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

Speaker indexing of an audio database consists in organizing the audio data according to the speakers present in the database. It is composed of three steps : (1) segmentation by speakers of each audio document ; (2) speaker tying among the various segmented portions of the audio documents ; and (3) generation of a speaker-based index. This paper focuses on the second step, the speaker tying task. The result of this task is a classification of the segmented acoustic data by clusters ; each cluster should represent one speaker. This paper investigates on hierarchical classification approaches for speaker tying, and proposes two discriminant dissimilarity measures using the information provided by the segmentation. The experiments are conducted on a subset of the Switchboard database, a conversational telephone database, and show that the proposed method allows a satisfying speaker tying among various audio documents.

1. INTRODUCTION

L'indexation automatique en locuteur d'une collection de documents sonores est le processus aboutissant à la création d'un index identifiant 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 consiste à indexer chaque document indépendamment. L'index obtenu définit les locuteurs du document et référence leur interventions. La seconde tâche consiste à identifier les locuteurs apparaissant dans plusieurs documents en utilisant les informations contenues dans les index produits à la première étape. Cette tâche revient à construire un "index d'index". La clé de l'index produit est un identifiant de locuteur. Cet identifiant référence les documents dans lesquels ce locuteur parle. La dernière tâche engendre un index de la collection adapté à une exploitation dans un système de recherche documentaire.

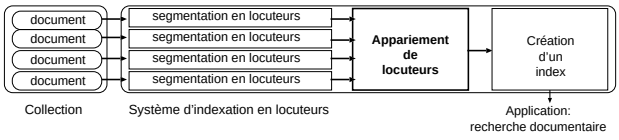

FIG. 1: Diagramme d'un système d'indexation en locuteurs

Le problème de l'indexation en locuteurs d'un document sonore (Figure 2) est généralement abordé par une des méthodes suivantes. La première (décrite dans [1][2][3]) applique une détection de ruptures (soit sur les locuteurs, soit sur la parole et le silence). Puis les segments définis entre les ruptures sont regroupés par locuteurs. La seconde méthode (voir [4][5]) effectue les phases de détection de ruptures et de classification simultanément. Dans cette méthode, la conversation est modélisée 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 leur noms, ni des données d'apprentissage spécifiques aux locuteurs à détecter ne sont disponible. La clé de l'index est un couple de valeurs composé du nom de document et du libellé d'un locuteur [1]. Chaque valeur de la clé référence les débuts et fins des segments du locuteurs (ses interventions).

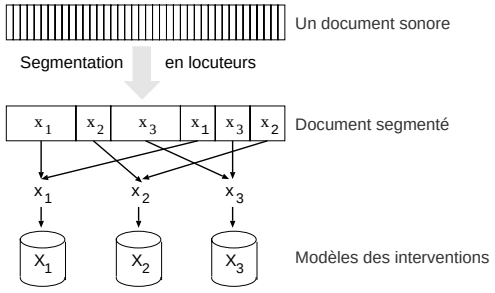

FIG. 2: Segmentation et modèles des interventions de locuteurs

L'indexation d'une collection de documents sonores est un problème de classification proche de la classification des segments [2][3]. La classification des segments est appliquée à un document donné, alors que l'indexation d'une collection groupe les locuteurs intervenant dans plusieurs documents (Figure 3). Nous parlerons "d'appariement de locuteurs". Il est à noter que les locuteurs intervenant dans le même document ne peuvent pas être regroupés sans remettre en cause l'indexation d'un document. La variabilité du canal de transmission pour un même locuteur est une difficulté pour l'appariement des locuteurs, alors que dans la première tâche les systèmes tirent partie de cette différence de canal. La clé de l'index met en relation les libellés de locuteur et l'ensemble des interventions du locuteur pour chacun des documents.

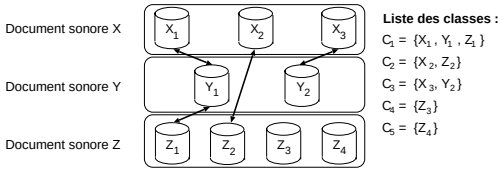

FIG. 3: Exemple d'appariement de locuteurs

La dernière étape du processus adapte l'index de la collection à une tâche visée. Dans le cadre des sytèmes de recherche documentaire, il existe au moins deux possibilités d'applications. Soit le système de recherche utilise des exemples comme clé de recherche. Le système propose alors tous les documents ou toutes les parties de documents les plus similaires à l'exemple sonore du locuteur recherché. Soit le système de recherche utilise l'identité des locuteurs. Le mot-clé recherché est alors le nom du locuteur. Ce type de recherche implique de disposer de données externes pour attribuer un nom aux locuteurs de la collection.

Dans ce papier, nous nous intéressons en particulier à l'indexation d'une collection de documents sonores. Nous

* Projet RAVOL : support financier du Conseil Générale de la région Provence Alpes Côte d'Azur et de DigiFrance. www.digifrance.fr

[1] Le libellé est généré en automatique par le système, il ne correspond pas au nom du locuteur.

supposons que l'index de chaque document existe et qu'il est correct et précis. Nous voulons nous affranchir des erreurs commises lors de la première tâche. En conséquence, nous utilisons les index de références utilisés pour le calcul de l'erreur de segmentation.

Les conditions de l'appariement de locuteurs sont proches des conditions de l'indexation en locuteurs d'un document sonore. Le nombre de locuteurs présents dans la collection est inconnu, mais il est supposé élevé. Les identités des locuteurs ne sont pas disponibles. Un modèle de voix est disponible pour chaque groupe d'interventions généré à la fin de la première tâche. Par contre, le processus d'appariement ne peut pas recalculer l'ensemble des modèles à chaque étape, ou à chaque ajout de nouveau document à la collection.

Les méthodes pour l'indexation en locuteur d'un document sonore (décrites en particulier dans [1][2][3]) sont applicables à l'indexation de collection. La classification hiérarchique est la principale méthode proposée dans la littérature. C'est une méthode itérative. La classification initiale est composée d'un ensemble de classe, où chaque classe contient un seul objet (ici les interventions extraites d'un document). A chaque étape, l'algorithme groupe les deux classes les plus proches, selon une mesure de dissimilarité. L'algorithme s'arrête quand toutes les classes ont été regroupées en une seule. Le résultat d'une classification est généralement présenté sous la forme d'un dendrogramme qui illustre les associations successives (Figure 4).

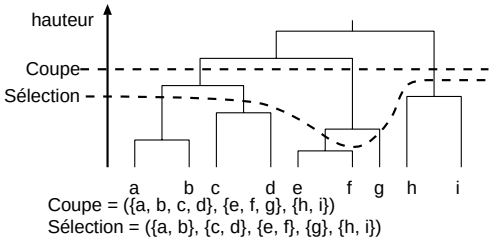

FIG. 4: Exemple de dendrogramme : méthodes d'élagage

Dans ce papier, nous proposons une approche utilisant une classification hiérarchique adaptée aux contraintes de l'indexation d'une collection. Cette technique est appliquée à une base de données composée de conversations téléphoniques (un sous-ensemble issu de Switchboard II)

2. APPARIEMENT DE LOCUTEURS

Une méthode de classification hiérarchique est définie par :
• une mesure entre les classes pour séléctionner les classes à fusionner ;
• une règle d'agglomération des classes. Après la fusion de deux classes, il est necessaire de réévaluer les mesures entre cette nouvelle classe et les autres ;
• une méthode d'élagage du dendrogramme pour sélectionner l'ensemble final de classes.

2.1. Mesures de dissimilarité

Une mesure de dissimilarité exprime la non-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 symétrique : $d(u, v) = d(v, u)$.

Différentes mesures de dissimilarité sont proposées dans la littérature :
• Des mesures qui nécessitent d'apprendre des modèles pour chaque dissimilarité comme le rapport de vraisemblance généralisé (generalized likelihood ratio, GLR [2]) ou comme le critère d'information bayésien (Bayesian information criterion, BIC [1]). Ces deux méthodes ont des coûts de calcul importants, mais elles mènent à des classifications de bonne qualité.
• Des mesures qui utilisent seulement les modèles des interventions comme le rapport de vraisemblance croisé (the cross likelihood ratio, d_{clr} [3]) ou la distance de KullbacK-Leibler symétrique (KL2, [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 X_i le modèle correspondant au données x_i.
• Soit $\overline{x_i} = S_x - \{x_i\}$ les interventions ne correspondant pas au locuteur i dans le document X.
• Soit $l(v|U)$ la vraisemblance des données v suivant le modèle U. $l(v|U)$ est normalisé par le nombre de trames contenues dans les données v.
• Soit $l(v|\overline{X_i}) = \max_{Z \in \overline{X_i}} l(v|Z)$.
• Soit $l(\overline{x_i}|U) = \max_{z \in \overline{x_i}} l(z|U)$.
• Soit W un modèle du monde.
• Soit $r(v|U)$ le rapport de vraisemblance entre $l(v|U)$ et $l(v|W)$.

Mesure de dissimilarité classique

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

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

Cette mesure n'utilise pas les informations apportées par la segmentation. Ni $\overline{x_i}$ et $\overline{y_j}$, ni $\overline{X_i}$ et $\overline{Y_j}$ ne sont pris en compte.

Nouvelles mesures proposées

Nous proposons deux mesures qui utilisent explicitement les données des index des documents X et Y. Si les interventions x_i et y_j sont produites par le même locuteur, alors le locuteur i ne produit pas $\overline{y_j}$ et le locuteur j ne produit pas $\overline{x_i}$. Nous supposons ici que $\overline{y_j}$ et $\overline{x_i}$ sont porteurs d'informations utilisables dans la classification.

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

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

f représente soit une vraisemblance (l) soit un rapport de vraisemblance (r).

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

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

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

Matrice de dissimilarité

Une matrice de dissimilarité est produite à partir de l'une de ces mesures. Cette matrice est composée des mesures entre toutes les paires d'interventions.

Les interventions d'un même locuteur sont regroupées lors de l'étape précédente. Le système ne remet pas en cause ce résultat et on a : $d(X_i, X_j) = +\infty$ pour tous i, j d'un document X donné.

2.2. Méthodes agglomératives

Les méthodes d'agglomération de classes sont abondantes dans la litérature [2][6][7]. Nous rappelons juste les formules des deux algorithmes retenus dans nos expériences.

• Soit $P_n = \{C_1, \ldots, C_i, \ldots, C_j, \ldots, C_n\}$ la partition composée de n classes.
• Soit c_i^a un élement de la classe C_i.
• Soit c_j^b un élément de la classe C_j.

Dans la méthode "single link", la dissimilarité entre deux classes est le minimum de dissimilarité entre toutes les paires d'éléments pris dans les deux classes : $d(C_i, C_j) = \min_{a,b} d(c_i^a, c_j^b)$.

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

2.3. *Élagage du dendogramme*

A la fin de l'algorithme de classification hiérarchique, un dendrogramme est construit dans lequel chaque noeud correspond à une classe. L'élagage du dendrogramme produit une partition composée de toutes les interventions. Plusieurs techniques existent dans la littérature [2][7] pour sélectionner la partition. Ces techniques consistent à couper le dendrogramme à une hauteur donnée ou à sélectionner un ensemble de classes à différentes hauteurs (voir Figure 4).

Méthode d'élagage classique

Dans nos expériences, nous constuisons la partition par sélection de classes à différentes hauteurs. La méthode appelée $Best$ décrite dans [3], est basée sur l'estimation de la pureté des classes[2] $\hat{p}_i$ de la classe i (voir [2][3]). Le $score_i = \hat{p}_i - \frac{Q}{n_i}$ est calculé pour chaque noeud i correspondant à la classe i composée de n_i éléments. Le $score_i$ le plus élevé est sélectionné. Les descendants et les parents sont supprimés du dendrogramme. La classe i est retenue pour la partition finale. L'algorithme continu tant qu'il reste des noeuds dans le dendrogramme.

Nouvelle méthode proposée

Nous proposons une nouvelle méthode, appelée Asc, issue de la méthode précédente. Les $score_i$ sont calculés pour chaque noeud. Le dendrogramme est parcouru des feuilles jusqu'à la racine suivant l'ordre d'agrégation des classes obtenu lors de la classification hiérarchique. Quand le $score_i$ n'augmente plus entre le noeud i et ses fils, ceux-ci sont ajoutés à la partition. Le noeud i et ses ancêtres sont supprimés du dendrogramme. L'algorithme s'arrête quand il n'y a plus de feuille dans le dendrogramme.

Nous remarquerons que la méthode $Best$ favorise plus la création de classes composées de nombreux éléments que la méthode Asc.

3. EXPÉRIENCES ET RÉSULTATS

3.1. *Base de données*

L'approche proposée a été expérimentée sur un sous-ensemble de conversations à 2 locuteurs ("*2-speakers*") utilisé durant la campagne d'évaluation NIST 2001 [8]. Les index de référence sont disponibles pour chacun de ces tests. Ce sous-ensemble est composé de 408 conversations téléphoniques extraites du corpus Switchboard II. Le nombre de locuteurs est de 319 (132 hommes et 187 femmes). Chaque locuteur apparaît dans 1 à 4 tests (voir Tableau 1). Chaque locuteur intervient en moyenne 31 secondes (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

TAB. 1: Nombre de locuteurs apparaissant dans 1 à 4 tests.

[2]appelée "Nearest Neighbor Purity Estimator"

Pour information, dans les évaluations NIST 2001, le meilleur système de segmentation a obtenu respectivement un taux d'erreur ∼10% et ∼20% pour la tâche "*2-speakers*" et pour la tâche "*n-speakers*".

Le modèle du monde W est entraîné sur un corpus indépendant des données de test. Ce corpus est composé de 472 tests Switchboard II prononcés par 100 locuteurs (homme et femme).

3.2. *Système de reconnaissance du locuteur*

La paramétrisation acoustique (16 coefficients cepstraux et 16 coefficients Δ) est calculée par le module SPRO développé par le consortium ELISA [9]. Les modèles et les vraisemblances sont calculés par le système de reconnaissance automatique du locuteur AMIRAL développé au LIA [10]. Les interventions des locuteurs sont modélisées par un modèle de mixture de gaussienne (GMM) à 128 composantes et à matrices de covariances diagonales [11]. Ces modèles sont adaptés depuis un modèle du monde par la méthode du *maximum a posteriori* (MAP).

3.3. *Expériences en vérification du locuteur*

Évaluation en vérification du locuteur

La première évaluation proposée mesure la précision des dissimilarités. Des tests de vérification du locuteur, proche des conditions de la tâche *1-Speaker* NIST, sont calculés 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 : $s(u, v) = -d(u, v)$. Les dissimilarités d_{clr}, d_1 et d_2 (avec $f = l$ ou $f = r$) sont calculées sur les interventions provenant uniquement de différents documents (pour chaque $d(u, v) \neq +\infty$).

Résultats et discussion

Les resultats des 332112 tests "*1-Speaker*" sont reportés dans les courbes DET (Figure 5).

Comme pour les résultats classiquement observés en vérification du locuteur, la normalisation des scores par le modèle du monde réduit les taux d'erreur.

Les mesures étudiées obtiennent des taux d'erreurs supérieurs aux résultats obtenus lors des évaluation NIST 2001 (EER ∼10%). Cette différence est en particulier due à la durée d'apprentissage des modèles qui est en moyenne de 31 secondes, alors que pour les campagnes NIST, 2 minutes de parole sont disponibles. Mais aussi lors des évaluations NIST, une normalisation des scores de type H-norm ou HT-norm est appliquée.

La mesure s_1 (avec $f = r$) obtient un meilleur résultat que la mesure s_2 (avec $f = r$). Utiliser les données des autres locuteurs présents dans les deux documents augmente donc les performances.

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

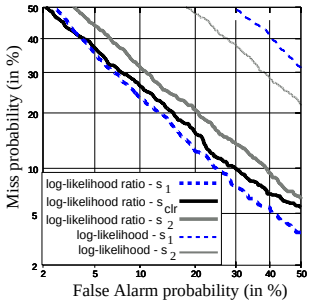

FIG. 5: Courbes DET pour l'expérience en vérification du locuteur utilisant les mesures de similarité.

3.4. *Expériences d'appariement de locuteurs*

Évaluation

L'évaluation de l'appariement de locuteur est réalisée à partir de la partition obtenue après l'élagage du dendrogramme. Le calcul de deux taux d'erreurs est 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 principal[3] de la classe i.
- Soit OUT_i le nombre d'interventions en dehors de la classe i du locuteur principal de la classe i.

Les taux d'erreurs de type I et II s'expriment par : $e_I = \frac{1}{N_c} \sum_{i \in P} \frac{n_i - IN_i}{n_i}$; $e_{II} = \frac{1}{N_c} \sum_{i \in P} \frac{OUT_i}{IN_i + OUT_i}$

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.

Le tableau 2 présente les résultats des deux méthodes d'agglomération. Pour ces deux méthodes, le score e est proche, mais le nombre de classe N_c est plus proche du nombre de classes réelles (319) avec la méthode "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

TAB. 2: Complete vs. 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 résultats des différentes mesures de dissimilarité. La mesure du rapport de vraisemblance croisé d_{clr} est la dissimilarité qui produit le meilleur score e et qui propose le nombre de classes le plus proche du nombre réel. Le nombre de classes ($N_c = 596$) est important comparé au nombre réel (319) mais le taux d'erreurs de type I e_I est faible ($\sim 7\%$). Prendre une valeur du paramètre Q plus grande produit moins de classes (voir tableau 2). Bien que la mesure d_1 donne de meilleurs résultats que la mesure d_{clr} pour les expériences "*1Speaker*", la dissimilarité d_1 est moins performante que la mesure d_{clr} pour les expériences d'appariement.

Les résultats sur les méthodes d'élagage donnent pour la méthode Asc de plus petites classes que pour la méthode $Best$. Cependant, le score e est très proche pour un nombre donné de classes quelque soit la méthode d'élagage.

	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

TAB. 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 du rapport de vraisemblance*.

[3] l'indentité du locuteur qui minimise e_I.

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 une mesure de dissimilarité usuelle 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 présenté un nouvel algorithme de classification hiérarchique bottom-up. Les performances obtenues sur une base de données composée de conversations téléphoniques (Switchboard) sont satisfaisantes.

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.

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, 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*, 1998.

[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*, 1998.

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

[5] L. Wilcox, D. Kimber, and F. Chen, "Audio indexing using speaker identification," *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," www.nist.gov/speech/tests/spk/2001/doc/2001-spkrec-evalplan-v53.pdf, Mar. 2001.

[9] Ivan Magrin-Chagnolleau, 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 identification and verification using gaussian mixture speaker models," *Speech Communication*, pp. 91–108, Aug. 1995.