs pour l'apprentissage de la PLDA	47	130	35	190
Nombre de sessions moyen par classe-locuteur	7,31	5,25	5,45	4,34
Durée moyenne d'une session	1m10s	1m25s	2m50s	2m07

TABLE 3 – Composition des données d'apprentissage de l'UBM, TV et PLDA, pour les deux corpus $cible$.

5.4 Adaptation au domaine

Nous avons constaté que les classes produites par la SRL à l'expérience $PLDA_{sup}$ ne sont pas suffisantes pour apprendre une PLDA dans tous les cas. L'expérience $PLDA_{adapt}$ consiste à utiliser conjointement le corpus supervisé $train_{sup}$ et le corpus cible non supervisé $cible_{nonsup}$. On constate une amélioration de la performance, avec une baisse du DER inter-enregistrements de 16,58% à 15,60% et de 12,60% à 11,38% pour les deux corpus. Les locuteurs récurrents étiquetés après l'expérience $PLDA_{sup}$ permettent donc d'améliorer la qualité de la modélisation de la variabilité inter-locuteurs.

Dans (Shum *et al.*, Proc Odyssey 2014), la question de l'adaptation au domaine est abordée par une combinaison de données annotées externes et de données non annotées acoustiquement proche des données cibles, mais sans utiliser les données cibles elles-mêmes. Les travaux cités consistent à introduire une variable de pondération entre les données internes et externes pour l'estimation des paramètres de la PLDA. Notre approche est plus simple, elle consiste à concaténer deux corpus, la pondération dépendant de la taille relative des corpus. Les résultats de nos travaux sont similaires à ceux de l'article cité, avec une amélioration des performances lorsqu'on utilise l'adaptation, ceci malgré la différence de qualité des données utilisées.

La sortie du système $PLDA_{adapt}$ peut être utilisée pour apprendre un nouveau modèle PLDA. Cependant, nos expériences montrent que lorsqu'on itère après l'expérience $PLDA_{adapt}$, le système ne s'améliore plus. L'idée d'apprentissage non supervisé itératif de la PLDA a été introduite dans (Liu *et al.*, 2014), mais il s'agissait d'appliquer un regroupement en locuteurs sur le corpus d'entraînement lui-même, initialisé par une première itération de SRL *i-vector/cosine*. Dans notre approche, les paramètres de la PLDA sont initialisés sur des données externes annotées, puis estimées à nouveau avec l'ajout des données cibles. Si les travaux antérieurs montraient une amélioration des performances en vérification du locuteur avec un apprentissage itératif, notre approche n'a pas été aussi efficace.

Nos données étant des enregistrements non segmentés et multi locuteurs, les *i-vectors* utilisés pour l'apprentissage itératif pourrait ne pas être suffisamment précis en termes de représentation du locuteur. De meilleurs résultats pourraient être obtenus en modifiant le seuil de regroupement inter-émissions. Un seuil plus strict améliorerait la pureté des classes-locuteurs et pourrait permettre une meilleure modélisation PLDA.

6 Conclusion

Dans cet article, nous avons proposé un système de Segmentation et Regroupement en Locuteurs inter-enregistrements utilisant une stratégie d'autoapprentissage pour le traitement de petites collections d'archives audiovisuelles non annotées. Alors que des travaux antérieurs montraient que l'apprentissage non supervisé d'un système *i-vector/PLDA* sur de telles données était valide, la petite taille des corpus cibles ne permet pas un autoapprentissage performant. L'utilisation de données externes pour un premier apprentissage supervisé reste indispensable.

Nous avons appliqué avec succès une technique d'adaptation au domaine, qui s'était avéré efficace pour la tâche de vérification du locuteur sur des données mono locuteurs, afin d'améliorer le système de SRL $référence$. L'utilisation de l'information sur les locuteurs récurrents apportée par une première itération de SRL sur la collection cible a permis d'améliorer le DER sur les deux corpus.

Les travaux futurs seront consacrés à l'amélioration de la méthode d'apprentissage, en introduisant une pondération dans les paramètres de l'adaptation au domaine entre les données externes et les données cibles. Nous prévoyons également d'approfondir l'aspect itératif de l'apprentissage, la littérature ayant montré que des améliorations étaient possibles.

Références

BARRAS C., ZHU X., MEIGNIER S. & GAUVAIN J. (2006). Multi-stage speaker diarization of broadcast news. *IEEE Transactions on Speech and Audio Processing*, **14**(5), 1505–1512.

BOUSQUET P.-M., LARCHER A., MATROUF D., BONASTRE J.-F. & PLCHOT O. (2012). Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis. In *Speaker Odyssey Workshop*, p. 157–164.

DUPUY G., , MEIGNIER S. & ESTÈVE Y. (2014a). Is incremental cross-show speaker diarization efficient to process large volumes of data ? In *Proceedings of Interspeech*, Singapore.

DUPUY G., MEIGNIER S., DELÉGLISE P. & ESTÈVE Y. (2014b). Recent improvements towards ILP-based clustering for broadcast news speaker diarization. In *Speaker Odyssey Workshop*.

FERRÀS M. & BOURLARD H. (2012). Speaker Diarization and Linking of Large Corpora. In *Proceedings of IEEE Workshop on Spoken Language Technology*, Miami, Florida (USA).

GALIBERT O. (2013). Methodologies for the evaluation of speaker diarization and automatic speech recognition in the presence of overlapping speech. In *INTERSPEECH*, p. 1131–1134.

GALIBERT O. & KAHN J. (2013). The first official repere evaluation. In *Proceedings of Interspeech satellite workshop on Speech, Language and Audio in Multimedia (SLAM)*.

GALIBERT O., LEIXA J., GILLES A., CHOUKRI K. & GRAVIER G. (2014). The ETAPE Speech Processing Evaluation. In *Conference on Language Resources and Evaluation*, Reykyavik, Iceland.

GALLIANO S., GRAVIER G. & CHAUBARD L. (2009). The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. In *Proceedings of Interspeech*, Brighton, Royaume Uni.

GHAEMMAGHAMI H., DEAN D., VOGT R. & SRIDHARAN S. (2012). Speaker attribution of multiple telephone conversations using a complete-linkage clustering approach. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, p. 4185–4188 : IEEE.

KARAM Z. N. & CAMPBELL W. M. (2013). Graph embedding for speaker recognition. In *Graph Embedding for Pattern Analysis*, p. 229–260. Springer.

KENNY P. (2010). Bayesian speaker verification with heavy tailed priors. In *Speaker Odyssey Workshop*.

LE LAN G., MEIGNIER S., CHARLET D. & DELÉGLISE P. (2016). Speaker diarization with unsupervised training framework. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on* : IEEE.

LIU W., YU Z. & LI M. (2014). An iterative framework for unsupervised learning in the plda based speaker verification. In *Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on*, p. 78–82 : IEEE.

MEIGNIER S., BONASTRE J.-F. & MAGRIN-CHAGNOLLEAU I. (2002). Speaker utterances tying among speaker segmented audio documents using hierarchical classification : towards speaker indexing of audio databases. In *INTERSPEECH* : Citeseer.

SHUM S. H., CAMPBELL W. M., REYNOLDS D. *et al.* (2013). Large-scale community detection on speaker content graphs. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, p. 7716–7720 : IEEE.

SHUM S. H., REYNOLDS D. A., GARCIA-ROMERO D. & MCCREE A. (Proc. Odyssey 2014). Unsupervised clustering approaches for domain adaptation in speaker recognition systems.

TRAN V.-A., LE V. B., BARRAS C. & LAMEL L. (2011). Comparing multi-stage approaches for cross-show speaker diarization. In *INTERSPEECH*, volume 201, p. 1053–1056.

VAN LEEUWEN D. A. (Proc. Odyssey 2010). Speaker linking in large data sets.

YANG Q., JIN Q. & SCHULTZ T. (2011). Investigation of cross-show speaker diarization. In *INTERSPEECH*, volume 11, p. 2925–2928.