

SEGMENTATION EN LOCUTEURS DE DOCUMENTS AUDIO: UNE NOUVELLE APPROCHE BASÉE SUR LES MÉTHODES A VECTEURS SUPPORT MONO CLASSE

Belkacem Fergani¹, Manuel Davy² and Amrane Houacine¹

¹ LCPTS - USTHB, B.P. 32, El Alia, Bab Ezzouar, Alger, ALGERIE

² LAGIS/CNRS, BP 48, Cité Scientifique, 59651 Villeneuve d'As cq cedex, FRANCE

bfergani@msn.com, Manuel.Davy@ec-lille.fr, ahouacine@lycos.fr

SOMMAIRE

Avec l'augmentation récente et continue du volume d'archives sonores (radio, TV, Web, ...), il devient désormais indispensable de trouver des méthodes efficaces qui permettent de faciliter la recherche d'informations dans les grandes bases de données. Ainsi, on associe aux fichiers audio numérisés des fichiers textuels (fichiers index), de moindre complexité que le fichier signal original, mais contenant néanmoins un résumé des informations recherchées dans ce signal. 45 minutes de parole, 1 minute de musique, 10 locuteurs (6 hommes et 4 femmes) sont un exemple d'informations pertinentes. Ces fichiers d'index stockés en même temps que le signal original, seront d'un apport considérable lors de l'étape de recherche d'informations, permettant alors un accès direct et immédiat à l'information recherchée. Dans le cas où l'on voudrait savoir qui parle et quand dans un document sonore, la clé d'indexation est alors le locuteur. Un système d'indexation par locuteurs peut servir également comme étape préliminaire à des tâches de transcription ou de suivi de locuteurs et représente souvent un facteur important pour l'amélioration des performances des systèmes de reconnaissance automatique de la parole. Une étape préalable indispensable d'un système d'indexation par locuteurs est la segmentation en locuteurs. Celle-ci consiste à réaliser deux tâches séquentielles: la première étape permet de découper le signal paramétré en intervalles ou segments correspondants à des tours de parole de locuteurs, c'est à dire obtenir des segments les plus longs possibles homogènes en termes de locuteur, puis la deuxième étape consiste à regrouper les segments appartenant à un même locuteur.

Cet article présente une nouvelle approche originale pour la segmentation en locuteurs d'enregistrements audio. Il s'agit de l'application d'un algorithme basé sur l'utilisation des Méthodes à Vecteurs Support mono-classe (SVM-1), introduit récemment par l'un des auteurs, pour la détection de ruptures puis le regroupement de segments correspondant à l'intervention des différents locuteurs. Nous montrons à travers de nombreuses expériences sur deux bases de signaux d'enregistrements radiodiffusées, la pertinence de cette méthode et son comportement comparés par rapport à la méthode classique à base du rapport de vraisemblance généralisé utilisant le critère d'information bayésien (RVG-BIC).

ABSTRACT

With recent and continued increases in the number of available sound archives (radio, TV, Web,...), effective methods must be established to facilitate the process of searching for information within massive databases. Of less complexity than the original sound file but nevertheless containing a summary of important information pertaining to the signal, text files (index files) are linked to the digital sound files. An example of relevant information found in the text file is as follows: 45 minutes of speech, 1 minute of music, 10 speakers (6 men and 4 women). These index files, stored with the original signal, will contribute considerably to the information retrieval process, allowing an immediate and direct access to the information sought. If one would like to know who speaks and when in a sound file, the index key is hence the speaker. A preliminary stage of a speaker indexing system is speaker diarization. State-of-the-art speaker diarization techniques require two main steps: speaker turn detection which consists of detecting speaker turn times, that is boundaries of audio file segments where only one speaker is present, followed by a clustering step which consists of labelling the previous segments in terms of speakers. These two stages require a metric to be defined in order to compare and group speech segments. This paper presents a novel approach for the speaker diarization of audio recordings. The proposed approach uses a metric based on one-class Support Vector Machines (SVM-1), introduced recently by one of the authors, for the speaker change detection and clustering tasks. Through many experiments using two databases of broadcast recordings, we demonstrate the relevance and superiority of this approach compared to the traditional method based on the generalized likelihood ratio using bayesian information criterion (RVG-BIC).

1 INTRODUCTION

Avec la multiplication de sources de données multimédia et le développement des techniques de numérisation de l'information, nous assistons à une explosion des bases de données d'archivage. De ce fait se pose un problème crucial: Comment accéder facilement et rapidement à l'information recherchée ? Ces deux critères (rapidité et facilité d'accès) sont indispensables pour toute requête d'utilisateur.

L'indexation en locuteur d'un signal sonore consiste à structurer ce signal selon l'information véhiculée par le locuteur. Dans ce contexte, la segmentation en locuteur constitue une étape préalable et déterminante pour la suite du processus d'indexation. Elle consiste à d'abord découper le signal audio en zones homogènes contenant uniquement les informations relatives à un seul locuteur. Cette étape est suivie du regroupement (clustering) de ces segments afin d'assembler les zones appartenant à un seul locuteur.

Ce problème a déjà fait l'objet de nombreuses études (voir par exemple [9, 10, 14, 16] et dans les références qui y sont indiqués). Les techniques standard utilisent généralement des descripteurs acoustiques (souvent des MFCC et leurs dérivées) puis appliquent deux fenêtres d'analyse glissantes sur les données de part et d'autre de l'instant courant. Etant donné les vecteurs acoustiques $\mathbf{x}(n)$, $n = 1, 2, \dots$, la fenêtre d'analyse située avant l'instant d'analyse n définit l'ensemble passé immédiat $X_p(n) = \{\mathbf{x}(n - m_p), \dots, \mathbf{x}(n - 1)\}$ de m_p vecteurs acoustiques, tandis que l'autre fenêtre contient m_f vecteurs acoustiques représentant l'ensemble futur immédiat $X_f(n) = \{\mathbf{x}(n + 1), \dots, \mathbf{x}(n + m_f)\}$. L'objectif de ces techniques classiques est de comparer les ensembles $X_p(n)$ et $X_f(n)$ en évaluant une distance (ou une mesure de similarité) entre ces deux ensembles de données. Ceci est réalisé au moyen de méthodes à base de rapport de vraisemblance généralisé (RVG) soit directement [10] ou indirectement comme dans l'approche par critère d'information bayésien (BIC) notée RVG-BIC et adoptée comme référence dans ce papier [4]. Les méthodes à base de RVG-BIC nécessitent la connaissance d'un modèle de la distribution de probabilité des données $\mathbf{x}(n)$. Les modèles gaussien ou mélange de gaussiennes ont été largement exploités dans ce cadre. Bien que ces méthodes aient souvent donné des résultats satisfaisants, l'adéquation du modèle aux données réelles constitue néanmoins une hypothèse très forte et non généralisable. D'autre part, les méthodes inspirées du RVG-BIC souffrent d'un grand taux de fausse alarme pour de courtes interventions ($\leq 2-5$ secondes) et sont aussi gourmandes en temps de calcul [14]. Les méthodes indirectes [4], appelées aussi méthodes pas à pas, opèrent en deux étapes :

1. Détection de ruptures acoustiques: Cette étape consiste à segmenter le signal (ou sa version paramétré) en plages acoustiquement homogènes, dans le sens que chaque segment contient un seul et unique locuteur. A l'issue de cette étape on obtient les instants correspondants aux frontières de ces segments.
2. Regroupement de segments homogènes : Cette étape

consiste à regrouper les segments appartenant à un même locuteur. Autrement dit, le cluster correspondant à un locuteur donné ne doit contenir que les interventions (segments) appartenant à ce locuteur et tous les segments relatifs à ce locuteur doivent se trouver dans ce même cluster.

Les méthodes intégrées [8] consistent à fusionner les deux étapes dans le sens qu'elles consistent à détecter puis regrouper les locuteurs au fur et à mesure du déroulement du signal.

Dans cet article, nous suivrons le formalisme des méthodes pas à pas et nous proposons d'appliquer l'algorithme basé sur une méthode à noyau introduite dans [5] aux tâches de détection de ruptures et de regroupement dont le résultat d'ensemble est connu sous le vocable de segmentation en locuteurs [6]. Différemment des approches citées précédemment, notre approche exploite un algorithme à base de Méthodes à Vecteurs de Support mono-classe (SVM-1) dont la finalité est de comparer les ensembles $X_p(n)$ et $X_f(n)$ à chaque instant d'analyse au moyen d'une mesure de similarité. Dans ce sens, notre méthode reste semblable aux méthodes classiques à base de RVG, néanmoins notre approche exploite les informations extraites de l'entraînement de deux (SVM-1), dont l'avantage principal est de contrôler la complexité du modèle ajusté aux données et de prendre en compte l'information paramétrée selon diverses configurations et tailles des vecteurs acoustiques.

Le reste du papier est structuré de la façon suivante: Dans la section suivante, nous rappelons le principe de l'algorithme de détection de rupture basé (SVM-1), puis la section 3 présente la méthodologie d'application à la segmentation en locuteurs. La section 4 est particulièrement privilégiée puisqu'elle présente les résultats d'application de notre méthode aux signaux réels issus de deux bases de données radio-phoniques en comparaison avec la méthode RVG-BIC. Nous y détaillons le choix des paramètres de détection de ruptures et de regroupement. Finalement la dernière section 5 présente les conclusions et perspectives.

2 LA DETECTION DE RUPTURE BASEE SUR SVM-1

Nous partons de l'hypothèse que les vecteurs acoustiques $\mathbf{x}_1, \dots, \mathbf{x}_m$ sont générés identiquement et indépendamment par une distribution de probabilité (ddp) inconnue $p(\mathbf{x})$. Le principe de l'algorithme Kernel Change Detection (KCD) développé dans [5] est de comparer les ensembles $X_p(n)$ et $X_f(n)$ au travers de la comparaison de leur support de ddp. On définit le support d'une ddp S^λ par l'ensemble des points de l'espace des vecteurs acoustiques $\mathcal{X}$ telle que $p(\mathbf{x}) \geq \lambda$, avec λ une constante positive quelconque. Lorsque $\mathcal{X} = \mathbb{R}^d$ et $p(\mathbf{x})$ est une distribution gaussienne multivariée de dimension d , S^λ est une ellipsoïde de dimension d . Dans le cas le plus général, le support de densité peut avoir une configuration géométrique quelconque (généralement lisse). L'algorithme KCD applique un modèle SVM-1 aux données afin d'estimer leur support de ddp, ce que nous détaillons dans la section

suivante.

2.1 Les Méthodes à Vecteurs Support mono-classe (SVM-1)

Soit une fonction réelle symétrique appelée noyau définie dans $\mathcal{X}$. Dans la suite, nous considérons un noyau de forme gaussien, comme suit:

$$k(\mathbf{x}_1, \mathbf{x}_2) = \exp -\frac{1}{2\sigma^2} \|\mathbf{x}_1 - \mathbf{x}_2\|_{\mathcal{X}}^2 \quad (1)$$

avec $\|\cdot\|_{\mathcal{X}}^2$ est une norme définie sur $\mathcal{X}$. Le modèle SVM-1 estime le support de la ddp comme suit :

$$S^\lambda = \{\mathbf{x} \in \mathcal{X} | f^\lambda(\mathbf{x}) + b \geq 0\} \quad (2)$$

Ce problème d'estimation du support de la ddp revient à estimer une fonction dans l'espace augmenté $\mathcal{H}$ (hilbertien et à noyau reproductible induit par $k(\mathbf{x}_1, \mathbf{x}_2)$), proche du support de la ddp recherchée. On montre dans [12] que les fonctions minimisant le risque régularisé s'écrivent en fonction de $\mathbf{x}$ comme :

$$f^\lambda(\mathbf{x}) + b = \sum_{i=1}^m \alpha_i k(\mathbf{x}, \mathbf{x}_i) + b \quad (3)$$

les coefficients de pondération α_i sont dénommés les multiplicateurs de lagrange. Le paramètre ν (réel positif) joue un rôle de contrôle des vecteurs supports. Ainsi, choisir $\nu = 0.2$ équivaut à admettre 20% de vecteurs acoustiques dans $\mathcal{X}$ comme marginaux¹. A ce problème d'optimisation correspond un problème dual plus simple à résoudre puisque quadratique avec des contraintes linéaires:

$$\begin{aligned} &\text{minimiser} && \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \\ &\text{par rapport à} && \{\alpha_1, \dots, \alpha_m\} \\ &\text{avec les contraintes : a)} && 0 \leq \alpha_i \leq \frac{1}{\nu m} \text{ t.q. } i = 1, \dots, m \\ &\text{et b)} && \sum_{i=1}^m \alpha_i = 1 \end{aligned} \quad (4)$$

le modèle SVM-1 admet une simple interprétation géométrique dans l'espace augmenté $\mathcal{H}$: premièrement, les vecteurs acoustiques dans $\mathcal{X}$ sont projetés vers $\mathcal{H}$ au moyen de l'application : $\mathbf{x} \rightarrow k(\mathbf{x}, \cdot)$. Deuxièmement, les vecteurs acoustiques dans $\mathcal{H}$ sont de norme unitaire lorsque le noyau gaussien est choisi, car $\|k(\mathbf{x}, \cdot)\|_{\mathcal{H}}^2 = \langle k(\mathbf{x}, \cdot), k(\mathbf{x}, \cdot) \rangle_{\mathcal{H}} = k(\mathbf{x}, \mathbf{x}) = 1$ (propriété du noyau reproductible dans l'espace hilbertien), ainsi ces vecteurs sont situés sur la surface d'une hypersphère de rayon unité. Troisièmement, la résolution de Eq. (4) peut se ramener à trouver dans $\mathcal{H}$ un hyperplan orthogonal à $f(\cdot)$ tel que celui-ci serait le plus éloigné de l'origine, séparant ainsi les données d'apprentissage $k(\mathbf{x}_i, \cdot)$ entre deux classes -voir figure 1.

¹en anglais:outliers

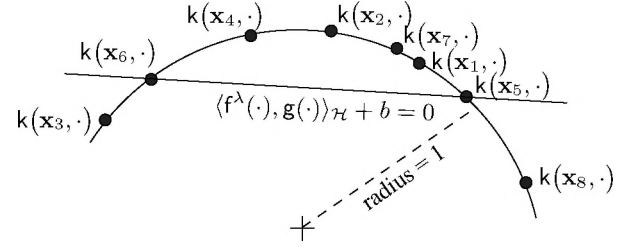


Figure 1: Interprétation géométrique du modèle SVM-1 dans $\mathcal{H}$. Les vecteurs acoustiques projetés $k(\mathbf{x}_i, \cdot)$, $i = 1, \dots, m$ sont situés sur une hypersphère de rayon unité. La fonction $f^\lambda(\cdot)$ et b définissent un hyperplan d'équation $\langle f^\lambda(\cdot), g(\cdot) \rangle_{\mathcal{H}} + b = 0$. La majorité des données est située du côté de l'hyperplan ne comprenant pas l'origine de l'hypersphère. Les coefficients α_i correspondants sont nuls ($i \in \{1, 2, 4, 7\}$), tandis que les points marginaux $k(\mathbf{x}_3, \cdot)$ et $k(\mathbf{x}_8, \cdot)$ sont situés du côté de l'hyperplan comprenant l'origine $\alpha_3 = \alpha_8 = 1/\nu m$. Les points situés sur l'intersection de l'hyperplan et de l'hypersphère vérifient $0 < \alpha_i < 1/\nu m$ ($i \in \{5, 6\}$).

2.2 Une mesure de similarité basée sur SVM-1

Cette mesure est construite sur le principe que les ensembles $X_p(n)$ and $X_f(n)$ sont similaires si et seulement si les supports de densités estimés sont similaires selon un certain critère de similarité. Notons que, dans l'espace initial des données $\mathcal{X}$, la forme des contours de décision représentant $S_p^\nu(n)$ et $S_f^\nu(n)$ peuvent être complexes et discontinus, rendant ainsi la définition d'une mesure de similarité dans cet espace très difficile. Heureusement, l'interprétation géométrique des SVM-1 dans l'espace augmenté $\mathcal{H}$ permet d'en déduire une mesure très intuitive et simple à mettre en oeuvre : les quantités $S_p^\nu(n)$ et $S_f^\nu(n)$ correspondent géométriquement aux hypercercles résultants de l'intersection de l'hypersphère avec l'hyperplan, voir figure 1. Ainsi, la comparaison des quantités $S_p^\nu(n)$ et $S_f^\nu(n)$ dans l'espace initial des données $\mathcal{X}$ se ramène à une comparaison dans $\mathcal{H}$ en comparant les hypercercles correspondants dont les centres sont notés $c_p^\nu(n)$ et $c_f^\nu(n)$ et les arcs de cercle $r_p^\nu(n)$ et $r_f^\nu(n)$, voir figure 2.

La mesure de similarité basée sur notre algorithme est définie comme suit [5]:

$$D^\nu(n) = \frac{d(c_p^\nu(n), c_f^\nu(n))}{r_p^\nu(n) + r_f^\nu(n)} \quad (5)$$

Pratiquement, $D^\nu(n)$ est calculée à partir des coefficients α_i de chaque support de densité $S_p^\nu(n)$ et $S_f^\nu(n)$, -voir l'article de F. Desobry et al. dans [5] pour les détails de calcul et de développement de cette mesure.

3 DETECTION DE CHANGEMENT DE LOCUTEURS

3.1 La détection de ruptures

Le processus de détection de ruptures basé sur l'algorithme KCD effectue séquentiellement les étapes suivantes :

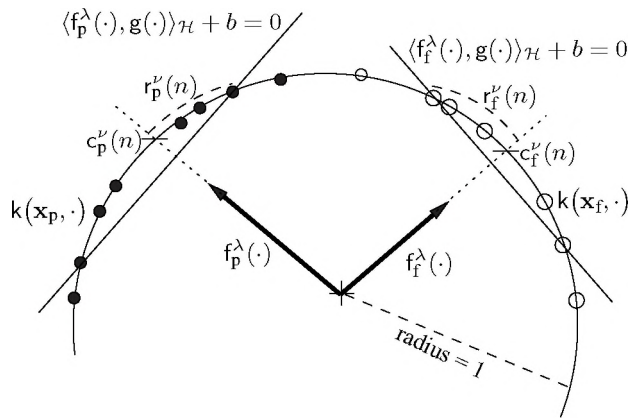


Figure 2: Interprétation géométrique de l'algorithme basé sur SVM-1. La situation représentée ici correspond à une détection de ruptures, car les hyperplans représentés par $f_p^\lambda(\cdot)$ (correspondant à l'ensemble passé immédiat – cercles pleins) et $f_f^\lambda(\cdot)$ (correspondants à l'ensemble futur immédiat – cercles vides) sont distinctement séparés, et la distance $d(c_p^\nu(n), c_f^\nu(n))$ est grande par rapport aux arcs $r_p^\nu(n)$ et $r_f^\nu(n)$.

1. **Paramétrisation acoustique:** Celle-ci est effectuée au moyen d'outils conventionnels tel que HTK Tools [15]. Cette étape est effectuée pour l'ensemble du signal.
2. **Entraînement des SVM-1 :** Pour chaque instant n et consécutivement à la formation des ensembles $X_p(n)$ et $X_f(n)$ deux modèles SVM-1 sont entraînés en résolvant le problème (4) pour les deux ensembles. Pour réduire les charges de calculs et accélérer la procédure, la technique développée dans [3] peut être utilisée.
3. **évaluation de la mesure de similarité:** Comme dans toute technique de segmentation (détection de rupture), une rupture est détectée lorsque l'indice $D^\nu(n)$ définie dans Eq. (5) dépasse un seuil prédéterminé. Une autre approche consiste à calculer toute la courbe des distances puis choisir un seuil donné permettant d'estimer les instants de ruptures.

3.2 Le regroupement hiérarchique

Suite à la détection de ruptures, nous sommes en présence d'une collection d'objets (segments de paroles homogènes/locuteurs) et nous devons regrouper ces objets par classe (les locuteurs). Nous avons choisi de mettre en oeuvre le regroupement hiérarchique agglomératif qui consiste à considérer au départ chaque segment comme étant une classe et à chaque itération on réunit deux classes les plus proches au sens d'un critère, appelé critère de regroupement [4]. Dans ce cas ce critère est la mesure de similarité définie dans la section 2.2. Ce processus est réitéré jusqu'à l'obtention d'une classe unique. Nous obtenons à l'issue du regroupement un arbre de classification appelé dendrogramme. C'est la manière de parcourir l'arbre qui définit la partition finale.

4 EXPERIENCES ET RESULTATS

L'objectif de ces expériences est de conduire une étude comparative des performances de notre méthode par rapport à la méthode RVG-BIC prise comme référence. Ces expériences sont menées pour deux bases de données différentes. Le but poursuivi est d'abord de réaliser une étude comparative préliminaire avec une base de données locale (ALSIG) puis de mener une étude détaillée dont l'objectif est l'étude des performances de notre méthode en fonction du réglage de ses différents paramètres, c'est ce que nous avons réalisé avec la base de données (NISTDB) issue de la campagne d'évaluation NIST RT'03 [11].

4.1 Bases de données

Les signaux utilisés sont issus d'enregistrements d'émissions d'informations radiodiffusées (Broadcast News) de diverses stations locales algériennes et une base de données de stations radio américaines.

Pour la base de données locale (ALSIG) [13], les fichiers ont été enregistrés à partir du web au moyen du logiciel praat [1]. La segmentation de référence est obtenue également par annotation manuelle, une option offerte par le logiciel praat. Ces fichiers sonores sont structurés comme suit:

- 6 enregistrements radiodiffusées de la Radio Algérienne désignées par bndz1 à bndz6 de 10 mn chacun, en langues arabe et française.
- 4 fichiers sonores de durées 3 minutes chacun et pour lesquelles l'intervention moyenne de chaque locuteur est de 3 secondes en moyenne.

Les fichiers de la base NISTDB [11] se divisent en deux catégories: 6 fichiers de développement (dry run files) de 10 minutes environ chacun dédiés au réglage des paramètres et 3 fichiers d'évaluation (Eval files) de 30 minutes chacun.

4.2 Critères de Performances

Pour la base de données (ALSIG) utilisée pour une première évaluation de notre méthode, le critère de performance retenue est la mesure des erreurs d'insertions et d'omissions. Une erreur d'insertion se produit lorsqu'un changement de locuteur est détecté alors que celui-ci n'existe pas dans le fichier réel analysé. Une erreur d'omission correspond au fait de manquer de détecter un changement de locuteur alors que celui-ci existe. On l'appelle aussi l'erreur de détection manquée (DM). Notre méthode de segmentation en locuteurs étant conçu pour être insérée dans un système d'indexation global, dans ce cas une erreur d'insertion (ou fausse alarme) (FA) due à une sursegmentation s'avère toujours moins critique qu'une erreur de détection manquée due à une soussegmentation. Ceci s'explique par le fait que l'étape de détection de rupture est suivi séquentiellement par l'étape de regroupement (clustering) des segments appartenant au même locuteur, et par conséquent une erreur de fausse alarme est souvent rattrapée et corrigée lors de l'étape de regroupement.

Dans ce cadre, une segmentation de référence est absolument nécessaire pour évaluer ce genre d'erreurs. Cependant la précision de détection des instants de changement dépend de la construction de la segmentation de référence qui est entachée d'une erreur systématique due à la subjectivité d'évaluation d'un instant de changement de locuteur selon l'expérimentateur. Ainsi, un instant candidat est déclaré erreur d'insertion ou erreur de fausse alarme si on ne trouve pas un instant de référence qui l'entoure dans un intervalle de confiance prédéfini. De même, l'absence d'un instant candidat (généré par notre système) autour d'un instant de référence correspond à une erreur d'omission ou détection manquée.

Pour la base (NISTDB), le critère de performance établi et fourni par l'institut NIST est le Speaker Diarization Error ou Diarization Error Rate (DER) obtenu par un script en langage perl (voir [9] pour d'amples détails).

Selon ce protocole d'expérimentation, Il s'agit de présenter le résultat de la segmentation issu du système proposé, selon le format défini par NIST. Ce fichier constitue une entrée au script fourni conjointement au fichier de référence. Le résultat est un fichier texte fournissant le DER (en %).

4.3 Résultats d'expériences sur la base ALSIG

Les résultats reportés dans la table 1 mettent en évidence les performances comparées de notre algorithme KCD et la méthode de référence basée sur le Rapport de Vraisemblance Généralisé utilisant le critère d'information bayésien (RVG-BIC). Les conditions d'expérience (initialisation des paramètres) communes aux deux méthodes sont : La paramétrisation acoustique est effectuée au moyen de l'outil HTK et les descripteurs utilisés sont les coefficients MFCC au nombre de 16 (pas de coefficient C_0) avec une fenêtre de Hamming de 20 ms et un recouvrement de 50%. La taille des fenêtres glissantes sur les données est fixée à $m = m_p = m_f = 1.5$ secondes avec un pas de progression $\Delta_n = 0.1$ secondes. Les paramètres spécifiques à chaque méthode sont les suivantes: Pour la méthode RVG-BIC, les données passées et futurs par rapport à l'instant d'analyse sont modélisés comme une mono-gaussienne avec une matrice de covariance diagonale et le paramètre de regroupement est fixe à $\lambda_{reg} = 1.5$ [4]. Pour notre méthode KCD, il y a lieu de préciser le paramètre du noyau gaussien σ que nous avons initialisé à $\sigma = 1$. Les fichiers sonores bndz1 à bndz6 contiennent respectivement de 15 points à 42 points de ruptures relatifs à des tours de parole des locuteurs intervenant dans l'enregistrement audio. Les résultats montrent que la méthode KCD réalise de meilleurs performances en terme de taux de DM. Cependant le taux de FA reste comparable par rapport à la méthode de référence.

L'objectif de l'expérience suivante est de montrer que notre algorithme permet de détecter des changements de locuteurs, même dans des situations où la méthode RVG-BIC ne donne pas de bons résultats [14]. L'expérience dont les résultats sont portés dans la table 2 tendent à démontrer la supériorité des performances de notre méthode pour des fichiers sonores pour lesquelles l'intervention moyenne de

chaque locuteur est très courte (3 secondes). La paramétrisation est la même que pour l'expérience précédente. Les autres paramètres communs restent les mêmes, alors que les paramètres spécifiques de chaque méthode sont tels que: Pour RVG-BIC: $\lambda_{reg} = 1$. Pour KCD, $\sigma = 1$. Les résultats préliminaires, montrent néanmoins des performances supérieures de notre méthode par rapport à la méthode de référence. Le Taux de FA reste cependant important même pour notre méthode.

Ces expériences préliminaires montrent la nécessité d'ajuster les paramètres spécifiques à chaque méthode en fonction des données à traiter et des expériences. Pour cela nous avons considéré une base de données plus importante et reconnue par la communauté scientifique active dans ce domaine pour mener une étude détaillée de notre méthode en fonction de ses paramètres et tenant compte de diverses paramétrisations acoustiques.

4.4 Expériences sur les signaux de développement de la base NIST

Afin de régler les paramètres de notre algorithme pour les tâches de détection de ruptures et de regroupement nous utilisons les six fichiers de développement pris séparément et le score global est la moyenne sur l'ensemble des fichiers. Pour les expériences concernant cette catégorie de fichiers, la paramétrisation utilisée est de 16 coefficients MFCC. Ce choix initial, à priori arbitraire est motivé par le fait que le type de paramétrisation et leur nombre est souvent utilisé dans la méthode de référence (RVG-BIC) et donne des résultats satisfaisants.

Sélection de paramètres

Les paramètres pertinents de notre algorithme objet de cette étude de sélection sont : $m = m_p = m_f$ (taille de l'ensemble $X_p(n)$ et de $X_f(n)$ que nous supposons égaux) et le pas de progression Δ_n de ces ensembles appelés communément fenêtres glissantes précédent et succédant à l'instant d'analyse n . Ces deux paramètres m et Δ_n sont communs avec la méthode de référence RVG-BIC. Aux fins de comparaison des deux méthodes nous avons fixé les mêmes valeurs pour ces deux paramètres.

Notre algorithme utilise un paramètre additionnel relatif au noyau, assurant le contrôle de la corrélation des points voisins dans l'espace augmenté $\mathcal{H}$. Dans le cas du noyau gaussien ce paramètre est l'écart-type σ .

La table 3 montre l'évolution du DER en fonction de la variation de σ . Le minimum d'erreur est atteint pour $\sigma = 0.51$, et vaut 10.73%. Les autres paramètres utilisés pour cette expérience sont fixés comme suit: $m = m_p = m_f = \text{win} = 1.5$ s et $\Delta_n = 0.2$ s, qui sont des paramètres adéquats pour les deux méthodes en comparaison. Le paramètre $\nu = 0.1$ traduit une tolérance maximale de 10% des vecteurs support.

Les tables 4 et 5 résument les variations de la taille des fenêtres adjacentes glissantes $m = m_p = m_f$ et leurs pas de progression Δ_n . On confirme, que les valeurs de $m = 1.5$ s et $\Delta_n = 0.2$ s sont des réglages adéquats au regard de la l'erreur obtenue (10.73%).

fichier	RVG-BIC		KCD	
	FA	DM	FA	DM
<i>bndz1</i> (15 pts)	8	4	9	2
<i>bndz2</i> (19 pts)	8	5	7	2
<i>bndz3</i> (21 pts)	10	7	11	4
<i>bndz4</i> (35 pts)	18	8	17	5
<i>bndz5</i> (39 pts)	12	9	14	5
<i>bndz6</i> (42 pts)	15	8	13	5

Table 1: Expériences sur signaux composites

fichier	RVG-BIC		KCD	
	FA	DM	FA	DM
<i>file1</i> (55pts)	10	9	11	4
<i>file2</i> (63pts)	9	12	12	5
<i>file3</i> (59 pts)	12	15	16	4
<i>file4</i> (62 pts)	10	8	9	4

Table 2: Expériences pour Interventions courtes

Evaluations de stratégies de paramétrisation acoustique

Nous présentons dans cette sous-section l'impact des diverses stratégies de paramétrisation acoustiques décrites dans la table 6. Le but poursuivi est d'évaluer l'effet de combinaison des descripteurs sur les performances de notre méthode. Cette combinaison est réalisée en concaténant plusieurs vecteurs acoustiques résultant de diverses paramétrisations.

Pour chaque configuration, nous avons optimisé la sélection des paramètres, telle que mentionnée dans la section ci-dessous. Nous reportons dans les tables 7 et 8 les erreurs minimales obtenues avec le jeu de paramètres sélectionné.

L'estimation du nombre de locuteurs présents dans la conversation analysée est obtenue en parcourant le dendrogramme selon une coupe horizontale, ce qui revient à faire une hypothèse sur le nombre de locuteurs désiré puis de vérifier celle-ci selon la performance obtenue. Les travaux de synthèse reportés dans [9] offrent une explication claire sur la détermination automatique du nombre de locuteurs. Nous observons que les résultats confirment que la combinaison de diverses paramétrisations acoustiques (C_1 , C_2 et C_5) améliorent globalement les résultats pour les deux méthodes avec une nette supériorité pour la méthode KCD. Ce résultat confirme aussi les conclusions portées dans les travaux [7] et [2] à savoir qu'une paramétrisation conjointe peut mettre en évidence des caractéristiques du signal de parole qui peuvent être cachées en utilisant une paramétrisation unique.

4.5 Validation sur les fichiers d'évaluation

La table 9 montre que notre méthode obtient des taux d'erreurs DER bien inférieurs à la méthode de référence RVG-BIC. Le meilleur résultat obtenu est la moyenne sur ces trois fichiers, soit un taux d'erreur de 11.30%. Ces résultats

sont d'autant plus prometteurs car comparés à ceux publiés récemment dans la littérature constituant l'état de l'art [9] et [14] dans laquelle les méthodes présentées ont été optimisées indépendamment de notre algorithme.

5 CONCLUSIONS ET PERSPECTIVES

Les résultats présentés dans cette étude montrent clairement que notre approche basée sur un algorithme à base des méthodes à vecteurs support mono-classe ouvre une voie de recherche très prometteuse pour l'indexation en locuteurs de discours multi-locuteurs. Nous avons montré comment intégrer la mesure de similarité basée sur le modèle des SVM mono-classe pour des tâches de détection de ruptures et de regroupement en locuteurs. Un autre résultat intéressant concerne l'étude comparée des méthodes RVG-BIC et SVM-1 en fonction des diverses stratégies de paramétrisation. Ainsi, il apparaît, qu'une paramétrisation même redondante améliore les résultats globalement pour les deux méthodes mais que c'est notre méthode qui assure une nette supériorité. Un travail en cours concerne la sélection des descripteurs acoustiques selon les performances obtenues par le modèle SVM-1 en faisant varier ses paramètres (m_p , m_f et σ). Une autre perspective intéressante est l'automatisation de la sélection du seuil de détection qui reste un problème ouvert pour les méthodes RVG-BIC et KCD.

REFERENCES

- [1] Paul Boersma and David Weenink. Praat: Doing phonetics by computer. <http://www.praat.org>.
- [2] H. Christensen. *Speech recognition using heterogenous information extraction in multi-stream based systems*. PhD thesis, Aalborg University, Denmark, 2002.

σ	DER
1.29	33.12
1	43.78
0.85	12.57
0.75	21.01
0.707	13.60
0.67	14.88
0.62	16.41
0.58	15.75
0.54	12.87
0.51	10.73
0.49	12.46
0.45	23.86

Table 3: Evolution du DER (%) en fonction de σ

m (s)	0.5	0.7	0.9	1.5	2.5	3.5
DER (%)	14.70	13.83	12.71	10.73	16.65	25.36

Table 4: Evolution du DER en fonction de la taille des fenêtres glissantes m pour $\nu=0.1s$, $\sigma=0.51$ et $\Delta_{\tau} = 0.2s$

- [3] M. Davy, F. Desobry, A. Gretton, and C. Doncarli. An online support vector machine for abnormal events detection. *Signal Processing*, 86:2009–2025, 2006.
- [4] P. Delacourt and C. Wellekens. Distbic: a speaker-based segmentation for audio data indexing. *Speech Communication*, 32(1):111–126, September 2000.
- [5] F. Desobry, M. Davy, and C. Doncarli. An online kernel change detection algorithm. *IEEE Trans. Sig. Proc.*, 53(5), August 2005.
- [6] B. Fergani, M. Davy, and A. Houacine. Unsupervised speaker indexing using one-class support vector machines. In *European Signal Processing Conf. 2006*, Florence, Italy, September 2006.
- [7] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde. Significance of joint features derived from the modified group delay function in speech processing. *EURASIP Journal on Audio, Speech, and Music Processing*, 2007:13 pages, January 2007.
- [8] S. Meignier. *Indexation en locuteurs de documents sonores: Segmentation d'un document et Appariement d'une collection*. PhD thesis, Université d'Avignon et des pays de vaucluse, Laboratoire Informatique d'Avignon, 2002.
- [9] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier. Step-by-step and integrated approaches in broadcast news speaker diarization. *Computer Speech and Language*, 20, 2006.
- [10] D. Moraru, S. Meignier, C. Fredouille, L. Besacier, and J.-F. Bonastre. The ELISA consortium approaches in broadcast news speaker segmentation during the NIST 2003 rich transcription evaluation. In *IEEE ICASSP'04*, Montreal, Canada, 2004.
- [11] NIST RT03S. The rich transcription spring 2003 (rt-03s) evaluation plan <http://www.nist.gov/speech/tests/rt/rt2003/spring/docs/rt-03-spring-eval-plan-v4.pdf>. 2003.
- [12] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, USA, 2002.
- [13] ALSIG SPLAB. Algerian speech database. 2006.
- [14] Sue E. Tranter and Douglas A. Reynolds. An overview of automatic speaker diarization systems. *IEEE Trans. Audio, Speech, and Language Proc.*, 14(5):1157–1565, September 2006.
- [15] S. Young and all. *The HTK Book (for HTK Version 3.2.1)*. cambridge.
- [16] X. Zhu, C. Barras S. Meignier, and J.-L. Gauvain. Combining speaker identification and bic for speaker diarization. In *Proc. Eur. Conf. Speech Commun. Technol.*, pages 2441–2444, Lisbon, Portugal, September 2005.

Δ_n (s)	0.1	0.2	0.3	0.4	0.5	0.6
DER (%)	24.42	10.73	11.93	15.27	12.85	22.68

Table 5: Evolution du DER en fonction du pas de progression Δ_n , pour $m = 1.5s$, $\nu = 0.1$ et $\sigma = 0.51$

Configuration	composition du vecteur acoustique
C_0	16 MFCCs
C_1	16 MFCCs et 10 LPCCs
C_2	C_1 et 10 coefficients de reflection
C_3	C_2 et 10 coefficients banc de filtre
C_4	16 MFCCs et 16 Δ MFCCs
C_5	C_4 et 16 $\Delta\Delta$ MFCCs

Table 6: Paramétrisations acoustiques testées.

Config	DERmin (%)		estim. # loc.	
	KCD	RVG-BIC	KCD	RVG-BIC
C_0	10.73	26.38	17	9
C_1	8.37	21.18	17	9
C_2	7.95	15.08	17	11
C_3	10.90	15.26	9	9
C_4	11.44	20.91	13	11
C_5	8.63	14.30	19	11

Table 7: Performances comparées (KCD/RVG) en fonction des paramétrisations acoustiques décrites dans Table 6.

Configuration	Jeu de paramètres sélectionné
C_0	$m = 1.5s, \sigma = 0.51, \Delta_n = 0.2s$.
C_1	$m = 2.0s, \sigma = 1, \Delta_n = 0.3s$
C_2	$m = 1.5s, \sigma = 1, \Delta_n = 0.3s$
C_3	$m = 2.5s, \sigma = 0.707, \Delta_n = 0.2s$
C_4	$m = 0.7s, \sigma = 0.707, \Delta_n = 0.2s$
C_5	$m = 0.9s, \sigma = 0.85, \Delta_n = 0.3s$

Table 8: Jeu de paramètres en fonction des paramétrisations acoustiques testes.

Fichier	RVG-BIC	KCD	Estimation du Nbre de Locuteurs
(a)	25.60	11.25	21
(b)	20.17	10.55	20
(c)	22.69	12.11	28

Table 9: Résultats obtenus sur les signaux d'évaluation. Les paramètres sont $m = 1.5s$, $\nu = 0.1$, $\sigma = 0.51$ et $\Delta_n = 0.2s$. La paramétrisation choisie est C_2 . Les fichiers sont (a) 20010228.2100-2200-MNB-NBW, (b) 20010217.1000-1030-VOA-ENG and (c) 20010220.2000-2100-PRI-TWD.