Appariement de locuteurs entre documents sonores préalablement segmentés en utilisant la classification hiérarchique

Sylvain Meignier, Jean-François Bonastre, Ivan Magrin-Chagnolleau

▶ **To cite this version:**

Sylvain Meignier, Jean-François Bonastre, Ivan Magrin-Chagnolleau. Appariement de locuteurs entre documents sonores préalablement segmentés en utilisant la classification hiérarchique. JEP 2002, 2002, Nancy, France. pp.5. hal-01434575

HAL Id: hal-01434575
https://hal.science/hal-01434575

Submitted on 22 Mar 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

APPARIEMENT DE LOCUTEURS ENTRE DES DOCUMENTS SONORES PRÉALABLEMENT SEGMENTÉS EN UTILISANT LA CLASSIFICATION HIÉRARCHIQUE

Sylvain Meignier[*(1)], *Jean-François Bonastre*[(1)], *Ivan Magrin-Chagnolleau*[(1)(2)]

[(1)] LIA / CERI

Université d'Avignon - Agroparc - BP 1228 - 84911 Avignon Cedex 9 - France

[(2)] Laboratoire Dynamique Du Langage

Université Lumière Lyon 2 & CNRS UMR 5596 - 14, avenue Berthelot - 69363 Lyon Cedex 07 - France

{sylvain.meignier, jean-francois.bonastre}@lia.univ-avignon.fr - ivan@ieee.org

RÉSUMÉ

L'indexation par locuteurs d'une collection de document sonore consiste à organiser ces données sonores en fonction des locuteurs présents dans la base de données. Cette indexation se fait selon trois étapes : (1) la segmentation par locuteurs de chaque document sonore de la base ; (2) l'appariement de locuteurs entre les diverses portions segmentées des documents ; (3) la génération d'un index basé sur les locuteurs. Ce papier se focalise sur la deuxième étape, c'est-à-dire l'appariement de locuteurs, qui n'a été que très peu abordée jusqu'à maintenant. Le résultat de cette tâche est une classification des différentes portions segmentées en classes correspondant chacune à des locuteurs différents. Dans cet article, nous étudions l'intérêt d'approches de type classification hiérarchique pour l'appariement de locuteurs. Nous proposons deux nouvelles mesures de dissimilarité discriminantes et un nouvel algorithme "bottom-up" que nous comparons avec des approches plus classiques en classification hiérarchique. Les expériences sont réalisées sur un sous-ensemble de la base de données Switchboard, une base contenant des conversations téléphoniques. Les approches proposées permettent un appariement de locuteurs satisfaisant avec un bon niveau de purité pour chacune des classes, mais le nombre de classes tend à être supérieur au nombre réel de locuteurs.

1. INTRODUCTION

L'indexation en locuteur d'une collection de documents sonore est le processus aboutissant à la création d'un index identifant les locuteurs de la collection et leurs interventions respectives dans chacun des documents. Ce processus se décompose en trois tâches (Figure 1). La première tâche est l'indexation de chaque document. L'index obtenu définit les locuteurs du document et chaque locuteur référence leur intreventions. La seconde tâche consite à identifier les locuteurs intervenant dans plusieurs documents en utilisant les informations contenues dans les indexes produits à la première étapes. Cette tâche revient à construire un index d'indexes. L'index produit sera constitué d'un identifiant de locuteur qui référence les documents où ce locuteur parle. La dernière tâche correspond à la construction d'un index de la collection adapté à une exploitation dans un système de recherche documentaire.

* RAVOL project: financial support from Conseil général de la région Provence Alpes Côte d'Azur and DigiFrance.

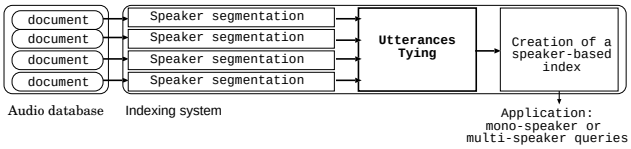

Fig. 1. Block-diagram of a speaker indexing system

Le problème de l'indexation en locuteur d'un document sonore (Figure 2) est généralement abordé par l'utilisation d'une des méthodes suivantes. La première (décrite dans [1][2][3]) applique une détection de rupture (soit par les locuteurs, soit par parole/silence). Puis, les segments définis entre les ruptures sont classifiés par locuteurs. La seconde méthode (voir [4][5]) effectue les phases de détection de rupture et de classification en simultanées en modèlisant la conversation par un modèle de Markov. Quelque soit la méthode, aucune information *a priori* sur les locuteurs n'est disponible. Ni le nombre de locuteurs, ni des données d'apprentissage spécifique aux locuteurs à détecter. La clé de l'index est un couple de valeurs composé du nom de document et du libellé de locuteurs. Chaque valeur de la clé référence les debut et fin des segments du locuteurs.

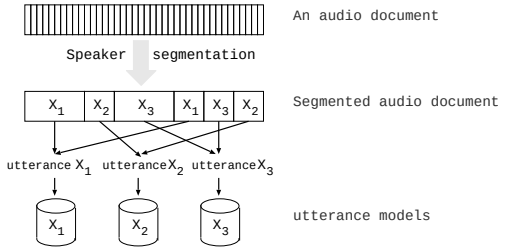

Fig. 2. Speaker segmentation and generation of speaker utterance models

L'indexation d'une collection de document sonore est un problème de classification proche de la classification des segments [2][3]. La classification des segments est appliqué à un document donné, alors que l'indexation d'une collection groupe les locuteurs intervenant dans plusieurs documents (Figure 3). Nous parlerons d'appariment de locuteurs. Il est a noter que les locuteurs intervenant dans le même document ne peuvent pas être regrouper sans remettre en cause l'indexation d'un document. Par ailleur, pour cette tâche, nous diposons des interventions de chacun des locuteurs, ce qui représente plus de données que pour l'indexation d'un documents ou la classification est faite sur les segments. La varia-

bilité du cannal de transmission pour un même locuteur est une difficulté suplementaire pour l'appariment des locuteurs, alors que la permière tâche tire partie de cette différence de canal. La clé de l'index prend comme valeur les libellés de locuteurs et référence des couples composé d'un nom de document et de ces interventions.

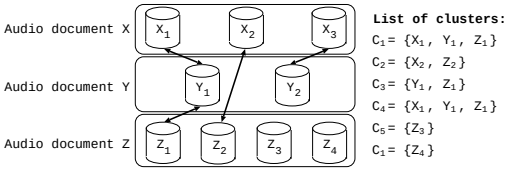

Fig. 3. Example of speaker utterances tying

La dernière étape du processu adapte l'index de la collection à une tâche visée et à un système de recherche documentaire. Il y a peu de papier sur le sujet. Il existe au moins deux possibilités d'applications. Soit un système de recherche basé sur l'exemple. L'utilisateur demande tous les documents ou toutes les parties de documents les plus similaire à un enregistrement sonore du locuteur recherché. Soit un système de recherche basé sur l'identité des locuteurs. Les mots-clé recherché seront alors les noms de locuteur. Ce type de recherche implique de disposer de donnée pour identifier les locuteurs de la collection.

Dans ce papier, nous nous interessons particulièrement à l'indexation de collection de documents sonore. Nous supposerons que l'index de chaque document est correcte et précise. Nous utiliserons donc les indexes de références utilisé lors de l'évaluation de ces documents. Nous disposons pour chaque ensemble d'interventions d'un modèle. L'indexation de collection consiste à grouper les couples (nom de document, interventions d'un locuteur) en classe identifé par un libellé de locuteur. Le nombre de locuteurs présent dans la collection est inconnu. Cependant, dans un cadre applicatif, le nombre de locuteur est important. Cela implique que le processus ne pourra pas recalculer l'ensemble des modèles à chaque étapes de la classification, ou à chaque ajout de nouveau document à la collection. La technique retenu doit prendre en compte cette dernière remarque.

Les methodes pour les tâches d'indexation en locuteur d'un document sonore (décritent en particulier dans [1][2][3]) sont applicable à l'indexation de collection. La classification hérachique est la principale méthode proposée dans la littérature. C'est une méthode itérative et aglomérative. La classification initiale est composée d'un ensemble de classe, où chaque classe contient un seul object. A chaque étape, l'algorithme goupe les deux classes les plus proche, selon une mesure de dissimilarité fixée. Le résultat d'une classification est généralement présenté sous la forme d'un dendrogramme qui illustre les associations sucessive (Figure 4).

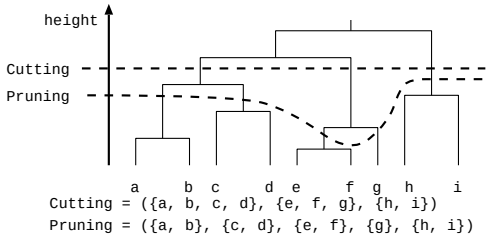

Fig. 4. Example of dendrogram : cluster selection methods

Dans ce papier, nous proposons une approche utilisant une classification hérachique adaptée aux contraintes de l'indexation de collection. Cette technique est appliqué à une base de données composé de conversations télephoniques (un sous ensemble issu de Switchboard II)

2. ALGORITHME D'APPARIEMENT DE LOCUTEURS

Une methode de classification hierarchique est definie par : • une mesure entre les classes ; • une regle d'aglomeration des classes. Pour une nouvelle classe, il est necessaire d'évaluer les mesures entre cette classe est les autres ; • une methode d'élagage du dendrogramme pour selectioner l'ensemble finale de classes.

2.1. Mesures de dissimilarité

Une mesure de dissimilarité exprime la proximité entre deux classes contenant des interventions de locuteurs. Soit $d(u, v)$ une mesure de dissimilarité entre les classes u et v. d est une mesure symetrique : $d(u, v) = d(v, u)$.

Un certain nombre de mesures de dissimilarité ont été proposé dans la littérature :

• des mesures qui necessite d'apprendre des modèles sur chaque classe comme le rapport de vraissamblance généralisé (generalized likelihood ratio [2]) ou comme le citere d'information bayésien (Bayesian information criterion, BIC [1]). Ces deux méthodes ont des coups de calcul important, car a chaque aglomeration des modèles doivent être réévluer. Mais ces methodes généralement donnent les meuilleurs résultats.

• des mesures qui ne necessite pas la réevaluation des modèles Methods which do not require models to be trained at each utterances grouping as the cross likelihood ratio (d_{clr}) [3] or the symmetric Kullbach-Leiber distance [2].

Notation

Soit X_i l'ensemble des interventions du locuteur i dans le document X.

Soit $S_X = \{X_1, \ldots, X_i, \ldots, X_n\}$ l'ensemble des X_i dans le document X.

Soit $\lambda(X_i)$ le modèle correspondant au données X_i.

Soit $\overline{X_i} = S_X - \{X_i\}$ les interventions ne corespondant pas au locuteur i dans le document X.

Soit $l(v|\lambda(u))$ la vraissemblance des données v suivant le modèle $\lambda(u)$. $l(v|\lambda(u))$ est normalisé par le nombre de trames contenu dans les données v.

Soit $l(v|\lambda(\overline{X_i})) = \max_{x \in \overline{X_i}} l(v|\lambda(x))$.

Soit $l(\overline{X_i}|\lambda(u)) = \max_{x \in \overline{X_i}} l(x|\lambda(u))$.

Soit $\lambda(W)$ un modèle du monde.

Soit $r(v|\lambda(u))$ le rapport de vraissamblance entre $l(v|\lambda(u))$ et $l(v|\lambda(W))$.

Mesures de dissimilarité classiques

Le rapport de vraissemblance croisé [3] exprimé en terme de dissimilarité est :

$$d_{clr}(X_i, Y_j) = \frac{l(Y_j|\lambda(W))}{l(Y_j|\lambda(X_i))} \cdot \frac{l(X_i|\lambda(W)}{l(X_i|\lambda(Y_j))}$$

Cette mesure n'est pas discriminante entre les modèles X_i et Y_j par rapport aux autres modèles des documents ($\overline{X_i}$ et $\overline{Y_j}$).

Nouvelle mesure proposée

Nous proposons deux mesures qui utilise explicitement les informations donnée par le segmentation des documents X et Y (itreventions ou modèles). Si les interventions X_i et Y_j sont produite par le même locuteur, alors le locuteur i ne produit pas $\overline{Y_j}$ et le locuteur j ne produit pas $\overline{X_i}$.

La première dissimilarité proposée utilise les données des autres intreventions des documents X et Y :

$$d_1(X_i, Y_j) = \frac{f(\overline{Y_j}|\lambda(X_i)) + f(\overline{X_i}|\lambda(Y_j))}{f(Y_j|\lambda(X_i)) \cdot f(X_i|\lambda(Y_j))}$$

f repésente soit une vraissamblance (l) soit un rapport de vraissemblance (r).

La seconde utilise les modèles des autres interventions (c'est à dire $\lambda(\overline{Y_j})$ and $\lambda(\overline{X_i})$) presentes dans les documents X et Y :

$$d_2(X_i, Y_j) = \frac{f(Y_j|\lambda(\overline{X_i})) + f(X_i|\lambda(\overline{Y_j}))}{f(Y_j|\lambda(X_i)) \cdot f(X_i|\lambda(Y_j))}$$

De même, f représente une vraissemblance (l) ou un rapport de vaissamblance (r).

Matrice de dissimilarité

Une matrice de dissimilarité est produite à partir d'une de ces messure. Cette matrice est composé des mesure entre toutes les paire d'interventions.

Par hypothèse nous supposons que la segmentation a été correctement réalisée, alors les intervention du même document ne sont pas grouper. Par consequent, $d(X_i, X_j) = +\infty$ pour tous i, j d'un document X donnée.

2.2. Méthodes agglomératives

Agglomerative methods are abundantly documented in the literature [2][6][7]. We just remind the formulae of the two algorithms that we used in our experiments.

Let $P_n = \{C_1, \ldots, C_i, \ldots, C_j, \ldots, C_n\}$ be the partition composed of n clusters.

Let c_i^a be an element of the cluster C_i.

Let c_j^b be an element of the cluster C_j.

Dans la methode "single link", la dissimilarité entre deux classes est le minimum de dissimilarité entre tous les paires d'element prises dans les deux classe. la formule est : $d(C_i, C_j) = \min_{a,b} d(c_i^a, c_j^b)$.

Dans la methode "complete link", la dissimilarité entre deux classe est le maximum parmi toutes les paires de dissimilarité : $d(C_i, C_j) = \max_{a,b} d(c_i^a, c_j^b)$.

2.3. Élagage du dendogramme

A la fin de l'algorithme de classification herachique, un dendrogramme est construit dans lequel chaque noeud corresponds a une classe. L'élagage du dendrogramme produit une partition composée de toutes les interventions. Plusieurs technique exist dans la literature [2][7] pour selectionner la partition. Ces techniques consistent à couper le dendrogramme à une hauteur donnée ou a sélectionner un ensemble de classe a différente hauteur (see Figure 4).

Méthode d'élagage classique

Dans nos expérience, nous construions la partitions par séléction de classe à différente hauteur. La methode appellé $Best$ décrite dans [3], est basé sur l'estimation de la purté des classes[1] $\hat{p}_i$ de la classe i (voir [2][3]). Le score $score_i = \hat{p}_i - \frac{Q}{n_i}$ est calculer pour chaque noeud i correspondant à la classe i composée de n_i éléments. le meilleurs $score_i$ est séléctionné. Les descendants et les parents sont supprimés du dendrogramme. La classe i est retenu pour la partition finale. L'algorithme continu tant qu'il reste des noeuds dans le dendrogramme.

Nouvelle m'éthode proposée

[1] appellé "Nearest Neighbor Purity Estimator"

Nous proposons une nouvelle methode appellé Asc, elle est une adaptation de la methode précédante. Les $score_i$ sont calculer pour chaque noeud. Le dendrogramme est parcouru des feuilles jusqu'à la racine suivant l'ordre d'agregation des classes donné par la classification herachique. Quand le $score_i$ n'augmente plus entre le noeud i et ces fils, les deux fils sont ajouter à la partition. le noeud i et ces ancêtre sont supprimer du dendrogramme. L'algorithme s'arrete quand il n'y a plus de feuille dans le dendrogramme.

Nous remarquerons que la première methode ($Best$) encourage les classes composées de plus d'éléments que la seconde méthode (Asc).

3. EXPÉRIENCES ET RÉSULTATS

3.1. Base de données

L'approche proposé a été expérimentée sur un sous-ensemble des données "2-speaker" utilisées durant la campagne d'évaluation NIST 2001 [8]. La segmentation de référence est disponible pour chaque test. Ce sous-ensemble est composé de 408 conversation téléphonique extratite du corpus Switchboard II. Le nombre de locuteurs est de 319 (132 hommes et 187 femmed). Chaque locuteur est présent dans 1 à 4 tests (voire Table 1). Chaque locuteur intervient en moyenne 31 seconde (Min =∼14 s, Max =∼53 s.). La durée totale des tests est proche de 422 minutes.

Loc. apparaissant dans	1 test	2 tests	3 tests	4 tests
Nombre de loc.	72	90	64	93

Table 1. Nombre de locuteurs apparaissant dans 1 à 4 tests.

Pour information, dans les évaluation NIST 2001, le meuilleur système de segmentation a obtenu respectivement un taux d'erreur ∼10% and ∼20% pour la tâche *2-speakers* et la task *n-speakers*. Dans nos expérience, nous utiliserons la segmentation de référence *2-speakers* donné donnée par NIST.

Un modèle du monde entrainé sur un coprus différnt des donnée de tests est utilisé pour l'adaptation des modèles et pour le calcule des rapport de vraissemblance. Ce corpus est composé de 472 tests Switchboard II prononcé par 100 locuteurs (homme et femme).

3.2. Système de reconnaissance automatique du locuteur

La paramétrisation accoustique (16 coefficients cepstrals et 16 Δ coefficients) est calculée par le module SPRO developpé par le consortium ELISA [9]. Les modèles et les vraissemblance sont calculé par le système de reconnaissance automatique du locuteur AMIRAL developed au LIA [10]. Les locuteurs sont modèlisé par un modèle de mixture de gaussienne à 128 composantes et à matrices de covariances digonale [11] adapté depuis un modèle du monde. L'adapatation est faite par la méthode du *maximum a posteriori* (MAP).

3.3. Expériences en vérification du locuteur

Évaluation en vérification du locuteur

La première évaluation proposé mesure la précision des dissimilarités proposé. Des tests de vérification du locuteur (*1-Speaker*) sont calculer entre les différentes interventions. Le score de similarité, utilisé comme score de vérification, s'exprime comme l'opposé de la mesure de dissimilarité. Ce score est calculé entre chaque paire d'interventions de locuteur : $s(u, v) = -d(u, v)$. Ce score est claculé pour chaque dissimilarité (d_{clr}, d_1 avec $f = l$ ou $f = r$, d_2 avec $f = l$ ou $f = r$). Le score est claculé uniquement pour les interventions u et v provenant de differents documents. C'est à dire pour $d(u, v) \neq +\infty$.

Résultats et discussion

Les resultats des 332112 tests *1-Speaker* sont reporté dans les courbe DET 5.

Comme pour les résultats classiquement observé pour les tâche de verification du locuteurs, les résultats des similarité normalisé par le modèles du monde donnent un taux d'erreur plus faible.

Tous les taux d'erreur de mesures obtiennent des taux d'erreurs supérieurs aux tâches de vérification obtenus lors évaluation NIST 2001 ($EER \sim$10%). Les deux différences majeurs existent entre les résultats donnée et les résultats des évaluations NIST. Premièrement, dans nos expériences, la durée d'apprentisage des modèles est en moyenne de 31 secondes, alors que pour les campagnes NIST 2 minutes de parole sont disponible pour chaque modèle. Deuxièmement, lors des évaluation NIST, un système de normalisation des scores H-norm ou HT-norm est appliqué.

les resultats s_1 (avec $f = r$) donne de meuilleur résultats que s_2 (avec $f = r$). Utilisé les données des autres locuteurs présant dans le deux documents augmente les performances.

Finallement, s_1 (avec $f = r$) surpasse s_{clr} à l'EER.

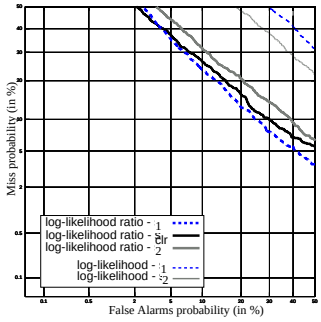

Fig. 5. Courbes DET pour l'expérience en vérifi cation du locuteur utilisant les mesures de similarité.

3.4. Expériences d'appariement de locuteurs

Évaluation

L'évaluation de l'appariment de locuteur est faite sur la partition obtenu après l'étape d'élagage par le calcule de deux taux d'erreur proposé dans [1].

Soit N_c le nombre de classes dans la partition P.

Soit n_i le nombre d'interventions dans la classe i.

Soit IN_i le nombre d'interventions du locuteur principale[2] de la classe i.

Soit OUT_i le nombre d'interventions en dehors de la classe i du locuteur principale de la classe i.

Le taux d'erreur de type I s'exprime par :

$$e_I = \frac{1}{N_c} \sum_{i \in P} \frac{n_i - IN_i}{n_i}$$

Le taux d'erreur de type II s'exprime par :

$$e_{II} = \frac{1}{N_c} \sum_{i \in P} \frac{OUT_i}{IN_i + OUT_i}$$

Les taux d'erreurs de type I et II sont sommés : $e = e_I + e_{II}$

Résultats et discussion

On notera que les valeurs de e et de N_c sont liées à la valeur du paramètre Q utilisé dans l'élagage.

[2]l'indentité du locuteur qui minimise r.

Le tableau 2 présente les résultas des deux methodes d'aglomérations. Entre les deux méthodes, le score e est proche, mais le monbre de classe N_c est plus proche du monbre de classes réelr (319) avec la methode "complete link".

	Complete Linkage				Single Linkage			
	Asc		Best		Asc		Best	
	e	N_c	e	N_c	e	N_c	e	N_c
0.0	63.7	596	68.6	519	64.3	643	65.1	636
0.5	67.5	537	72.8	449	70.0	598	70.6	588
1.0	70.0	505	83.5	350	77.0	544	84.5	484

Table 2. Complete v.s. Single linkage ; pour la mesure d_{clr} avec $Q \in \{0, 0.5, 1\}$; $e = e_I + e_{II}$ est en %.

Le tableau 3 présente les résultas des différentes mesures de dissimikarité. La mesure "cross likelihood ratio" d_{clr} est la dissimilarité qui produit le meuilleur score e et qui produit le monbre de classe le plus proche du nombre réelle. Le nombre de classe ($N_c = 596$) est important comparé au nombre réel (319) mais les erreurs de type I e_I est faible ($\sim 7\%$). Prendre une valeur du paramètre Q plus grande produit moins de classe (voir tableau 2). Bien que la mesure d_1 donne de meilleur resultats que la mesure d_{clr} pour les expérience *1Speaker*, les résultats de ma mesures d_1 sont moins bon que ceux de d_{clr} pour les expérience d'appariement des interventions.

Les résultats sur les methode d'évagage donnent pour la méthode Asc de plus petite classe que pour la methode $Best$. Cependant, le score e est très proche pour un nombre donnée de classe quelque soit la methode d'élagage. Le choix de la methode d'élagage dependra donc du type d'application visé.

	Élagage Asc				Élagage Best			
	e_I	e_{II}	e	N_c	e_I	e_{II}	e	N_c
d_2 llk	1.4	66.5	67.9	790	27.7	63.0	90.7	540
d_1 llk	0.6	67.2	67.8	808	28.1	63.5	91.6	546
d_2 llr	6.6	63.8	70.4	708	21.4	60.3	81.7	563
d_1 llr	5.5	59.7	65.2	646	14.6	54.6	69.2	529
d_{clr}	7.4	56.3	63.7	596	15.1	53.5	68.6	519

Table 3. Résultats des différentes mesures de dissimilarité pour le complete linkage. $Q = 0$ et e, e_I, e_{II} en %. *llk = log-vraisemblance, llr = log de rapport de vraisemblance*.

4. CONCLUSION ET PERSPECTIVES

Dans cet article, nous avons évalué l'intérêt de l'utilisation de la classification hiérarchique en appariement de locuteurs, dans le cadre de l'indexation par locuteurs. Nous avons utilisé les mesures de dissimilarité classiques en classification hiérarchique. Nous avons également proposé deux nouvelles mesures de dissimilarité dans le but de prendre en compte toute l'information disponible dans un document sonore préalablement segmenté par locuteurs. Nous avons aussi présenté un nouvel algorithme de classification hiérarchique bottom-up. Les performances obtenues sur une base de données composée de conversations téléphoniques (Switchboard) sont satisfaisantes, et les classes obtenues ont une très bonne purité mais avec un plus grand nombre de classes qu'il n'y a de locuteurs en réalité.

Ce travail étant l'un des premiers sur le thème de l'appariement de locuteurs entre documents segmentés, il reste bien entendu

de nombreuses voies d'amélioration. En particulier, nous nous focaliserons sur la normalisation des mesures de dissimilarité discriminantes. Nous souhaitons également associer un système réel de segmentation par locuteurs avec un système d'appariement de locuteurs afin d'évaluer l'importance des erreurs de segmentation sur les performances globales d'un tel système. La génération d'un index basé sur les locuteurs sera la dernière étape d'un système d'indexation par locuteurs, mais il reste encore plusieurs problèmes à résoudre avant d'en arriver là, comme la gestion d'un grand volume de données ainsi que le format de représentation des index afin de permettre une recherche rapide et un accès efficace aux données indexées.

5. RÉFÉRENCES

[1] S. Chen, J.F. Gales, P. Gopalakrishnan, R. Gopinath, H. Printz, D. Kanevsky, P. Olsen, and L. Polymenakos, "IBM's LVCSR system for transcription of broadcast news in the 1997 HUB4 english evaluation," in *DARPA speech recognition workshop*, 1998, http ://www.nist.gov/speech/publications/darpa98/html/bn20-/bn20.htm.

[2] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, "Clustering speakers by their voices," in *Proceedings of ICASSP 98*.

[3] D.A. Reynolds, E. Singer, B.A. Carlson, J.J. McLaughlin G.C. O'Leary, and M.A. Zissman, "Blind clustering of speech utterances based on speaker and language characteristics," in *Proceedings of ICSLP 98*.

[4] Meignier Sylvain, Jean-François Bonastre, and Stéphane Igounet, "E-HMM approch for learning and adapting sound models for speaker indexing," in *2001 : a Speaker Odyssey*, 2001, pp. 175–180.

[5] L. Wilcox, D. Kimber, and F. Chen, "Audio indexing using speaker identifi cation," *SPIE*, pp. 149–157, 1994.

[6] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering : A review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264 – 323, Sep. 1999.

[7] B.S. Everitt, *cluster analysis*, Oxford University Press Inc., New York, third edition, 1993.

[8] NIST, "The NIST 2001 speaker recognition eveluation plan," http ://www.nist.gov/speech/tests/spk/2001/doc/2001-spkrec-evalplan-v53.pdf, Mar. 2001.

[9] Ivan Magrin-Chagnoleau, Guillaume Gravier, and Raphaël Blouet for the ELISA consortium, "Overview of the ELISA consortium research activities," in *2001 : a Speaker Odyssey*, Jun. 2001, pp. 67–72.

[10] C. Fredouille, J.-F. Bonastre, and T. Merlin, "Amiral : a block-segmental multi-recognizer approach for automatic speaker recognition," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 172–197, Jan.-Apr. 2000.

[11] D. A. Reynolds, "Speaker identifi cation and verifi cation using gaussian mixture speaker models," *Speech Communication*, pp. 91–108, Aug. 1995.