Segmentation et Regroupement en Locuteurs d'une collection de documents audio

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

RÉSUMÉ

Nous proposons d'étudier la segmentation et le regroupement en locuteurs dans le cadre du traitement d'une collection de documents audio. L'objectif est de détecter les locuteurs qui apparaissent dans plusieurs émissions. Dans notre approche, les émissions sont traitées indépendamment les unes des autres avant d'être traitées globalement, afin de regrouper les locuteurs intervenant dans plusieurs émissions. Deux méthodes de regroupement sont étudiées pour le traitement global de la collection : l'une utilise la métrique NCLR et l'autre s'inspire des techniques à base de i-vecteurs, employées en vérification du locuteur, et est exprimé sous la forme d'un problème de PLNE. Ces deux méthodes ont été évaluées sur deux corpus de 15 émissions issues d'ESTER 2. La méthode basée sur l'utilisation des i-vecteurs réalise des performances légèrement inférieures à celles obtenues par la méthode NCLR, cependant le temps de calcul est en moyenne 17 fois plus rapide. Cette méthode est, par conséquent, adaptée au traitement de grandes quantités de données.

ABSTRACT

Cross-show speaker diarization

We propose to study speaker diarization from a collection of audio documents. The goal is to detect speakers appearing in several shows. In our approach, shows are processed independently of each other before being processed collectively, to group speakers involved in several shows. Two clustering methods are studied for the overall treatment of the collection: one uses the NCLR metric and the other is inspired by techniques based on i-vectors, used in the speaker verification field, and is expressed as an ILP problem. Both methods were evaluated on two sets of 15 shows from ESTER 2. The method based on i-vectors achieves performance slightly lower than those obtained by the NCLR method, however, the computation time is on average 17 times faster. Therefore, this method is suitable for processing large volumes of data.

MOTS-CLÉS : SRL, traitement de collection, i-vecteurs, regroupement PLNE.

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

1 Introduction

La tâche de segmentation et de regroupement en locuteurs (SRL) a été définie par le NIST 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 s'applique

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 433–440,
Grenoble, 4 au 8 juin 2012.

individuellement sur chacun des enregistrements audio du corpus sans utiliser de connaissances *a priori* sur les locuteurs.

La plupart des systèmes de SRL proposés jusqu'à très récemment ont suivi cette définition de la tâche, où les émissions sont traitées et évaluées individuellement (SRL d'émissions). Dans ce cadre, les locuteurs détectés par les systèmes sont identifiés par des étiquettes anonymes propres à chaque enregistrement. Un même locuteur intervenant dans deux émissions est donc identifié par deux étiquettes différentes.

La segmentation et le regroupement en locuteurs joue un rôle prépondérant dans de nombreuses applications de traitement automatique de la parole, telles que la transcription automatique, la détection des entités nommées, la détection du rôle des locuteurs. En considérant la quantité toujours croissante de ressources multimédia disponibles, il devient intéressant et nécessaire de considérer la SRL dans un contexte plus global. L'inconvénient majeur de l'approche traditionnelle en SRL est la non prise en compte des interventions récurrentes de certains locuteurs dans plusieurs émissions. Cette situation est très fréquente dans les émissions journalistiques où, généralement, les présentateurs, journalistes et autres invités qui les animent apparaissent régulièrement. (Tran *et al.*, 2011) et (Yang *et al.*, 2011) introduisent la notion de SRL sur une collection d'émissions provenant d'une même source. Les auteurs présentent différentes approches pour détecter et regrouper globalement les locuteurs sur l'ensemble des émissions de la collection (SRL de collection). Ainsi, un locuteur intervenant dans plusieurs émissions est identifié par la même étiquette dans chacune de ces émissions.

Nous présentons et comparons dans cet article deux méthodes de regroupement en locuteurs adaptées au traitement de collections d'émissions journalistiques françaises. Nous utilisons une architecture à deux niveaux qui combine à la fois une SRL d'émissions, dans laquelle les émissions sont traitées individuellement, et une SRL de collection, où les émissions de la collection sont regroupées pour être traitées de manière globale.

Dans les paragraphes suivants, nous décrivons le système de SRL d'émissions du LIUM [1] ainsi que l'architecture et les méthodes proposées pour la SRL de collection. Nous présentons ensuite les corpus de données utilisés, la configuration des systèmes de SRL de collection et nos résultats expérimentaux.

2 Système de SRL d'émissions

Le système utilisé lors de nos expériences, le *LIUM_SpkDiarization* [2] (Meignier et Merlin, 2009), a été développé pour la campagne d'évaluation française ESTER 2 (Galliano *et al.*, 2009), 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. Les limites des segments sont ensuite ajustées au moyen d'un décodage de Viterbi utilisant des GMM (Gaussian Mixture Model) à 8 composantes apprises sur les données de chaque locuteur via l'algorithme EM (Expectation-Maximization). Une segmentation en

1. Laboratoire d'Informatique de l'Université du Maine
2. `http://www-lium.univ-lemans.fr/fr/content/liumspkdiarization`

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 normalisé (NCLR) (Le *et al.*, 2007) 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. Cet UBM à 512 composantes correspond à la concaténation de quatre GMM à 128 composantes, dépendantes du genre (homme ou femme) et du canal (studio ou téléphone).

3 Architectures pour la SRL de collection

Un système de SRL d'émissions permet de détecter les interventions des locuteurs au sein d'une émission. Un système de SRL de collection doit être, en plus, capable de détecter les locuteurs qui apparaissent dans plusieurs émissions. (Tran *et al.*, 2011) et (Yang *et al.*, 2011) ont expérimenté trois architectures différentes (figure 1) :

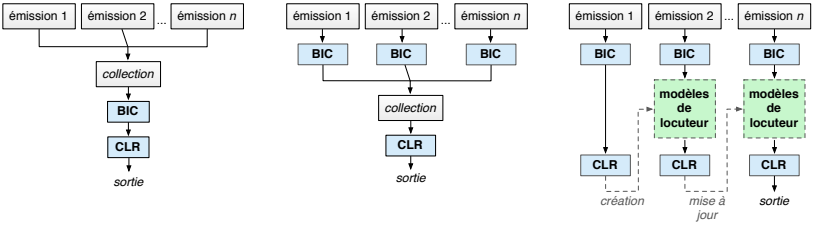

FIGURE 1 – Les trois architectures de SRL de collection proposées par (Tran *et al.*, 2011) : la *concaténation* des émissions à gauche, le système *hybride* au centre, le système *incrémental* à droite.

1. une *concaténation* de toutes les émissions de la collection, sur laquelle est utilisé un système classique de SRL d'émissions (proche du système présenté dans la section 2),
2. un système *hybride*, dans lequel une classification BIC est réalisée individuellement sur chaque émission, et dont la concaténation des sorties est utilisée pour une classification BIC globale (Yang *et al.*, 2011) ou CLR globale (Tran *et al.*, 2011),
3. un système *incrémental*, qui traite les émissions individuellement les unes après les autres. Seules les informations provenant des émissions déjà traitées peuvent aider la SRL de l'émission en cours. Les modèles de locuteurs appris sur chaque émission sont utilisés et mis à jour au fil du traitement de la collection.

Les performances des systèmes par *concaténation* et *hybride*, présentés par (Tran *et al.*, 2011), sont comparables. Le système *incrémental* se démarque par la rapidité avec laquelle le traitement de

la collection est réalisé. Cette architecture est la plus adaptée à l'insertion de nouvelles émissions dans la collection, mais elle présente deux inconvénients : les résultats en termes de taux d'erreur en reconnaissance de locuteur sur l'ensemble de la collection sont supérieurs à ceux obtenus par les deux autres systèmes, et l'ordre dans lequel les émissions sont traitées influence les résultats. Ces expériences ont montré que les meilleurs résultats sont obtenus au détriment du temps de traitement, et *vice-versa*.

Nous avons considéré une approche différente en choisissant de mettre en œuvre un système approprié au traitement de grandes quantités de données. Un tel système se doit d'être à la fois performant en termes de taux d'erreur et raisonnable en temps de calcul, ainsi qu'en consommation mémoire. Nous nous sommes inspirés de l'architecture du système *hybride* en testant deux méthodes de classification différentes pour le traitement global de la collection. Les schémas de la figure 2 présentent les deux méthodes de regroupement testées : la première met en oeuvre une classification NCLR (schéma de gauche) et la seconde est formulée par un problème de Programmation Linéaire en Nombre Entier (PLNE), basé sur l'utilisation de i-vecteurs (schéma de droite). Dans les deux cas, chaque émission est traitée individuellement, en utilisant le système de SRL décrit dans la section 2, avant de chercher à détecter les locuteurs communs à la collection. La collection est obtenue par concaténation des sorties des traitements locaux aux émissions. L'utilisation de la méthode de classification globale par PLNE présente un double avantage par rapport à sa variante NCLR : le temps de calcul est plus rapide et la quantité de mémoire utilisée est réduite, alors que les résultats en termes de taux d'erreur restent similaires.

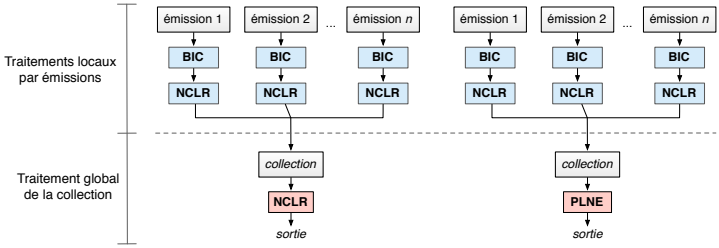

FIGURE 2 – Architecture de SRL pour une collection avec deux variantes : classification globale NCLR sur le schéma de gauche et classification globale par PLNE sur le schéma de droite.

SRL de collection par NCLR

Avec cette variante, le traitement global de la collection est réalisé par une classification NCLR. Cette architecture est proche du système *hybride* proposée par (Tran *et al.*, 2011), la seule différence notable étant la présence d'une classification NCLR au niveau du traitement individuel des émissions.

SRL de collection par PLNE

Les i-vecteurs, utilisés principalement dans le domaine de la vérification du locuteur (Dehak *et al.*, 2011), 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 (Shum *et al.*, 2011).

Ici, le nombre de locuteurs est inconnu. Un i-vecteur j est extrait à partir de chacune des classes j issues de la classification BIC en utilisant un UBM-GMM à 1024 composantes et 19 paramètres MFCC complétés de l'énergie, avec leurs dérivées première et seconde. Les N i-vecteurs résultants sont ensuite normalisés dans un processus itératif (Bousquet *et al.*, 2011). Le problème de classification consiste, d'une part, à minimiser le nombre K de classes centrales choisies parmi les N i-vecteurs et, d'autre part, à minimiser 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

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

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

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

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 (Bousquet *et al.*, 2011). 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).

Des expériences préliminaires ont montré que la résolution d'un problème PLNE offre une meilleure classification qu'un regroupement agglomératif hiérarchique, quelque soit le critère de liaison utilisé. Cette méthode de classification par PLNE a d'abord été adaptée à la SLR d'émissions dans un travail parallèle (Rouvier et Meignier, 2012).

4 Expériences

4.1 Données

Les données choisies pour réaliser nos expériences constituent un sous-ensemble du corpus d'apprentissage de la campagne d'évaluation ESTER 2. Les données sélectionnées correspondent aux enregistrements de deux émissions de *Radio France International* (RFI) sur trois semaines

du mois d'octobre 2000. Il y a deux enregistrements différents pour chaque jour ouvré, l'un sur la plage horaire 9h30 - 10h30, l'autre sur la plage horaire 11h30 - 12h30. Nous avons choisi de constituer deux corpus indépendants en fonction de l'heure à laquelle les émissions ont été enregistrées :

1. Le corpus n° 1 est constitué des 15 émissions correspondant à la plage horaire 9h30 - 10h30. Il totalise 358 locuteurs parmi lesquels 203 sont formellement identifiés par leur nom et prénom. Parmi ces 203 locuteurs, 47 apparaissent dans au moins deux émissions.
2. Le corpus n° 2 est constitué des 15 enregistrements de la plage horaire 11h30 - 12h30. Il totalise 298 locuteurs parmi lesquels 142 sont formellement identifiés par leur nom et prénom. Parmi ces 142 locuteurs, 41 apparaissent dans au moins deux émissions.

Pour évaluer la tâche de SRL de collection, les locuteurs apparaissant dans plusieurs émissions doivent nécessairement être identifiés par la même étiquette dans toutes les émissions. Nous évaluerons uniquement les locuteurs formellement identifiés par leur nom et prénom. Les autres étiquettes (Christelle, speaker#151, ...) ne fournissent aucune garantie sur l'identité du locuteur : un même locuteur peut être identifié par des étiquettes différentes dans plusieurs émissions.

4.2 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. L'outil d'évaluation que nous avons utilisé est celui développé par le LNE [3] dans le cadre de la campagne REPERE [4]. Cet outil permet de distinguer deux différents taux d'erreur : d'une part, le DER d'émissions *(DER-emi)*, lorsque l'évaluation est réalisée en considérant les émissions indépendamment les unes des autres, et d'autre part, le DER de collection *(DER-col)*, lorsque l'évaluation est réalisée simultanément sur toutes les émissions de la collection. Le *DER-emi* correspond à la moyenne des DER mesurés sur chaque émission, pondérés par leurs durées. Le *DER-col* tient compte de la réapparition des locuteurs dans plusieurs émissions.

4.3 Configuration des systèmes de SRL de collection

Le modèle du monde (UBM) a été appris sur le corpus de test distribué lors de la campagne d'évaluation ESTER 1 (Galliano *et al.*, 2009). Les modèles de locuteur sont obtenus en effectuant une itération de l'algorithme MAP. Le corpus d'apprentissage utilisé durant l'étape de normalisation des i-vecteurs est également celui de ESTER 1. Le programme d'optimisation linéaire utilisé pour résoudre le problème p-centre est le GNU Linear Programming Toolkit [5].

Le seuil optimal de classification NCLR pour le traitement individuel des émissions, réalisé par le système de SRL d'émissions décrit dans la section 2, est de 0,97. Ce seuil a été fixé à partir d'une évaluation individuelle des émissions du corpus n° 1 *(DER-emi)*. Le seuil optimal de classification NCLR et la distance optimale δ de regroupement des i-vecteurs (eq. 4), pour le traitement de la collection, sont respectivement de 0,82 et 120. Ces deux valeurs ont été déterminées à partir

3. Laboratoire National de métrologie et d'Essais
4. `http://www.defi-repere.fr/`
5. `http://www.gnu.org/software/glpk/`

d'une évaluation sur l'ensemble des émissions du corpus n° 1 *(DER-col)*. Ces trois seuils ont été appliqués tels quels sur le corpus n° 2.

4.4 Résultats et discussion

Nous présentons dans le tableau 1 les résultats obtenus en termes de *DER-emi* et *DER-col*, avec le système de SRL d'émissions, décrit dans la section 2, et les deux systèmes de SRL de collection, sur les corpus n° 1 et n° 2.

Nous avons évalué les références et les sorties du système de SRL d'émissions avec la mesure DER de collection *(DER-col)*. L'évaluation des références permet de mesurer la difficulté de la tâche. Dans ce cas, l'hypothèse évaluée correspond à la référence, dans laquelle les étiquettes des locuteurs ont été préalablement préfixées par le nom des émissions. Comme nous pouvions nous y attendre, les taux d'erreur *DER-col* sont très élevés : 53,14% pour l'évaluation des références sur le corpus n° 1 et 52,26% pour l'évaluation des sorties du SRL d'émission. Le corpus n° 2 donne des taux d'erreur similaires. Nous constatons que les taux d'erreur *DER-col* des références et des sorties du systèmes de SRL d'émissions sont relativement proches : sur le corpus n° 1, le taux d'erreur n'augmente que de 2,94% en absolu avec le système automatique.

Systèmes	Corpus n° 1		Corpus n° 2	
	DER-emi	*DER-col*	*DER-emi*	*DER-col*
Référence	0,00%	53,14%	0,00%	52,26%
SRL d'émissions	9,65%	56,08%	13,37%	54,30%
SRL de collection - NCLR	8,91%	14,91%	12,29%	19,97%
SRL de collection - PLNE	8,50%	15,06%	12,58%	21,52%

TABLE 1 – Résultats d'évaluation en termes de *DER-emi* et *DER-col* sur les corpus n° 1 et n° 2, avec l'évaluation des références, le système de SRL d'émissions et les deux systèmes de SRL de collection.

Les deux variantes du système de collection proposé obtiennent des taux d'erreur au niveau de la collection *(DER-col)* d'environ 15% pour le corpus n° 1 et environ 21% pour le corpus n° 2.

– Le système de SRL de collection par NCLR obtient des *DER-col* de 14,91% sur le corpus n° 1 et 19,97% sur le corpus n° 2. De plus, l'influence du traitement global de la collection sur les *DER-emi* est positive, avec un gain absolu de 0,74% sur le corpus n° 1 et 1,08% sur le corpus n° 2.

– Le système de SRL de collection par PLNE donne des résultats légèrement inférieurs à ceux obtenus par NCLR, mais très proches (15,06% pour le corpus n° 1 et 21,52% pour le corpus n° 2). La différence de 0,15% entre les deux systèmes représente environ une minute de signal sur les 10h évaluées du corpus n° 1. De la même manière qu'avec la méthode NCLR, on observe un faible gain au niveau de l'évaluation par émissions *(DER-emi)* : 1,15% en absolu pour le corpus n° 1 et 0,79% en absolu pour le corpus n° 2.

On peut supposer que les faibles gains observés au niveau de l'évaluation par émissions *(DER-emi)* sont dus au fait que les systèmes de collection disposent de plus de données pour apprendre les modèles de locuteur, les rendant plus discriminants.

Si les performances des systèmes de SRL de collection sont similaires en termes de *DER*, ce n'est pas le cas en termes de temps de calcul. Les deux méthodes peuvent être décomposées en plusieurs étapes, réalisées soit au niveau des émissions, soit au niveau de la collection. Les traitements réalisés pour chaque émission peuvent être mémorisés pour une utilisation ultérieure. Les traitements réalisés sur l'ensemble de la collection sont à réitérer si de nouveaux documents sont ajoutés à la collection.

Nous avons mesuré le temps de calcul nécessaire à l'étape de classification des deux méthodes testées, sur les deux corpus. La durée du calcul des modèles de locuteurs n'a pas été prise en compte, cette étape étant facilement parallélisable. La classification par PLNE a été réalisée en 03:16 heures sur les données du corpus n° 1, contre 39:28 heures pour la variante NCLR. Sur le corpus n° 2, les durées mesurées sont respectivement de 04:35 pour la classification par PLNE et 81:21 heures pour la variante NCLR. En moyenne sur ces deux corpus de 15 heures, la classification par PLNE est 17,67 fois plus rapide que la classification par NCLR.

5 Conclusions

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 PLNE sur la distance entre les i-vecteurs. Les performances du système implémentant cette approche sont comparables, en termes de *DER*, à celles du système implémentant la classification globale par NCLR. Néanmoins, le regroupement par PLNE est plus efficace que le regroupement par NCLR en termes de rapidité, tout en restant raisonnable au niveau de la quantité de mémoire consommée. Cette méthode est particulièrement appropriée pour le traitement de collections volumineuses.

Références

Bousquet, P.-M., Matrouf, D. et Bonastre, J.-F. (2011). Intersession compensation and scoring methods in the i-vectors space for speaker recognition. *In Proceedings of Interspeech'11*, Florence, Italie.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P. et Ouellet, P. (2011). Front-end factor analysis for speaker verification. *In Proceedings of IEEE TASLP*, volume 19, pages 788–798.

Galliano, S., Gravier, G. et Chaubard, L. (2009). The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. *In Proceedings of Interspeech*, Brighton, UK.

Le, V. B., Mella, O. et Fohr, D. (2007). Speaker diarization using normalized cross-likelihood ratio. *In Proceedings of Interspeech*, Antwerp, Belgique.

Meignier, S. et Merlin, T. (2009). LIUM SpkDiarization: an open-source toolkit for diarization. *In CMU SPUD Workshop*, Dallas, Texas (USA).

Rouvier, M. et Meignier, S. (2012). Nouvelle approche pour le regroupement des locuteurs dans des émissions radiophoniques et télévisuelles. *In 29e Journées d'Études sur la Parole*, Grenoble, France.

Shum, S., Dehak, N., Chuangsuwanich, E., Reynolds, D. et Glass, J. (2011). Exploiting intra-conversation variability for speaker diarization. *In Proceedings of Interspeech*, Florence, Italie.

Tran, V.-A., Le, V. B., Barras, C. et Lamel, L. (2011). Comparing multi-stage approaches for cross-show speaker diarization. *In Proceedings of Interspeech*, Florence, Italie.

Yang, Q., Jin, Q. et Schultz, T. (2011). Investigation of cross-show speaker diarization. *In Proceedings of Interspeech*, Florence, Italie.