وزارة التعليم العالي والبحث العلمي

BADJI-MOKHTAR UNIVERSITY– ANNABA-

UNIVERSITÉ BADJI MOKHTAR–ANNABA

جامعة باجي مختار
-عنابة-

Faculté : Sciences de l'ingénieur – Année 2018

Département : Informatique

THÈSE

Présentée en vue de l'obtention du diplôme de doctorat en sciences

Recherche d'information parlée

Option
S.I.C

Par
Mr BENDIB Issam

Soutenue le : 25/09/2018

DEVANT LE JURY

Pr.GHOUALMI Nacira	Pr. Université Badji-Mokhtar - Annaba	Président
Pr.LAOUAR Mohamed Ridda	Pr. Université Larbi Tebessi - Tébessa	Directeur de Thèse
Pr. KAZAR Okba	Pr. Université Mohamed Khider -Biskra	Examinateur
Dr. AZIZI Nabiha	MC. Université Badji-Mokhtar - Annaba	Examinatrice
Dr. BOURAMOUL Abdelkrim	MC. Université A. Mehri, Constantine 2	Examinateur
Pr. BENDJENNA Hakim	Pr. Université Larbi Tebessi - Tébessa	Examinateur

Dédicace

Je dédie ce travail :

À mes parents

À ma femme

A mes enfants : Ala errahmene, Med Taha et Aroua

À tous les membres de ma famille,

À tous ceux qui, de près ou de loin, ont soutenu mon travail par leurs encouragements et conseils.

Remerciements

Merci Allah, le Tout Puissant.

Je tiens à exprimer ma profonde reconnaissance au directeur de cette thèse, Mr. Pr LAOUAR Mohammed Ridda, pour m'avoir dirigé, guidé, conseillé, pendant tout le déroulement de ce travail, dont j'espère être à la hauteur de ses attentes.

Mes remerciements vont également, Mme Pr BAHI Halima, pour sa proposition de cette piste de recherches et ses encouragements pendant le parcours de cette thèse.

Mes remerciements vont également à mon cher collègue, Mr KHELIFA Boudjemaa, pour son soutien et son aide.

Mes remerciements vont également à mon beau père, Mr ZAIET Ahmed, pour ses efforts linguistiques.

Mes remerciements vont également à ma femme Dr Lilia. ZAIET pour son encouragement et soutien

Je tiens à remercier aussi tous les membres du jury qui me font l'honneur d'examiner mon travail.

ملخص

إن التطور الكبير في التقنيات المعلوماتية كالتسجيلات المرئية والمسموعة أدى إلى توفر كم هائل من المعلومات والوسائط الرقمية على الشبكات العنكبوتية ومنصات التواصل والتبادل المختلفة. هاته المعلومات والوسائط الرقمية تختلف وبشكل كبير وخاصة من حيث تركيبتها عن الملفات الرقمية البسيطة كالملفات النصية وقواعد المعطيات حيث أنها تعتمد تقنيات مركبة ومعقدة كتقنيات الصورة والموجات الصوتية.

إن الأنظمة وتقنيات البحث المعلوماتي المستخدمة حاليا موجهة بشكل كبير لاستغلال الوسائط الرقمية البسيطة كالملفات النصية ونحوها، وهذا نضرا للتطور والفعالية الكبيرة المسجلة في هذه الأنظمة من حيث نتائج البحث وخاصة فيما يتعلق بعمليات البحث في المحتوى وهذا ما تؤكده نسبة الاستخدام المرتفعة لهاته الأنظمة كمحركات البحث في الشبكات العنكبوتية.

وبالمقابل فإننا نلاحظ تزايد استخدام منصات التواصل والتبادل الالكترونية وارتفاع حجم الوسائط الرقمية المتداولة مما ينجر عنه ضرورة إيجاد حلول وطرائق فعالة للبحث في محتوى هذه الوسائط. علما أن جل الحلول المستخدمة حاليا لا تتوفر على إمكانيات فعالة للبحث في محتوى هذه الوسائط أ الملفات وإنما تكتفي باستخدام المعلومات المدمجة كاستخدام الكلمات المفتاحية أو كلمات الشرح المرفقة معها.

في هذا الإطار تتقدم هذه الاطروحة بالتركيز على إشكالية البحث في محتوى الوسائط الرقمية وبالتحديد بالبحث في محتوى الملفات الصوتية من خلال اقتراح وتقديم طريقة جديدة للفهرسة الآلية على أساس معاني المفردات للملفات الصوتية باستخدام تقنيات المعالجة الآلية. هذه الطريقة تؤسس لنظام يسمح لنا بفهرسة ومن ثم البحث في المحتوى الصوتي باستخدام تقنيات البحث عن طريق المنطوق وهذا استنادا على تقنيات التعرف الآلية للكلام مع ادماج المعارف المبينة على المعاني اللفظية باستخدام الشبكات الدلالية وتقنيات تشابه وتماثل الألفاظ والمعاني.

من الناحية التقنية، إن استخدام الطرق الآلية للتعرف على الكلام مكلفة جدا من ناحية الحوسبة وزمن التنفيذ مما يحد من استخدامها على مستوى الشبكات، ولهذا نقدم في هذه الأطروحة تصور لحل عملي لهذا الاشكال باستخدام الطرق الآلية للتعرف على الكلام على أجزاء من محتوى الملفات الصوتية من أجل الحصول على مجموعة من الكلمات الممثلة جزئيا لمحتوى الملف الصوتي ثم نقوم بعد ذلك بإثراء هذه المجموعة على أساس المعنى اللفظي للكلمات باستخدام الشبكات الدلالية WordNet على أساس قياس والتشابه اللفظي.

في هـذا الإطـار نقـدم في هاتـه الأطروحـة كـل المراحـل المتبعـة مـن أجـل إنجـاز نظـام للبحـث في محتـوى الملفـات الصـوتية وفـق المنهجيـة المقترحـة والـتي تتكـون مـن ثلاثـة وحـدات متكاملـة فيمـا بينهـا وهـي: محلـل الدلاليـة والمفـردات المبـني علـى تقنيـات التعـرف الآليـة للكـلام باسـتخدام اسـتراتيجيات تقسـيم المحتـوى، محلـل المفـاهيم والمعـاني المبـني علـى اسـتخدام تقنيـات تكنولوجيـا المعلومـات والمعـارف ومحـرك البحـث والكشـف عـن الكلمـات المفتاحيـة والفهـارس باسـتخدام تقنيـات الكشف عن المصطلحات المنطوقة والكلمات المفتاحية.

أمـا بالنسـبة لمرحلـة التجسـيد والتحقـق فقـد قمنـا باسـتخدام مجموعـة الملفـات الصـوتية TEDLIUM والمسـتخرجة مـن مـؤتمر TED حيـث ملتقـى التكنولوجيـا والترفيـه والتصـميم، ويغطـي اليـوم تقريبـا جميـع المواضـيع. هـذا الاختيـار مبـني علـى تنـوع المواضـيع والمفـردات المسـتخدمة بالإضـافة إلى الصـبغة التقنيـة في اللغـة المسـتخدمة خـلال دورات المـؤتمر. النتـائج المحصـل عليهـا في مختلـف الوحـدات مشـجعة وتـدفعنا إلى مواصـلة العمـل في المنحـى البحثـي مـن أجـل تحسينه وتطويره واستخدامه في إطار الحوسبة السحابية.

الكلمـات المفتاحيـة: البحـث في المحتـوى المنطـوق، الملفـات الصـوتية، الكشـف عـن المصـطلحات المنطوقة، التعرف التلقائي على الكلام، الفهرسة، التشابه اللفظي، المعنى اللفظي

Résumé

L'évolution importante dans les outils et les ressources informatiques a permis de générer une masse importante des ressources multimédias éparpillées sur le web et les différentes plateformes de communications et de partages. Cependant, ces ressources possèdent des structures complexes et très différentes à celles des ressources classiques comme les fichiers textes et les bases de données. Cette complexité nécessite la recherche des solutions et des approches efficaces pour la recherche dans le contenu de ces ressources, sachant que la majorité des solutions utilisées actuellement ne permettent pas l'accès directement au contenu de ces ressources, mais elles se limitent d'utiliser les métas donnés et les annotations manuelles diffusées conjointement par ces ressources.

Dans ce manuscrit, nous mettons l'accent sur la problématique de recherche sur le contenu des documents multimédia et précisément dans le contenu parlé par la proposition d'une nouvelle approche pour l'indexation automatique a base sémantique pour le contenu parlé en utilisons les techniques de reconnaissance automatique. Cette approche permette de construire un système qui permet l'indexation ainsi la recherche dans le contenu parlé par le biais des techniques de détection des termes parlés avec l'intégration des distances et similarités sémantiques.

De point de vue technique, l'utilisation des techniques de reconnaissance automatique de la parole est très couteuse de point de vue de calculs et temps d'exécutions, ce qui limite leurs utilisations dans les réseaux et le web. Pour cela, ce manuscrit présente une proposition d'une solution pour cette problématique par l'utilisation des transcriptions partielle du contenu pour extraire les termes les plus représentatifs, puis on enrichit cet ensemble sémantiquement par l'ontologie *WordNet* à l'aide des distances et les similarités sémantiques.

Dans ce contexte, nous décrivons une démarche pour la construction d'un système de recherche sur le contenu parlé. Cette dernière est structurée en trois modules qu'ils exploitent les techniques de reconnaissance automatique de la parole, les techniques de segmentation, les techniques de représentation des connaissances, les mesures des distances sémantiques et les techniques de détection de termes parlés et *Word Spotting*.

Dans la phase de validation et d'expérimentation, nous avons utilisé le corpus TEDLIUM extrait de la conférence TED. Cette conférence représente le forum de technologies, science et il couvre actuellement la majorité des disciplines. Ce choix est jugé par la diversité du contenu des ressources parlées ainsi que la qualité des termes de la langue utilisés lors de ces conférences. Les résultats obtenus dans les différentes phases sont encourageants et nous stimulons à continuer dans cette voie afin de la développer et de l'utiliser dans le cadre de cloud.

Mots-clés : recherche sur le contenu parlé, document parlé, indexation sémantique, détection de termes parlés, LVCSR, WordNet, ontologie, mesures de similarités

Abstract

The rapid development of information technology, such as the growing of video and audio recordings, has led to the availability of a large amount of information and digital media on the various networks and communication platforms. This information and digital media vary greatly, especially in terms of its composition from simple digital files such as text files and databases, as it adopts complex techniques such as image processing and sound waves.

The information systems and techniques used today are highly oriented to exploit simple digital media such as text files and so forth. This is the result of the great development and efficiency of these systems in terms of search results, especially with regard to the content retrieval. This is confirmed by the high utilization rate of these systems such as search engines on the web.

On the other hand, we note the increasing use of electronic communication and exchange platforms and the high exchange of digital media in circulation, resulting in the need to find solutions and effective methods based on content extraction of these multimedia resources. Note that most of the solutions in use are limited in terms of their ability to search within the content of visual and spoken files, but simply use the integrated information such as the use of keywords or the accompanying words.

In this context, we interest in this to the problem of searching digital media content and specifically searching the content of audio files by proposing and introducing a new method of automatic indexing based on the vocabulary meanings of audio files using automated processing techniques. This method establishes a system that allows us to index and then search the audio content using search techniques by means of the operative and this is based on automatic recognition techniques of the words with integration of the knowledge side shown on the verbal meanings using the techniques of similarity and parity of words and meanings.

Technically, the use of automated speech recognition methods is very expensive in terms of computing and execution time, which limits its use at the network level. Therefore, we present a practical solution to these formats using automated methods of speech recognition on parts of the audio file content in order to get a set of partially represented words for the audio file content and then we enrich this set based on the verbal meaning of the words using WordNet based on the distance and the verbal similarity.

In this framework, we present all stages of the thesis in order to achieve a system to search the content of audio files according to the proposed methodology, which consists of three units integrated among them: semantic analyser and vocabulary based on techniques of automatic recognition of speech using the strategies of content division, On the use of information technology, knowledge technology, search engines and the detection of keywords and indexes using techniques for the detection of spoken terms and keywords.

In the Implementation and validation step, we used the TEDLIUM audio file collection extracted from the TED conference where the Forum of Technology, Entertainment and Design, covering almost all subjects today. This choice is based on a variety of topics and vocabulary used in addition to the technical nature of the language used during the sessions of the conference. The results obtained in various units are encouraging and prompt us to continue to work in the research direction for its improvement, development and use in the framework of cloud computing.

Keywords: Content search, spoken documents, semantic content indexing, spoken term detection, LVCSR, Keyword Spotting, WordNet Ontology, Similarities measure.

Liste de Figures

Liste de Tableaux

Table des matières

Introduction Générale

1. Introduction

Les travaux de recherche dans le domaine d'indexation et recherche d'informations « *RI* » ont écoulés beaucoup d'encre ces dernières années et font l'objet de plusieurs recherches et projets scientifiques. Ces activités ont permis de manipuler divers concepts tels que : les ressources numériques, les documents multimédias, la transcription de la parole, les ontologies informatiques, les métriques de distances sémantiques, l'annotation des ressources, l'indexation conceptuelle, l'indexation automatique, ...etc.

Dans l'histoire des ressources numériques nous trouvons que l'humanité a connu les premiers enregistrements sonores depuis environ un siècle et demi. Exactement en 1860 par le typographe français, *Edouard-Léon Scott*, la masse de ressources audios a connu depuis cette année une énorme croissance à travers le monde.

Ces ressources audios deviennent plus répandues dans les différents médias comme : la radio, la télévision et l'internet. Entre autres, l'avènement du web a provoqué aussi ces dernières années un véritable raz de marée en matière de production des ressources multimédia par les biais des plateformes de partages et discussions, comme *: YouTube* ou *Dailymotion*. Le contenu de ces ressources peut-être de la parole, de la musique ou d'autres flux sonores. En général, ils contiennent de l'information qui est d'une importance cruciale pour les individus et les sociétés. En revanche, ces ressources ont permis aux individus et organismes un panorama d'informations, de compétences et des connaissances qui favorisent la créativité dans tous les domaines de la société.

D’autres parts, actuellement des nouveaux systèmes et applications informatiques ont permis aux utilisateurs le pouvoir d'exploiter et d'échanger n'importe quel type de données ou ressources comme : les documents textuels, les images, les flux parlés …etc. via des modèles et structures spécifiques. Dans ce contexte, des conséquences néfastes sur la qualité d'organisation utilisée pour ces énormes volumes d'informations, dans lesquelles il est devenu impossible de rechercher d’une manière exhaustive et manuelle une information donnée. Devant ces difficultés, il est devenu nécessaire de reproduire la capacité des systèmes automatiques à organiser les données et surtout l’information qu’elles contiennent, de façon à pouvoir y accéder rapidement et directement à leurs contenus. Il s’agit donc d’une problématique de recherche sur les techniques et méthodes d’indexation de données multimédia. Cette piste de recherche est accentuée par l'évolution rapide dans les systèmes informatiques et ces applications ces dernières années. Ils ont été accompagnés par l'admission des nouvelles orientations qui favorisent le partage et la réutilisabilité des ressources informationnelles non seulement à l'échelle des communautés restreintes, mais aussi à l'échèles planétaires.

Historiquement, la croissance importante du volume de données textuelles et multimédias comme les livres, les ressources documentaires et les ressources numériques dans les bibliothèques à imposer de trouver des mécanismes efficaces pour les exploiter. Les premières techniques, comme l'abstraction, l'indexation et l'utilisation des catalogues de classification ont marqué la naissance de concept « Recherche d'Information » comme étant une discipline de recherche prometteuse. D’étonnants efforts ont été déployés pour le développement des approches et des techniques permettant de retrouver l'information voulue effectivement et efficacement à partir de vastes collections de données. Depuis les années 1990, notamment avec l'avènement d'internet, la recherche d'information est devenue plus d'actualité et plus exigeante que jamais. Même si les efforts de recherches continus pour doter le domaine d'un ensemble riche d'outils et des protocoles intelligents et performants comme les protocoles de transmission, les systèmes d’acquisition, les supports de stockage …etc. Malgré l’importance de cette amélioration de ces outils et protocoles, reste la nécessité majeure et incontournable de chercher et concevoir des approches et techniques pour la gestion et l’exploitation de ces flux numériques produits.

Actuellement, des techniques avancées pour l'indexation et la recherche d'information ont été développées, comme celles utilisées par les moteurs de recherches connus comme : *Google, Bing, Yahoo, Yandex,* … etc. Cependant, malgré les performances enregistrées dans ces moteurs de recherches, qui répondent parfaitement au contenu textuel, ils présentent des insuffisances pour le traitement du contenu multimédia et le flux parlés. Ils sont toujours très loin d'accompagner l'immense trafic multimédia échangé et disponible dans le web et les systèmes de stockage numérique.

En effet, la discipline de « *Recherche d'information »* repose sur les techniques de : représentation, de stockage, d'organisation et des modalités d'accès aux informations. Ces systèmes permettent aux utilisateurs la recherche et la restitution des informations par le biais des requêtes utilisateurs. Dans ce contexte, les travaux de recherches actuelles se focalisent sur le développement des solutions fiables et efficaces pour les problématiques engendrées par ces techniques, ainsi que pour traiter l'évolution rapide enregistrée dans les techniques de production des ressources numériques et parlés.

2. La recherche dans le contenu parlé

Jusqu'à la fin du moyen âge, l'information a été transmise par voie orale ou sous forme manuscrite. L'invention de l'imprimerie a permis d'offrir l'information aux personnes de manière radicale. Cependant, l'internet est une extension de l'imprimerie puisqu'elle permet la publication de ces informations à grande échelle, en minimisant de nombreux obstacles pratiques qui prévalaient dans les jours anciens. En revanche, l'augmentation de la quantité d'informations véhiculées dans les différents structures et organismes comme les bibliothèques, internet, intranet, …etc. à émerger la nécessité de trouver des moyens et structures d'accès pour la restitution des informations pertinentes par rapport aux besoins spécifiques.

Les premières solutions ont été basées sur l'utilisation des techniques de catalogage à l'aide d'un processus d'assignement manuelle des mots-clés pour référencer et restituer les ressources informationnelles. Cette procédure est souvent combinée avec des index locaux pour raffiner la recherche à des pages ou sections au sein d'une ressource. Au fur à mesure, l'information textuelle a été numérisée et les systèmes informatiques sont généralisés, l'indexation manuelle devient quasi impossible par rapport au volume croissant des ressources ce qui émerge la voie vers la recherche des techniques automatiques d'indexation de ces ressources.

Cependant, l'introduction de portails du partage vidéo tel que *YouTube* et *Vimeo* ont élargissant les possibilités de publication et de partage des ressources numériques en donnant la possibilité d'hébergement des ressources d'informations plus complexes telle que le multimédia « les ressources audiovisuelles -AV » et les flux parlés. Du point de vue de modalité de recherche d'information, l'accès à ce type de contenu est souvent fait le recours aux méthodes classiques d'assignement manuelles des indexes « *mots-clés* » et d'annotations manuelles. L'indexation automatique efficace du contenu reste toujours un problème de recherche non résolu. Actuellement, le moyen le plus fiable pour trouver les segments pertinents dans les documents parlés est par le biais d'étiquetages manuels personnalisés [Marlow, 2006], ou en utilisant des informations contextuelles de commentaires ou de sites de référence.

Des études sur les anciennes collections de discours, tels que les collections les interviews et les archives de radio ont donné que ces derniers ne sont pas généralement complètement annotés et indexés. En effet, l'ajout manuel de descripteurs et index d'une façon rétroactive pour ces ressources est souvent impossible en raison de l'énorme quantité de discours.

Dans ce contexte, nous citons comme exemple les projets de recherches : *CHoral* et les ressources collectées à partir des archives de *Radio Rijnmond*. Dans le deuxième projet, ils ont effectué un étiquetage pour toutes les diffusions audios archivées par des descripteurs contenant la date de diffusion. Mais aucune métadonnée supplémentaire n'a été créée ou ajoutée pour cette collection. En effet, ces ressources représentent un vrai trésor pour les historiens intéressés par la région et ses habitants, la collection est surtout reste inexploitée [Wer, 2012]. Malheureusement, c'est le cas toutes les collections des discours parlés archivées dans le monde. En effet, l'accès à ces collections et ressources est extrêmement limité, d'où la nécessité potentielle de trouver des approches pour exploiter le contenu informationnel de ses trésors.

Entre autres, les ressources multimédias sont formées d'images, de vidéos et de bandes sonores. Bien que l'utilisation des techniques de traitement d'image pour l'extraction des méta informations sur les images fixes ou animées est un axe de recherche très actif. L'utilisation des systèmes de reconnaissance de la parole pour transcrire et indexer le flux parlé est aussi est un axe de recherche très encouragent et surtout avec le développement de standards et outils performants qui peuvent être utilisés dans les systèmes de recherche dans le flux parlé. Néanmoins, ses ressources et standards souffrent du problème de la charge et complexité des calculs et même voir sera incapable de gérer les grands volumes de flux parlé. Cependant, les méthodes et techniques de l'indexation du flux sont des pistes de recherche très actives, notamment avec la collaboration de la campagne d'évaluation *NIST*[1] « *National Institute for Science and Technology* ».

Dans ce contexte, nous pouvons citer à titre d'exemple les travaux réalisés sur la collection de « *Radio Rijnmond* » qui illustrent l'importance d'une solution d'indexation automatique pour les collections de discours [Wer, 2012]. À cet effet, les auteurs concluent que la modélisation d'une solution de recherche dans les documents parlés *SDR* « *Spoken Document Retrieval* » est possible pour les collections parlées numérisées et accessibles par les systèmes informatiques. En effet, l'approche classique consiste à générer automatiquement une transcription textuelle littérale du contenu du flux parlé en utilisant les systèmes de reconnaissance à large vocabulaire pour la parole continue « *LVCSR* ».

Ensuite, les résultats du processus de la transcription automatique obtenus sont traités comme n'importe quelle autre source textuelle en utilisant les techniques habituelles de recherche d'information pour détecter les segments pertinents. La facilité d'utilisation de ces systèmes est souvent considérée inférieure à celles pour la recherche de texte. Aussi, les travaux réalisés à grandes échelles sur les collections des discours d'Actualités en langue anglaise ont montré que les performances sont moins à celle pour le texte [Garofolo, 2000].

Cependant, les performances des systèmes de recherche d'information dans les ressources parlées dépendent étroitement par la qualité de la transcription automatique de ces derniers. Dans ce contexte, les programmes radio diffusés en langue anglaise sont des applications

[1] http://www.nist.gov/

presque idéales vis-à-vis les performances des systèmes de reconnaissance de la parole qui donnent des taux d'erreurs inférieurs à 10%. La plupart d'approches de RI doivent être suffisamment robuste au bruit et aux conditions d'enregistrement et de transmissions. Toutefois, si le type de discours, la qualité de l'enregistrement ou la phonétique de la langue parlée utilisée ne sont pas fiables, le bruit de la transcription peut rapidement atteindre des valeurs non acceptables. Par exemple, des expériences effectuées sur la collection « *Radio Rijnmond* » contenant un mélange de discours répété et spontané dans diverses conditions, ont indiqué que les taux d'erreur de transcription à dépasser les 50%, bien pire que le taux d'erreur de 20% qui a été généralement atteint par ce système sur la diffusion des nouveaux discours [Wer, 2012].

En effet, il est important que les systèmes de transcription des documents parlés doivent être capables de traiter les problèmes de bruits. Afin de permettre un accès optimal de ses ressources. Toute évaluation des approches pour la recherche des documents parlés « *SDR* » basées sur la transcription du contenu par des systèmes de reconnaissance automatique de la parole doit tenir compte des conséquences engendrées par les erreurs de reconnaissance et de leurs influences sur les performances du système de recherche d'information. Cette vérification n'est pas souvent automatisable et elle nécessite des ressources humaines immenses [Wang, 2009], la figure 0-1 présente l'architecture générale de ces systèmes. Actuellement, plusieurs ressources parlées sont inexploitées et cela dut aux erreurs de transcription du contenu qui influe sur les performances des systèmes des recherches des documents parlés. Dans ce contexte, notre contribution principale dans cette thèse est de proposer une approche d'indexation sémantique du flux parlé des ressources multimédias basé sur la recherche dans leurs contenus.

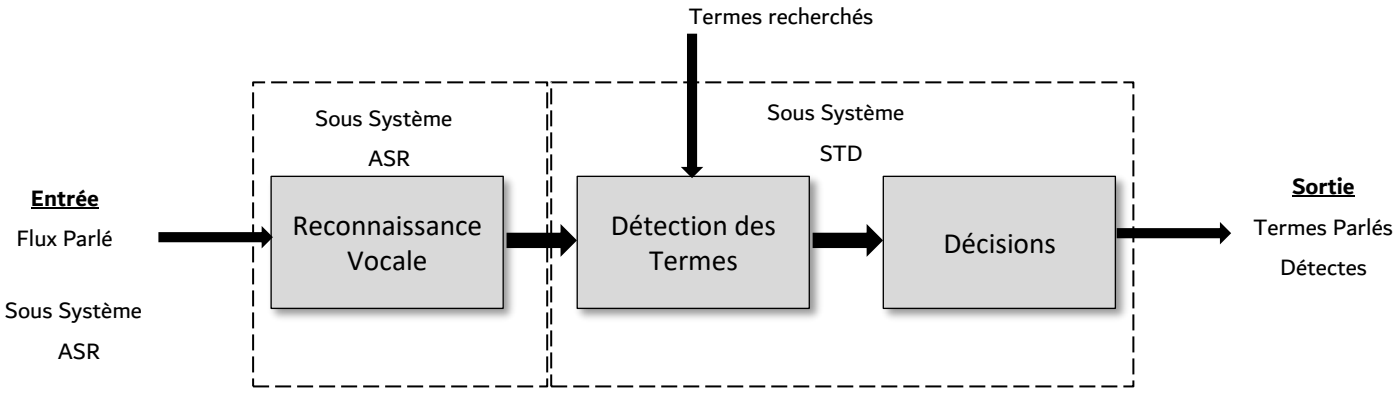

Figure 0-1 : Architecture standard d'un système STD d'après [Wang, 2009]

3. Motivations et problématiques

L'évolution technologique rapide marquée dans le monde de l'informatique ces dernières années a été accompagnée par l'admission de nouvelles orientations qui favorisent le partage, la diffusion et la réutilisabilité des ressources informationnelles non seulement à l'échelle des communautés restreintes, mais aussi à l'échèles planétaires. Entre autres, les progrès de la technologie numérique et de l'informatique sont à l'origine d'une croissance cruciale dans le domaine de traitement automatique de l'information et de l'acquisition de connaissances. Cependant, ce développement technologique ne cesse pas a engendré des gigantesques volumes d'informations multidisciplinaires qui se mesure de l'ordre des milliards voir plus de ressources numériques qui intègrent les textes, images, vidéo et les flux parlés.

D'où, la nécessité de trouver de nouvelles solutions qui améliorent les systèmes de recherche et d'indexation du contenu numérique, afin d'accompagner le progrès enregistré par l'introduction des nouvelles technologies dans le tous les domaines de la vie quotidienne (éducation, médecine, communication, divertissement, cinéma, économie, vie sociale, etc.).

D'autre part, le multimédia représente la partie majeure du contenu numérique et les stratégies courantes d'indexation du contenu présentent des insuffisances engendrer par la diversité des sources de données, la qualité d'enregistrement et le volume important des documents multimédia. Dans ce contexte, nous visons dans cette thèse de contribuer dans l'amélioration des systèmes de recherches dans les flux parlés qui permettent la restitution des segments du flux parlés en se basant sur les techniques de détection des termes parlés via un processus d'indexation sémantique automatique.

En outre, le volume des ressources parlées disponibles : nouveaux ou archivés et non annoté est très important, en plus de temps nécessaires de traitement de l'annotation rend l'indexation manuelle fastidieuse. C'est pour ces causes-là, qu'on fait appel à proposer une conception d'une démarche d'indexation automatique sur le contenu, via les techniques de gestion du contenu numérique et de gestion des connaissances via les ontologies et les mesures de similarités.

Cependant, la plupart des méthodes d'indexation actuelles demandent, de plus en plus, la prise en compte du contenu sémantique des ressources parlées en complément d'autres formats des ressources structurés. Par ailleurs, il existe plusieurs outils, d'extraction des connaissances sémantique, à partir des ressources multimédias, qui sont opérationnelles.

D'autre part, les systèmes d'indexation actuels présentent des insuffisances sérieuses. En effet, ils ne permettent que l'indexation sémantique du contenu structuré sous forme de texte. Par conséquent, une bonne partie des ressources parlées ne sont pas exploitées et prises en considération parce qu'ils ne sont pas référencés ou annotés. Les approches d'indexation sémantique actuelles offrent uniquement une indexation à base des textes en associés aux ressources parlées. En effet, ils ne permettent que d'exploiter les structures

textuelles disponibles conjointement avec le flux parlé. Entre autres, elle ne tente pas d'accéder et d'explorer le contenu réel de ce genre de ressources.

C'est dans ce contexte, et dans le cadre de cette thèse que nous avons choisi de nous pencher sur ces problèmes, par une contribution qui apporte une amélioration dans le domaine d'indexation sémantique du contenu du flux parlé. En effet, nous nous intéressons aux données numériques multimédias et en particulier les ressources parlées, qui peuvent contenir des informations nécessaires et utiles.

Cependant, afin d'exploiter ces ressources parlées, sachant que ces derniers sont enrichis avec des annotations manuelles constituées généralement de titre, mots-clés, nom d'auteur et/ou un résumé décrivant le contenu. Ces informations permettent une classification et indexation sommaire afin de faciliter l'accès et la recherche de ces ressources. Néanmoins ces annotations, souvent réduites au strict minimum et elles dépendent étroitement par les compétences de l'expert humain et peuvent ne pas décrire fidèlement le contenu de ces ressources parlées. Ces descripteurs restent insuffisants pour les utiliser pour développer un système de recherche pour le contenu parlé efficace et performant.

À cet effet, notre contribution se développe autour des problématiques suivantes :

- Comment contribuer à l'amélioration des méthodes et techniques d'exploitation du contenu des ressources parlées ?
- Comment accéder et repérer des segments spécifiques pour une requête donnée et non pas la totalité du document parlé ?
- Comment extraire les indexe ou les descripteurs discriminants du contenu des ressources parlées d'une façon automatique sans recourir aux experts des domaines.
- Comment pallier les problèmes liés au processus de reconnaissance automatique de la parole tels que les fausses alarmes, les fausses acceptations et les mots hors vocabulaire ainsi que les termes techniques du domaine ? On note dans cette thèse qu'on s'intéresse seulement sur les mots hors vocabulaire de point de vue documents et non pas du point de vue requêtes.

4. SOMI : une nouvelle approche d'indexation du contenu parlé

Notre contribution dans la problématique de recherche dans le flux parlé est par la proposition d'une approche d'indexation sémantique à base ontologique pour la primitive parole dans les documents multimédias que nous appelons « *Semantic Ontologies for Multimedia Indexing -SOMI* ». Cette approche intègre les techniques utilisées dans les traitements automatiques du flux parlé avec celles relatives aux technologies de l'information et de connaissances. Cette combinaison a pour objectif de surmonter les problèmes engendrés par les systèmes de reconnaissance automatiques de la parole par le biais des traitements linguistiques sémantiques. Tandis que, pour l'aspect recherche, nous avons fait recours aux techniques des détections des termes parlés.

En effet, la problématique de cette thèse est la recherche dans le flux parlé. Cette problématique concerne deux domaines intrinsèques : le domaine de recherche d'information et le domaine de traitement de la parole. Dans ce contexte, notre contribution est développée autour de ces objectifs. Nous avons proposé une approche composée de trois modules afin d'assurer la recherche et l'accès dans le contenu du flux parlé. Le premier module est dédié pour le traitement du contenu parlé à l'aide des techniques de reconnaissance automatique de la parole et les techniques de segmentation de flux parlé. L'objectif de ce module est le passage du contenu avec sa structure complexe qui est le signal vers une représentation textuelle la plus proche possible syntaxiquement. Notre stratégie est d'éviter la transcription intégrale du contenu qui nécessite des charges et des puissances de calculs immenses par une transcription partielle enrichie sémantiquement.

Le deuxième module a pour vocation l'enrichissement sémantique des résultats du premier module en deux sens : premièrement, passer du l'intégrale vers le partiel et deuxièmes comment assurer que la partiel doit couvrir sémantiquement l'intégrale. Dans le dernier module, nous avons travaillé sur les techniques de détection des termes parlés et les techniques de détection phonétiques des mots clés pour avoir une idée sur les modalités d'accès dans le contenu des ressources parlées. La figure 0-2 présente une brève description des techniques et méthodes utilisées pour chaque module. Entretemps, les descriptions détaillées de ces modules ainsi que les stratégies utilisées sont détaillées dans le troisième chapitre de cette thèse.

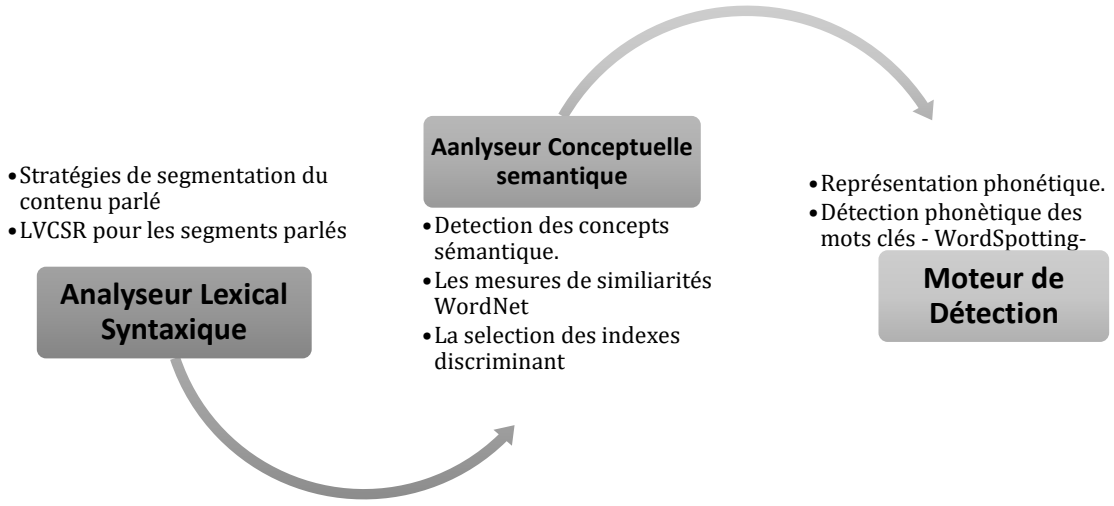

Figure 0-2 : Stratégie de l'approche d'indexation proposée [Bendib, 2018]

Dans ce contexte, cette contribution est basée sur l'hypothèse suivante :

« Si le système de recherche dans le flux parlé est suffisamment doté des connaissances nécessaires pour l'identification des concepts relatifs à un domaine ou un contexte quelconque. Ces informations peuvent-elles servir en collaboration avec d'autres ressources sémantiques (ontologies) une indexation automatique de ces ressources par un processus d'enrichissement basé sur une transcription automatique du contenu ».

Cette approche d'indexation sémantique présente plusieurs avantages avec une réduction importante des ressources utiles pour l'indexation du contenu multimédia dû à :

- *La transcription partielle du contenu multimédia :* il est évident que la tâche de transcription automatique du contenu multimédia est très couteuse. Alors qu'il est possible de déterminer le domaine ou le contexte général du contenu parlé des ressources multimédias à travers d'un processus de traitement linguistique sémantique sur des segments extraits du ces ressources.

- *L'utilisation de l'ontologie du domaine pour l'annotation du contenu de la primitive parole des ressources multimédias :* généralement les ontologies du domaine sont souvent des ontologies légères, ce qui nous aide à accélérer le processus de recherche des concepts qui correspond au contenu parlé de ces ressources.

- *L'enrichissement automatique des concepts décrivant le contenu parlé des ressources multimédias* : bien que la transcription du contenu parlé soit partielle pour des raisons d'optimisation du temps et de charge de calculs. Ainsi que pour des raisons liées aux difficultés et limites des systèmes de reconnaissance automatique à large vocabulaire de la parole. Il est souvent utile d'avoir un processus d'enrichissement de l'ensemble des termes obtenus lors de la phase de la transcription automatique partielle du contenu parlé. Ce processus exécute aussi des mécanismes d'alignement des termes relatives aux contenus parlé avec les distances sémantiques et hiérarchiques de l'ontologie utilisée. Cette phase a pour vocation de trouver les termes représentatifs candidat pour l'annotation du contenu du flux parlé des ressources multimédias.

- *L'application des outils d'indexation sémantique* : lors de la phase de rattachement des mots provenant de la transcription partielle du contenu parlé des ressources multimédias, dans le cas où le mot ne figure pas sur la hiérarchie de l'ontologie on tente la recherche par d'autres synonymes du mot par l'utilisation de l'ontologie *WordNet*[1].

[1] https://wordnet.princeton.edu/wordnet/

5. Organisation du manuscrit

Ce manuscrit est composé de quatre chapitres en incluant la présente introduction générale. Le premier chapitre introduit brièvement les concepts relatifs aux systèmes de recherche d'information, les techniques de représentation de l'information et des connaissances ainsi que les particularités des caractéristiques et structure des documents parlés.

Dans le deuxième chapitre, nous définissons les grandes lignes des techniques de reconnaissance automatique de la parole à large vocabulaire. Et nous présentons une étude bibliographique sur les recherches et les travaux effectués dans la discipline de détection de termes parlés à base phonétique. Entre temps, nous citons les travaux de recherches liées aux distances sémantiques dans le contexte de représentation des connaissances.

Le troisième chapitre est consacré à notre contribution dans cette thèse par la proposition d'une approche d'indexation sémantique sur le contenu des documents parlés (SOMI). Cette approche permet d'assurer une bonne qualité de recherche dans le contenu parlé à base d'une méthode d'indexation automatique. Nous détaillons dans ce chapitre les trois modules constituons notre approche : SLE, SCA et KDE.

Cependant, dans le quatrième chapitre nous présentons les stratégies utilisées pour l'implémentation et la validation de notre approche. Nous définissions les ressources parlées utilisées ainsi que les différents environnements et routines développés et les résultats obtenus avec les analyses et les synthèses appropriées. Enfin nous clôturons ce manuscrit par une conclusion générale sur les travaux réalisés ainsi que nos perspectives futures.

Chapitre 1

Vers la recherche dans les ressources parlées

1.1. Introduction

Actuellement, la quantité et le volume des documents hétérogènes comme les documents parlés, les documents numériques et les documents multimédias disponibles dans les plateformes de communication et réseaux informatiques sont très importants. Ces ressources numériques représentent une grande masse de connaissances et elles seront exploitées seulement lorsqu'elles sont structurées et accessibles. À cet effet, l'indexation et la recherche d'information sont devenues des tâches primordiales pour réaliser ces objectifs.

Entre autres, l'importante avance en matière de techniques de recherche d'information textuelle réalisées dans les dernières années. Ainsi que l'apparition de différents types des ressources numériques comme les ressources multimédias et les ressources parlées ainsi que l'évolution des techniques de stockage, le débit et de la puissance de calcul ont poussé l'émergence des nouvelles problématiques pour le développement des approches et techniques pour l'indexation et la recherche dans le contenu de ces ressources.

L'information multimédia est formée d'images, de vidéos et de bandes audio en plus du texte. Alors que l'extraction d'informations de haut niveau à partir des images fixes et animées est la technique utilisée souvent. Nous trouvons que le traitement de l'information parlée ; qui est l'objectif de ce manuscrit ; par l'utilisation des techniques de reconnaissance automatique de la parole pour transcrire et indexer le contenu parlé ne sont pas largement abordés. Ce qui rend cette piste de recherche très prometteuse pour la

recherche d'information dans le contenu parlé. En effet, l'indexation, la recherche et la détection des termes parlés sont des problématiques très étudiées, notamment lors des campagnes d'évaluation NIST[1] « *National Institute for Science and Technology* ».

1.2. Recherche d'information

Le concept « recherche d'information » a était utilisé pour décrire un domaine de recherche dédié aux techniques de représentation, stockage, organisation, et 'accès à des éléments d'information [Salton et al, 1983]. Ce domaine s'intéresse sur les méthodes, modèles et techniques utilisées pour la recherche d'information dans le contenu des documents et leurs descripteurs comme les métadonnées et les annotations.

Généralement, le scénario général de la recherche d'information consiste à restituer un ensemble de ressources ou de documents à partir d'une grande collection de données qui reflète un besoin utilisateur donnée. Ce besoin est représenté à l'aide d'un langage spécifique qui s'appelle langage requêtes et il est généralement cette représentation est une spécification incomplète de besoins utilisateurs. En effet, la concrétisation du besoin utilisateur par le langage de requêtes n'est pas évidente, car le besoin est souvent vague est non restrictif. Donc, l'objectif des systèmes de recherches d'information est de restituer toutes les ressources qui répond au mieux aux besoins utilisateurs classés en ordre de pertinence système.

1.2.1. Système de recherche d'information

Les Systèmes de Recherche d'information sont des tâches multidisciplinaires qui utilisent plusieurs domaines de traitement d'information. Les plus importants systèmes d'information automatisés sont les systèmes de gestion d'information, les systèmes de gestion de bases de données, les systèmes de support à la décision, les systèmes de question-réponse, ainsi que les systèmes de recherche d'information [Salton et al, 1983]. Les évolutions récentes des technologies d'information ont modifié ce paysage et multiplié les banques de données en texte intégral, les applications de la gestion électronique de documents et surtout les systèmes d'information multimédias.

1.2.2. Architecture des systèmes de recherche d'information

Les systèmes de recherche d'information « *SRI* » sont des systèmes qui assurent le stockage, l'indexation, la représentation et l'accès aux contenus des ressources d'information (les documents). Ils sont définis par un langage de représentation du contenu des documents, un langage de requêtes qui permet d'exprimer les besoins utilisateurs et une fonction de mise en correspondance du besoin de l'utilisateur et du corpus de documents en vue de fournir comme résultats des documents pertinents pour l'utilisateur [Tambellini,2007]. La figure 1.1 présente l'architecture globale d'un *SRI*.

[1] NIST: https://www.nist.gov/about-nist

De point de vue de l'architecture, il est composé de trois modules principaux : (Tambellini, 2007).

- Module d'indexation : l'indexation des documents et des requêtes.
- Module d'expression besoin utilisateur : la mise en correspondance requête-documents avec un ordonnancement des documents quand le modèle le permet.
- Module de recherche : la restitution des documents reconnus pertinents par rapport à la requête.

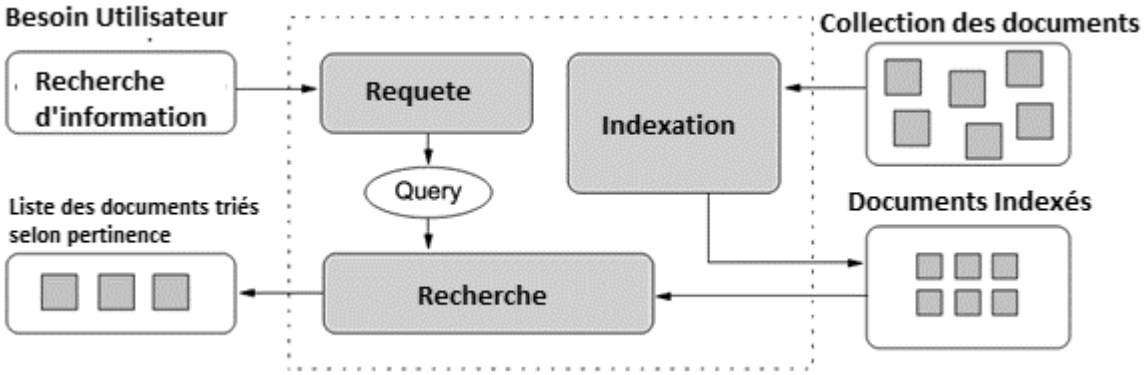

Figure 1-1 : Schéma illustrant les principaux modules d'un système de recherche d'information selon [Tambellini, 2007]

1.2.3. Le processus d'indexation

Dans les systèmes de recherche d'information, les documents sont définis comme des volumes d'information auto-explicatives. À cet effet, la première difficulté se réside sur les techniques utilisées pour trouver une définition précise pour les termes constituant ces documents afin de maximiser l'expressivité sémantique des termes. L'indexation consiste à trouver les termes qui représentent synthétiquement le contenu sémantique des documents, ces termes sont appelés « *index* ».

Les index sont représentés sous un format et un modèle spécifique. Le formalisme qui permet de représenter syntaxiquement un index est appelé le langage d'indexation ainsi que les techniques utilisées pour représenter l'expressivité des index sont appelées modèles de représentation. La complexité de ces formalismes peut variée d'une simple représentation d'un ensemble de termes jusqu'à l'utilisation des graphes conceptuels où les nœuds représentent des concepts et les arcs entre les nœuds représentent les relations existantes entre les différents indexes et concepts. Les mots ou les unités linguistiques qui représentent ces concepts sont soumis aux techniques de comparaison pour répondre aux besoins utilisateurs [Champclaux, 2010].

Afin qu'un système de recherche soit performant, il est important que les index utilisés représentent et reflètent au maximum le contenu des ressources de la collection originale. Indexer un document c'est d'extraire ses termes représentatifs afin de générer la liste des termes d'indexation susceptibles et construire avec ces indexes une liste de références de collection. Le terme référence représente le rôle de l'identifiant, c'est-à-dire un moyen de

rechercher et de retrouver d'une façon non ambiguë un ou plusieurs documents ou même des segments des documents dont ils contiennent des index ou des termes utilisés lors de la formulation d'une requête à partir d'un besoin utilisateur.

Cependant, le choix de ces termes n'est pas implicite, car nous trouvons parfois que les termes interdépendants sont sensibles au contexte et ne sont pas forcément des bons représentants dans un contexte différent [Champclaux, 2010]. Pour cela, afin d'extraire ces index, il faut qu'on exécute plusieurs étapes de prétraitements. Nous résumons ces étapes dans les sections suivantes.

1.2.3.1. Le prétraitement lexical

L'objectif de cette étape est l'extraction des termes à partir du contenu d'un document textuel. Cette extraction est réalisée après quelques prétraitements classiques comme l'élimination des signes de ponctuation et la casse.

Cependant, pour capter les termes représentatifs de contenu, plusieurs techniques peuvent être utilisées comme l'élimination des mots vides ou les mots fonctionnels de la langue qui sont appelés souvent « stop word » [Luhn, 1959]. Car ces mots qui ont tellement communs qu'ils sont inutiles de les indexer ou de les utiliser comme des références dans le processus de recherche. Outre ces mots de liaison, nous citons aussi l'élimination des mots fréquemment utilisés qui deviennent aussi des mots vides à cause de leur fréquence d'apparition et ces densités élevées. Il est évident que le rôle d'un mot dans un document dépend de contexte dans lequel il est employé, et qu'il peut avoir un pouvoir d'information différent dans un autre contexte. Cependant, cette liste de mots doit être dépendante de la collection utilisée et son contexte. En pratique, nous utilisons souvent le concept anti dictionnaire pour représenter cette liste de mots. Parmi eux, nous citons comme exemple : « *Snowball stop word*[1] » , « *Terrier stop word*[2] », « *XPO6*[3] ».

1.2.3.2. Indexation par radicaux « Stemming »

Dans les sciences de l'information et linguistique, un mot peut avoir deux possibilités de variations : soit morphologiques [Frakes, 1992] ou sémantiques [Paice,1996]. Les variantes morphologiques des mots ont la plupart du temps un sens très proche. Par exemple, il peut être utile à retrouver des documents contenant les mots « *détection* », « *détecteur* », « *détecter* », « *détectable* » à partir d'une requête comportant le mot « *détection* ». Pour cela, il est possible d'éliminer les différences non significatives et de garder la partie commune qui présente son radical.

Cependant, dans le processus d'indexation par radicaux, il faut représenter plusieurs variantes d'un mot sous une forme unique appelée racine ou radical. Nous trouvons dans la littérature que la différence morphologique entre la racine et le radical est faite : la racine est la forme abstraite servant de base de représentation de tous les radicaux communs. En

[1] http://snowball.tartarus.org/algorithms/english/stop.txt
[2] https://github.com/RxNLP/text-mining-and-nlp-apis/blob/master/terrier-stop-word-list.txt
[3] http://xpo6.com/list-of-english-stop-words/

effet, le radical d'un mot est une simple réduction du nombre de lettres de ce mot, et celui-ci peut différer que la racine morphologique correcte. Par exemple, le mot « *computation* » peut être représenté par plusieurs radicaux « *computa* », « *comput* », « *compu* », sa racine linguistiquement correcte étant « *compute* ».
En revanche, les algorithmes qui permettent de transformer un mot vers son radical sont appelés les algorithmes de radicalisation « *Stemming* ». En effet, ces algorithmes de radicalisation peuvent être linguistiques, par exemple l'algorithme de « Porter » [Porter, 1980]. En pratique, les études effectuées dans cette piste ont montré qu'il est difficile d'évaluer et de comparer les différents algorithmes de radicalisation pour les besoins de la RI. Les travaux effectués utilisent souvent un seul algorithme de radicalisation, ce qui ne permet pas de prouver l'impact de l'algorithme utilisé sur les résultats obtenus [Champclaux, 2010]. Néanmoins, l'idée résultante est que l'usage de la radicalisation est bénéfique pour les performances des systèmes de recherche.

1.2.3.3. Indexation par lemmes « Lemmatization »

L'indexation par lemmes est une méthode plus fine qui se concentre exclusivement sur l'impact des catégories grammaticales sur les termes, en amenant tous les mots vers leur lemme. L'avantage de cette méthode par rapport à celle basée sur les radicaux concerne surtout les formes courtes et/ou irrégulières, et les verbes. Notons aussi que cette méthode utilise les concepts du lexique morphologique et nécessite une opération préalable de catégorisation des unités lexicales comme par exemple le terme « détectable » sera lemmatisé en « détecter » parce qu'il est son adjectif. À cet effet, plusieurs types d'erreurs peuvent être générées dans ce processus, notamment sur les mots inconnus du lexique de référence utilisé et les unités ambiguës sur le plan catégoriel.

Entre autres, l'indexation par lemmes entraîne apriori une augmentation le nombre de documents pertinents restitués par rapport à l'indexation par mots simples, mais dans une moindre mesure que l'indexation par radicaux. En pratique, les algorithmes qui permettent de convertir un mot à son radical racine sont appelés les algorithmes « *lemmatisers* ». Ils peuvent être automatiques quand ils se basent sur des méthodes statistiques par exemple les *n-grammes* [Adamson, 1974] ou être hybrides comme les algorithmes [Paice, 1996]. Ils peuvent également se baser sur des lexiques afin de valider ou d'invalider une tentative de transformation d'un mot en radical [Savoy, 1993].

1.2.3.4. La pondération des termes

La pondération des termes dans le processus d'indexation permet d'associer un poids de représentativité $w_{i,j}$ pour chaque terme t_j d'un document d_i. De manière générale, les formules de pondération utilisées sont basées sur la combinaison d'un facteur de pondération local quantifiant la représentativité locale du terme dans le document, et d'un facteur de pondération global quantifiant la représentativité globale du terme vis-à-vis de la collection de documents. Les différentes approches et techniques de pondération sont détaillées dans les travaux effectués par Salton [Salton, 1993].

Dans ce contexte, plusieurs formules existent, nous citons :

- *La formule tf-idf :* Cette mesure représente une bonne approximation de l'importance d'un terme dans un document par rapport au corpus cible. Elle est particulièrement utilisée dans des corpus de documents de tailles intermédiaires. Elle est calculée comme suit [$Salton, 1987$]:

$$w_{i,j} = \frac{tf_{i,j}}{df_j} = tf_{i,j} \times \frac{1}{df_j} = tf_{i,j} \times idf_j \quad (1\text{-}1)$$

Où :

$tf_{i,j}$: est la fréquence d'occurrences du terme t_j dans le document d_i

df_j : est la fréquence documentaire du terme t_j

idf : est la fréquence documentaire inverse

- *La normalisation pivotée de Singhal :* Cette mesure présente l'adaptation de la mesure précédente pour les documents les plus long, elle est calculée par la formule suivante [$Singhal\ et\ al., 96$]:

$$w_{i,j} = \frac{tf_{i,j} \times idf_j}{1 + \frac{slope}{(1 - slope) * pivot} * \sqrt{\sum_j (tf_{i,j} * idf_j)^2}} \quad (1\text{-}2)$$

Où :

$tf_{i,j}$: est le nombre d'occurrences du terme t_j dans l'unité documentaire d_i

idf_j : est la fréquence documentaire inverse, elle est définie par $log(n/N_j)$ avec N_j représente le nombre de documents indexés par le terme t_j

$pivot$: est un paramètre qui représente la valeur d'écart entre les probabilités de pertinence et de sélection

$slope$: est un facteur de normalisation fixé pour minimiser l'écart entre la pertinence et la sélection.

- *L'adaptation de Robertson :* Cette adaptation permet le calcul de la pondération des termes en se basant sur des petites quantités d'informations pertinentes, elle est calculée par la formule suivante [$Robertson\ et\ al., 97$]:

$$w_{i,j} = \frac{tf_{i,j} \times (K_1 + 1)}{K_1 \left[(1 - b) + b \times \frac{dl_i}{\Delta l}\right] + tf_{i,j}} \quad (1\text{-}3)$$

Où :

K_1 : est une constante qui permet de contrôler l'impact de la fréquence du terme t_j dans le document d_i.

b est constante qui permet de contrôler l'effet de la longueur du document.
dl_i est la longueur du document d_i.
Δl est la longueur moyenne des documents dans la collection entière.

1.2.4. Les approches d'indexation

L'indexation des ressources est une étape primordiale dans les systèmes de recherche d'information. Cependant, cette tâche comme la décrit la figure 1-2 est difficile à réaliser puisqu'elle permet de trouver la définition des index qui seront considérés comme des termes représentatifs du contenu des ressources. Techniquement, l'indexation peut être manuelle, automatique ou semi-automatique, dans les littératures nous trouvons trois formes d'indexation, qui seront détaillées dans les sections suivantes :

.

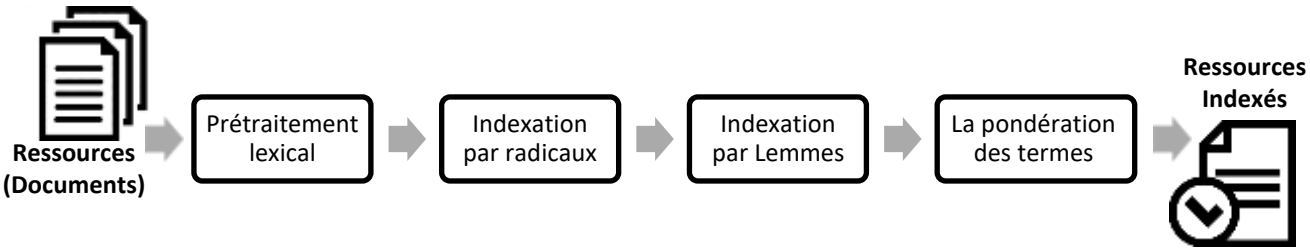

Figure 1-2 : Schéma illustratif des étapes de processus d'indexation d'un document

1.2.4.1. L'indexation manuelle

L'indexation manuelle est une tâche humaine de traitement intellectuel du contenu des documents à l'aide d'une représentation compacte basée sur un langage d'indexation. Cette représentation facilite l'accès ainsi que la recherche d'informations dans le contenu de ces ressources. Ce mode d'indexation est pratiqué généralement par les professionnels des bibliothèques et les experts des domaines en se basant sur des outils linguistiques spécifiques tels que les thésaurus[1] et les répertoires. Il s'agit donc d'une indexation à la fois humaine, manuelle contrôlée par des langages documentaires.

Dans ce contexte, nous trouvons dans les littératures des travaux basés sur une indexation manuelle limitée à des termes simples qui sont choisis manuellement et fournissent de bonnes performances comparons à celles qui utilisent les listes de vocabulaire contrôlé (Thésaurus) [Savoy, 2005]. Cependant, l'utilisation des thésaurus repose sur des règles d'indexation qui rendent l'indexation spécialisée et exhaustive, mais la construction et la maintenance des thésaurus posent des problèmes de coût, tout comme l'indexation. En revanche, le nombre important de documents stockés sur les supports électroniques ne permet pas de recourir aux telles techniques d'indexation, ce qui impose de passer vers les méthodes automatiques.

[1] https://fr.wikipedia.org/wiki/Thésaurus

1.2.4.2. L'indexation automatique

L’indexation automatique résout les problèmes de coût de point de vue des ressources humaines et les listes de vocabulaire contrôlé qui sont indispensables dans le processus d’indexation manuelle. L’indexation automatique repose sur l’extraction à base statistique des termes d’indexation à partir de leurs contenus. Ce type d’indexation est composée de deux étapes : la recherche des termes caractérisant le contenu et l'évaluation du pouvoir de caractérisation de ces termes. Cependant, les techniques d’indexation se diffèrent selon certain nombre de paramètres définis lors des étapes de l’indexation telle que :

- Le choix du type d'élément qui constituera les unités d'indexation (radical, mot simple, groupe de mots),
- La définition des règles d’équivalence entre termes issus des documents et termes d’indexation (radicalisation, lemmatisation, troncature, etc.),
- Le principe de sélection des termes représentatifs du document et ceux qui ne le sont pas, en fonction du contenu du document (termes d’indexation),
- La fonction déterminant le pouvoir de caractérisation des termes d’indexation : certains termes sont plus importants que d'autres dans la caractérisation du contenu.

Le résultat de ce type d'indexation est un ensemble de couples (terme d'indexation, poids) associés à chaque document. Ce type d’indexation à l’avantage de sa facilité d’implémentation et d’intégration dans les systèmes automatiques. Néanmoins, cette approche doit confronter quelques problèmes comme :

- La correspondance entre les termes d'indexation et les unités sémantiques.
- Les métriques d’équivalences utilisées entre les termes.
- Les ambigüités entre les termes engendrés par le traitement automatique.
- L’impact de l’assomption que les termes sont considérés comme indépendants dans le processus de pondération
- L’exhaustivité et la spécificité de l’indexation ne sont pas garanties, car les termes d’indexation sont choisis par rapport à leur pouvoir discriminant et en fonction de leurs fréquences d’apparition.

1.2.4.3. L'indexation linguistique

L’objectif de cette stratégie est l’amélioration de la qualité de l’indexation par l’utilisation des informations sémantiques obtenues à l'aide des techniques de traitement de langage naturel conjointement avec celles obtenues par les techniques statistiques. Cette stratégie permet l’amélioration de la précision des recherches, en réduisant le nombre de fausses alertes. L'information sémantique est extraite à partir d’un traitement du langage plutôt que de traiter chaque mot comme une entité statistique indépendante. La sortie la plus simple de ce processus est la génération des indexes sous forme de phrases pour les ressources. Aussi, une analyse plus complexe génère une représentation thématique des ressources plutôt que des phrases. Contrairement aux approches statistiques qui utilisent la proximité comme base pour déterminer la force des relations de mots dans la génération des indexes,

l'indexation a base syntaxiques et sémantiques générées par des algorithmes de traitement du langage naturel améliorent la spécification de l'indexation en fournissent un autre niveau de désambiguïsation tel que les distances sémantiques. Le traitement du langage naturel peut également combiner les concepts par de agrégations vers des concepts de niveau supérieur parfois appelés représentations thématiques. Ces traitements permettent de représenter les indexes sous formes conceptuelles comme les triplets *« concept-relation-concept »* [Kowalski, 2010].

1.2.5. L'interrogation et la recherche

Le processus crucial d'un système de recherche d'information est la comparaison de la représentation du contenu de la requête avec les modèles de représentations des documents L'utilisateur exprime son besoin en information par une requête dans la forme imposée par le système (SRI), ce processus est connu par la formulation de la requête. Comme dans le cas des représentations de documents, la requête doit être capable de capturer les informations importantes contenues dans la requête d'origine sous une forme qui lui permettent d'être comparée à la représentation du document. La requête est utilisée par le système de recherche pour sélectionner les documents pertinents de la collection.
La réponse du système est un ensemble de références triées à des documents qui obtiennent une valeur de correspondance élevée. Cet ensemble est généralement présenté sous la forme d'une liste ordonnée suivant la valeur de correspondance. Cette réponse est réalisée par le biais de la sélection et le classement des ressources via des mesures de similarité qui calculent la similarité entre les éléments de la requête et les ressources indexées. En revanche, l'utilisation de techniques d'expansion de requêtes permet l'amélioration de la pertinence des résultats en se basant sur les résultats des recherches antérieurs. En effet, le système de recherche traduit la requête dans son langage d'indexation. Ce processus est similaire à l'indexation des ressources. Il utilise certaines mesures statistiques telles que : la pondération, la fréquence du document et la fréquence totale pour un terme spécifique. Fréquemment dans un système d'indexation de concepts ou les indexes sont déterminés en appliquant des algorithmes statistiques à un échantillon représentatif du corpus [Kowalski, 2010]. Entre autres, les techniques d'indexation basées langage naturel ont tendance à utiliser les algorithmes les plus indépendants des corpus et des techniques d'expansion par des ressources sémantiques comme les thésaurus et les ontologies. Le tableau 1-1 illustre les différentes phases du processus de recherche. La parenthèse est utilisée dans la deuxième étape de liaison pour indiquer une expansion par un thésaurus.

Données	Action
« Find me information on the impact of the oil spills in Alaska on the price of oil »	*Requête de recherche d'utilisateur formuler avec son propre vocabulaire*
impact, oil (petroleum), spills (accidents), Alaska, price (cost, value)	*Représentation du requête utilisateur avec le langage système (Langage d'indexation)*
impact (.308), oil (.606), petroleum (.65), spills (.12), accidents (.23), Alaska (.45), price (.16), cost (.25), value (.10)	*Pondération des termes de la requête basée sur la fréquence de document inverse et les fréquences des termes*

Tableau 1-1 : Exemple de représentation d'une requête utilisateur dans un système de recherche d'après [Kowalski, 2010]

1.2.6. L'évaluation des performances des SRI

Les performances des systèmes de recherche d'information sont mesurées et évaluées sur plusieurs dimensions. En effet, les facteurs de déploiement tels que le coût de mise en œuvre et la maintenance, la stratégie d'indexation des nouvelles ressources et la performance de recherche sont importantes.

Mesure	Description	Définition
Rappel	Rappel exact par rapport à l'ensemble des documents retrouvés. Le rappel mesure la capacité du système à restituer l'ensemble des documents pertinents (en lien avec le silence documentaire).	$R = \frac{Nb\ de\ Doc\ pertinents\ retrouvés}{Nb\ de\ Doc\ Trouvés}$
Precision	Précision moyenne non interpolée par rapport à l'ensemble des documents pertinents. Mesure la capacité du système à ne restituer que des documents pertinents	$P = \frac{Nb\ de\ Doc\ pertinents\ retrouvés}{Nb\ de\ Doc\ pertinents}$
MAP	Elle calcule la précision et le rappel à chaque position dans la séquence classée de documents pour les systèmes qui renvoient une séquence classée de documents	$MAP = \frac{1}{\lvert R\rvert}\sum_{i=1}^{n} precision(i).relevance(i)$
F1-Score	Le score F1 est la moyenne harmonique de la précision et du rappel, où un score F1 atteint sa meilleure valeur à 1 (précision parfaite et rappel) et le pire à 0.	$F1 = 2 * \frac{P \times R}{P + R}$
Fall-Out	Une mesure liée directement liée à la détection des documents non pertinents. Elle présente la proportion de documents non pertinents par rapport au total des documents non pertinents.	$Fallout = \frac{Nb_{Doc} Non\ Pert\ \ trouvés \cap Nb_{Doc}\ trouvée}{Nb_{Doc}\ Non\ Pertinent}$

Tableau 1-2 : Tableau récapitulatif des critères d'évaluation les plus populaires dans les SRI

Cependant, les critères utilisés pour l'évaluation de performance des SRI les plus populaires sont basés sur des mesures statistiques relatives au nombre de ressources restituées par rapport aux requêtes utilisateurs. Dans le tableau 1-2 nous présentons les métriques les plus souvent utilisées avec une brève description ainsi que leurs règles de calculs.

Cependant, ces mesures définies nécessitent des standards et des ressources qui permettent l'évaluation ainsi que la comparaison des différents systèmes de recherches d'information. En effet, le jugement concernant la pertinence des ressources est orienté utilisateurs. Bien que la pertinence soit la métrique primordiale pour évaluer un système, il est quasiment évident qu'il faut construire des corpus et de ressources pour valoriser ces mesures. À cet effet plusieurs institutions de recherches ont travaillé sur l'élaboration des ressources et des normes pour cette fin. Dans le tableau 1-3, nous présentons quelques ressources et collections utilisées pour l'évaluation des SRI

Nom	Contenu	Description
Cranfield Collection	Résumés d'articles extraits à partir des revues d'aérodynamiques	Pilote pour les tests élémentaire des SRI pour les jugements de pertinence des paires (requête, document).
TREC Text Retrieval Conference NIST	Enregistrement des actualités diffusé dans les médias	La collection de données les plus utilisés pour les systèmes de recherche ad hoc
CLEF Cross Language Evaluation Forum	Collection d'actualités diffusées par la chaine médiatisée *Reuters.*	Dédié pour les Traitements automatique du Langage Naturel
ClueWeb	Collection des pages Web dans la langue Anglaise	La plus grande collection Web facilement disponible à des fins de recherche
RCV1 Reuters	Un corpus d'articles de presse récemment mis à disposition par Reuter	Les meilleurs ensembles de données de test pour la classification (catégorisation)
NTCIR NII Test Collections	Basé sur le modèle TREC mais pour les langues de l'Asie de l'Est	Recherche d'information en multi-langue
LDC Linguistic Data Consortium	Des collections des ressources linguistiques et logiciels et matériels pour les ressources parlés	La création et le partage des ressources linguistique et normes pour le traitement automatique de la parole

Tableau 1-3 : Un panorama des collections et ressources utilisées pour l'évaluation des SRI

1.3. Les ressources multimédias

Le terme ou le concept multimédia s'est apparu et utilisé dans la fin des années 1980 pour désigner une création innovante des modèles de représentation pour les ressources médiatisées tel que : les images fixes ou animées, la musique et la parole et les séquences vidéo. Ensuite, il a englobé les concepts : logiciels, matériels et contenus interactifs mettant en œuvre l'image fixe ou animée, le son, le texte et l'hypertexte. Le terme vient du pluriel du mot latin « medium » (en anglais Cross-media). Les documents multimédias peuvent se présenter sous différents formats adaptés au contenu et au contexte. Les types de documents multimédias les plus populaires sont : l'image, la parole et la vidéo.

1.3.1. Image

L'image[1] est une représentation visuelle et mentale d'un objet observable. Elle peut être naturelle (ombre, reflet) ou artificielle (peinture, photographie), conceptuelle ou palpable. Entretemps, elle est dite numérique lorsque sa sauvegarde est obtenue sous forme binaire par des techniques de photographie, du vidéo ou l'utilisation des logiciels spécialisés. Techniquement, l'image numérique est constituée d'une matrice avec un nombre fini de lignes où chaque ligne contient un nombre de points appelés « pixel ». Cette matrice contient les caractéristiques da la couleur attribuée aux pixels selon la nature d'image : couleur ou nuance de gris. Cependant, une image peut avoir plusieurs tailles différentes selon la nature de la codification utilisée pour sa sauvegarde telle que : la taille, la résolution, la profondeur, type de compression …etc.

Entres temps, nous trouvons que le concept « les nouvelles images » : qui présente les images de synthèse produites par les systèmes informatiques. Cette présentation utilise les techniques de trois dimensions et les standards de cinéma ainsi que les techniques de la réalité virtuelle et augmentée. Des logiciels de plus en plus puissants et performants permettent la création de tel univers virtuel et peuvent traiter et manipuler les images pour basculer dans les univers du « réel » et « visuel ».

1.3.2. Son, Signal et parole

L'alternative liée aux images dans le multimédia est l'événement sonore. Ce dernier est modélisé par les concepts : signal, son et parole. Le concept « signal » désigne la représentation physique de l'information qu'il convoie de sa source à son destinataire et il est considéré comme une grandeur électrique. Le concept « son » représente l'ébranlement élastique de l'air, d'un fluide ou d'un solide qui se manifeste par des variations de pression autour de la pression moyenne du milieu [Bellanger, 2008]. Ainsi, le concept « parole » représente l'action volontaire et coordonnée d'un certain nombre de muscles du système articulatoire afin de produire un signal de parole.

1 http://fr.wikipedia.org/wiki/Image

Cependant, ces concepts fournissent des ondes sonores qui seront représentées par des courbes graphiques modélisant les variations de la pression de l'air en fonction du temps comme la présente la figure 1-3. Entre autres, l'utilisation de ce type de multimédia (les ondes sonores) nécessite un traitement ou une numérisation afin de les exploiter dans les systèmes automatiques. Dans les sections qui se suivent, nous présentons une brève description sur ces techniques.

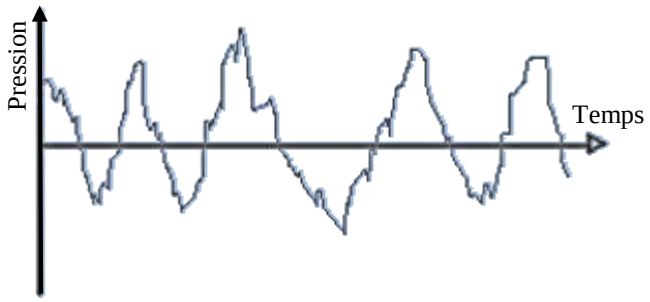

Figure 1-3: Exemple d'une représentation graphique des ondes sonores.

1.3.2.1. Echantillonnage du son

La représentation numérique des signaux sonores et paroles s'effectue par le biais d'une conversion ou discrétisation vers des valeurs numériques. Cette étape permet de relever des petits échantillons du signal dans des intervalles de temps réparti uniformément. On appelle cette action l'échantillonnage ou la numérisation du son. L'intervalle de temps entre deux échantillons est appelé taux d'échantillonnage. En effet, pour arriver à stimuler un signal continu audible à l'oreille il faut des échantillons tous les quelques 100.000èmes de seconde. Pratiquement, le nombre d'échantillons par seconde est exprimé en Hertz (Hz). Dans le tableau 1-4 nous présentons quelques exemples de taux d'échantillonnage standards utilisés pour quelques types des fichiers sonores utilisés.

Taux d'échantillonnage	Qualité du son
44 100 Hz	Qualité CD
22 000 Hz	Qualité radio
8 000 Hz	Qualité téléphone

Tableau 1-4 : Exemples de taux d'échantillonnage et de qualités de son associée

1.3.2.2. Représentation physique des signaux sonores

Lors de l'opération d'échantillonnage, chaque échantillon (correspondant à un intervalle de temps) est associé une valeur qui détermine la valeur de la pression de l'air à ce moment, le son n'est donc plus représenté comme une courbe continue qui présente les variations, mais il est présenté par une suite de valeurs pour chaque intervalle de temps, la figure 1-4 présente un exemple d'échantillonnage.

Figure 1-4 : Exemple d'une représentation de l'échantillonnage d'un signal sonore

Pratiquement, la numérisation et l'utilisation du support informatique signifient une représentation binaire. Donc, il faut fixer l'intervalle des valeurs d'échantillonnage. Ceci revient à fixer le nombre de bits sur lequel on code les valeurs des échantillons.

- Avec un codage sur 8 bits, on a 2^8 possibilités de valeurs, c'est-à-dire 256 valeurs possibles.
- Avec un codage sur 16 bits, on a 2^{16} possibilités de valeurs, c'est-à-dire 65536 valeurs possibles.

En effet, avec la deuxième configuration, on aura bien évidemment un signal sonore avec une qualité bien meilleure, mais aussi un espace de stockage plus élevé.

Un signal numérique est donc représenté physiquement par plusieurs paramètres comme :

- La fréquence d'échantillonnage
- Le nombre de bits d'un échantillon
- Le nombre de voies (une seule correspond à du mono, deux à de la stéréo, et quatre à de la quadriphonie)

1.3.3. Vidéo :

La vidéo[1] est une série successive d'images avec certaines cadences. Physiquement, la capacité de l'œil humain permet de distinguer autour de 20 images différentes par seconde. Donc, le balayage de plus de 20 images par seconde permet de stimuler l'animation et la création des séquences vidéo. En pratique, nous caractérisons la fluidité d'une vidéo par le nombre d'images balayer par secondes (en anglais frame rate), exprimées par FPS (*Frames Per Second*), en français se traduit par (*Trames Par Seconde*). Entre autres, le concept « vidéo » reflète le sens du terme multimédia et il est généralement composé des séquences vidéo accompagnées par des signaux sonores, c'est-à-dire des données audios.

[1] http://www.commentcamarche.net/contents/1493-introduction-a-la-video-numerique

1.3.3.1. Les caractéristiques du vidéo numérique

Généralement, tous les formats vidéo utilisés possèdent le même objectif de base. Ils servent à stocker les informations relatives sur les deux modes : noir et blanc ou couleur sous la forme numérique constituant une trame vidéo. Le nombre de trames vidéo enregistrées par seconde varie en fonction de la norme vidéo prise en charge par le format utilisé comme : le format NTSC (*29 FPS*) ou le format PAL (*25 FPS*). Cependant, nous trouvons plusieurs caractéristiques du format de la vidéo numérique. Nos citons parmi eux :

- Le support utilisé pour le stockage : Il s'agit principalement d'une bande vidéo, mais il peut également s'agir d'un disque optique, d'un disque dur …etc.
- La norme vidéo prise en charge : Par exemple, NTSC, PAL, ATSC (HDTV 1080i ou 720p), …etc.
- Le rapport d'aspect de la trame vidéo : Le rapport entre la largeur du cadre et la hauteur du cadre.
- Les dimensions de la trame vidéo : Le nombre de pixels par ligne, et le nombre de lignes par trame.
- Méthode d'enregistrement des couleurs : RVB, composant (YUV), S-Vidéo (Y/C) ou composite.
- Profondeur de bits : Le nombre de bits utilisés pour stocker chaque échantillon de la vidéo numérique. Il détermine la capacité du format à capturer précisément l'intensité de la lumière de chaque échantillon (ou pixel).
- Le type de compression : La technique de compression permet de réduire la quantité de données numériques requises pour stocker chaque image sans compromettre la qualité de l'image.

1.3.3.2. Les Types de compression

Il existe plusieurs méthodes de compression pour les fichiers vidéo qui varie entre une simple réduction de taille de chaque image vers l'utilisation des algorithmes de compression « codec[1] ». Les codecs sont fournis soit par des ressources matérielles soit par des logiciels avec des taux de compression fixe ou variable. Aussi, certains codecs permettent une souplesse de configuration par la définition d'un attribut de qualité qui contrôle le débit autorisé. Le tableau 1-5 présente quelques types de codecs vidéo ainsi que leurs domaines d'application.

[1] Un codec est un dispositif ou un programme informatique pour coder ou décoder un flux ou un signal de données numériques.

Format	Résolution	Type de compression	Débit	Applications
MJPEG	720 x 486	Intra- trame	0,5 – 25 Mo/s	Générale
MPEG-1	352 x 240	Intra- trame	0,01 – 0,06 Mo/s	CD-ROM, Internet
MPEG-2	720 x 480	Intra trame et inter trame	0,01 - 2 Mo/s	DVD, TV satellite
DV	720 x 480	Intra- trame	3,5 Mo/s	Grand public, professionnelles, Télédiffusion
D1	720 x 486	Aucun	25 Mo/s	Télédiffusion

Tableau 1-5 : Exemples de quelques codecs vidéo avec domaines d'application

1.4. Les documents multimédias

1.4.1. Caractéristiques des documents multimédias

Généralement, le terme « documents multimédias » désigne les documents qui sont constitués d'images ou des signaux sonores et des textes. Bien que la vidéo soit le type le plus complexe, car il peut englober les autres types. En effet, l'exploitation de ce type complexe nécessite l'intégration des méta données sur leurs contenus, ces méta données sont souvent appelées signature. Ces signatures extraites ou intégrées présentent les caractéristiques des images, du signal sonore ou du texte utilisés pour l'annotation. D'une manière générale, la représentation d'un document multimédia consiste à lui attribuer une signature numérique qui le décrit d'une façon précise. Dans le cas d'une recherche, c'est cette signature qui permettra la localisation des séquences. Dans le cas d'une navigation, les indexes permettent la bonne classification et dans le cas d'un résumé, elle permettra la recherche dans leurs contenus [Youssoufou, 2009].

Dans les littératures, nous pouvons identifier trois grandes familles de caractéristiques utiliser pour l'exploitation des documents multimédias [Scuturici, 2002] :

- Les caractéristiques physiques du document : Sa taille, sa date de création, sa durée, type d'encodage …etc.
- Les caractéristiques sur le contenu (appelées aussi caractéristiques de haut niveau) caractérisant la description ou la sémantique du contenu du document multimédia ; elles sont généralement extraites manuellement du texte associé à ce dernier.

- Les caractéristiques visuelles (appelées également des caractéristiques de bas niveau) : les traits visuels que l'on peut extraire du multimédia tel que la couleur, la texture, les locuteurs, dialecte utilisé …etc. Elles permettent d'effectuer l'indexation sur le contenu de ces ressources.

1.4.1.1. Les caractéristiques physiques

Similairement à toute ressource numérique ou fichier informatique, les documents multimédias sont dotés d'un ensemble de caractéristiques physiques basique tel que : le nom du document, le chemin relatif, la taille, la date de mise à jour …etc. Ainsi que des caractéristiques particulières telles que le format de compression utilisé, la longueur de la séquence en secondes, la résolution des images, la qualité de son, la fréquence d'échantillonnage…etc. Ces caractéristiques sont très importantes dans toutes les opérations effectuées sur les documents multimédias comme : l'enregistrement, la restitution, l'organisation, le classement, l'indexation, la recherche, la visualisation…etc.

1.4.1.2. Les caractéristiques relatives au contenu

Les caractéristiques relatives au contenu ou les caractéristiques sémantiques et sont appelé aussi les caractéristiques de haut niveau sont les informations et les données susceptibles de décrire d'une manière fidèle et efficace le contenu de ces ressources. Généralement, ces caractéristiques sont obtenues à l'aide des éditeurs de ressources elles-mêmes ou par le biais des experts par un processus de transcription manuelle selon le domaine cible de ces ressources.

Généralement, ces caractéristiques sont obtenues à partir de plusieurs sources comme :

- Le texte encapsulant le multimédia : ce texte porte généralement une description du contexte ou même une transcription adéquate du contenu de la ressource multimédia.
- Les annotations : on peut attribuer une sémantique à une ressource multimédia en l'annotant. Le processus d'annotation est un procédé indispensable dans l'indexation et la recherche dans le contenu des ressources multimédias. Il permet l'affectation à chaque ressource multimédia un certain nombre de mots-clés ou de phrases qui décrivent au maximum son contenu, et cela de façon manuelle, automatique ou semi-automatique. Le texte utilisé pour l'annotation peut à son tour servir aux procédés d'indexation de ces ressources [Youssoufou, 2009].
- La parole et les textes extraits des ressources multimédias : dans le cas où la ressource qui contiennent de la parole et du texte à l'intérieur (ex. vidéo dont certaines scènes sont décrites avec du texte). On peut extraire ces informations pour les exploiter dans le processus d'indexation.

1.4.2. L'indexation automatique des documents multimédias

Actuellement, il y a un besoin commercial croissant pour les systèmes de recherche d'information d'aider les utilisateurs à s'organiser et à trouver des informations dans les grandes quantités des ressources multimédias qu'ils stockent sur leurs systèmes locaux et dans le Web. En effet, les utilisateurs s'attendent à pouvoir obtenir dans les systèmes de recherche d'informations pour les ressources multimédias une précision similaire à celle dans les ressources textuelles. Les demandes de recherches et d'extraction des informations dans le contenu des ressources multimédias continueront à croître en tant qu'une nouvelle tendance dans les SRI en tenant compte l'évolution et le développement des nouvelles techniques de recherche.

1.4.2.1. Indexation des ressources multimédias

L'indexation des ressources multimédias présente quelques particularités par rapport aux techniques d'indexation dans les ressources classiques. En effet, elle n'est pas une simple conversion directe à la structure classique d'indexation, mais elle est beaucoup plus l'application des algorithmes aux structures numérique complexe pour extraire l'unité de traitement des différentes modalités (Son, vidéo …etc.) qui seront utilisées pour représenter le contenu de ces derniers.

Pour les images les premiers travaux effectués ont constaté que la recherche du texte associé aux images était plus précise que la recherche des images elles-mêmes. Ensuite, la recherche de ces modalités commence à devenir plus réalisable et fournira de meilleurs résultats que le texte [Kowalski, 2010]. Tandis que, la recherche dans la ressource parlée prend de plus en plus d'importance. Ainsi que, la vidéo est considérée comme une modalité qui synchronise les séquences d'images et les pistes audio.

1.4.2.2. Indexation des ressources audio

Généralement, l'indexation des ressources audio s'effectue après un processus de prétraitement sonore, elle utilise souvent des techniques de reconnaissance automatique de la parole basée sur les phonèmes, les modèles de Markov Cachés et les modèles de langages (ces techniques seront détaillés dans le chapitre suivant). Pratiquement, plusieurs ressources et corpus ont été collectés et traités par le consortium LDC[1] afin de permettre le développement des systèmes de recherches dont les indexes sont des unités phonétiques comme les mots et les phonèmes.

En effet, Il y a deux approches principales d'indexation et de recherche dans les ressources parlées : la première dite « *Text based Continuous Speech Recognition* » et la deuxième dite « *Phonetic Search* ». Dans le première approche, on effectue une transcription intégrale du contenu de la ressource parlé on exploitants les performances offertes par les systèmes de reconnaissance automatique a large vocabulaire de la parole continue « *LVCSR* ». Une fois la ressource parlée est convertie en texte, nous appliquons les

[1] https://www.ldc.upenn.edu/

techniques d'indexation utilisées. Cependant, le problème majeur de cette approche se réside dans les performances des *LVSCR* utilisées et surtout pour les langues moins dotées. En effet, les erreurs de reconnaissances soient par ajout « insertion » ou suppression « omission » influente sur la qualité d'indexation. En pratique, nous trouvons des taux d'erreurs de 10% pour la transcription des infos radiodiffusées « *Broadcast news* » et 40% pour la transcription des discours conversationnels ou plus de mots erronés [Kowalski, 2010]. Pour cela, des mécanismes ont été utilisés pour réduire l'impact de ce phénomène par l'utilisation des ressources manuelle ainsi que les méthodes statistiques.

Tandis que pour la deuxième approche, elle est basée sur la détection des modèles phonétiques dans le contenu de la ressource parlée. En effet, l'aspect exclusif d'un système de recherche à base phonétique est le modèle de phonème utilisé pour la création des indexes, stockage et la recherche. Ce dernier est un dictionnaire phonétique qui assure la correspondance entre les indexes et termes parlés et textuels dans le processus de recherche.

Cependant, de point de vue avantages et des inconvénients, nous trouvons que les deux approches sont sensibles à la qualité et à la quantité des ressources utilisée pour la construction des modèles linguistique et phonétique. Entre autres, le style de parole utilisé influe sur la qualité des systèmes de recherches, nous trouvons des systèmes avec un taux de précision de 90% pour les ressources d'émissions médiatisées tandis que la précision se diminue vers 65% pour les ressources qui contiennent des discours conversationnels [Kowalski, 2010]. En termes de charge de calculs, l'approche de l'indice phonétique est moins gourmande a celle basée sur *LVCSR* mais elle souffre du problème de limitation des termes d'indexation. En effet, les nombres termes d'indexation et de recherche est limitée par le modèle phonétique utilisé, par contre l'utilisation des *LVCSR* permettent l'ajout des nouveaux termes lors de la phase de décodage.

1.5. Les documents parlés

1.5.1. Les caractéristiques des documents parlés

Les documents parlés partagent plusieurs caractéristiques avec les documents textes classiques. Cependant, leurs contenus sont souvent complexes et même incomplets et dépend étroitement par le style de parole utilisée, par exemple dans la parole instantanée où nous trouvons plusieurs phrases tronquées, incomplètes …etc. Donc, la nécessité de trouver et développer des nouvelles techniques et méthodes efficaces pour manipuler ses contenus [Dahlbäck, 1997].

Entre temps, de point de vue structurel, il existe une grande différence entre la structure des documents texte et celles parlées qui soulèvent de nouveaux challenges qui doivent être abordés et traités dans le processus de recherche d'information.

Parmi ces challenges, nous citons que le flux parlé est un support plus riche et plus expressif que le texte [Schmandt, 1994]. En plus, il contient plus d'informations que les mots parlés comme : les caractéristiques des locuteurs, style de parole, dialecte …etc. Avec telle ressource parlée, des informations supplémentaires telles que l'identité de la langue ou dialecte parlée, l'identité du locuteur, l'humeur ou le ton des locuteurs sont capturées parallèlement avec les mots parlés. Cependant, ces informations supplémentaires peuvent être utiles dans le développement des systèmes d'indexation et de recherche d'informations et elles offrent de nouvelles voies d'exploitation et d'application dans différents contextes.

1.5.2. Les documents parlés versus documents textes

En effet, le concept « sujet du document » ou en anglais « *Topic* » est une caractéristique primordiale pour les documents texte et parlée à la fois. Dans ce contexte, nous intéressons dans cette thèse sur le contenu des documents parlés de point de vue de l'accès, de l'indexation et les modalités de recherche ; l'exploitation des autres informations acoustiques des ressources parlées telle que : l'identité de locuteur où ses caractéristiques ne sont pas traitées dans cette thèse et elles seront parmi nos perspectives futures.

À cet effet, il est important de chercher des méthodes et démarches pour l'extraction et la représentation du contenu des documents parlés sous une forme adéquate pour éventuels systèmes d'indexation et de recherche dans ces flux parlés. Bien que ces objectives sont similaires à celles visées pour les documents de textes, le passage de la modalité texte vers la modalité parole engendre une nouvelle dimension de complexité et d'incertitude. En effet, la méthode ou l'approche sollicitée doit confronter un ensemble de défis comme : la capacité de traiter l'aspect multi locuteurs, l'impact des bruits dans les flux parlés, le style de parole utilisé (parole conversationnelle, spontanée, discussions …etc.) et le traitement des langages a large vocabulaire ainsi que les langages moins dotés. Dans ce contexte, nous détaillons dans le troisième chapitre les techniques et les outils qui permettent de surmonter ces problèmes en introduisant les concepts liés au langage elle-même comme les syllabes et la modélisation phonétique ainsi que les techniques et modalités de recherches et d'accès dans le contenu du flux parlés.

Entre autres, l'amélioration de la robustesse des modèles de recherche dans le contenu des documents parlé au milieu bruité et aux erreurs de transcription est fortement sollicitée. Dans ce contexte, la plupart des méthodes d'indexation et de recherche qui ont été développées pour les documents texte supposent implicitement que les transcriptions se génèrent sans erreur. Avec le texte, les mots dans les documents sont supposés être connus avec une certitude élevée. Par conséquent, il n'y a pas de mécanisme explicite dans les modèles pour le traitement des erreurs dans la représentation du document. Cependant, avec la parole, il n'y a pas actuellement des méthodes de transcription automatique parfaite et surtout pour certains styles de parole et il y aura implicitement des erreurs de transcriptions générées par les systèmes de reconnaissances automatiques de la parole.

Dans ce contexte, l'utilisation des dictionnaires linguistiques pour la vérification des résultats de transcription est largement sollicitée. Entre autres, l'aspect sémantique est une nécessité majeure pour surmonter les problèmes de transcription. À cet effet, l'utilisation des modèles de représentation sémantique comme les thesaurus ou les ontologies avec la combinaison de mots à leurs radicaux pour l'amélioration de la qualité de la transcription automatique sont des approches prometteuses. La contribution par la proposition d'un modèle d'indexation et de recherche en utilisant des fonctions de correspondance plus complexe pour permettre la correspondance sémantique approximative des termes d'indexation, ainsi que d'autres techniques pour traiter les erreurs des représentations de documents ainsi que les techniques d'accès et de recherche dans le contenu du flux parlé sont des objectifs essentiels traités dans cette thèse.

1.6. La Représentation des connaissances

Les tendances actuelles pour la représentation des connaissances se déroulent autour des ontologies. Ils ont la puissance de modélisation de l'univers en tenant compte des concepts avec ces relations entre eux. Les problèmes rencontrés auparavant dans les systèmes de recherche d'information de point de vue le traitement de l'aspect sémantique sont dégagés. Cependant d'autres difficultés résident toujours dans les SRI comme les ambigüités des mots lors de la création des indexes fiables et discriminants ou la formulation des besoins utilisateur.

Plusieurs systèmes sont mis à la disposition aux utilisateurs pour la présentation l'extraction des informations à partir des ressources parlées comme les systèmes de reconnaissance automatique de la parole (*SRAP*). Nous pouvons citer comme exemple le projet « *Transcriber*[1]», qui est un outil d'aide à la transcription de corpus oraux développés par la DGA (*Claude Barras, Direction Générale de l'Armement d'Amérique*). Cet outil « *Freeware* » permet d'éditer la plupart des formats de signaux de parole et offre une interface interactive très bien conçue pour écouter et transcrire en parallèle ces corpus oraux.

Dans la littérature, le terme transcription à l'origine est employé pour décrire le phénomène de copie ou de reproduction par écrit, de tout ce qui a été écrit, parlé, vue, réfléchi…, il a désigné aussi la reproduction des notes musicales écrite pour un instrument pour les adaptés à un autre. Actuellement et comme c'est le cas du notre contexte on se limite à la transcription de la parole qui représente pour nous l'entité la plus porteuse de connaissance sémantique dans ressources multimédias.

Dans la pratique il est difficile de joindre parallèlement une transcription à son document multimédia d'origine, car il n'existe pas dans la majorité des formalismes actuelle une association physique (dans un même fichier) de ces derniers. Même les normes les plus adaptées pour les ressources multimédias comme le « *Codec MPEG-7* » qui offre la possibilité d'intégrer quelque descripteur qui facilitant l'indexation de la vidéo, ne permet

[1] http://trans.sourceforge.net/en/presentation.php

pas d'incorporer la vidéo et sa transcription dans un seul fichier. Cependant il est plus pratique d'associer aux documents multimédias des annotations pour faciliter l'indexation et la recherche.

Les annotations ou les indexes des documents multimédias et les ressources parlées doivent respecter les règles sémantiques issues du domaine d'ontologies, où les concepts peuvent servir comme entités pour la représentation du contenu parlé. Bien que l'utilisation des thesaurus ou des ontologies lors de la phase d'indexation permettrait de surmonter les problèmes d'ambigüités lexicales des termes utilisés et de mieux représenter les connaissances inhérentes dans le contenu de ces ressources. En termes d'indexation sémantique, les indexes et les concepts de l'ontologie sont associés au contenu des documents selon la sémantique encapsulée [Hubert, 2009].

1.6.1. Les ontologies

Les ontologies offrent une modélisation des connaissances d'un domaine basée sur une hiérarchie des concepts et termes d'un domaine. Ils ont souvent utilisé dans les Systèmes de Recherche d'Information (SRI) dans les tâches d'indexation du contenu des ressources et l'enrichissement du langage d'interrogation requête. En effet, ils permettent notamment de surmonter les problèmes d'ambiguïtés lexicales et sémantiques du langage des ressources dans les SRI classiques [Sy, 2012]. Ainsi, les ontologies ont pris une grande part dans le domaine de représentation de connaissance. La figure 1-5, présente une représentation simplifiée des concepts d'une ontologie.

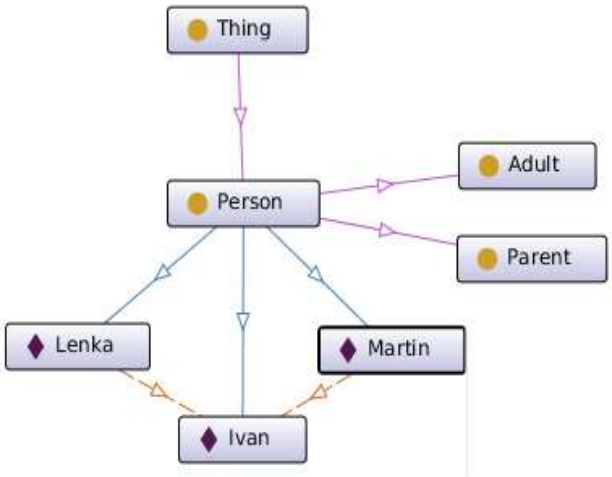

Figure 1-5 : Exemple d'une représentation simplifiée d'une ontologie d'après [Sy, 2012]

1.6.1.1. Définitions

En philosophie l'ontologie est une branche de la métaphysique qui s'intéresse à l'étude de l'être en tant qu'être. Elle présente l'étude des propriétés générales de ce qui existe. En informatique une ontologie, selon *Tom Gruber* [Gurber, 1993] qu'il définit initialement comme étant une spécification explicite d'une conceptualisation « *An explicit specification of a conceptualization* », elle permet la spécification dans un langage formel les concepts relatifs au domaine avec ces relations. Cette définition a été étendue par *Nicola Guarino* [Guarino, 1995] qui a accentué le caractère formel pour assurer l'exploitation et la compréhension des composantes des ontologies par les machines et les systèmes informatiques. Ces systèmes devront être capables d'interpréter la sémantique intégrée dans les informations fournies. Ainsi, une ontologie est une représentation explicite de la sémantique d'un domaine.

Entre temps, les travaux de *Sean* [Sean, 2003] ont permis la définition de l'ontologie comme étant un artefact de l'ingénierie constitué et développé autour d'un vocabulaire spécifique. Elle est utilisée pour décrire une certaine réalité avec un ensemble d'hypothèses explicites et de règles sur le sens exprimé des concepts du vocabulaire. Ainsi, une ontologie décrit une spécification formelle d'un certain domaine. C'est une modélisation de connaissances spécifiques partagée et un modèle formel d'inférence exploitable par les systèmes informatiques.

Cependant, nous trouvons dans les travaux de *Tom Gruber* [Gruber, 1993] la précision de la définition de point de vue du contexte de l'informatique et des sciences de l'information. Il a défini que l'ontologie est un ensemble de primitives de représentation pour modéliser un domaine de connaissance. Les primitives de représentation sont généralement des ensembles, des classes, des attributs, des propriétés) et des relations entre les classes. Les définitions des primitives de représentation incluent des informations cohérentes sur les aspects : significations, contraintes et applications. En pratique, de point de vue de structure et modèles, les modèles et langages de description utilisés dans les ontologies ont une puissance expressive plus proche de la logique du premier ordre que celle des langages utilisés pour à celles des bases de données. À cet effet, nous trouvons que les ontologies sont considérées comme un niveau « sémantique », tandis que les modèles et les schémas de bases de données sont considérés comme des niveaux « logique » ou « physique ».

Aussi, nous trouvons les travaux de *Fankam* [Fankam, 2009] qui ont caractérisé les ontologies comme une représentation formelle, référençable et consensuelle de l'ensemble des concepts partagés d'un domaine. En effet, les ontologies sont des modèles de représentation des connaissances formelles puisqu'elles permettent des raisonnements automatiques ayant pour objectif d'inférer de nouveaux faits et d'effectuer des vérifications de consistance. Entre autres, elles sont consensuelles, c'est-à-dire admises par l'ensemble des membres et des systèmes d'une communauté. De plus, chaque entité ou relation décrite dans l'ontologie peut être directement référencée par un symbole, à partir de n'importe quel contexte.

En effet, la représentation des connaissances par les ontologies est une alternative très prometteuse pour la modélisation de la sémantique emboitée dans le contenu des ressources informationnelles. Ces ontologies permettent de fournir :

- Une structure conceptuelle de base à partir de laquelle il est possible de développer des systèmes à base des connaissances qui seront partageables et réutilisables.
- L'interopérabilité entre les ressources d'information et de connaissances.

1.6.1.2. Composantes d'une ontologie

Malgré les différentes définitions sur les ontologies, on trouve qu'il y a un consensus sur ses composants. Parmi les travaux de recherches effectués sur les ontologies et leurs composants, nous trouvons les travaux de Gomez Perez [Gomez Perez, 1999]. Ces travaux confirment que l'exploitation des connaissances traduites par les ontologies s'effectue à l'aide des composantes suivantes : concepts, relations, fonctions, axiomes et instances.

- *Les concepts ou les classes* : qui présente l'abstraction pertinente du domaine. Elles sont retenues en fonction des objectifs et les connaissances pour l'ontologie et ces applications. Ces concepts peuvent être classés selon plusieurs dimensions telles que le niveau d'abstraction : concret ou abstrait, l'atomicité : élémentaire ou composée et le niveau de réalité : réel ou fictif.
- *Les relations* qui traduisent les associations pertinentes qui existent entre les concepts. Ces relations incluent les associations de généralisation/spécialisation (sous-classe-de), d'agrégation ou de composition (partie-de), …etc. En effet, elles permettent la modélisation de sa structure ainsi que les relations inter et intra concepts.
- *Les fonctions* qui sont des cas particuliers des relations. Dans lesquelles, l'élément de relation est défini en fonction des éléments précédents.
- *Les axiomes* qui sont des expressions qui sont toujours vraies. Leurs utilisations dans les ontologies permettent plusieurs objectifs comme : la définition de la signification des composants, la définition des restrictions sur les valeurs des attributs, la définition des arguments des relations et la vérification de la validité des informations spécifiées où en déduire de nouvelles.
- *Les instances* qui constituent la définition extensionnelle de l'ontologie. Ces objets couvrent les connaissances statiques ou factuelles à propos du domaine des connaissances cibles.

1.6.2. Classification des ontologies

Premièrement, nous trouvons dans les littératures des distinctions entre les ontologies dites légères « *Light weight ontologies* » et celles dites lourdes « *heavy weight ontologies* » [Gomez Perez, 2004]. La définition de « ontologies légères » couvre les concepts comprenant des propriétés et ils sont organisées en taxonomies avec des relations conceptuelles comme par exemple : *Yahoo! Directory*. En effet, nous trouvons que certains auteurs considèrent les taxonomies et les thésaurus comme des ontologies parce qu'elles fournissent des conceptualisations partagées pour les connaissances des domaines donnés. Cependant, la définition de « ontologies lourdes » est une extension pour les ontologies légères par des axiomes et des restrictions modélisant l'aspect sémantique entre les concepts du domaine. Entre autres, les ontologies peuvent être classifiées selon plusieurs dimensions. Chacun d'entre eux se base sur différentes dimensions comme : l'objet de conceptualisation, le niveau de détail de l'ontologie, le niveau de complétude et le niveau de formalisme) pour la classification des ontologies. Parmi ces classifications, nous citons :

1.6.2.1. Classification basée sur la richesse de la structure

Cette classification est proposée dans les travaux de *Lassila et McGuinness* [Lassila, 2001]. Ils ont proposé une classification des ontologies selon le besoin l'information de l'ontologie et la richesse de sa structure interne. Cette classification est réalisée sous forme d'un palier allant des ontologies légères vers les ontologies lourdes :

- Les vocabulaires contrôlés : liste de termes.
- Les glossaires : liste de termes avec leur sens spécifié en langage naturel.
- Les thésaurus : glossaire contenant des descriptions sémantiques entre les termes.
- Les « *Frame* » : ontologies incluant des classes avec propriétés pouvant être héritées.
- Les ontologies avec restrictions de valeur : ontologies pouvant contenir des restrictions sur les valeurs des propriétés.
- Les ontologies avec contraintes logiques : ontologies pouvant contenir des contraintes définies dans un langage logique entre ces constituants comme les relations.

1.6.2.2. Classification basée sur l'expressivité du langage

C'est une classification réalisée à base de l'expressivité des ontologies ou du langage de représentation des connaissances). En effet, différents types de composants d'ontologies peuvent être définis comme : les concepts, les propriétés, les instances, les axiomes …etc. Dans ce contexte, les auteurs de cette classification définissent quatre types de classes ontologiques [Roussey, 2011].

- *Ontologies d'information* : ce sont composer de schémas et structures pour clarifier et organiser les idées et les connaissances de collaborateurs dans le développement d'un projet. Ces ontologies ne sont utilisées que par les humains.
- *Linguistiques / terminologiques ontologies* : ce sont des ontologies linguistiques qui peuvent être des glossaires, dictionnaires, vocabulaires contrôlés, taxonomies, thesaurus ou bases de données lexicales. Ce type d'ontologie se concentre principalement sur les termes et leurs relations.
- *Ontologies Logiciel* : ils fournissent des schémas conceptuels dont l'objectif principal est orienté vers le stockage et la manipulation des données et connaissances. Ils sont utilisés pour des activités de développement de logiciels, dans le but d'assurer la cohérence des données.
- *Ontologies formelles* : Ils exigent une sémantique claire pour la langue utilisée pour définir les concepts, ainsi que des règles strictes pour la définition de ces derniers ainsi que leurs relations. Ces règles sont obtenues en utilisant la logique formelle ou la logique de premier ordre pour modéliser la sémantique formelle.

1.6.2.3. Classification basée sur le domaine d'application

D'autres classifications ont été proposées dans les travaux de Gómez-Pérez et Roussey [Gomez Perez, 1999], [Roussey, 2011]. Ils ont proposé une classification des ontologies selon deux critères : le premier est basé sur la quantité et le type de structure utilisée dans le processus de conceptualisation. Le deuxième critère est basé sur le sujet de la conceptualisation. A cet effet, ils proposent la classification suivante :

- *Les ontologies de représentation de connaissances* : Elles modélisent les représentations primitives utilisées pour la formalisation des connaissances sous un paradigme donné.
- *Les ontologies générales ou communes* : Elles modélisent les connaissances de sens commun réutilisables dans plusieurs domaines. Ces ontologies intègrent un vocabulaire relatif aux différents concepts comme : les évènements, le temps, l'espace, la causalité, le comportement …etc.
- *Les ontologies de niveau supérieur* « *Top-level, Upper-model* » : Elles modélisent les concepts très généraux auxquels les racines des ontologies de plus bas niveau devraient être liées. Cependant, il existe plusieurs ontologies de niveau supérieur et qui sont divergentes. Afin de résoudre ce problème, l'organisation de standardisation IEEE tente de développer un standard pour les ontologies de niveau supérieur.
- *Les ontologies de domaine* : Elles modélisent les connaissances réutilisables dans un domaine précis. Ces ontologies fournissent les concepts et les relations permettant de couvrir les vocabulaires, activités et théories du domaine cible. Les concepts des ontologies de domaine sont souvent des spécialisations de concepts définis dans des ontologies de niveau supérieur.

- *Les ontologies de tâches* : Elles modélisent les vocabulaires relatifs à une tâche ou une activité générique en spécialisant certains termes des ontologies de niveau supérieur.
- *Les ontologies de tâches de domaine* : Ce sont des ontologies de tâches réutilisables dans un domaine spécifique, mais pas d'un domaine à l'autre et qui sont indépendantes de l'application.
- *Les ontologies de méthodes* : Elles modélisent les définitions des concepts et des relations pertinentes pour le processus de raisonnement afin d'effectuer une tâche spécifique.
- *Les ontologies d'applications* : Elles modélisent les connaissances requises pour des applications spécifiques. Les ontologies d'applications spécialisent souvent le vocabulaire des ontologies de domaine et des ontologies de tâches.

En résumant, La figure 1-6 présente une schématisation de la classification basée sur le champ d'application de l'ontologie. Par exemple, la portée d'une ontologie locale est plus étroite que la portée d'une ontologie de domaine ; ontologies de domaine ont des concepts plus spécifiques que les ontologies de référence de base, qui contient le concept fondamental d'un domaine.

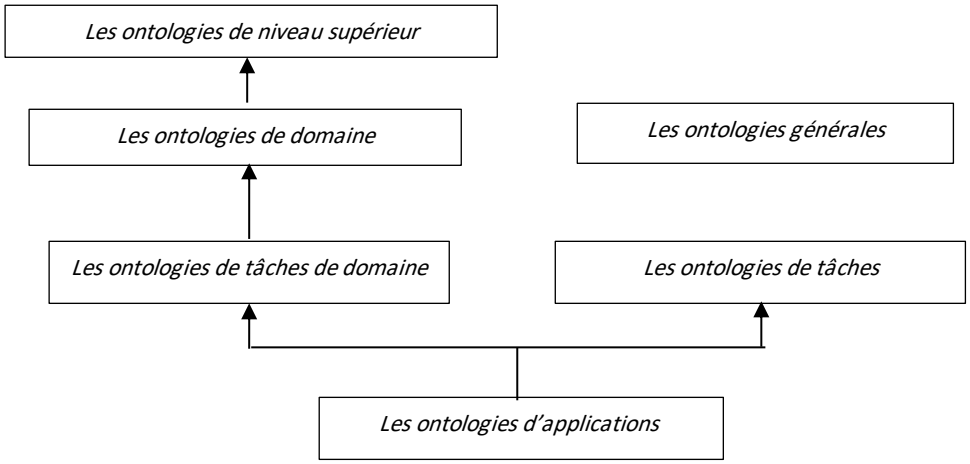

Figure 1-6 : Classification des ontologies selon son domaine d'application d'après [Roussey, 2011]

1.6.3. WordNet comme ontologie générale pour l'indexation sémantique

WordNet est une base de données lexicale créée par un groupe de psychologues et de linguistes du laboratoire de sciences cognitives de l'université de Princeton. Initialement, l'objectif de ce projet est la construction d'une ressource lexicale qui permette l'exploitation des mots en tenant compte de leurs relations dans sans contexte conceptuel ou ontologique. Ainsi, ce réseau lexical de *WordNet* est modélisé sous la représentation conceptuelle « *Lexical Conceptual Graph-LGC* » proche du contexte ontologique [Guarino, 1999].

En effet, le contenu de *WordNet*, couvre la majorité des noms, verbes, adjectifs et adverbes de la langue anglaise. Sa dimension ainsi que le domaine de la langue générale qu'il traite lui permettent souvent de couvrir les sujets traités dans les collections de tests conventionnelles de la RI comme TREC, CLEF, …etc.

WordNet est un réseau composé de 155 287 structuré autour de 121 012 termes ou concepts appelés « *Synsets* ». Le Tableau 1-6 présente une description statistique sur les concepts dans la base de données de WordNet dans sa version 3.1.

Catégorie	Mots	Concepts	Total Paires « Mots-Sens »
Nom	117 798	82 115	199 913
Verbe	11 529	13 767	25 296
Adjectif	21 479	21 509	42 988
Adverbe	4 481	3 621	8 102
Total	**155 287**	**121 012**	**276 299**

Tableau 1-6 : Description statistique des concepts WordNet (Ver. 3.1)[1]

Dans *WordNet*, les entrées sont des concepts représentés par des *Synsets* contenant l'ensemble des termes synonymes qui peuvent désigner un concept. En effet, les concepts sont reliés sémantiquement par des relations affectées aux Synsets et elles sont représentées par des classes. Les relations de base définie dans *WordNet* entre les termes sont la *Synonymie*. Entre autres, nous trouvons d'autres relations pour les *Synsets* comme les relations de type hyponyme-hyperonyme « *is-a* », et les relations de type méronymie-holonymie, la figure 1-7 présente une illustration simple de ces relations.

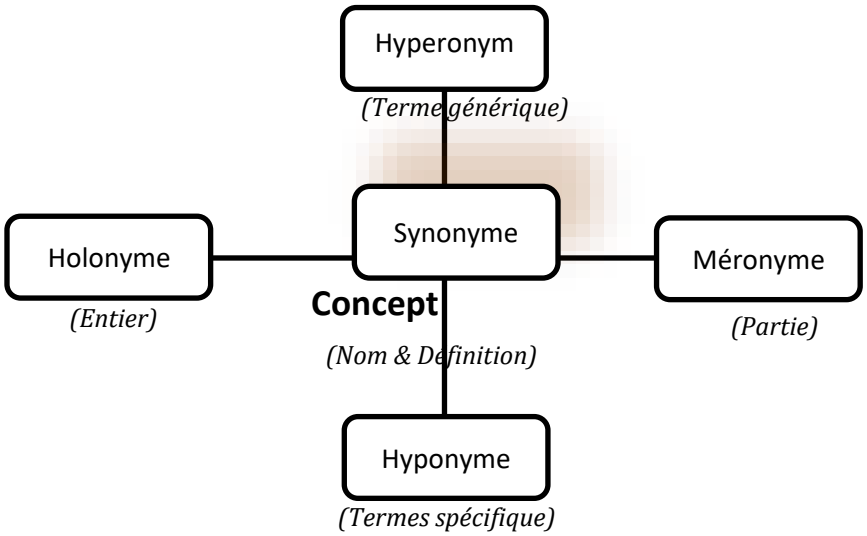

Figure 1-7 : Schéma illustratif des principaux relations sémantiques dans l'ontologie WordNet

[1] *WordNet* Ver. 3.1 disponible à l'adresse http://wordnetcode.princeton.edu/wn3.1.dict.tar.gz

Cependant, différents types de relations entre les concepts sont définies dans *WordNet* , parmi eux nous citons :

- *La Synonymie* : c'est l'association d'un mot à un concept.
- *L'hyperonymie* : C'est le terme générique utilisé pour désigner une classe englobant des instances de classes plus spécifiques
- *L'hyponymie* : C'est le terme spécifique utilisé pour désigner un membre d'une classe.
- *La Méronymie* : Le nom d'une partie constituante, substance ou membre d'une autre classe.
- *L'holonymie* : Le nom de la classe globale dont les noms *Méronymes* font partie.

En effet, il existe autres relations moins utilisées comme : la relation « *Domain* », la relation « *Antonymy* » qui exprime les sens opposés pour les *Synsets*, la relation « *Troponymy* » qui représente la similitude partielle entre les verbes …etc.

1.7. Conclusion du chapitre

Dans ce chapitre nous avons essayé de mettre en évidence les concepts et les disciplines utilisés dans notre contribution proposée dans cette thèse. Pour cela nous avons présenté d'abord la discipline principale de notre problématique : les systèmes de recherche d'information sur le contenu. Ensuite nous avons présenté les caractéristiques des documents parlés dont on a présenté ses différentes caractéristiques physiques et ces particularités structurelles. En terminons ce chapitre, par la présentation de la discipline de représentation de l'information et des connaissances par les ontologies. Cette discipline sera utilisée ultérieurement dans notre contribution pour assurer une bonne qualité de recherche sur le contenu parlé.
D'après l'étude effectuée sur les caractéristiques des ressources parlées, nous constatons que l'utilisation des ressources linguistiques intelligentes ; basée sur la sémantique ; peuvent surmonter les problèmes liés aux erreurs de reconnaissance et aux mécanismes de décodage. Entre autres, nous trouvons que l'automatisation de processus d'indexation permet l'amélioration de la qualité de recherche dans le contenu du flux parlé.

Dans le chapitre suivant, nous étudions en détail les techniques actuelles utilisées dans notre contribution pour la recherche dans le flux parlé. Nous étudions les techniques liées à la structure complexe du flux parlé, les techniques linguistiques pour réaliser une indexation automatique ainsi que les modalités d'accès pour le moteur de recherche.

Chapitre 2

Techniques de reconnaissance et détection dans les flux parlés

2.1. Introduction

La motivation principale dans la discipline de la recherche des documents parlés « SDR » réside autour de l'hypothèse qui suppose que la qualité de ces systèmes de recherches se dégrade si nous utilisons des systèmes de reconnaissance automatique de la parole continue au lieu de l'utilisation des transcriptions manuelles par des experts du contenu. Cependant, vu la masse importante des ressources parlées stockées, il est fortement souhaitable que ces ressources doivent être transcrites et indexées par des mécanismes automatiquement.

Dans ce contexte, la mise en œuvre des systèmes de recherches dans le contenu parlés « *SDR* » nécessite une intégration intrinsèque des différentes disciplines telles que : la reconnaissance automatique de la parole, la recherche d'information et aussi les technologies d'informations et de connaissances. En Particulier, le domaine de détection des termes parlés a motivé plusieurs communautés et équipes de recherches, par exemple : *TREC SDR* collection [Bertoldi, 2003], *LIMSI*[1], *CMU Sphynx*[2],… etc.

[1] https://www.limsi.fr/fr/

[2] https://cmusphinx.github.io/

En effet, il existe plusieurs codes sources et routines qui peuvent être utilisés dans les systèmes de reconnaissance de la parole continue à large vocabulaire. Mais ces ressources nécessitent une maitrise des différentes étapes et modèles utilisés durant ce processus de transcription automatique. Pour cela, nous avons choisi dans ce chapitre de tracer une description des différentes étapes des systèmes de reconnaissances automatiques. En plus, nous présentons dans les sections suivantes de ce chapitre un état de l'art sur les techniques de détection des mots clés à base phonétique. Entre autres, nous présentons aussi une étude bibliographique sur les techniques des mesures de similarités. Ces techniques seront utilisées dans notre contribution pour capter et enrichir sémantiquement les résultats de la transcription automatique des ressources parlées.

2.2. Fondements théoriques sur les SRAP

La transcription des documents parlés vers le texte est un sujet à recherche sollicité depuis l'invention des ordinateurs, mais il a commencé de s'améliorer d'une façon significative seulement depuis les années soixante-dix du vingtième siècle. En effet, cette évolution est marquée par la puissance de calculs qui permettent d'exécuter des calculs complexes et expériences intenses dans des délais acceptables. Sachant que, la modélisation stochastique du flux de parole est la base de la majorité des systèmes de reconnaissance automatique « ASR ». Cependant, la qualité de ces systèmes dépend étroitement par la qualité et la disponibilité des ensembles de données étiquetés pour l'apprentissage, qui sont appelés souvent « corpus ». Le développement des corpus riches qui peuvent être utilisés comme fondement de l'apprentissage était une partie primordiale pour accroitre les performances des systèmes automatiques de la reconnaissance de la parole.

En pratique, nous trouvons que l'institut national de la science et la technologie « *National Institute of Standards and Technology*[1] *-NIST* » et le consortium linguistique des données « Linguistic Data Consortium[2] - LDC » sont les plus répondus pour la création, l'annotation et la distribution des corpus et des ressources linguistiques pour la communauté scientifique.

En revanche, plusieurs outils sont hormis disponibles peuvent être utilisés dans les projets de recherches scientifiques. Parmi eux, nous trouvons : la plateforme HTK[3], elle utilise la modélisation markovienne « *Hidden Markov Model Toolkit* » et elle intègre un module de décodage à grand vocabulaire « *HDecode* » [Young, 2015], la plateforme *PocketSphinx* [Huggins-Daines, 2006] et *Sphinx-4*[4] [Lamere, 2003] qui contiennent des systèmes de reconnaissance automatique des différents langages et la plateforme *Kaldi*[5] [Povey, 2011]. En effet, l'utilisation de ces outils et plateforme permettent la construction d'un système de reconnaissance vocale complet sans la nécessité de développement des ressources

[1] http://www.nist.gov/
[2] https://www.ldc.upenn.edu/
[3] http://htk.eng.cam.ac.uk/
[4] http://cmusphinx.sourceforge.net/wiki/tutorialsphinx4
[5] https://sourceforge.net/projects/kaldi/

supplémentaires ou de procéder à la programmation de bas niveau.
Dans la suite de cette partie, nous nous présentons une vue d'ensemble sur les concepts les plus importants dans le processus de la reconnaissance automatique de la parole ainsi que ces métriques d'évaluation. Nous ne détaillons pas des fonctionnalités comme les techniques d'extraction des paramètres de flux parlé comme : *Mel-frequency cepstrum* « MFCC » et *Linear Predictive Coding* « LPC ». En effet, cette étape est importante dans les systèmes de reconnaissance automatique de la parole, mais que nous jugeons moins importante pour notre problématique de recherche dans le contenu des flux parlés.

2.2.1. Les systèmes de reconnaissance de la parole

Généralement, l'objectif des systèmes de reconnaissances de la parole est de générer une séquence des mots étant donné un flux parlé « *ondes sonores* ». La structure générale typique de ces systèmes est présentée par la figure 2-1. La première étape dans le processus de reconnaissance de la parole est l'extraction des paramètres acoustiques représentant au maximum le signal vocal. Le vecteur acoustique extrait doit contenir les informations essentielles avec une représentation compacte pour une reconnaissance efficace. Cette phase généralement appelée : Extraction des paramètres « *Features Extraction* ». En effet, l'objectif du processus du décodage est de trouver la séquence plus probable étant donné une séquence d'observations.

Généralement, pour détecter la séquence de mots la plus probable il faut passer par la construction de trois modèles d'information. Le premier c'est le « lexique » appelé souvent le « dictionnaire ». Il est utilisé dans le système de reconnaisse de la parole continue à large vocabulaire LVCSR pour mapper les sous-unités d'un mot du modèle acoustique avec celle définit dans le vocabulaire de la langue cible. Le deuxième c'est le modèle de langage qui représente l'information lexicale et syntaxique de la séquence à reconnaitre. Il contient toutes les configurations possibles pour un mot donné. Le troisième c'est le modèle acoustique qui sert à segmenter la séquence observée en unité parlée élémentaire comme les syllabes et les phonèmes.

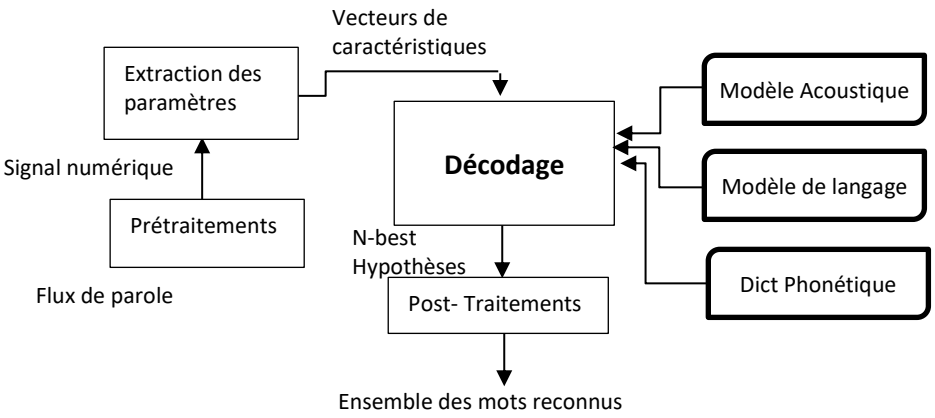

Figure 2-1 : Structure Générale d'un système de reconnaissance automatique de la parole d'après [Gruhn, 2011]

2.2.2. Formulations mathématiques

Le problème de reconnaissance de la parole peut être décrit comme étant une fonction de correspondance d'une observation acoustique $X = (x_1, x_2, x_3, \dots, x_t)$ à l'instant t avec une séquence probable $W = (w_1, w_2, w_3, \dots, w_n)$ de n mots.

De point de vue statistique, le système de reconnaissance sélectionne la séquence de mots la plus probable étant donné une observation acoustique. On note $P(W/X)$ la probabilité de prononciation des séquences de mots W sachant l'observation acoustique X. Donc, on cherche la séquence de mots $\widehat{W}$ satisfaisant la règle de Bayes suivante :

$$\widehat{W} = \underset{W}{\operatorname{argmax}}\, P(W/X) \tag{2-1}$$

En appliquant la règle de Bayes sachant que la séquence de mots la plus probable est indépendante de la probabilité de séquence d'observation $P(X)$, la règle de décision peut être écrite comme suit :

$$\widehat{W} = \underset{W}{\operatorname{argmax}} \left(\frac{P(X/W)P(W)}{P(X)} \right) \tag{2-2}$$

$$\widehat{W} = \underset{W}{\operatorname{argmax}} (P\,X/W)\, P(W) \tag{2-3}$$

Tel que P(W) est la probabilité a priori fournit par le modèle de langage. P(X⁄W) est la probabilité conditionnelle calculée en utilisant le modèle acoustique. On note que les Modèles de Markov Cachés (MMC) sont les plus utilisés à ce jour [Kai, 2006]. Dans la suite de cette thèse, nous adaptons l'utilisation de la modélisation markovienne.

2.2.3. Le processus de reconnaissance « Recognition Engine »

L'utilisation de la modélisation statistique à l'aide des modèles markovien dans le processus de reconnaissance de la parole ; comme la présente la figure 2-2 ; nous amène a traité les problèmes suivants :

- *Le problème de modélisation acoustique* : comment calculer la probabilité conditionnelle $P(X/W)$? Il est nécessaire d'avoir plusieurs modèles acoustiques des différents locuteurs pour construire le modèle de séquences de mots W. Ces derniers sont liés étroitement avec le type de l'application qu'ils l'utilisent (commande vocale, dictée vocale ou parole continue). Généralement, plusieurs contraintes sont définies pour ce calcul. Dans la section suivante, on décrit les modèles de Markov cachés qui sont les plus utilisés pour l'estimation des modèles acoustiques [Adami, 2010].
- *Le problème de modélisation du langage* : comment calculer la probabilité à priori $P(W)$ pour une séquence de mots W ? La technique la plus utilisée est basée sur l'hypothèse que le $N^{iéme}$ mot dans la phrase dépend seulement du *N-1* précédent. Cette hypothèse est appelée *N-gram* qui sera définie dans la section 2.3.2.1

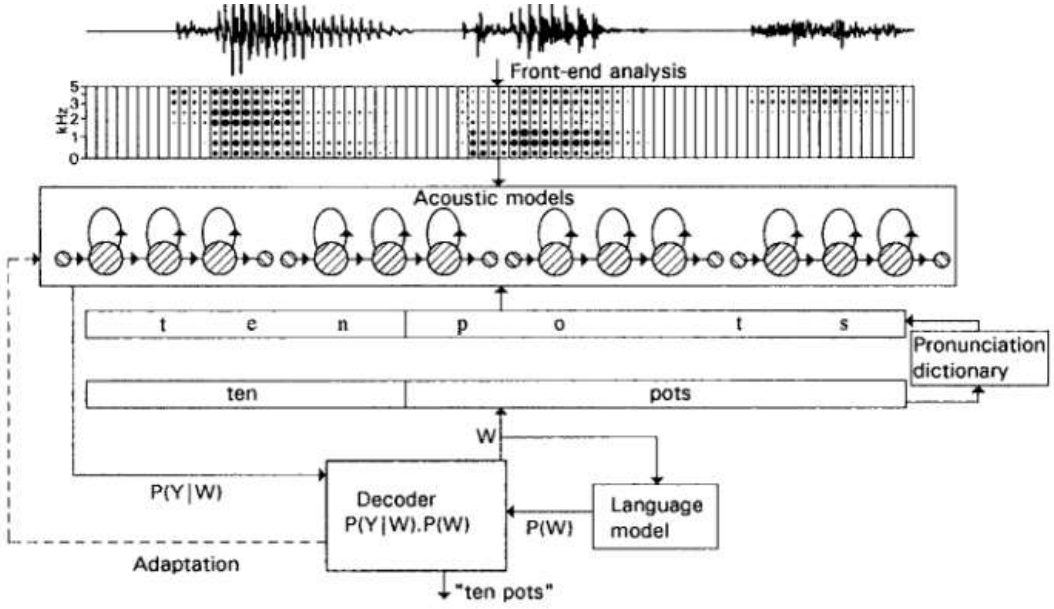

Figure 2-2 : Framework de décodage d'un signal de parole d'après [Holmes, 2001]

- *Le problème de décodage ou de recherche* : comment trouver la meilleure transcription W de l'observation X, étant donné le modèle acoustique et le modèle de langage ? En revanche, il est quasiment impossible de rechercher d'une manière exhaustive toutes les séquences possibles des mots. Dans la section 2.3.2.2, nous présentons quelques méthodes utilisées pour effectuer cette recherche.

2.2.3.1. La modélisation acoustique

Les modèles acoustiques $P(X/W)$ sont utilisés pour calculer la probabilité d'observation acoustique X lorsque le locuteur prononce la séquence de mots W. L'estimation efficace de tel modèle est l'un des défis du processus de reconnaissance de la parole. En tenant en compte les facteurs de la variabilité du signal de la parole tels que : la prononciation, le contexte phonétique, les caractéristiques psychologiques du locuteur …etc. La modélisation acoustique la plus efficace est basée sur une structure dénommée Modèles de Markov cachés « *MMC* ». Dans la reconnaissance de la parole, plusieurs aspects doivent être définis avant l'utilisation des *MMC*. Parmi ces aspects : La fonction de discrimination, le choix d'unité de la parole, la topologie du modèle, le modèle de distribution de l'observation, l'initialisation des paramètres et quelques techniques d'adaptation.

a. La fonction de discrimination

Le maximum de vraisemblance « *MV* » a pour vocation de maximiser la probabilité d'observation d'une séquence O étant donné le modèle *MMC* λ définit par :

$$\lambda_{MV} = \underset{\lambda}{argmax}\, P(O/\lambda) \qquad (2\text{-}4)$$

Cependant, le problème de reconnaissance de la parole est défini comme étant une tâche de classification où l'on définit pour chaque classe acoustique $c \in 1..C$ un modèle *MMC* correspondant λ_c. Donc le *MV* permet d'estimer le modèle λ_c par rapport à la séquence acoustique O^c de la classe C, par la formule suivante :

$$(\lambda_c)_{MV} = \underset{\lambda}{argmax}\, P(O^c/\lambda) \tag{2-5}$$

En effet, le critère *MV* estime chaque modèle à part, donc il ne garantit pas une solution optimale pour minimiser l'erreur de probabilité de reconnaissance. Il ne prend pas en compte la capacité de discrimination de chaque modèle. Entre autres, la capacité de distinguer les observations générées par le modèle correct de ceux générés par les autres modèles. Un autre critère qui permet de maximiser cette dernière est le critère de Maximum d'information Mutuel « *MIM* ». L'information mutuelle entre l'observation de la séquence O^c et la classe c, paramétrée par : $\Lambda = \{\lambda_c\}$, $c = 1,2,\dots,C$ calculé par :

$$I_\Lambda(O^c, c) = log \frac{P(O^c/\lambda_c)}{\sum_{w=1}^{C} P(O^c/\lambda_w, w)P(w)} \tag{2-6}$$

$$= log\, P(O^c/\lambda_c) - log \sum_{w=1}^{C} P(O^c/\lambda_w, w)P(w) \tag{2-7}$$

Le critère *MIM* cherche à trouver l'ensemble de modèles Λ qui maximise l'information mutuelle.

$$\Lambda_{MIM} = \underset{\Lambda}{max}\left\{\sum_{c=1}^{C} I_\Lambda(O^c/c)\right\} \tag{2-8}$$

En pratique, le *MIM* est basé sur une variante de l'algorithme *Baum-Welch* appelé *Extended Baum-Welch* qui maximise ce critère. En bref, l'algorithme calcule probabilités avant-arrière pour les séquences d'apprentissage. Puis, un autre passage avant-arrière est calculé sur toutes les autres expressions possibles. On note aussi que la deuxième étape est extrêmement calcul intensif. Dans la littérature, nous trouvons les travaux de recherches effectués par *Woodland* et *Paovey* qui montrent que l'utilisation de MIM peut améliorer les performances cohérentes par rapport aux systèmes similaires formés avec *MV* [Woodland, 2002].

b. Le choix d'unité de parole

Pour les systèmes de reconnaissance de la parole avec un vocabulaire de petite taille, soit inférieur à 1*K* mots, l'unité de modélisation utilisée est le mot, comme les systèmes de reconnaissance des chiffres. Cependant, pour la reconnaissance vocale avec un vocabulaire de taille moyenne, soit dans l'intervalle [1k, 10K], il est impossible d'avoir des modèles *MMC* pour chaque mot et même il n'existe pas les ressources nécessaires pour la

construction et l'apprentissage de chaque modèle de mot. Pour pallier ce problème, l'unité utilisée sera les sous-mots « *sub-word* » comme unité de modélisation au lieu d'utiliser les mots entiers.

En effet, le phonème est l'unité de modélisation la plus utilisée dans ces situations. C'est l'élément palpable le plus petit dans le flux parlé. En plus, la disponibilité des standards et des règles de passage du phonème vers le mot permet l'utilisation des phonèmes comme unités de modélisation. Pratiquement, le nombre de phonèmes est beaucoup moins inférieur au nombre de mots dans le vocabulaire. Par exemple, dans les flux parlés en langue anglaise, il n'y a que 46 phonèmes[1], ce qui nécessite un nombre de modèles incomparable par rapport un vocabulaire de taille moyenne. Par conséquent, il est généralement possible d'avoir des corpus de données suffisants pour la construction et l'estimation des paramètres des modèles MMC robustes et efficaces. Entre temps, pour la modélisation en phonème, l'utilisation d'un modèle lexical « dictionnaire » pour le mappage des mots vers ses transcriptions phonétiques est indispensable. L'apprentissage et la reconnaissance s'exécutent au niveau phonétique. À la fin du processus, les séquences phonétiques détectées sont reconverties en séquence de mots.

De point de vue phonétique, il existe deux types de modélisation : mono phonème indépendant du contexte et phonème dépendant du contexte. Dans le premier, chaque unité phonétique est indépendante par rapport aux phonèmes adjacents. Elle ne tient pas en compte le problème d'articulation qui influe sur la prononciation de ce phonème. Ainsi, l'utilisation de mono phonème indépendant du contexte dans le processus de reconnaissance ne donne pas des résultats encouragent [Holmes, 2001].

Dans la modélisation en phonème dépendant du texte, nous trouvons les triphonèmes qui sont largement utilisés dans les systèmes de reconnaissance de la parole. Elles prennent en considération les phonèmes adjacents : le précèdent et le suivant. Ainsi, nous trouvons des modèles qui intègres des informations du contexte plus élargi tel que les deux précédents et le deux suivant qui sont appelé cinq-phonème « *quin-phones* » [Hain, 2004], [Prasad, 2005]. Par exemple pour le mot « book », son écriture phonétique est « bʊk », on a les triphonèmes suivants :

$$\boldsymbol{book} = |\mathbf{b}\upsilon\mathbf{k}| = \{\boldsymbol{sil} - \boldsymbol{b} + \upsilon, \mathbf{b} - \upsilon + \mathbf{k}, \upsilon - \mathbf{k} + \mathbf{sil}\}$$

Cependant, il y a deux possibilités pour la segmentation des mots d'une phrase en triphonèmes. Soit, par le croisement des triphonèmes entre deux mots : triphonème croisé « *Cross word triphones* » ou sans intersection entre deux mots. Donc, ils utilisent les bi-phonèmes pour marquer le début et la fin d'un mot. En effet, les systèmes actuels utilisent les triphonèmes croisés dans la phase de reconnaissance des documents parlés, car elle fournit des performances intéressantes.

[1] http://www.studyenglishtoday.net/english-phonetics.html

Entretemps, l'utilisation des triphonèmes augmente le nombre d'unités acoustique à modéliser. Par exemple, pour les 44 phonèmes de langue anglaise, le nombre de triphonèmes croisés est environ 100000. D'où, il est très difficile d'avoir les corpus nécessaires pour le processus d'apprentissage. Pour remédier ce problème, des techniques d'attachement « *Tying* » ou de segmentation « *Clustring* » sont utilisés [Young, 1993], [Young, 1994], [Yu, 2004]. Leurs principes sont basés sur la recherche d'un ensemble d'unités qui partagent les mêmes valeurs spectrales. En apprentissage, tout l'ensemble est utilisé pour l'estimation des paramètres partagés. L'approche la plus utilisée est appelée « *State clustring* » [Young, 1994] ou la distribution d'émission est partagée pour tous les éléments de l'ensemble, comme la montre la figure 2-3.

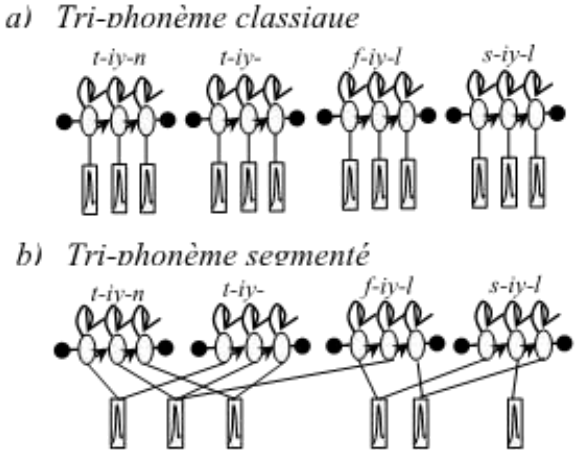

Figure 2-3 : Illustration des techniques de segmentation et rattachements des états. (a) l'état initial où chaque état a sa propre distribution. (b) Montre les états après attachements et segmentation ou quelques états partagent les mêmes distributions [Young, 1994].

Cependant, pour l'implémentation de ces techniques nous trouvons des approches comme « *Buttom-up Clustring* ». Elle cherche un regroupement de bas en haut. Elle calcule une distance pour chaque pair de triphonème observé dans le corpus d'apprentissage. Le problème de cette approche qu'elle est liée étroitement au contexte du corpus d'apprentissage. Elle ne peut pas gérer des contextes loin de celle de l'apprentissage. Pour cela, nous trouvons des approches de regroupement basées sur les arbres de décision phonétique pour remédier ces problèmes [Young, 1993], [Young, 1994]. En effet, l'arbre de décision phonétique est un arbre binaire des reposes sur le contexte des adjacents : le précèdent et le suivant de chaque phonème marqué par une des primitives vrai ou faux. Ensuite, la segmentation est réalisée en explorant l'arbre. Initialement, tous les états sont dans la racine. Ensuite, les réponses aux questions du contexte des adjacents permettent la création des branches. Ce processus de division s'arrête dès qu'on atteint le seuil d'apprentissage défini auparavant. Néanmoins, le choix de la question de contexte phonétique est primordial. La question utilisée pour chaque division permet de maximiser la probabilité locale. Bien que, la segmentation avec l'arbre de décision est une tâche de recherche binaire optimale. Elle permet aussi de traiter efficacement le problème de

triphonème omis, car tous les contextes existent dans les nœuds de l'arbre. Par conséquent, cette technique est celle utilisée dans les systèmes actuels.

c. Topologie de modèle

Parmi les défis d'utilisation des *MMC*, nous trouvons le choix optimal de nombre d'états et les transitions entre eux. Sachant que le flux de la parole est un signal temporel non stationnaire, l'utilisation des topologies gauche-droite permet de capturer mieux la dynamique temporelle. Un *MMC* gauche-droite peut être décrit comme un automate probabiliste à états finis comportant deux processus : un processus caché de changement d'état et un processus d'émission. Le premier processus est dit « caché » car non observable, alors que le deuxième génère des observations pour chaque transition entre les états du modèle. Un exemple de modèle de Markov gauche-droite à trois états est présenté en Figure 2-4.

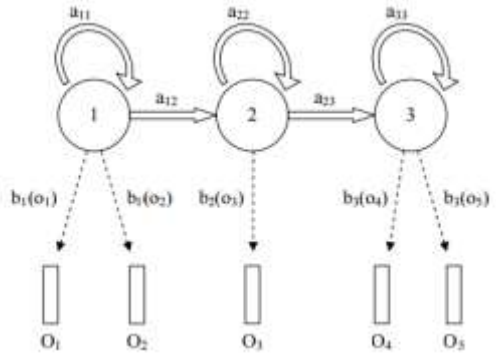

Figure 2-4 : Exemple d'un MMC gauche-droite à trois états

La réalisation d'un processus de Markov caché se traduit par l'existence d'une séquence $Q = (q_0, \dots, q_T)$ d'états de l'automate. Le processus d'émission du modèle de Markov caché associe à Q une séquence de T observations $O = (o_0, \dots, o_T)$.

D'une façon plus formelle, il peut être décrit par :

- Un ensemble fini d'états

$$Q = \{q_1, \dots, Q_N\} \tag{2-9}$$

- Les probabilités des transitions

$$a(i,j) = P(q_j/q_i) \tag{2-10}$$

Qui peuvent êtres écrits sous forme d'une matrice $A[N \times N]$

- Un ensemble χ des symboles d'émission possibles x (discret ou continu).
- Les probabilités d'émission

$$b(t,j) = P(x_j/q_t) \tag{2-11}$$

d. Le modèle de distribution de l'observation

Généralement, les probabilités d'émission sont soit des valeurs dans espace discret à l'aide des techniques de la quantification vectorielle, ou sont des valeurs dans un espace continu calculé avec une fonction de densité continue. Les observations discrètes sont rarement utilisées, car la nature de signal de la parole lui-même est continue et multidimensionnelle. Cependant, les observations continues avec des fonctions de densités gaussiennes ou même avec les réseaux de neurones sont largement utilisées. La majorité des systèmes de reconnaissance vocale utilisent des observations générées par des mixtures gaussiennes décrites par une matrice de covariance Σ un vecteur de moyenne μ. Cette mixture de gaussienne permet de modéliser l'observation et estime la probabilité d'émission pour chaque état par :

$$b_i(o) = \sum_{k=1}^{M} C_{jk} \mathrm{N}(o, \mu_{ik} \Sigma_{ik}) \tag{2-12}$$

Tel que O est le vecteur d'observation cible, $N(o, \mu_{ik}\Sigma_{ik})$ est une fonction de gaussienne simple avec le vecteur moyenne μ_{ik} et la matrice de covariance Σ_{ik} pour l'état i, M représente le nombre de gaussienne et C_{ik} est le poids du $k^{éme}$ gaussienne.

Cependant, malgré l'utilisation des mixtures de gaussienne, le nombre de paramètres reste important. Entre autres, certains états peuvent partager des densités d'observation similaires. Afin d'améliorer l'estimation des paramètres, les distributions des états similaires peuvent être attachées ou regroupées selon une règle. Les techniques les plus utilisées pour sélectionner les états à attacher sont l'utilisation des arbres de décision à base des modèles de triphonèmes [Young, 1994].

Un arbre de décision est un arbre binaire dans lequel une question est attachée à chaque nœud. Les questions sont liées au contexte phonétique adjacent à gauche ou à droite. Par exemple, sur la Fig. 2-5, la première question dans l'arborescence « nœud racine » est : *Est-ce le phonème contextuel gauche est nasale ?* Un arbre de décision est construit pour chaque phonème pour regrouper tous les états correspondants de tous les triphonèmes. Chaque groupe d'état dans les nœuds feuilles de l'arbre formera un seul état.

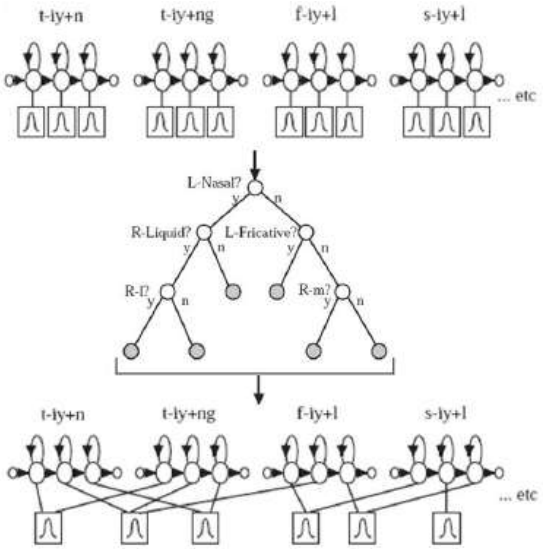

Figure 2-5 : Exemple des états attachés d'un MMC avec un arbre de décision phonétique [Young, 1994]

Entre autres, les réseaux de neurones artificiels « *ANN* » peuvent être utilisés pour estimer les probabilités d'émission. Dans les littératures, nous trouvons l'épreuve que la sortie d'un classifieur *ANN* peut être interprétée comme l'estimation de la probabilité a posteriori de la classe de sortie relative aux données d'entrées [Bishop, 1995]. Ensuite, la probabilité de sortie d'état peut être estimée en appliquant la règle de Bayes aux sorties [Bourlard, 1994]. L'approche hybride *HMM/ANN* a été utilisée dans un grand nombre de systèmes de reconnaissance tel que les travaux de *Gold* [Gold, 2011].

e. Les estimations initiales

L'algorithme largement utilisé dans cette étape est l'algorithme de « *Baum-Welch* ». Il utilise des estimations initiales des probabilités des transitions et d'observations. En effet, il tend à chercher un maximum local de la fonction de vraisemblance en basant sur les valeurs d'estimation initiale la plus proche possible du maximum global.

En effet, nous trouvons que les travaux de recherches enregistrés dans ce contexte ont montré que les applications vocales et en particulier celles qui utilisent les HMM discrets peuvent fonctionner bien avec des estimations initiales aléatoires ou estimations initiales uniformes [Rabiner, 1993]. Toutefois, lorsque les observations sont dans l'espace continues, ils appliquent des méthodes plus développées pour calculer l'estimation initiale par exemple les techniques de segmentation avec *k-means* pour l'extraction de la fonction de densité de probabilité pour chaque état [Fukunaga, 1990]. Entre autres, nous trouvons d'autres méthodes pour l'estimation des paramètres de la fonction de densité de probabilité. Elles repartirent la séquence d'observations sur tous les états du modèle, puis elles exécutent une segmentation sur le maximum de vraisemblance de la séquence jusqu'à un critère d'arrêt est atteint [Young, 2008].

Aussi, nous citons que les modèles de mixture de gaussiennes peuvent être estimées avec une division incrémentale des densités gaussiennes pour chaque itération.

2.2.3.2. La modélisation du langage

Le langage est l'aptitude de mettre en œuvre un système de symboles linguistiques permettant la communication et l'expression de la pensée. Cette aptitude peut être mise en œuvre par un langage avec des moyens vocaux, graphique ou gestuel. Cependant, pour la reconnaissance de la parole continue, la seule information acoustique est insuffisante pour la transcription d'une suite de mots correctement. Dans ce contexte, elle nécessite l'utilisation un modèle de langage qui permet le choix des termes candidats lors du processus du décodage du flux parlé. En effet, nous trouvons deux types d'approches pour la construction des modèles de langages qui sont principalement utilisés : les approches à grammaire formelle guidées par des experts linguistiques et les approches probabilistes qui utilisent des corpus pour la construction des modèles stochastiques pour l'estimation automatique des probabilités des suites des mots du langage cible.

a. L'approche formelle

Les approches à syntaxe formelle sont exprimées par un ensemble de règles à l'aide des graphes des grammaires non contextuelles « *GNC* ». En effet, la génération d'un ensemble de règles décrivant un langage est un processus long et difficile. Dans la littérature nous trouvons que les limites de ces approches est l'impossibilité de décrire de façon exhaustive un langage au travers d'une grammaire à base de règles [Lefevre, 2000]. Aussi, la limitation importante des approches formelle réside dans leurs incapacités à reconnaître des messages contiennes des erreurs grammaticales comme la parole spontanée avec ses faux départs, hésitations …etc. En revanche, la nécessité de la recherche des approches permettant la modélisation d'un langage d'un point de vue strictement probabiliste ce qui permet une modélisation la plus proche à la nature du discours humain.

b. L'approche Probabiliste

Le principe d'un modèle de langage probabiliste est de capturer des séquences ou des modèles réguliers dans un ensemble de suites de mots. En réalité, la probabilité qu'un certain mot soit prononcé dépend de tous les mots l'ayant précédé dans la phrase. Cette modélisation consiste à donner la probabilité d'un mot à partir de la séquence de mots qui le précède. La probabilité d'une séquence de N mots, $P(M)$ est le produit des probabilités conditionnelles d'un mot sachant les mots qui précèdent. Elle s'écrit comme suit :

$$P(m_1 \dots m_N) = P(m_1) \prod_{i=2}^{N} P(m_i / m_1 \dots m_{i-1}) \qquad (2\text{-}13)$$

En pratique, la manipulation des probabilités de toutes les suites de mots possibles est quasiment irréalisable, ce qui implique de réduisez le nombre de précédents à utiliser pour un terme. Dans ce contexte, nous trouvons que le choix de se limiter de trois ou quatre précédents est un choix consistant. Donc, nous trouvons des systèmes de reconnaissance qui utilisent les modèles trigrammes ou quadri-grammes [Lefevre, 2000].

c. Le modèle n-gram

Cette approche est l'approche la plus commune dans l'état de l'art pour les *SRAP*. Elle ne prend pas en compte la structure complexe de la langue. Elle est basée uniquement sur une classification d'équivalence très simple, qui utilise uniquement l'historique des *n-1* mots qui précèdent le mot cible. Cette approche est appelée le modèle de langage *n-gram*. La probabilité $P(M)$ est approximée avec le modèle *n-gram* par la formule suivante :

$$P(M) = P(m_1\ m_2 \dots m_N) \approx \prod_{i=1}^{N} P(m_i / m_{i-n+1} \dots m_{i-1}) \qquad (2\text{-}14)$$

En effet, le choix de la valeur de $\boldsymbol{n}$ est un compromis entre la stabilité de l'estimation et sa performance. Le *tri-gram* « n=3 » est un choix commun pour les grands corpus d'apprentissage. Alors que, le *bi-gram* « n=2 » est souvent utilisé avec des corpus de petite taille. Cependant, l'augmentation de la valeur de n accroître la difficulté de l'estimation de la probabilité à priori. Cette probabilité peut-être estimer par l'approche de fréquence relative définie par :

$$P(m_i / m_{i-n+1} \dots m_{i-1}) = \frac{F(m_{i-n+1} \dots m_{i-1} m_i)}{F(m_{i-n+1} \dots m_{i-1})} \qquad (2\text{-}15)$$

Pratiquement, les fréquences sont estimées à base des corpus de textes de taille importante et elles sont estimées par leurs fréquences d'apparition. Le problème essentiel est la valeur exponentielle du nombre des *n-grammes* obtenus par rapport à la valeur de *N* utilisé. En effet, pour un vocabulaire simple de mille mots nous pouvons avoir un milliard de trigrammes distincts. Cependant, de point de vu statistique, la majorité de ces trigrammes apparaissent rarement dans le corpus utilisé et génèrent des problèmes pour le calcul des probabilités apriori [Lefevre, 2000].

d. Le Lissage du modèle n-gram

Les techniques de lissage sont conçues pour remédier les problèmes des *n-grammes* rares. Ces techniques consistent à prendre de la masse de probabilité des *n-grammes* observés, pour donner une valeur non nulle aux probabilités des n-grammes non observés ou rarement observés. L'une des techniques de lissage les plus utilisées est la technique dite de « *Kneser-Ney* » [Kneser, 1995]. Dans cette technique, les probabilités des *n-grammes* rarement observés sont estimées comme avec les autres techniques de lissage, en faisant un repliement « *Back off* » sur les observations d'ordre moins grand.

Pour un tri-gramme par exemple, le bi-gramme puis l'uni-gramme si nécessaire sont utilisés. L'originalité de cette technique est de ne pas prendre la même distribution de probabilité pour les ordres les plus petits que n. Au lieu de prendre la fréquence d'observation d'ordre *n−1* à savoir m_{i-n+1}^{i-1} , c'est le nombre de contextes différents dans lesquels se produit m_{i-n+1}^{i-1} qui est consulté. L'idée est que si ce le nombre est faible alors la probabilité accordée au modèle d'ordre « *n-1* » doit être petite et ceci même si m_{i-n+1}^{i-1} est fréquent. Ainsi le biais potentiel introduit par la fréquence d'observation est évité [Pellegrini, 2008].

e. Approches plus évoluées

Bien que l'estimation robuste des modèles *n-grammes* reste un problème complexe malgré les techniques introduites. En effet, l'intégration des modèles de langage lors du processus du décodage conduit à des gains significatifs de point de vue performance. Toutefois, l'étude et la recherche des solutions plus évoluées pour surmonter les limites de ces approches comme l'exploitation des dépendances plus profonde telle que les accords sujet-verbe.

Dans ce contexte, nous trouvons un panorama d'approches de modèles de langage évolués développées comme : *Tree-Based Models, Cache Models, Treillis Models, Trigger Models*, … etc. [Lefevre, 2000]. En effet, leurs utilisations sont limitées par la difficulté de leurs intégrations dans les algorithmes de décodage classiques ainsi que la généralisation de leurs usages dans les systèmes de reconnaissance automatique.

f. Evaluation des modèles de langage

En effet, la question primordiale est comment valoriser l'impact de ces approches dans les performances des systèmes de reconnaissance de la parole. Entre autres, comment peut-on comparer les différentes approches des modèles de langages. Une manière de mesurer la qualité de ces approches est d'estimer la probabilité de séquences de mots qui ne font pas partie du corpus d'apprentissage du modèle. La probabilité d'un texte $M = m_1 m_2 \dots m_n$ appelée vraisemblance « *Likelihood* » et notée lh.

$$lh(M) = \hat{P}(m_1 m_2 \dots m_n) \tag{2-16}$$

Cependant, plus la valeur de vraisemblance est élevée, plus le modèle de langage est capable de prédire les mots contenus dans le corpus. En effet, la valeur de $L\widehat{P}$ représente la valeur estimée de la probabilité par le modèle de langage [Pellegrini, 2008].

Ainsi, la grandeur la plus utilisée pour caractériser les performances d'un modèle de langage est la perplexité, souvent notée pp. Elle est définie par :

$$pp = {}^{1}/_{\hat{P}(m_1^N)^{1/N}} \tag{2-17}$$

Cette mesure est équivalente à la vraisemblance mais elle fait intervenir une normalisation sur le nombre de mots du corpus utilisé. Plus la probabilité de la séquence de mots est grande, plus la vraisemblance est grande, plus la perplexité est petite.

2.2.3.3. Les techniques de décodage

La tâche de reconnaissance de la parole continue avec les modèles stochastiques est effectuée par un décodage statistique. Son rôle est de chercher dans un espace d'hypothèses très grand, le meilleur chemin qui donnera la séquence de mots la plus probable. Le processus de reconnaissance suppose que les symboles que l'on cherche à reconnaître s'enchaînent sans que l'on sache pour un segment donné ni leur nombre ni leur localisation. Dans le cadre de la SRAP pour la parole continue, l'approche du décodage contraint à envisager le décodage de tous les messages possibles. Une double problématique doit être alors résolue : la génération d'une structure de recherche représentant tous les messages possibles et le décodage à l'aide de cette structure d'une suite d'unités correspondant au message prononcé.

Il existe de nombreuses stratégies et techniques de décodage, leurs utilisations dépendent étroitement par les contraintes de temps et puissance de calculs et de la taille de vocabulaire utilisée. Entre temps, n point commun entre ces différentes stratégies et techniques est le compromis nécessaire entre la taille des modèles en nombre de paramètres et les réductions de l'espace de recherche par les techniques d'élagage « *Pruning* » [Chou, 2003].

a. Espace de recherche

L'espace de recherche peut être défini comme étant une automate à états finis, où les états sont les mots et leurs transitions définis dans le modèle de langage. La figure 2-6 présente un exemple d'une automate d'un modèle de langage bi-gramme. La définition de l'espace de recherche est effectuée par la combinaison du modèle de langage avec le modèle acoustique pour toutes les séquences possibles des mots du vocabulaire. Généralement, l'espace de recherche est modélisé par les modèles Markoviens *MMC*. La séquence la plus probable d'un mot donné peut être trouvée par l'algorithme de « *Viterbi »* qui sera détaillé dans la section suivante. La complexité de décodage est liée étroitement par la complexité de l'espace de recherche. Dans les systèmes de reconnaissance à grand vocabulaire, le nombre de mots est assez important. En plus, l'unité de modélisation est les phonèmes ou les syllabes, ce qui rend l'espace de recherche très complexe. À cet effet, le décodage exhaustif dans cet espace est impossible. D'où la nécessité d'adaptation des techniques de recherche plus efficaces.

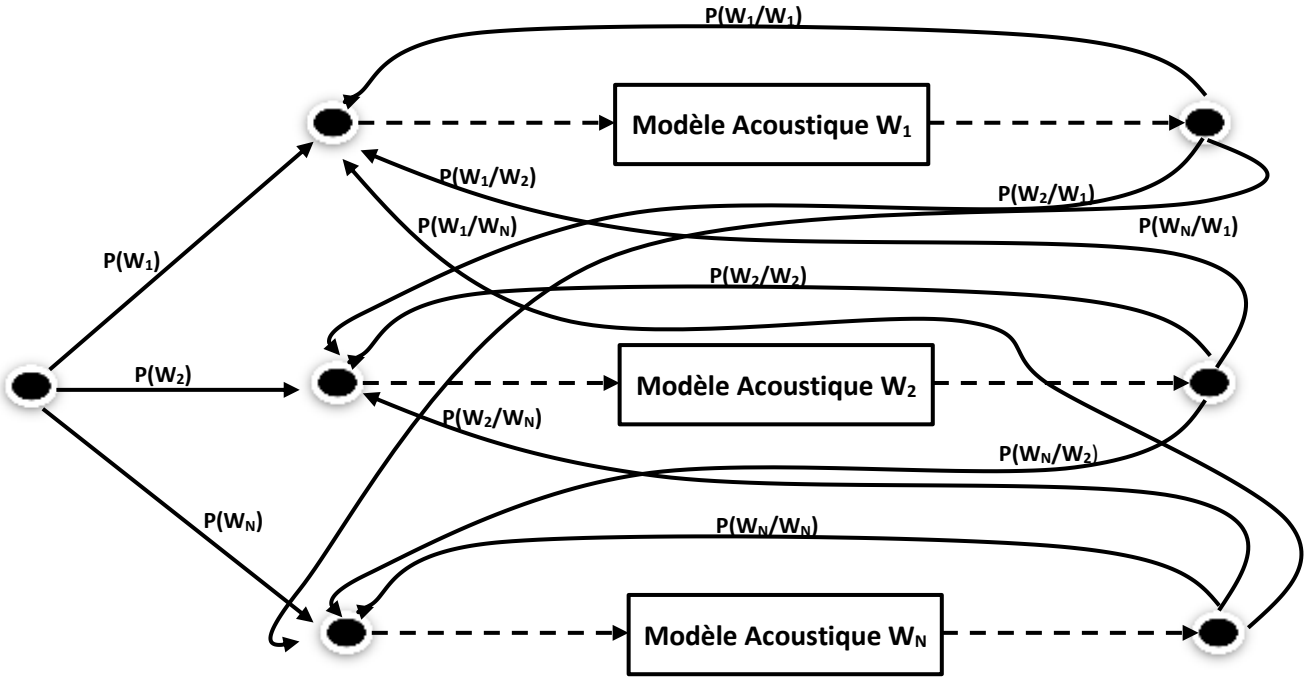

Figure 2-6 : Exemple d'un automate du modèle de langage bi-gramme [Adami, 2010]

b. L'algorithme Viterbi à un seul passage pour le décodage de LVCSR

L'algorithme *Viterbi* nécessite un espace de recherche qui contient tous les chemins possibles des phonèmes construisant les différents mots de l'automate du modèle du langage utilisé. Cependant, pour rendre cette automate exploitable, elle est représentée sous forme de structure arbre, comme la présente la figure 2-7. Ce réseau d'arbre est établi dynamiquement au besoin, où l'on partage les modèles avec les différentes hypothèses qui commencent par les mêmes ordres des sous-mots ou tri-phonème.

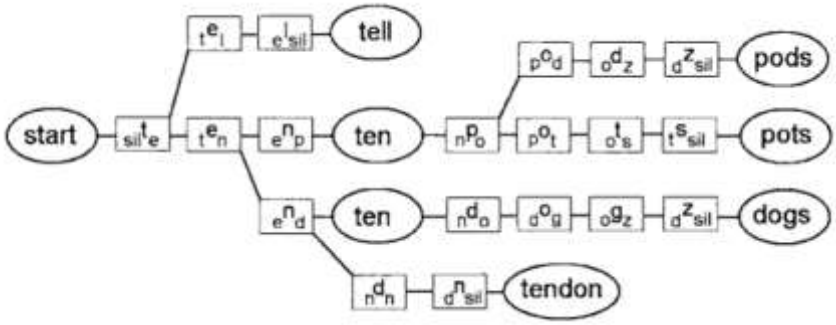

Figure 2-7 : Un fragment très petit d'un réseau pour décodage dans une automate de modèle de langage trip-honème. Ce fragment présente la séquence « ten pots » avec quelques chemins possibles dans le réseau de décodage. Noter que différents nœuds dans le réseau, selon si le mot suivant est des « pots » ou des « dogs », représentent le mot « dix », [Holmes, 2001].

Cependant, l'élagage efficace de l'arbre est primordial pour les systèmes de LVCSR. La stratégie habituelle est la recherche vectorielle, par lequel à instant *t* de la séquence recherchée, on élimine tous les chemins qui ont une probabilité qui n'appartient pas aux probabilités de meilleur chemin par la technique de « *Best-scoring path* ». En pratique, généralement c'est tous les chemins sauf ceux qui ont des tendances de probabilités très faibles [Holmes, 2001].

Entre temps, le vocabulaire du langage utilisé permet de cerner et limiter l'ensemble de mots qui sont probables à n'importe quel point donné dans une expression ou une séquence acoustique. Donc, l'intégration du modèle de langage pour élaguer les chemins moins probables durant le décodage d'une séquence acoustique par l'algorithme de *Viterbi* conventionnel.

c. Décodage Viterbi à passes multiples

Parmi les solutions qui permettent d'accroitre la qualité de reconnaissance avec le décodage de Viterbi à une seule passe est d'utiliser ce dernier on plusieurs passe en multi-passe. À cet effet, Il exécute un décodage simple pour extraire un nombre limité d'hypothèses probables, puis il exécute un autre décodage plus profond pour chercher l'hypothèse la plus probable. Par exemple, la première passe utilise une modélisation en tri-phonèmes simple pour les mots avec un modèle de langage bi-gram. Tandis que, la deuxième passe utilise une modélisation en tri-phonèmes interconnectés pour les mots avec un modèle de langage tri-gram.

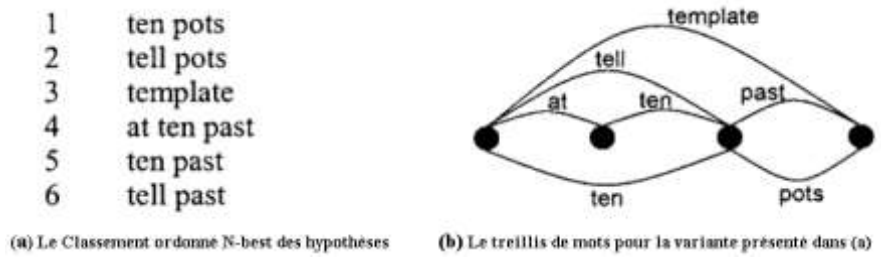

Figure 2-8 : Les alternatives possibles qui peuvent être générés pour un court énoncé d'après (Holmes, et al., 2001)

La sortie de la première passe de décodage est souvent exprimée par classement ordonné *N-best* de toutes les séquences possibles du mot, ou par un réseau de graphe ou treillis de mots qui contient toutes les séquences probables. La figure 2-8 présente un exemple de *N-best* séquences possibles et le treillis de mots correspondant.

Plusieurs algorithmes peuvent être adaptés pour fournir la liste des N meilleures hypothèses [Huang, 2001]. Parmi eux nous trouvons l'algorithme de décodage à pile « *Stack-decoding* » qui fournit une phrase complète en choisissant la meilleure hypothèse partielle et l'algorithme progressif-rétrogressif « *Forward- Backward* »

L'algorithme progressif-rétrogressif utilise une recherche approximative temporelle synchrone dans la direction vers l'avant pour faciliter une recherche plus complexe et coûteuse dans la direction arrière. Il utilise un modèle acoustique ou de langage simplifié pour effectuer une recherche passe-avant « *Forward* » rapide et efficace, dans laquelle il stocke le score de tous les chemins partiels qui surmontent le seuil d'élagage défini. Ensuite, il effectue une recherche en arrière « Backward » pour générer la liste de N meilleurs hypothèses « N-best ». La recherche arrière donne un score élevé sur une hypothèse seulement s'il existe aussi une bonne séquence conduisant à la fin d'un mot à cet instant du temps (Holmes, et al., 2001).

2.2.4. Évaluation des systèmes de reconnaissance

Les SRAP sont souvent évalués à l'aide d'une mesure appelée : Taux d'erreur de mots « *Word Error Rate*, WER ». Cette mesure est effectuée à l'aide d'une comparaison dynamique entre la transcription manuelle de référence et celle de l'hypothèse. Dans ce contexte, il existe trois types d'erreurs. Les substitutions (S) correspondent aux mots qui ont été reconnus au lieu d'un mot de la transcription manuelle. Les insertions (I) sont les mots reconnus qui se sont insérés par erreur entre deux mots corrects de la transcription de référence. Ainsi que les suppressions (D) qui correspondent aux mots de la référence qui ont été non reconnus dans l'hypothèse de reconnaissance.

$$\boldsymbol{WER = \frac{S + I + D}{N}} \qquad (2\text{-}18)$$

Cependant, nous trouvons dans la littérature d'autres métriques plus spécifiques comme celle introduite pour estimer la fidélité sémantique des transcriptions réalisées par *Sarikaya* pour des systèmes d'interprétation de dialogue, d'indexation, …etc. [Sarikaya, 2005].

Entre temps, nous trouvons que pour l'évaluation de la fiabilité de ces mesures statistiques, il convient de calculer un intervalle de confiance relatif au nombre d'échantillons et d'erreurs. Cet intervalle de confiance est calculé en considérant que l'apparition d'une erreur de reconnaissance sur un mot ou sa non-reconnaissance est associée à une variable aléatoire binomiale, dont la distribution dépend des couples des mots reconnus et les mots prononcés [Lecouteux, 2008].

2.3. Recherches actuelles sur la détection des termes parlés

La détection des termes parlés « *Spoken Term Detection* -STD » est une tâche très différente par rapport aux tâches de reconnaissance automatique, car elle est basée sur le principe de probabilité d'existence d'un segment donnée dans un flux de parole au lieu de décrypter ou transcrire la totalité du flux parlé pour vérifier l'existence d'éléments recherchés.

Cependant, le principe de détection des segments ou des mots-clés parlés consiste à trouver les occurrences phonétiques des mots prononcés dans le contenu du flux parlé. En effet, les techniques STD sont appliqués pour la détection des termes ou même des séquences de mots dans le contenu du flux parlés. Cependant, la détection de mots-clés est considérée comme une partie de STD. Dans ce contexte, les défis importants dont elle cherche de résoudre les techniques de détection des termes parlés peuvent être résumés dans les axes suivants :

- L'amélioration de la performance de processus de détection.
- L'amélioration de la qualité de l'indexation.
- L'accélération du temps de recherche.
- L'amélioration du pouvoir de traitement des termes étranges du vocabulaire utilisé comme les mots hors vocabulaire « *Out of Vocabulary* – OOV » et les termes techniques.
- Le traitement des différents variantes et modes de prononciation acoustiques.

Dans les littératures, nous trouvons que ces défis ont été abordés en utilisant des approches différentes, qui seront décrites dans les sections suivantes [Hrtmam, 2016], [Lin-shan, 2015].

2.3.1. Approches générales de STD

La détection de termes parlés « STD » est considérée comme une variante du problème de la reconnaissance de la parole. Néanmoins, un bon nombre d'approches destinées à la reconnaissance de la parole trouvent leur applicabilité dans les systèmes *STD* avec des modifications appropriées. Dans les sections qui se suivent, nous présenterons une classification générale des différentes approches utilisées pour la détection de termes parlés, et nous nous concentrerons sur les trois premières approches supervisées : approche par détection de mots clés ; approche à base de LVCSR ; approche à base des sous-mots « *Subword* ».

2.3.1.1. Approches supervisées

Dans cette catégorie, nous trouvons les approches qui sont basées sur les notions d'apprentissage et de classifications. Parmi eux, nous pouvons citer :

- Les approches basées sur la détection des mots clés acoustique dans le contenu des flux parlés.

- Les approches basées sur des unités plus fines que le mot tel que les sous-mots « *sub-word* » avec des systèmes de reconnaissance de la parole continue à large vocabulaire
- Les approches de détection des termes parlés à base des exemples de requêtes « *Query-by-Example* : *Text based STD* »
- Les approches discriminatives de détection des mots clés comme « *Keyword Spotting* ».

2.3.1.2. Approches non supervisées

Pour cette catégorie, nous trouvons les approches et les techniques basées sur le principe « *Query-By-Exemple* -QBE » en utilisant l'appariement des modèles « Template Matching ». Parmi eux, nous citons :

- Les approches basées sur l'alignement dynamique de modèle des trames « *frame based template matching* ».
- Les approches basées sur la segmentation par le biais de l'alignement dynamique des modèles « *Segment based template matching* ».

2.3.2. Les principes des STD

Dans les littératures, nous trouvons qu'une grande partie des recherches antérieures effectuées dans ce domaine sont basées sur la détection acoustique des mots clés proposée par *Rose* [Rose, 1996]. En effet, dans cette thèse nous avons orienté vers l'étude de ces approches, car elles sont les plus largement utilisées dans les travaux de recherche pour la problématique de détection des termes parlés. Dans ce contexte, dans les sections qui se suivent, nous présentons un résumé des recherches effectuées dans le domaine de la détection acoustique de mot-clé.

En effet, beaucoup de systèmes STD actuels utilisent les techniques de reconnaissance de la parole continue à large vocabulaire « LVCSR ». Cette dernière nécessite un apprentissage supervisé de ses modèles, qui sont souvent des modèles probabilistes stochastiques : HMM-GMM. Cependant, les systèmes STD peuvent être appliqués aussi pour la majorité des langues humaines utilisées voire même sur quelques dialectes. A cet effet, des algorithmes de décodage ont était développés pour permettre le traitement automatique de ces langages et même pour les langages moins dotés ou celles qui ne disposent pas des ressources suffisantes pour le processus d'apprentissage de ses modèles de représentation, [Mandal, 2014], [Boves, 2009].

Cependant, même pour les langues bien dotées, les techniques et les approches de détection des termes parlés STD souffrent des limitations que celles des systèmes *LVCSR* comme les problèmes liés à la détection des mots hors vocabulaire « OOV ». Aussi, nous citons les problèmes de performance de ces systèmes vis-à-vis les termes techniques et les qui présente

des concepts de haut niveau [Novotney, 2009]. En effet, ces aspects rendent ces systèmes moins efficaces pour les tâches de STD et surtout pour les systèmes dont les ressources utilisées durant l'apprentissage des modèles de langage approprié ne sont pas assez riches. Cependant, ce problème a été abordé par les systèmes de détection de mots clés qui ont fait l'usage de systèmes de reconnaissance à base de phonétiques [James, 1994] ; [Thambiratnam, 2005] ; [Vergyri, 2006] ; [Mamou, 2007] au lieu de système de reconnaissance à base de mots entiers. La figure 2-9 synthétise les différentes approches utilisées dans les systèmes STD.

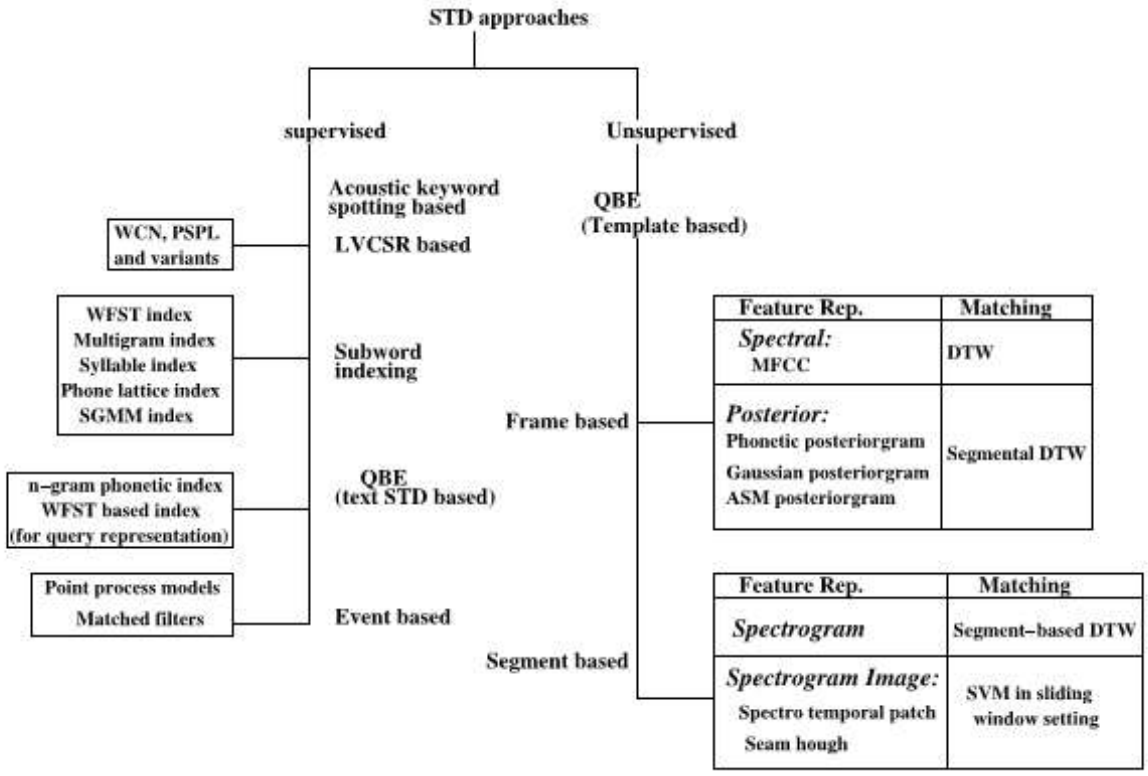

Figure 2-9 : Taxonomie des approches de STD d'après [Mamou, 2007]

2.3.3. Travaux existants sur les approches de STD

2.3.3.1. Détection de mots clés acoustiques

Les systèmes de détection de mot-clé acoustique sont basés sur une architecture a deux composants : les modèles mots clés « *Keyword Model* » et les modèles mots poubelles « *Filler Model* ». Ils utilisent un réseau parallèle de modèles des mots clés et un réseau des modèles poubelles [Rose, 1996]. En effet, le modèle construit pour un mot clé est obtenu par la concaténation des modèles de phonèmes constitutifs. Les modèles poubelles aussi sont construits en utilisant des boucles de phonèmes. Chaque phonème est modélisé par les modèles *HMM/GMM* en utilisant les modèles probabilistes stochastiques.

Les réseaux de neurones sont également utilisés pour la modélisation de phonèmes [Szoke, 2005]. Les résultats sont obtenus par le calcule d'un score de vraisemblance « *log-likelihood* » en utilisant les modèles de mots clés et les modèles poubelles conjointement « *Background filler model* ». Entretemps, la décision sur les mots clés obtenus est accordée par l'utilisation des techniques de classification binaire pour les scores de vraisemblances.

D'autre part, l'apprentissage des modèles *HMM/GMM* est généralement est effectué via des techniques basées sur la maximisation de probabilité. Cependant, nous trouvons dans des travaux de recherches effectués que l'objectif de la phase d'apprentissage est de maximiser la qualité de la transcription de paroles sans tenir compte les performances du processus de détection de ces mots clés [Grangier, 2009]. Dans ce contexte, nous trouvons que les approches d'apprentissage discriminant sont conçues essentiellement pour surmonter cette problématique. Elles maximisent durant le processus d'apprentissage les différents critères qui ont un impact direct sur la performance de la détection des mots clés.

Cependant, nous trouvons d'autres approches comme celle proposée dans les travaux de *Sukkur* qui visent à maximiser le rapport de vraisemblance « *log-likelihood* » entre les modèles de mots clés et les modèles poubelles pour les occurrences des mots clés. Entre temps, elle sert à minimiser cette vraisemblance pour les occurrences de fausses alarmes générées par les modèles des mots clés [Sukkar, 1996]. Entre autres, nous trouvons les travaux réalisés par *Sandness* qui utilisent les techniques d'erreur de classification minimum « *Minimum Classification Error – MCE* » pour surmonter les problèmes de détection des mots clés [Sandness, 2000].

En revanche, nous trouvons d'autres approches qui utilisent une combinaison de différents modèles *HMM/GMM* de détection des mots clés. Parmi eux, nous citons la combinaison de réseaux de neurones avec les scores de rapports de vraisemblances [Weintraub, 1997]. Ainsi que la combinaison des *SVMs* avec les scores de rapport de vraisemblances « *log-likelihood* » des phonèmes [Benayed, 2003]. Entre temps, nous trouvons les travaux qui proposent une méthode qui utilise une procédure d'apprentissage discriminante, dans laquelle la phase d'apprentissage vise à maximiser la surface sous la courbe ROC « *Receiver Operating Characteristic* » [Keshet, 2009].

En effet, la limitation majeure de ces systèmes acoustiques basés sur l'approche détection de mots clés réside dans la difficulté de traitement des nouveaux mots clés. Notamment, pour la détection d'un terme parlé, le système doit exécuter des itérations de décodage avec la nouvelle liste de mots clés pour chaque fois qu'un nouveau un modèle de mot clé est inséré dans le système. Ce qui implique un temps de recherche excessivement élevée. Cette limitation est traitée dans les systèmes STD par la solution d'utilisation des techniques basés sur LVCSRs et de reconnaissance de sous-mots comme décrits ci-après.

2.3.3.2. STD utilisant LVCSRs

Dans cette catégorie, nous trouvons des efforts de recherches considérables pour la détection des termes parlés dans flux parlé. Ils ont porté sur l'extension des techniques de recherche d'informations disponibles pour le texte aux documents parlés. Parmi ces travaux, nous citons celles ayant un système à base de LVCSR pour générer la transcription en niveau de mots correspondante au contenu du flux parlé en entrée. Ensuite, ils ont utilisé les techniques d'indexation et de recherche utilisées dans les systèmes de recherche d'information [Garofolo, 2000]. Cependant, souvent la transcription du mot généré par l'algorithme de décodage *1-Best* de l'LVCSR contient des erreurs de transcription. Ces erreurs, soit par insertion ou omission, affectent les performances des systèmes STD.

Par conséquent, l'utilisation des résultats du décodage basé sur les treillis de mots « *words lattices* » dans le processus de l'indexation au lieu de la sortie *1-Best* du LVCSR est fortement sollicité. Les treillis de mots sont des graphes acycliques dirigés où chaque nœud dans le treillis est associé à un « *timestamp* ». Ainsi que, chaque branche *(u, v)* est marquée avec un mot ou une hypothèse de phonème et la probabilité apriori qui est la probabilité du signal délimité par les *timestamps* des nœuds u e tv, qui forme l'hypothèse [Garofolo, 2000].

Entre autres, nous trouvons l'utilisation d'une représentation similaire au treillis de mots, mais plus compact d'un treillis de mot est appelé les réseaux de confusion à base de mots « *Word Confusion Network – WCN* ». Dans cette configuration, chaque branche *(u, v)* est marquée avec une hypothèse de mot et sa probabilité postérieure. Cette dernière présente la probabilité du mot donné par le flux parlé cible. La construction des *WCN* est basée sur les chemins de mots. Tous les chemins des mots qui se chevauchent temporellement sont regroupés dans des chemins respectifs, quelles que soient les positions de ces chemins. En effet, les *WCNs* fournissent un alignement temporel strict pour tous les mots dans le treillis [Hakkani-Tür, 2003], [Mangu, 2000].

En revanche, nous trouvons des travaux qui proposent une représentation plus compacte que celles de treillis de mot et des réseaux de confusion. Ces représentations sont appelées « *Posterieur Specific Position Lattice - PSPL* » [Chelba, 2005]. Elle est basée sur le calcul des probabilités postérieur de positions d'un mot dans le treillis des mots. La technique de représentation *PSPL* calcule la probabilité postérieure d'un mot *W* à une position spécifique dans un treillis. Tous les chemins dans le treillis sont énumérés, chacun avec son propre poids de chemin. La probabilité postérieure d'un mot donné à une position donnée est calculée en additionnant tous les poids de chemin qui incluent le mot indiqué à une position indiquée et puis elle est divisée par la somme des poids dans le treillis. La figure 2-10 montre un treillis de mot et sa représentation correspondante de *PSPL* et de *WCN*.

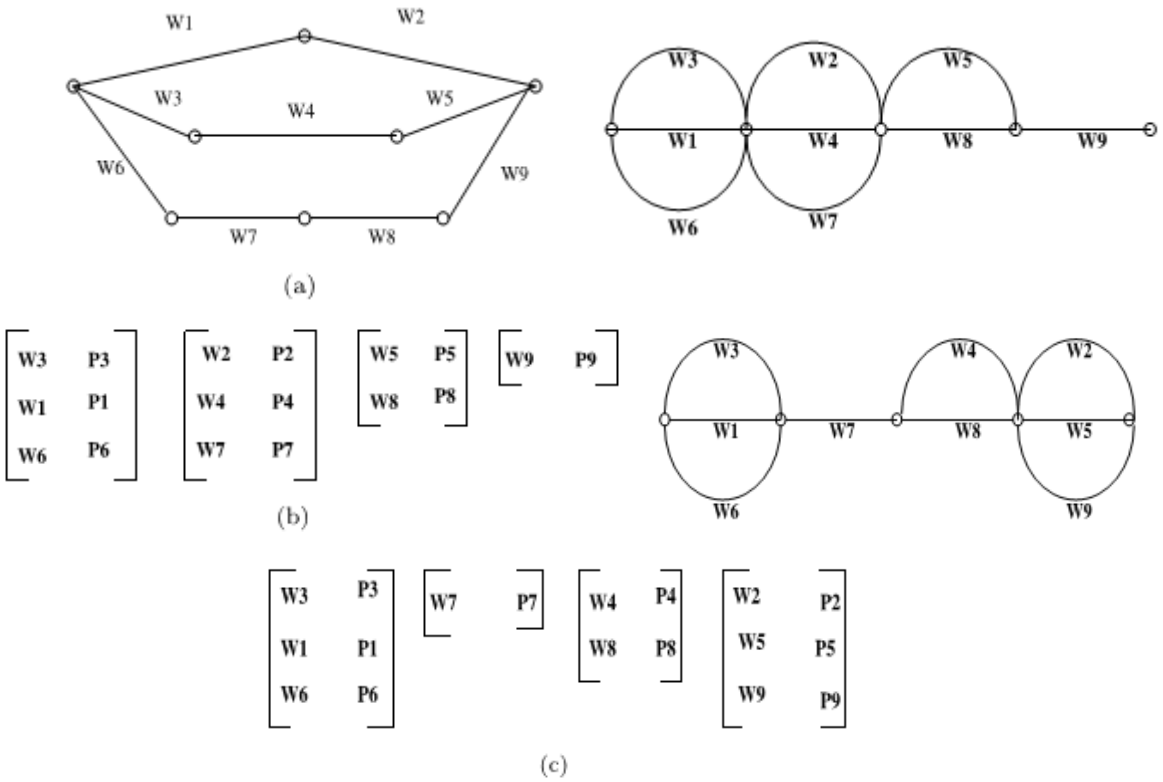

Figure 2-10 : (a) Treillis de mot avec sept mots et leurs (b) PSPL et (c) WCN correspondants respectivement. Les W_i représentent des mots-clés et $p_i{}'$ s représentent les probabilités postérieures correspondantes liées à chaque mot d'après [Chelba, 2005].

Les treillis de mot et ses variantes ont été employés avec succès pour améliorer le taux de détection des mots de vocabulaire « In vocabulary - IV ». Cependant, ils ne peuvent pas traiter les mots hors vocabulaire « OOV ». En effet, nous trouvons une comparaison détaillée des performances de PSPLs et WCNs [Pan, 2010]. Cette comparaison décrit que la performance de la représentation PSPLs est toujours meilleure à celle de WCNs, mais elle exige un espace de plus pour le stockage des index. Entretemps, l'utilisation des unités inférieure au mot comme les *sub-word* pour les deux modèles de représentation *S-PSPLs* et *S-WCNs* permettent l'amélioration de la technique par rapport à celles à base de mot. Les auteurs de cette comparaison constatent que les représentations S-PSPLs et S-WCNs, rapportent toujours une précision moyenne de performance « *Mean Average Precision-MAP* » bien mieux pour les deux types de requêtes OOV et IV tout en consommant beaucoup moins d'espace mémoire que les représentations PSPLs/WCNs.

2.3.3.3. STD utilisant des systèmes de reconnaissance à base de sous mot « *sub-word* »

Dans cette catégorie, nous présentons les approches qui impliquent l'utilisation des index avec différentes unités de sous-mots tels que des n-grammes, des phonèmes, des multi-grammes, des syllabes, des segments ou des représentations de treillis des unités phonétiques [Ng, 2000] ; [Szoke, 2008]. Ces index sont obtenus via les résultats de la transcription du contenu du flux parlé par les systèmes de reconnaissances de parole continue avec l'utilisation conjointement des techniques utilisées dans les systèmes de recherches d'information textuelles [Ng, 2000].

Dans ce contexte, Les transcriptions phonétiques du contenu du flux parlé sont obtenues par les systèmes de reconnaissance de large vocabulaire à base de phonèmes. En effet, elle utilise les transcriptions de niveau de phonèmes pour obtenir des unités de sous mots de complexité variable en matière de leurs longueurs. Les résultats obtenus montrent que la précision moyenne « *MAP* » pour les index tri-phonème est meilleure à celle de quinque-phonème [Ng, 2000].

Entre temps, nous trouvons d'autres travaux qui étudient l'impact des paramètres de multi-grammes ; à savoir sa longueur et le facteur d'élagage sur la taille de l'index sur l'efficacité des STD ; lors de son utilisation comme unités de sous mots pour traiter les mots hors vocabulaire « OOV » [Szoke, 2008]. Les résultats montrent que les performances de détection les plus élevées sont obtenues avec des unités de multi-grammes de longueur cinq. Cependant, l'impact de facteur d'élagage est minime sur les performances de détection de phonème. Cependant, ils proposent deux méthodes d'apprentissages de multi-grammes pour améliorer la performance de phonème et de la performance de STD. Ils constatent également que l'incorporation du modèle standard de langage de n-gramme sur des unités de multi-gramme est bénéfique, et avec le modèle de langage de trigramme donne la meilleure performance.

En revanche, nous trouvons des travaux présentent une amélioration de la précision de détection de mot exprimée par F-scores pour les requêtes du type IV et OOV par l'utilisation des deux types de treillis : à base phonétique et à base de mot [Saraclar, 2004]. Les auteurs proposent trois stratégies de détection. La première c'est une combinaison des résultats après un processus de recherche du mot et l'index phonétique. Dans la deuxième, ils suggèrent la recherche de l'index de mot pour les requêtes IV et l'index phonétique pour les requêtes de OOV. Ainsi que, dans la troisième stratégie ils recherchent l'index phonétique seulement si la recherche de l'index de mot est échouée.

Entre autres, nous trouvons des approches qui abordent le problème des requêtes hybrides [Mamou, 2007]. (Mamou, et al., 2007). Elles emploient deux index, le premier à base de WCN pour le stockage d'index et le deuxième à base de treillis de phonème pour les index phonétique. Ils stockent les « timestamps » correspondant au début et fin pour chaque unité de l'indexation. Ainsi que, pendant la recherche d'une requête sur un terme IV, une liste d'émission est extraite à partir de l'index de mot. Pour une requête sur un terme OOV, le

terme est converti en une séquence de phonèmes en utilisant un maximum d'entropie de modèles *N-gramme* commun. Pour une requête hybride de mot-clé impliquant les deux types de termes IV et OOV, l'index de mot pour les termes IV et l'index phonétique pour les termes OOV sont employés. Dans ce cas, les listes d'émission des termes IV extraites de l'index de mots sont fusionnées avec les listes d'émission des termes OOV obtenus à partir de l'index phonétique. Le résultat final de la requête hybride est obtenu par l'union ou l'intersection des résultats des requêtes individuelles en fonction de la relation entre les termes de la requête. Cette approche est plus performante que les méthodes basées uniquement sur l'index de mot ou l'index phonétique.

D'autre part, nous trouvons des travaux qui tentent de diminuer le taux d'omission et d'augmenter la vitesse de recherche pour la détection de mot-clé dans un vocabulaire illimité [Thambiratnam, 2005]. La stratégie de cette approche est la recherche de correspondance dynamique de treillis de phonèmes « Dynamic *Match Phone-Lattice keyword Spotting* – DMPLS », une extension de la recherche du treillis de phonèmes « *Phone-Lattice keyword Spotting – PLS* » qui peut gérer l'insertion, la suppression et la substitution des erreurs d'un système de reconnaissance à base de phonèmes. Ensuite, elle utilise une représentation phonétique de la parole en utilisant un décodage Viterbi multi passe *N-best*. Le treillis est alors décodé pour la séquence de phonèmes constituant le mot-clé. Entre temps, durant le processus de recherche, des scores de pénalités de coût appropriés sont imposés pour les erreurs de la reconnaissance de phonèmes. En effet, les résultats obtenus sont moins que celles obtenues par des systèmes classiques de détection de mot-clé basés sur les HMM [Rohlicek, 1995], mais la vitesse de recherche est largement meilleure.

Cependant, nous trouvons d'autres approches qui intègrent les transducteurs à états finis « *Weighted Finite State Transducers* -WFST » pour la construction d'une représentation d'index efficace et réduit les exigences de stockage et temps de la recherche [Mohri, 2008]. Dans ce contexte, nous trouvons des travaux qui décrivent un algorithme pour la création d'un index complet représenté par les techniques WFST. Son principe est le mappage de chaque sous-chaîne de terme de *x* à l'ensemble des indices dans les automates dans lesquels il apparait [Allauzen, 2004]. Pendant la recherche, il utilise des opérations de composition unique pour les index de la requête. Ces derniers sont représentés comme des transducteurs à états finis pondérés. Entre temps, nous trouvons une variante de la même structure d'index pour la tâche STD. Elle est nommée « *Timed Transducer Factor*- TFT » et elle utilise les informations de synchronisation pour l'évaluation des poids des chemins [Can, 2011]. En effet, le principe de la technique TFT est la représentation des indices temporels par WFST. Cette représentation mappe chaque facteur du terme *x* dans chaque automate ou il appartient avec les probabilités postérieures *x* qui se produisent réellement dans chaque automate pendant l'intervalle de temps correspondants. En effet, l'avantage de cette approche est que la complexité de la recherche est linéaire par apport à la longueur de la requête, et par conséquent elle est utile pour les requêtes longues.

Ainsi, elle est très souple pour les taches des STD comme la détection de toutes les relations à états finis d'un indice donné. La recherche de relations complexes entre les mots de la requête sans modifier les structures d'index.

Cependant, nous trouvons des approches qui exigent moins de ressources par rapport aux techniques décrites auparavant [Garcia, 2006]. Dans ces travaux, ils utilisent de petites quantités de données vocales ; autour de 15 minutes de flux parlé ; pour l'apprentissage d'un système de reconnaissance auto-organisationnelle. Cette stratégie permet la définition d'unités phonétiques propres au système pour le domaine spécifique. Ensuite, ils ont utilisé les transcriptions pour l'apprentissage d'un convertisseur graphème phonétique. La parole d'entrée est segmentée automatiquement et d'une manière non supervisée puis elle est modélisée par les modèles de segmentation des mixtures de gaussiennes « SGMMs ». Ensuite, les segments obtenus représentant les mots sont décodés en termes d'indices de SGMM. Ainsi, ils utilisent un modèle multi-grammes communs basé les transcriptions parallèles pour l'obtention d'une cartographie probabiliste entre des séquences de lettres dans les transcriptions au niveau mot et des séquences d'indices de SGMM. Ce modèle est utilisé pour prédire la prononciation d'un mot-clé donné en termes des unités de SGMM, au lieu d'utiliser un dictionnaire de prononciation. Enfin, ils effectuent une recherche par les techniques de programmation dynamique pour minimiser la distance entre la prononciation prédite d'un mot-clé et la transcription automatique obtenue.

Dans ce contexte, nous trouvons aussi des approches similaires qui cherchent l'amélioration des performances des systèmes LVCSR en se basant sur les techniques non supervisées pour améliorer l'apprentissage des modèles acoustiques et modèles de langue [Novotney, 2009]. Les résultats obtenus ont montré que le modèle acoustique est plus efficace par rapport à son équivalent supervisé ainsi qu'une amélioration enregistrée dans les taux d'erreurs.

Entre temps, dans toutes ces approches, le format de la requête est présenté sous la forme d'un texte. Par conséquent, ces méthodes supposent la disponibilité de l'expansion phonétique des mots clés, soit en utilisant des règles graphème-phonème ou par d'autres moyens.

2.3.4. Méthodologies d'évaluation des approches STD

L'objectif des systèmes STD est de détecter toutes les occurrences de chaque terme donné dans le corpus parlé. En effet, nous trouvons deux types d'erreurs permettant la caractérisation des performances des STD : les fausses alarmes et les omissions. Pratiquement, les systèmes de détection répondent à la question : « *Cette instance de données elle est un exemple des données d'apprentissage fournies* ». La réponse du cette question est obtenue par un alignement entre les occurrences détectées le système et les occurrences de référence d'évaluation [Tejedor, 2017]. Dans ce contexte, nous citons les méthodes d'évaluation largement utilisées dans les systèmes de détections.

2.3.4.1. Ponderation des Termes « Term Weighted Value - TWV »

Cette mesure est basée sur la probabilité d'omission ou les termes non reconnus P_{Miss} et la probabilité de fausse acceptation ou les termes insérés P_{FA} en fonction d'un seuil de détection fixé θ, ces deux probabilités sont calculées par les formules suivantes [Fiscus, 2007] :

$$P_{Miss}(term, \theta) = 1 - \frac{N_{correct}(term, \theta)}{N_{true}(term)} \quad (2\text{-}19)$$

$$P_{FA}(term, \theta) = \frac{N_{spurious}(term, \theta)}{N_{NT}(term)} \quad (2\text{-}20)$$

Ainsi, la Valeur pondérée par terme « TWV » représente le minimum des valeurs perdues moyennes par terme par le système. Cette valeur perdue par le système est une combinaison linéaire pondérée pour les valeurs de P_{Miss} et P_{FA}. Elle est définie par :

$$TWV(\theta) = 1 - average_{term}\{P_{Miss}(term, \theta) + \beta . P_{FA}(term, \theta)\} \quad (2\text{-}21)$$

Avec β est un poids qui tient en compte la probabilité préalable d'un terme et les poids relatifs pour chaque type d'erreur conjointement. Il est évalué par la formule suivante :

$$\beta = \frac{C}{V}(Pr_{term}^{-1} - 1) \quad (2\text{-}22)$$

Avec, $\frac{C}{V}$ représente le rapport cout/valeur et Pr_{term} présente la probabilité apriori d'un terme.

2.3.4.2. Figure de Mérite « FOM »

Cette métrique permet de mesurer le pourcentage moyen d'occurrences des termes correctement détectées tant que le système n'atteint pas dix fausses acceptations par deux heures [Wallace, 2010]. Pratiquement, étant donnée une requête Q composée de q termes : $q \in Q$. En suppose que les résultats du système de détection sont des événements $e \in E$. Chaque événement e est caractérisé par les attributs : $q_e, l_e \; et \; s_e$ qui représentent respectivement : q_e le terme référencé de la requête, $, l_e = 1$ si e est une fausse acceptation, 1 sinon et s_e est le score de l'événement.
Egalement, pour chaque $q \in Q$, il y aura deux ensembles : $E_q^+ = \{e \in E, \; l_e = 1 \; \cap \; q_e = q\}$ et $E_q^- = \{e \in E, \; l_e = 0 \cap \; q_e = q\}$, la valeur de *FOM* est alors définie par :

$$FOM = \frac{1}{A} \sum_{e_j \in E^+} h_{e_k} max\left(0, A - \sum_{e_j \in E^-} \left(1 - H(e_k, e_j)\right)\right) \quad (2\text{-}23)$$

Avec $A = 10T|Q|$, $h_e = \frac{1}{|Q||E_{qc}^+|}$ et

$$H(e_k, e_j) = \begin{cases} 1 & s_{e_k} > s_{e_j} \\ 0 & otherwise \end{cases}$$

2.3.5. Synthèse récapitulative des travaux existants

En résumons, nous présentons dans le tableau 2-1 une synthèse récapitulative des travaux sur STD existant dans la littérature qui sont étudiés dans les sections précédentes.

Travaux Réalisés	Description			
	Approches	**Techniques utilisées**	**Corpus d'expérimentation**	**Critères d'évaluation**
Szöke et al 2005	Approche basée KWS	HMM - GMM	TRAP_NN0477	FOM = 64.46
	Approche basée LVCSR	HMM-GMM Likelihood ratio confidence	ICST meeting	FOM = 66.95
	Approche basée Sub-word	Treillis de phonème	TRAP_NN0477	FOM = 58.90
Szöke et al 2008	Approche basée LVCSR	HMM- ML	NIST STD06 dev-set VTS data	UBTWV = 63.0
	Approche basée Subword			UBTWV = 65.0
Grangier et al 2009	Approche basée KWS	Discriminative	TIMIT	AUC = 0.99
		HMM-GMM		AUC = 0.96
Benayed et al 2003	Approche base KWS	HMM – GMM - SVM	Speech Data	EER = 26.7%
Keshet et al 2009	Approche basée KWS	Discriminative	TIMIT	AUC = 0.99
		HMM-GMM		AUC = 0.96
Mangu et al 2000	Approche basée LVCSR	Alignement de treillis - HMM	Switch Board Speech	WER = 37.3
			Broadcast news (DARPA Hub-4)	WER = 32.5%
Chelba et Acero 2006	Approche basée LVCSR	PSPL Treillis de phonèmes	MIT iCompus Data	WER- PSPL= 22% WER- 1-best = 45%
Mamou et al 2007	Approche basée LVCSR	Discriminative	NIST06 Broadcast News	MAP - OOV = 0.48
	Approche Basé Subword	Fuzzy phonetic search		
Can et al 2009	Approche basée LVCSR Subword	WFST - CN	Broadcast news (DARPA Hub-4)	ATW = 0.453
Vergyri et al 2006	Approche basée LVCSR	HMM	Broadcast news	WER = 10.7%
		4-gram LM	CTS	WER = 17.0%
		Cross word	ICST meeting	WER = 37.0%
Akbacak et al 2006	Approche basée LVCSR Subword	Term Weighted Value (TWV)	English Broadcast News	WER = 14.8%
Wallace et al 2010	Approche basée KWS - Subword	Discriminative training	Fisher CTS Corpus	FOM = 0.606
Thambiratnam et al 2007	Approche basée KWS	Dynamic Match Lattices Spoting (DMLS)	TIMIT	Miss rate = 10.2
			CTS	Miss rate = 13.9
Chan et Lee 2010	Approche basée LVCSR	Frame-Based et Segment-Based DTW	Mandarin broadcast news	MAP = 48.6%
Pan et Lee 2010	Approche basée LVCSR Subword	Sub-Word PSPL & CN	Mandarin Broadcast news	AUC = 86.12%
Fucks et Keshet 2017	Approche basée KWS	Prediction Discriminative avec deep network models	Wall Street Journal (WSJ)	AUC = 0.952
Alouazen et al 2004	Approche basée LVCSR & Sub-word	Partial index WFST	DARPA BCN	f-measure=86.0
Saraclar et Sproat 2004	Approche basée Subword	WFST	DARPA BCN	f-measure = 86.1
			Switch Board Corpus	f-measure=60.5
Can et saraclar 2011	Approche basée Subword	WFST	Nist06 English broadcast news	f-measure = 85.2

Tableau 2-1 : Synthèse récapitulative des travaux existant sur STD

2.4. Mesures de similarité sémantique

En effet, il est important de faire recours aux techniques de mesure de similarité utilisés dans les systèmes de recherche d'information. Cette importance devient une nécessité si nous utilisons les techniques de recherche et détection dans le flux parlé à base de systèmes LVCSR. En effet, l'objectif de ces techniques est le passage d'un flux composé des termes parlés vers une représentation textuelle avec certains degrés de confiance. C'est là où l'importance d'intégrer les mesures de similarités sémantiques pour surmonter les problèmes liés aux fausses acceptions et omissions des termes.

Entretemps, les mesures de similarité permettent de sélectionner parmi les différents concepts candidats pour un terme donné, celui qui représente au mieux le sens du terme dans un contexte local qui est défini par les autres concepts du document. En littératures, nous trouvons plusieurs variantes de mesures permettant le calcule la valeur de similarité sémantique entre deux concepts. Ces mesures ont fait l'objet de plusieurs études bibliographiques [Tchechmedjiev, 2012], [Dudognon, 2010]. Ces mesures sont calculées selon différentes stratégies et techniques et notions comme : plus court chemin, quantité d'information, traits lexicaux …etc. Nous citons, mais pas exclusivement, dans les sections suivantes quelques mesures répondues.

2.4.1. Mesures à base de traits lexicales

2.4.1.1. Similarité de Lesk

Cette mesure est basée sur un algorithme de désambiguïsation lexicale très simple. Il considère que la similarité entre deux sens est le nombre de mots en commun dans leurs définitions [Lesk, 1986]. En effet, cet algorithme dans sa version initiale, il ne tient pas compte l'ordre des mots dans les définitions d'un terme. Il considère que les concepts sont des sens de mots ou termes, et que les traits sont des mots de la définition des sens et D est une fonction qui retourne un ensemble de mots de définition d'un sens de mot c. La similarité sera calculée par :

$$Sim_{Lesk}(c_1, c_2) = |D(c_1) \backslash D(c_2)| \qquad (2\text{-}24)$$

En effet, l'avantage de cette mesure réside dans sa simplicité de calcul. Elle offre une désambiguïsation de qualité raisonnable entre 50 jusqu'au 70% de précision en utilisons seulement un simple dictionnaire. Cependant, il existe des variantes de cette mesure comme celle qui intègre la notion de fenêtre de contexte auquel appartient le sens. Elle correspond au recouvrement entre la définition du sens et tous les mots des définitions des mots du contexte [Tchechmedjiev, 2012], elle sera évaluée par :

$$Lesk_{var} = \left|contexte(w) \cap D(s_{w_n})\right|. \qquad (2\text{-}25)$$

Entretemps, le problème important de la mesure de *Lesk* est sa sensibilité aux mots présents dans la définition d'un terme ou sens. Dans ce contexte, nous trouvons plusieurs travaux d'améliorations de la mesure.

2.4.1.2. Extensions de la mesure de Lesk

Parmi les améliorations de la mesure de *Lesk*, nous trouvons une amélioration appelée « *Lesk étendu* » basée sur étapes. Premièrement, elle intègre via les relations taxonomiques de *WordNet* des définitions des sens du mot ou terme cible. Ensuite, elle calcule le recouvrement entre tous les mots définitions. En plus, elle fait l'extension vers les définitions de relations telles que : hyperonymes, méronymes, holonymes …etc.

Entre temps, afin de garantir que la mesure soit symétrique, ils proposent de grouper les évaluations de recouvrement entre les définitions de paires de relations. Ensuite, elle calcule le score de recouvrement par la loi de *Zipf* [1] via la formule :

$$Lesk_{etendu}(c_1, c_2) = \sum_{\forall (R_1,R_2) \in RELPAIRS^2} \left(\left|D\left(R_1(c_1)\right) \cap D\left(R_2(c_2)\right)\right|\right)^2 \quad (2\text{-}26)$$

Avec, $RELPAIRS$ est l'ensemble de relations considérées pour le calcul du recouvrement. Elle est définie par :

$$RELPAIRS = \{(R_1, R_2) | \forall (R_1, R_2) \in RELS^2, (R_1, R_2) \in RELPAIRS^2 \Rightarrow (R_2, R_1) \in RELPAIRS^2\} \quad (2\text{-}27)$$

2.4.2. Mesure à base de distance taxonomique

Dans cette catégorie, nous trouvons les techniques qui se reposent sur les techniques de comptage du nombre d'arcs qui séparent deux sens dans une taxonomie ou hiérarchie des concepts comme *WordNet* [Tchechmedjiev, 2012]. Dans ce contexte, la représentation d'une relation s'effectue via la hiérarchie définie dans la taxonomie dans les deux sens. Entre autres, elle est calculée par rapport au sens commun le plus spécifique et par rapport à la racine. La figure 2-11 présente une illustration qui sera utilisée dans la description des différentes mesures dans cette catégorie.

[1] https://fr.wikipedia.org/wiki/Loi_de_Zipf

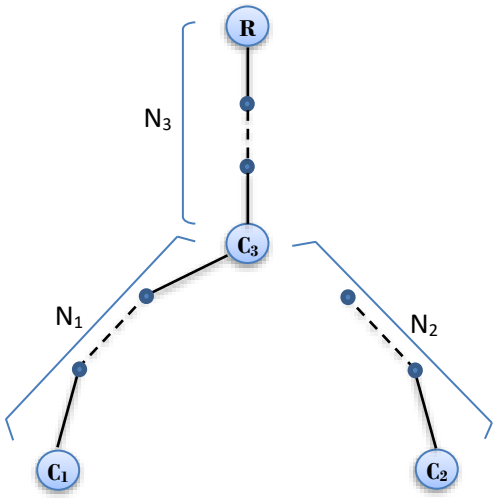

Figure 2-11. Illustration des sens et arcs utilisés dans une Taxonomie d'après [Tchechmedjiev, 2012]

2.4.2.1. Mesure de Rada

La mesure de Rada repose uniquement sur un principe de comptage d'arcs « *Edge counting* » à partir de la structure hiérarchique [Rada, 1989]. Cette mesure définie la distance entre deux concepts d'un réseau sémantique basé sur la relation « *is-a* » comme le nombre minimum d'arcs à parcourir pour aller de c_1 à c_2. Entre temps, cette mesure considère seulement les liens d'hyperonymie et d'hyponymie et e elle exprimé par la formule suivante :

$$Sim_{Rada}(c_1, c_2) = Dist_{edge}(c_1, c_2) = N_1 + N_2 \qquad (2\text{-}28)$$

Où

$Dist_{edge}(c_1, c_2)$: présente le nombre minimum d'arcs séparant c_1 à c_2.

2.4.2.2. Mesure de Wu et Palmer

Cette mesure repose sur la notion de plus petite généralisant commun, c'est-à-dire le concept généralisant commun à c_1 et c_2 le plus éloigné de la racine. Elle est définie par :

$$Sim_{WP}(c_1, c_2) = \frac{2.depth(s_3)}{depth(c_1) + depth(c_2)} = \frac{2.N_3}{N_1 + N_2 + 2.N_3} \qquad (2\text{-}29)$$

Où

- c_3 : est le concept qui généralise c_1 à c_2 et qui est le plus éloigné de la racine
- $depth(\ )$: est une fonction qui renvoie le nombre de nœuds entre un concept et la racine.

2.4.2.3. Mesure de Leacock et Chodorow

Cette mesure repose sur les techniques de valorisation des chemins. Elle est calculée en fonction de la longueur du plus court chemin entre concepts dans une hiérarchie « *is-a* » en se basant sur la mesure *Rada* [Leacock, 1998]. En effet, le plus court chemin est celui qui comprend le plus petit nombre de nœuds intermédiaires. Cette valeur est inversement proportionnelle à la profondeur maximale de l'arbre notée D qui représente la taille du plus long chemin de la feuille au nœud racine dans la hiérarchie. Elle est définie par :

$$Sim_{LCh} = -log\left(\frac{Dist_{edge}(c_1,c_2)}{2.D}\right) = -log\left(\frac{N_1+N_2}{2.D}\right) \quad (2\text{-}30)$$

Où

- $Dist_{edge}(s_1, s_2) = length(s_1, s_2)$: le plus court chemin entre deux nœuds
- D : la profondeur maximale dans la taxonomie.

2.4.2.4. Mesure de similarité de Hirst et St-Onge

Cette mesure est basée sur l'exploitation des chaines lexicales comme mesure de similarité sémantique en utilisant la structure de *WordNet*. Elle se repose sur l'assomption que l'enchainement des concepts dans un texte implique une forte probabilité de référence entre eux. Entre autres, l'enchainement des concepts forme des chaines cohésives [Hirst, 1998]. Dans ce contexte, pour chaque relation elle associe une direction : horizontale, ascendante et descendante et une qualité : relations fortes, très fortes et moyennement fortes respectivement. Ainsi que, les changements de direction constituant un élément de dissimilarité et la distance dans la taxonomie un élément de similarité. À savoir que plus la distance entre les sens sera grande, plus il y aura de changements de direction potentiels. Cette mesure est définie par :

$$Sim_{Hso} = C - N_1 + N_2 - k.virages(c_1, c_2) \quad (2\text{-}31)$$

Où

- $virages(c_1, c_2)$: le nombre de changements de direction entre les sens c_1 & c_2
- $C\ et\ k$: deux constantes.

2.4.3. Mesures à base de contenu d'information

Dans cette catégorie, nous trouvons les mesures qui intègrent des connaissances supplémentaires à celles de la structure hiérarchique afin de capter le contenu informationnel des nœuds ou des concepts

2.4.3.1. Mesure de Resnik

Cette mesure est basée sur le calcul de probabilité d'existence d'un concept par rapport aux caractéristiques de son contenu informationnel « *Information Content*- IC ». Cette caractérisation quantitative de l'information fournit une nouvelle façon de mesurer la similarité sémantique. En effet, plus deux concepts ont d'informations communes, plus ils sont plus semblables. Ainsi que, l'information partagée par les deux concepts est indiquée par le contenu informationnel des concepts qui les subsument dans la taxonomie [Rensik, 1995]. Elle est formulée par :

$$sim(c_1, c_2) = \max_{c \in S(c_1, c_2)} [-\log P(c)] \quad (2\text{-}32)$$

Où

- $S(c_1, c_2)$: l'ensemble des concepts qui subsument à la fois c_1 et c_2.
- $p(c)$: la probabilité de trouver c ou un de ses descendants dans le corpus en utilisant le sous-ensemble correspondant à la hiérarchie « *is-a* » ou hyperonymie de *WordNet*, avec :

$$p(c) = \frac{frec(c)}{N} \quad (2\text{-}33)$$

avec

$$freq(s) = \sum_{t \in words(s)} count(t)$$

Où

- $\mathrm{words(s)}$: l'ensemble des mots subsumés par le concept c.
- $count(t)$: le nombre d'occurrences d'un terme dans le corpus
- N : le nombre total d'occurrences des termes retrouvés dans le corpus.

2.4.3.2. Mesure de Seco

Cette mesure utilise une autre variante pour la définition du contenu informationnel des concepts « *Information Content*- IC ». Elle est fondée sur l'hypothèse que le nombre de descendants d'un concept affecte le contenu informationnel [Seco, 2004]. À cet effet, elle utilise les hyponymes des concepts pour calculer le contenu informatif par la formule suivante :

$$IC_{wn}(c) = \frac{log\left(\frac{hypo(c)+1}{max_{wn}}\right)}{log\left(\frac{1}{max_{wn}}\right)} = 1 - \frac{log(hypo(c)+1)}{log(max_{wn})} \quad (2\text{-}34)$$

Où

- $hypo(\ \ )$: fonction qui indique le nombre d'hyponymes d'un concept.
- max_{wn} : constante représente le nombre de concepts de la taxonomie.

2.4.3.3. Mesure de Lin

Cette mesure est basée sur celles de *Wu et Palmer* et *Resnik*. Elle mesure la similarité entre deux concepts par le rapport du contenu informationnel qu'il les décrit en commun « *Communality* » sur le contenu informationnel qui décrit chaque concept [Lin, 1998]. Elle est calculée par :

$$Sim_{Lin}(s_1, s_2) = \frac{2.IC\big(lcs(s_1, s_2)\big)}{IC(s_1) + IC(s_2)} = \frac{2.log\big(p(s)\big)}{log\big(p(s_1)\big) + log\big(p(s_2)\big)} \qquad (2\text{-}35)$$

Où

- $lcs(s_1, s_2)$: représente le contenu informationnel des (s_1, s_2) conjointement.

2.5. Synthèse

Dans ce chapitre, nous avons étudié les disciplines qui seront utilisées dans les différentes phases de notre approche proposée pour la recherche dans le contenu parlé. Nous avons abordé en premier lieu les techniques et les modèles utilisés dans les systèmes de reconnaissance de la parole continue à large vocabulaire « LVCSR ». Nous avons mis l'accent sur les différents modèles utilisés dans ces systèmes comme : le modèle de langage, le modèle phonétique, treillis de phonèmes, …etc. En effet, ces modèles ont un impact sur la qualité de la transcription de systèmes de reconnaissance ainsi que sur les exigences de calculs et espace mémoire. Dans ce contexte, nous constatons que les techniques et les modèles utilisés dans les routines de décodages comme *Viterbi* et *Wfst* sont performants, mais leurs complexités dépendent étroitement par la durée du flux traité. Entre autres, elles ne fournissent pas des bons résultats même ils ne seront pas capables de traiter les flux parlés de taille moyenne ou grande.

Ensuite, nous avons réalisé une étude bibliographique sur les techniques utilisées dans les tâches de détection de termes parlés. Nous avons présenté dans cette étude les deux stratégies utilisées : celles basées sur les techniques de détection phonétique comme : *Keyword Spotting* et PSPL et celles basées sur les systèmes LVCSR. Dans ce contexte, nous avons tracé une synthèse sur les différents travaux effectués dans ce domaine. Entre temps, nous constatons que les techniques basées sur les systèmes LVCSR ont prouvées leurs performances dans plusieurs études et surtout pour les langues bien dotées comme la langue anglaise. Ces résultats nous encouragent à exploiter cette stratégie dans notre contribution sur la recherche dans le flux parlé.

Enfin, pour surmonter les problèmes liés aux systèmes de reconnaissance automatique ainsi que les systèmes de détections. Nous avons proposé dans notre contribution d'intégrer l'aspect sémantique dans les sorties « Output » de ces systèmes. Dans ce contexte, nous avons cité avec de brèves descriptions les métriques utilisées pour évaluer les rapprochements sémantiques entre les termes du contenu du flux parlé. Dans ce contexte, nous avons focalisé sur les mesures basées sur les taxonomies comme : *Lesk, Rada, Wu & Palmer, …* etc. En effet, ce choix est effectué, car notre contribution propose d'utiliser les ontologies et les taxonomies pour surmonter les problèmes des traitements automatiques.

Chapitre 3

SOMI:

L'approche proposée

Semantic Ontologies for Multimedia Indexing contenent

3.1. Introduction

Avec l'expansion rapide de différentes sources de médias pour la diffusion de l'information, le volume des ressources multimédias ne cesse d'accroitre. Ce gigantesque volume de ressources numérique nécessite des mécanismes et systèmes de traitement automatiques de leurs contenus. Cependant, les méthodes actuelles utilisées pour la segmentation, la transcription et l'indexation du contenu parlé des ressources multimédias sont souvent manuelles ou guidées. Ces derniers sont réalisés par des experts humains qui sont chargés d'écouter et d'annoter le contenu, ainsi que la sélection des mots clés discriminants pour les exploiter dans les systèmes de recherche d'information. A cet effet, l'automatisation de certaines tâches de ce processus permettant de couvrir plus de sources d'informations numériques ainsi de faire réduire de manière significative les coûts de traitement tout en éliminant le travail fastidieux [Bendib, 2014].

Dans ce contexte, nous trouvons certaines applications existantes qui bénéficient et exploitent les nouvelles technologies comme la création et l'accès aux bibliothèques numériques multimédias comme le projet *OLIVE*[1] qui permet l'accès au contenu de ressources via des techniques d'indexation. Aussi, nous trouvons les services de surveillance des médias « *diffusion sélective de l'information basée sur la détection automatique des sujets d'intérêt* » ainsi que l'émergence des nouvelles applications telles que les informations et les multimédia sur demande « *News on Demand* » et « *Internet watch services* ».

[1] https://olivearchive.org/origins/

En revanche, les systèmes de reconnaissance automatique à large vocabulaire sont des technologies clés pour l'indexation des flux parlés des ressources multimédias. En effet, la plupart des informations linguistiques sont encapsulées dans la primitive « son » des ressources vidéo. Alors, une fois que le contenu du flux parlé est décrypté ou transcrit, la ressource multimédia devient accessible par les systèmes de recherche d'informations.

D'autre part, nous notons que les travaux effectués dans les systèmes de recherche dans le contenu parlé sont dans des domaines d'applications restreints tels que les émissions radio et télévisés comme « THISL[1] », les conversations téléphoniques et les lectures comme « MALACH[2] ». Ces systèmes sont réalisés d'une façon que nous pouvons qualifier comme des systèmes optimaux et fermés avec l'amélioration constante des technologies relatives au traitement de la parole comme la segmentation automatique, la transcription et les annotations. Entretemps, il existe une énorme masse de flux parlés dans les ressources multimédias éparpillées dans le web qui ne sont pas encore traitées et exploitées tels que : les livres audio « *Audio Books* », les plateformes éducatives comme les cours en ligne « *Massive Open Online Courses* - MOOCs » et les évènements scientifiques et culturels multimédias par exemple : TED[3], Google I/O[4], … etc.

Entre autres, l'avènement des bibliothèques numériques et ces systèmes de gestion a permis d'ouvrir des voies de recherches pour l'exploitation des documents multimédias, qui sont malheureusement limités aux documents numériques de types images ou des textes scannés. Donc, l'enrichissement de ces bibliothèques numériques par la primitive parole pour ces ressources multimédias éparpillées est fortement sollicité.

3.2. L'indexation au profil du recherche dans le contenu parlé

3.2.1. Motivation

Actuellement, la nécessité de trouver des solutions d'indexation pour le contenu du flux parlé des ressources multimédias est fortement sollicitée et surtout lors de la manipulation des documents multimédias de potentiel tailles. Dans ce contexte, nous proposons une approche pour améliorer l'efficacité des systèmes de recherches dans le contenu parlé. Entre temps, notre contribution est d'accroître les performances de ces systèmes pour la gestion du contenu parlé hétérogène et volumineux.

---

[1] Le projet THISL concerne l'intégration technologies des systèmes de reconnaissance de la parole continue à large vocabulaire (LVCSR), Recherche d'Information (IR) et Traitement naturel du Langage (NLP), axées sur une application cible constitué de l'indexation automatique et la récupération des diffusions radio et de télévision des programmes d'information.

[2] C'est un système d'accès multilingue aux Grandes Archives parlées réalisé sur une collection d'archive cohérente du monde d'histoires orales filmées assemblé par la « *Shoah de Visual History Foundation* ». Elee contient 116.000 heures d'entrevues numérisées en 32 langues à partir de 52.000 survivants, des libérateurs, des sauveteurs et des témoins de l'Holocauste nazi.

[3] TEDx : Evénement qui favorise le partage des idées dans les communautés du monde entier

[4] Google IO : Evènement qui rassemble tous les ans des développeurs du monde entier pour des discussions avec un apprentissage pratique avec des experts de Google et un premier aperçu des derniers produits de développement de Google.

En outre, le volume des documents multimédias disponible : nouveaux ou archivés et non annotés est très important, en plus du temps nécessaire de traitement de l'annotation rend l'indexation manuelle fastidieuse. Dans ce contexte, nous contribuons par la conception d'une démarche d'indexation sémantique automatique pour le contenu parlé de ces ressources, via les techniques de gestion du contenu parlé et de gestion des connaissances.

À cet effet, notre contribution permet de proposer des solutions pour pallier les défis suivants :

- Comment accéder et gérer le contenu du flux parlé des ressources multimédias d'une façon efficace et précise ?
- Comment extraire les index discriminants du contenu du flux parlé des ressources hétérogènes d'une façon automatique sans fait recourir aux experts des domaines ?
- Comment remédier les problèmes liés au processus de reconnaissance automatique de la parole telle que les fausses alarmes, les fausses acceptations et les mots hors vocabulaire – les termes techniques du domaine ? On note que nous nous intéressons dans cette thèse seulement sur les mots hors vocabulaire du point de vue documents et non pas du point de vue requêtes.
- Comment surmonter le problème de complexité de calcule engendrée par la taille de flux parlé ?

3.2.2. Vers une indexation automatique enrichi sémantiquement

Actuellement, nous trouvons des progressions importantes dans les systèmes de reconnaissance automatiques de la parole, qui sont actuellement capables de traiter les segments du flux de parole continu voire même la parole spontanée avec des tolérances d'erreurs acceptables. Néanmoins, il reste le souci majeur de ces systèmes est la complexité des automates de décodage et la surcharge des modèles de langage « *3-gram vers n-gram* ». A cet effet, ces systèmes sont incapables de gérer les flux de paroles volumineux et ses performances sont liées étroitement par la durée du flux de paroles traité.

Entre autres, nous trouvons plusieurs implémentations de ces systèmes soit dans des versions commercialisées ou libres « *Open Source* ». Dans notre contexte de recherche, nous nous intéressons en vertu des systèmes libres. Le tableau 3-1 présente quelques systèmes gratuits. Nous nous mentionnons que les routines sont certes gratuites tandis que les corpus d'apprentissage et de test sont payants, ce qui nous pose vraiment des problèmes lors de la phase de la validation de notre contribution [Bendib, 2018].

Nom	Description	Site Web	License	S.E	Langage de programmation
CMU Sphinx	HMM	CMU: Sourceforge	BSD style	Multi-platform	Java
HTK	HMM	HTK Web Site	HTK Specific License	Multi-platform	C
Julius	HMM trigrams	Julius Home page	BSD-like	Multi-platform	C

Tableau 3-1 : liste de quelques systèmes libres de Reconnaissance Automatique de la Parole[1]

Cependant, ces systèmes sont liés étroitement par la taille de flux de parole à traiter. Dans les littératures, nous trouvons que la tâche de segmentation est une tâche primordiale. Cette dernière est effectuée par des techniques d'extraction automatiques des intervalles de segments qui sont généralement interdépendants [Ostendorf, 2008]. Malheureusement, ces techniques ont prouvé leurs performances sauf en cas des phrases qui chevauchent. Tandis que la majorité du flux parlé de ces ressources multimédias actuelles contiennent de la parole continue. Ce qui nous amène dans cette thèse de proposer une solution qui utilise des scénarios de segmentation de ces flux parlés avec l'hybridation des transcriptions de ces segments avec des ressources sémantiques tels que les ontologies pour surmonter ces problématiques.

Entre d'autre, le développement des Interfaces Applicatives de Programmation « *Application Programming Interface* - API » dans le domaine de reconnaissance automatique de la parole ne cesser progresser. Ces APIs sont des moyens efficaces pour faire communiquer nos scripts de reconnaissances avec des ressources locales ou dans le cloud, le tableau 3-2 ci-dessous présente une étude descriptive de quelques APIs pour la reconnaissance du contenu parlé qui varient entre gratuites et payantes ainsi que leurs capacités de traitements.

En pratique l'utilisation de ces interfaces programmables qui sont souvent implémentées sous des plateformes de « *Cloud Computing* » et surtout pour les transcriptions en temps réel, nous fournissent des résultats encourageants pour de petits segments parlés. En contrepartie, les performances de ces interfaces sont meilleures pour les traitements des phrases par rapport aux textes. En plus, les APIs de licence libre ou des versions d'évaluation sont souvent dédiées en vue de fonctionner sur des segments des flux parlés et non sur sa totalité.

Dans ce contexte, notre contribution est conçue en s'accentuons sur l'exploitation des performances de ces APIs en surmontant le problème de traitements des flux parlés intégraux par des techniques de segmentation appliquées sur ces derniers.

[1] D'après le site : https://en.wikipedia.org/wiki/List_of_speech_recognition_software (DC : 23/09/2016)

Dans la section suivante, nous présentons l'architecture et la description de notre approche et dans les sections qui se suivent nous développons ainsi chaque module en détail.

Nom d'API	Description	Nature des traitements	Développeur	Langage	Licence
CLOUD SPEECH API	Conversion de la parole en texte optimisée par l'utilisation de l'apprentissage automatique	Courte parole en ligne	Google	Plus de 80 Langues	Gratuit
Speechmatics API	Reconnaissance de la parole dans le Cloud basée sur les dernières avancées en matière de réseaux de neurones et d'apprentissage profond	Moyenne durée de parole en ligne	Speechmatics	Anglais US et UK	Payant
Bing Speech API	Conversion de la parole en texte, en tenant compte l'intention et la reconversion de texte en paroles pour une réactivité naturelle.	Courte parole en ligne	Microsoft	07 langages	Payant
API.AI	Plateforme de compréhension du langage naturel	Courte parole en ligne	api.ai	15 langages	Payant
Speech to Text API	Suite de logiciels de transcription parole-texte spécifiques aux langues de *Vocapia Research* pour les plateformes Linux x86 et x86-64	Moyenne durée de parole en ligne	Vocapia Research	17 Langages	Payant
UWP Speech recognition	La reconnaissance vocale est composée d'un runtime vocal, d'API de reconnaissance pour la programmation du runtime, de grammaires prêtes à l'emploi pour la dictée et la recherche Web et d'une interface utilisateur système par défaut permettant aux utilisateurs de découvrir et d'utiliser les fonctions de reconnaissance vocale.	Moyenne et courte durée de parole en ligne	Microsoft	Anglais	Gratuit
Speech Engine	Un moteur de reconnaissance vocale caractérisé par un taux de reconnaissance élevé et construit avec la technologie de reconnaissance vocale de pointe.	Online short utterance	iFLYTEK	Anglais et Chinois	Payant

Tableau 3-2 : Liste des APIs les plus utilisés pour la transcription du contenu parlé.

3.3. SOMI : l'approche proposée

3.3.1. Description

Le processus de recherche sur le contenu des documents parlés dépend étroitement de deux contraintes. La première c'est le contenu lui-même qui varie entre des textes simples, des textes littéraires vers des textes purement scientifiques ou hétérogènes. Cette variété de contexte provoque l'utilisation des termes complexes, rares et même nouveaux comme les

termes techniques. Dans ce contexte, l'objectif de cette thèse consiste à contribuer par une démarche dédiée à l'exploitation des flux parlés des ressources multimédias qui couvrent généralement différents domaines techniques et hétérogènes. Actuellement, l'exploitation de ces ressources est basée sur des annotations manuelles qui sont souvent générales et non suffisantes pour la recherche sur le contenu de ces derniers.
Dans ce contexte, nous avons conçu dans notre approche SOMI un module qui permet de pallier ce problème par une annotation automatique à l'aide d'un système d'indexation qui intègre des ressources sémantiques pour mieux traiter tous les termes disponibles dans le contenu du flux parlé.

À cet effet, pour effectuer une indexation sur le contenu du flux parlé des ressources multimédias, nous affrontons deux problèmes : comment connaitre le domaine de la ressource parlée ainsi que ces termes discriminants. Aussi, comment repérer ces termes en tenant compte de la complexité de sa structure. Pour remédier le premier problème, nous proposons dans notre contribution une démarche pour l'extraction de ces termes sachant que nous estimons de passer d'une simple tâche manuelle basée sur les descriptions utilisateurs ou les connaissances des experts du domaine vers une tâche automatique à base des techniques de segmentation, systèmes de reconnaissance automatique et les techniques de gestion de connaissance à base des ontologies telle que *WordNet*.

Tandis que, pour la deuxième contrainte, nous proposons d'utiliser dans notre système d'indexation les techniques du domaine de traitement de la parole dans le volet de la détection des mots clés à base de flux phonétique de ces termes. On note aussi, que les techniques basées sur la transcription intégrale du flux parlé n'est pas toujours valide de point de vue de charge de calculs et d'erreurs de reconnaissances et surtout pour les termes étranges du vocabulaire qui sont appelés souvent mots hors vocabulaire. Dans les sections suivantes, nous présentons l'architecture proposée.

3.3.2. Architecture générale

Pour la recherche sur le contenu du flux parlé des ressources multimédias, nous contribuons par une architecture d'indexation sémantique et de détection ou recherche. Elle est décrite par la figure 3-1. Elle comporte trois modules :

- Le premier module sert à trouver les termes fréquents qui seront des candidats pour le processus d'indexation. À cet effet, nous utilisons des API pour les systèmes de reconnaissance à large vocabulaire pour assurer la transformation du contenu du flux parlé des ressources multimédias vers une représentation textuelle la plus proche possible. En effet, nous ne pouvons pas assurer une transcription automatique intégrale du contenu du long flux parlé ; et cela est dû au problème de complexité de calculs ainsi que pour la dégradation de la qualité des systèmes de reconnaissance de point de durée du segment parlé ainsi sa complexité de contenu. Pour cela, notre approche comporte un module analyseur syntaxique linguistique « SLA-SOMI » qui permet la segmentation automatique du contenu du flux parlé pour ces ressources afin de surmonter ces problèmes. Dans la section 3.4, nous présentons les détails de ce module.

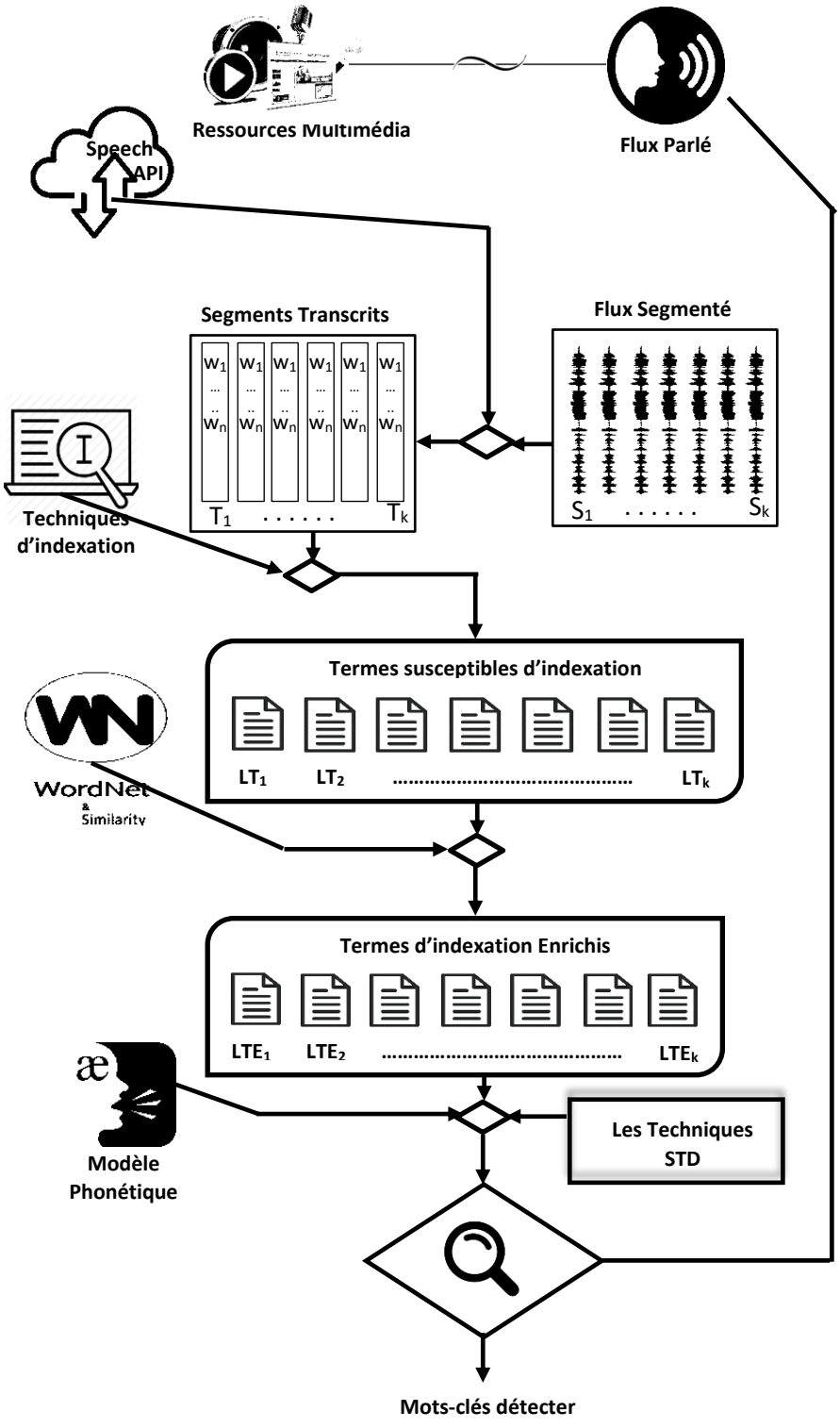

Figure 3-1 : Architecture générale de l'approche proposée [bendib,2014]

- L’objectif du deuxième module est de surmonter les problèmes de détections liées aux impacts du processus de segmentation et de reconnaissance automatique sur le contenu du flux parlé par un analyseur sémantique appliqué aux résultats de l’analyseur syntaxique linguistique. En effet, ce module a deux vocations :

  Premièrement, il élimine et traite les erreurs de reconnaissance et deuxièmement, il enrichit et traite les problèmes des termes hors vocabulaires « OOV » et les termes techniques avec l’utilisation des distances sémantiques via un processus d’enrichissement sémantique des résultats du premier module. Dans la section 3.5, nous présentons les détails de ce module.

- Enfin, le troisième module a pour vocation de définir les modalités d’accès au contenu des documents parlés. A cet effet, nous visons d’utiliser les représentations phonétiques pour la modélisation des index extraite à partir de la précédente étape. Ensuite, nous utilisons les techniques de détection des mots clés à base phonétiques « *word spotting* » avec les modèles de représentation les plus utilisées comme les réseaux de confusion et les probabilités a posteriori des treillis de phonèmes. Dans la section 3.6, nous présentons les détails de ce module.

3.4. L’analyseur Syntaxique Linguistique : SLA-SOMI

3.4.1. Description du SLA-SOMI

La recherche dans le contenu du flux parlé dépend étroitement de la qualité linguistique et sémantique de ces derniers ainsi que par ces caractéristiques physiques. Dans ce contexte, notre approche repose sur ces aspects afin d’améliorer les résultats de recherches. Généralement, les systèmes de recherches dans le contenu du flux parlé sont basés sur les annotations ou l’alignement avec des fichiers textuels qui sont souvent manuels.

À cet effet, l’objectif du module SLA-SOMI est de trouver une représentation linguistique automatique la plus proche de flux parlés qui couvrent au maximum le contenu de ces derniers, en palliant les problèmes de charge des calculs, les erreurs de reconnaissance ainsi que les termes étranges et les termes techniques. Ce module permet de trouver une représentation linguistique avec une couverture maximale du contenu du flux parlé de ces ressources, son architecture globale est présentée par la figure 3-2.

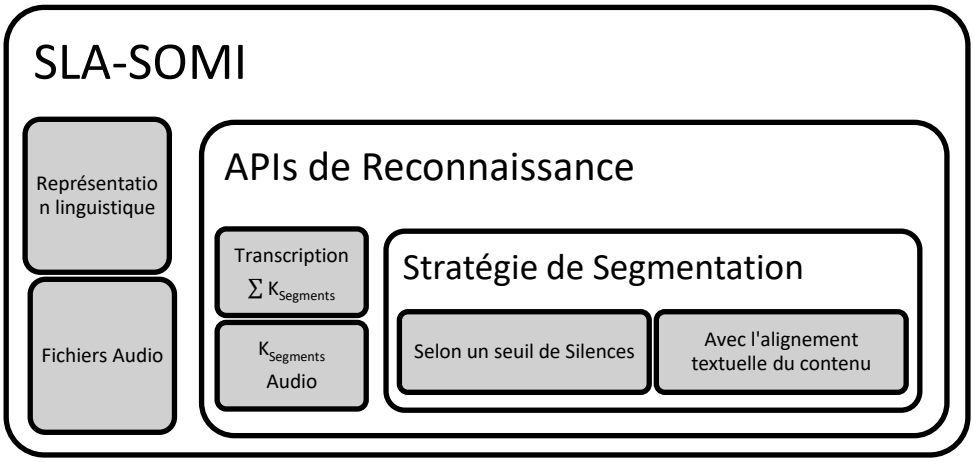

Figure 3-2 : Architecture du module SLA-SOMI [Bendib, 2018]

Le Module SLA-SOMI consiste à effectuer une transcription automatique du contenu parlé pour extraire un ensemble fini des mots significatifs à partir du contenu des documents parlés. A cet effet, nous recourons à un processus de segmentation qui doit couvrir le contenu du flux parlé de la ressource.

Ensuite, nous utilisons les APIs de reconnaissance automatique pour transcrire les segments obtenus. En effet, les performances de ces systèmes sont liées étroitement par les performances de la stratégie de segmentation utilisée. Dans ce contexte, nous avons mené une série d'expérimentation dans le quatrième chapitre pour évaluer la qualité de la stratégie de segmentation utilisée. Une fois le système recueille la représentation textuelle du contenu de flux parlé segmenté. Nous utilisons la mesure cosinus pour calculer les similarités entre les vecteurs de représentation des segments transcrits et le contenu des flux parlés correspondants. Cette mesure permet l'évaluation de la qualité des transcriptions par rapport aux systèmes de reconnaissance utilisés.

3.4.2. Les stratégies de segmentation

Généralement, les documents parlés ne sont pas souvent constitués de phrases bien formées. Par conséquent, la segmentation par la détection des « début et fin des phrases » n'est même pas bien définie. Dans ce contexte, le concept approprié est « énoncé » plutôt que « phrase ». Les énoncés sont définis par un critère acoustique. En outre, pour les documents parlés bien articulés qui consistent principalement en phrases bien formées, les énoncés correspondent à peu près aux phrases. En outre, la segmentation de la parole en unités de longueur d'énoncé est utile même si la parole ne forme pas de phrases.

Dans le flux parlé, il y aura des intervalles de paroles séparés par des pauses. Dans la parole naturelle, un locuteur ne s'arrête pas après chaque mot. Généralement, un orateur ne fera que s'arrêter là où il y aurait ponctuation dans la forme écrite. Il y a aussi d'autres pauses acoustiques, c'est-à-dire des interruptions plus courtes du son qui ne sont pas réellement des

pauses dans le flux de la parole, mais qui font partie intégrante de certains phonèmes, comme / p / et / t /. Généralement, la pause à la fin de la phrase sera plus longue que les pauses causées par les phonèmes ou par des ponctions de ponctuation moindres, comme la montre la figure 3-3. A cet effet, le concept « énoncé » est défini et détecté dans le domaine de reconnaissance de la parole par la durée des silences indépendamment que la phrase incomplète.

Figure 3-3 : Exemple d'une représentation acoustique de la parole et de silence

Techniquement, nous cherchons à séparer le contenu du flux parlé de tous les échantillons de silence plus le bruit de fond. En effet, il y a un chevauchement entre les bruits de fond plus forts et les bruits de voix plus silencieux. La parole a une gamme dynamique très large. Même si les voyelles accentuées sont plus bruyantes que le bruit de fond, certains sons de la parole plus silencieux ne seront pas aussi bruyants que le bruit de fond.

Cependant, pour structurer la parole en énoncés, il suffit de séparer les intervalles plus forts des intervalles plus silencieux. Donc, nous ignorons certains des bruits de voix plus silencieux et les bruits très brefs, celles moins de 30 millisecondes, même si elles sont bruyantes. Ensuite, nous cherchons les intervalles plus longs de silence. Dans ce contexte, le seuil sur la longueur d'intervalle du silence est en fonction de la langue utilisée dans la ressource, du débit de la parole et du style de parole. Généralement, tout intervalle silencieux supérieur à 300 millisecondes indique une rupture réelle dans la parole et peut être pris comme la fin d'un énoncé. Par exemple, nous utilisons un intervalle plus long si nous souhaitons distinguer les ruptures de phrase voire d'autres ponctuations, mais il peut y avoir un chevauchement. Entretemps, l'impact de la structuration du contenu des segments parlés est minime par rapport à leurs durées. En effet, ces systèmes favorisent des petites séquences parlées contre des grandes séquences avec un contenu structuré.

D'autre part, les systèmes de reconnaissance automatique de la parole basés sur la modélisation Markovienne sont robustes aux phénomènes de silence et bruit. Le processus de Markov modélise le silence et le bruit de fond comme faisant partie du même espace d'état de Markov que la parole. Le début de la parole est juste une transition d'état de Markov, comme toutes les transitions d'état qui se produisent pendant le flux parlé. Il n'est pas nécessaire de connaître le début du flux parlé pour entamer le processus de reconnaissance.

Dans ce contexte, nous avons utilisé dans cette contribution deux techniques de segmentation automatiques : la première est basée sur les seuils de silence et la deuxième à base de la structure des énoncés « *utterances* ».

3.4.2.1. Segmentation selon un seuil de silence

Actuellement, il existe plusieurs implémentations des systèmes qui permettent la manipulation du contenu des flux parlés selon leurs caractéristiques sonores et acoustiques comme : *PyAudioAnalysis*, *SoX* « *Sound eXchange* », *WavPad*, … etc. Entre autres, nous avons utilisé le système d'exploitation « *Ubuntu 16-04* » et le langage de développement « *Python* » pour la validation et les expérimentations effectuées dans cette contribution. Dans ce contexte, nous avons opté pour le choix d'utilisation de l'environnement *SoX*[1] puisqu'il est multiplateforme et il est exploitable dans les routines de programmation système « *Bash* » et les scripts « *Python* ». L'outil *SoX* nous permet de fragmenter des documents de paroles selon l'intensité de signal ou sur des durées des silences.

```
issam@issam-HP-Pavilion-15-Notebook-PC:~/Project$ sox -V3 Doc1.wav output.wav silence 1 0.50 0.1% 1 0.3 0.1% : newfile :
restart
Input File     : 'Doc1.wav'
Channels       : 1
Sample Rate    : 16000
Precision      : 16-bit
Duration       : 00:19:10.15 = 18402464 samples ~ 86261.6 CDDA sectors
File Size      : 36.8M
Bit Rate       : 256k
Sample Encoding: 16-bit Signed Integer PCM
Endian Type    : little
Reverse Nibbles: no
Reverse Bits   : no
Level adjust   : 1 (linear gain)

sox INFO sox: Overwriting `output057.wav'

Output File    : 'output057.wav'
Channels       : 1
Sample Rate    : 16000
Precision      : 16-bit
Sample Encoding: 16-bit Signed Integer PCM
Endian Type    : little
Reverse Nibbles: no
Reverse Bits   : no
Comment        : 'Processed by SoX'
```

Figure 3-4 : Exemple des commandes de segmentation d'un flux de paroles selon le silence avec SoX

On note que dans cette stratégie de segmentation est basée sur la valeur définie du silence, par exemple dans la figure 3-4, nous visualisons une commande de segmentation d'un flux parlé selon une durée de silence supérieur ou à égale à 0,3 seconde.

[1] http://sox.sourceforge.net/

3.4.2.2. Segmentation à base des énoncés

Une autre possibilité de fragmenter des fichiers parlés est d'exploiter les fichiers d'alignements textuels. Dans ce contexte, nous calculons les cartes de synchronisation qui expriment la correspondance entre le flux parlé et le fichier de transcription correspondant. Pour chaque fragment dans le fichier texte, il aligne l'intervalle de temps dans le flux parlé où le texte du fragment est prononcé. Ces cartes de synchronisation sont souvent de format « *json*[1] » ou « *xml* ».

À cet effet, nous utilisons la bibliothèque Python / C « *aeneas* » qui intègre un ensemble d'outils pour synchroniser automatiquement le flux parlé avec les fichiers « .stm » correspondante « *alias alignement forcé* ». Ensuite, avec des scripts python, nous effectuons la fragmentation physique du flux parlé selon la carte de synchronisation calculée. Les figures 3-5 et 3-6 simultanément présentent un exemple de calcul d'une carte de synchronisation et sa structure.

```
issam@issam-HP-Pavilion-15-Notebook-PC:~/Project/aeneas$ python -m aeneas.tools.execute_task \
>     BillGates_2010.wav \
>     BillGates_2010.txt \
>     "task_language=eng|os_task_file_format=json|is_text_type=plain" \
>     map.json
[INFO] Validating config string (specify --skip-validator to bypass)...
[INFO] Validating config string... done
[INFO] Creating task...
[INFO] Creating task... done
[INFO] Executing task...
[INFO] Executing task... done
[INFO] Creating output sync map file...
[INFO] Creating output sync map file... done
[INFO] Created file 'map.json'
```

Figure 3-5 : Exemple d'un script Python pour le calcul de carte de synchronisation


```
  },
  {
   "begin": "5.160",
   "children": [],
   "end": "16.320",
   "id": "f000002",
   "language": "eng",
   "lines": [
    "BillGates_2010 1 BillGates 15.861 19.986 <o,f0,male> i 'm going to talk today about energy and climate"
   ]
  },
  {
   "begin": "16.320",
   "children": [],
   "end": "29.520",
   "id": "f000003",
   "language": "eng",
   "lines": [
    "BillGates_2010 1 BillGates 19.986 34.89 <o,f0,male> and that might seem a bit surprising because my full time work
at the foundation is mostly about vaccines and seeds about the things that we need to invent and deliver to help the
poorest two billion live better lives"
   ]
```


Figure 3-6 : Extrait d'un fichier de carte de synchronisation « *json* »

[1] Json : Javascript Object Notation

3.4.3. Les routines de décodages :

Actuellement, il existe plusieurs solutions de reconnaissance automatique et de décodages des ressources parlées qui varient entre des systèmes fermés, plateformes de développements libres ou payantes, boites à outils et des interfaces de programmation en ligne ou hors ligne gratuite et payante et comme notre thèse est dans des fins de recherche nous avons opté pour l'utilisation des plateformes « boîte à outils » ou les interfaces programmables gratuites.

Dans ce contexte, les chercheurs dans le domaine de la reconnaissance automatique de la parole « ASR » ont mis plusieurs choix possibles de boîtes à outils open source pour la construction d'un système de reconnaissance. Parmi eux nous pouvons citer : HTK [Young, 2015], Julius développé en langage C [Lee, 2001], Sphinx-4 développé en Java [Walker, 2004], la version open source la boîte à outils RWTH ASR écrite en C [Rybach, 2009] et l'environnement Kaldi écrit en C++ [Povey, 2011]. Entre autres, nous présentons une étude comparative réalisée par *Gaida* dans le Tableau 3-3 [Gaida, 2014]. Cette étude consiste à décrire une évaluation à grande échelle des boites à outils de reconnaissance vocale open source. Les résultats obtenus sur les expériences effectuées leur permettent de présenter un ordre de ces systèmes évalués du point de vue de rapport calculs/performance.

Boite à Outils	VM11	WSJ12
HDecode v3.4.1	22.9	19.8
Julius v4.3	27.2	23.1
pocketsphinx v0.8	23.9	21.4
Sphinx4	26.9	22.7
Kaldi	12.7	6.5

Tableau 3-3 : Taux d'erreur de mots (WER) sur l'ensemble de test VM1 et l'ensemble WSJ1 selon [Gaida, 2014]

Cependant, l'utilisation et le développement d'un système de reconnaissance à l'aide de ces boites à outils nécessitent un grand volume de ressources de données pour l'apprentissage des modèles et des machines de calculs puissantes. Ces ressources sont généralement

[1] Corpus issus du projet German Verbmobil project (1993), inclus des dialogues vocaux en 3 langues (Anglais, Japonais et l'allemand).

[2] Corpus issus en 1994, inclus des lectures des articles de Wall Street Journal News.

payantes et sont mises aux droits de LDC[1] « *Linguistic Data Consortium* ». Entre autres, le développent et les performances réalisées par les interfaces programmables « APIs » nous encourage à les utilisés. Dans le tableau 3-4 nous présentons une étude comparative[2] des APIs les plus reconnues dans le contexte des phrases courtes comparées dans différentes conditions telles que le sexe, l'âge et le bruit de fond. Selon leurs critères de test de phrases exactes reconnues et le taux d'erreur de mot, l'étude montre que Google est de loin la meilleure solution. Ce n'est pas surprenant étant donné leur histoire de développement et de prouver la recherche vocale. Dans ce contexte, et notons que notre approche SOMI est basée sur la segmentation des ressources parlées vers des segments ; et surtout dans la segmentation à base des silences ; qui sont souvent court ce qui convient avec l'étude citée précédemment nous favorisons le choix d'utilisation des *APIs de Google Cloud speech* dans notre système.

Nom	Description d'API	Word Error Rate	Pourcentage de phrases reconnues exactes
Google	Google Speech API	15.8%	73.3%
Nuance	Nuance speech recognition REST API	39.7%	44.1%
IBM	IBM Watson REST API	42.3%	46.9%
AT&T	AT&T speech recognition REST API	63.3%	32.8%
WIT	WIT REST API	63.3%	29.5%

Tableau 3-4 : Etude comparative des APIs sur des petits segments Audio[2]

Actuellement, dans le volet programmation et développement nous trouvons que Python est un langage de script de haut niveau, interprété, interactif et orienté objet qui est remarquablement puissant [Beazley, 2009]. Il possède des caractéristiques clés qui le rendent unique par rapport d'autres langages orientés objet. Parmi eux, nous citons : la syntaxe très claire et perceptible ainsi que ses fortes capacités d'auto-analyse, … etc. Cependant, le choix d'utilisation de ce langage est jugé par les caractéristiques suivantes :

- *La portabilité* : Python est portable, non seulement sur les différentes variantes d'*Unix*, mais aussi sur OS et les différentes variantes de Windows. Aussi, un nouveau compilateur, baptisé *JPython*, est écrit en Java et génère du *bytecode Java*.
- La rapidité : La syntaxe de Python est très simple et conduit à des programmes à la fois très compacts et très lisibles. En effet, un programme Python est souvent de 3 à 5 fois plus court qu'un programme *C++* ou *Java* équivalent, ce qui représente en général un temps de développement de 5 à 10 fois plus court et une facilité de maintenance largement accrue.

[1] https://www.ldc.upenn.edu/language-resources

[2] Selon le site http://blog-archive.griddynamics.com/2016/01/automatic-speech-recognition-services.html

- La fonctionnalité : Python est orienté-objet. Il supporte l'héritage multiple et la surcharge des opérateurs. Dans son modèle objets, et en reprenant la terminologie de C++, toutes les méthodes sont virtuelles.
- L'accessibilité : la librairie standard de Python, et les paquetages contribués, donnent accès à une grande variété de services : chaînes de caractères et expressions régulières, services UNIX standard comme les signaux, les images, les threads…etc., les protocoles Internet comme Web, News, FTP, CGI, HTML..., bases de données, interfaces graphiques GUI.

De point de vue ressources, l'index des paquets Python « *Python Package Index* (*PyPI*) » héberge des milliers de modules tiers pour Python et parmi elles nombreux sont dans le domaine de traitements automatiques de la langue. Les bibliothèques standard de Python et les modules apportés par la communauté permettent des larges possibilités. Parmi eux, nous intéressons aux *SpeechRecognition Package*, *aeneas Package* et *NLTK Package*.

Le Speech Recognition *PyPl* dans sa version 3.5.0 est une bibliothèque *Python* pour la reconnaissance vocale, avec un support pour plusieurs moteurs et API, en ligne et hors ligne. Cette bibliothèque supporte les APIs suivantes :

- CMU Sphinx (fonctionne en mode hors ligne « offline ».
- Google Speech Recognition.
- Wit.ai.
- Microsoft Bing Voice Recognition.
- Houndify API.
- IBM Speech to Text.

À cet effet, nous avons utilisé cette bibliothèque dans ces deux formes : on ligne avec *Google-API-python-client* qui utilise le *Google Cloud Speech API* et hors ligne qui utilise *PocketSphinx* avec le module *Sphinx recognizer*. En revanche, nous trouvons que la transcription des deux méthodes est basée sur la taille de segments contrairement aux fichiers textes qui exploitent les signes de ponctuation. Pour cela, dans le chapitre suivant, nous réalisons une série d'expériences sur le corpus utilisé pour la simulation de notre approche afin de trouver le nombre de segments optimaux vis-à-vis les durées de silences ainsi que leurs débits phonétiques.

Dans ce contexte, nous étalons une présentation d'un exemple d'une ressource parlée *« /.sph/ »* avec son fichier de transcription *« /.stm/ »* extrait du corpus qui sera utilisée dans le chapitre suivant. Nous avons procédé à des tâches de segmentation ainsi des routines de transcription avec les deux formes de décodage : en ligne avec le service *Cloud Google Speech* et hors ligne avec la bibliothèque de *Pocket Sphynx* sous *Python*. Dans le tableau 3-5, nous présentons des extraits des différents fichiers de cette procédure.

Description du Fichier	Extrait du document
Fichier de Transcription StevenJohnson_2006S.stm (*segment time marked*)	*StevenJohnson_2006S 1 StevenJohnson_2006S 12.80 20.40 <o,f0,male> if you haven't(2) ordered yet {UM} i {COUGH} generally find the(2) rigatoni with the spicy {NOISE} tomato(2) sauce goes best with {BREATH} (StevenJohnson_2006S-12.80-20.40-F0_M-S1)* *StevenJohnson_2006S 1 StevenJohnson_2006S 131.12 139.34 <o,f0,male> and keep {SMACK} them in the attic until {NOISE} literally their milk ran out and(2) they died and then they would kind of drag them off {BREATH} to {UH} the {NOISE} bone boilers down the street {UM} (StevenJohnson_2006S-131.12-139.34-F0_M-S1)* *StevenJohnson_2006S 1 StevenJohnson_2006S 140.32 148.63 <o,f0,male> so {UH} you would just walk(2) around in london at this point and just(2) be {NOISE} overwhelmed(2) <sil> with this {NOISE} stench {SMACK} and what ended up happening(2) is that(2) an entire <sil> (StevenJohnson_2006S-140.32-148.63-F0_M-S1)*
Fichier de cartes de synchronisation StevenJohnson_2006S.json	*{* *"fragments": [* *……* *{* *"begin": "32.640",* *"children": [],* *"end": "38.120",* *"id": "f000004",* *"language": "eng",* *"lines": [* *"StevenJohnson_2006S 1 StevenJohnson_2006S 12.80 20.40 <o,f0,male> If you haven't(2) ordered yet {UM} i {COUGH} generally find the(2) rigatoni with the spicy {NOISE} tomato(2) sauce goes best with {BREATH} (StevenJohnson_2006S-12.80-20.40-F0_M-S1)"* *]* *}* *…..* *}*
Transcription de StevenJohnson_2006S.sph Avec *Google Cloud Speech API*	*If you haven't ordered yet I generally find the rigatoni with spicy tomato sauce goes best with diseases of the small intestine sorry I just feel like I should be doing standup here because I want to do is take you back 1854 in London for the next few minutes and can tell the story hymn proof of this of this outbreak which in many ways I think helped create the world that we live in today and the kind of city that we live in today this…* *…*
Transcription de StevenJohnson_2006S.sph Avec le module *Sphinx Recognizer*	*If you have been ordered yet found out by gentle if I'm the rigatoni with the spicy that Minnesota's that's what diseases of the small intestine had south so I just feel that it should be doing stand-up appear that this is I know at what I want you is take you back to that too eighteen fifty four among them of the next few minutes then and tell the story about you prefer this of this outbreak tom which in many ways I think help create the world we live in today and I the kind of city that we wouldn't that this period …* *…*

Tableau 3-5 : Extrait des transcriptions Online et Offline d'un document parlé

Entre autres, la transcription des flux parlés avec l'utilisation des segmentations à base des silences rend l'analyse lexicale et syntaxique des résultats de transcriptions automatiques est incontournable. Dans ce contexte, nous avons développé pour notre architecture un analyseur lexical syntaxique avec l'environnement Delphi, sous Windows pour les traitements des sorties de la phase de reconnaissance afin de surmonter les problèmes lexicaux et syntaxiques.

3.5. L'analyseur sémantique du contexte CSA-SOMI

3.5.1. Description du CSA-SOMI

Dans la première phase de notre approche, nous avons travaillé sur la recherche d'une représentation linguistique du contenu des documents parlés le plus proche du point de vue syntaxique. Cette dernière sera utilisée pour trouver les termes d'indexation les plus discriminants pour ces ressources parlées. Entretemps, nous étudions l'efficacité d'utilisation des transcriptions partielles d'un flux parlé avec un processus d'enrichissement sémantique dans le processus d'indexation. Aussi, nous évaluons les capacités de cette stratégie par rapport aux problèmes des erreurs de reconnaissance et les mots techniques et hors vocabulaire.

D'autre part, l'indexation et l'extraction des concepts du contexte des flux parlés sont des tâches lourdes. Elle consomme beaucoup de temps de calcul et de ressources, soit en mode manuel ou même en mode automatique. Pour surmonter ces problèmes, nous étudions l'efficacité d'un processus d'indexation sémantique basé sur des transcriptions partielles du contenu parlé dans un système de recherche.

À cet effet, l'objectif de ce module CSA-SOMI est de trouver les termes d'indexations discriminant avec un processus d'enrichissement sémantique pour les flux parlés qui couvrent la globalité de leurs contenus. Ce module permet de trouver une représentation sémantique avec une couverture globale du contenu des documents parlés, sa structure globale est présentée par la figure 3-7.

Notre solution est basée sur l'approche d'indexation sémantique du contexte en utilisant les résultats d'une transcription globale et partielle. En premier lieu pour détecter les topics abordés. Ensuite et en fonction de la disponibilité d'un modèle de représentation des connaissances comme les ontologies *WordNet*. Nous cherchons des correspondances et des similarités entre les concepts provenant des transcriptions automatiques du contenu parlé. Ainsi, nous exécutons un processus d'enrichissement sémantique par l'alignement de ces concepts avec le modèle de représentation de connaissance. Enfin, ces concepts sont soumis aux règles d'indexation pour la définition de l'ensemble des termes d'indexation qui représentent le contenu du flux parlé de ces ressources.

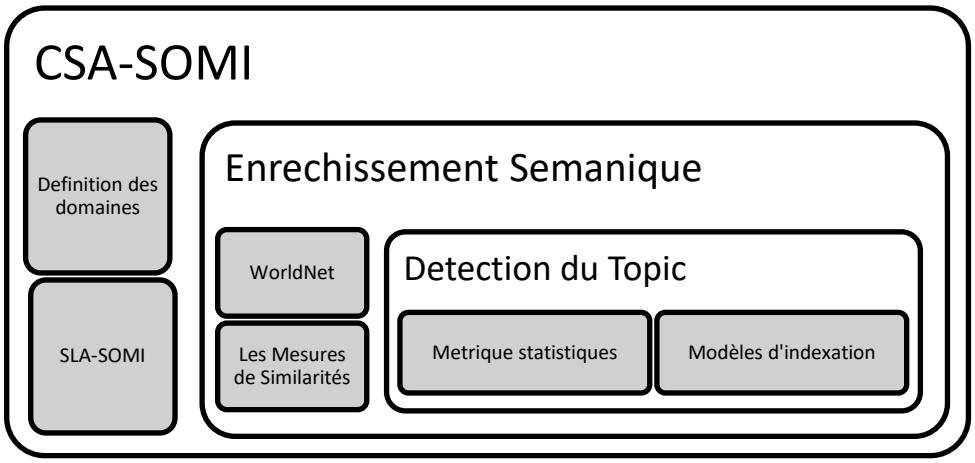

Figure 3-7 : Structure globale du module CSA-SOMI [Bendib, 2018]

3.5.2. La Détection des Topics

La détection des Topics du contenu des documents parlés nous permet d'avoir un ensemble de termes d'indexation homogène et discriminant. En effet, nous assignons par le terme « *Topic* » toutes partie limitée d'une discipline, phénomène, culture, branche scientifique…, caractérisée par un ensemble des termes fréquents susceptibles de le distinguer par rapport aux autres domaines.

En revanche, la détection des topics à partir des résultats des transcriptions automatiques de flux parlé n'est pas toujours implicite. Car les erreurs générées par ces systèmes de reconnaissance par insertion ou par omission ainsi que les problèmes de reconnaissance des mots techniques et les mots hors-vocabulaire influent sur la qualité du contenu du flux parlé obtenu. Aussi, le processus de segmentation automatique des ressources parlées engendre des troncatures dans le sens des phrases parlées. Ce qui affaiblit les performances de l'algorithme de décodage *Viterbi* dans les treillis de mots ou phonèmes par rapport aux modèles de langages utilisés.

3.5.2.1. Modèle de représentation des Topics

Nous utilisons le modèle vectoriel pour la modélisation des ressources et les Topics. Nous définissons un Topic T_i par un vecteur de fréquences des termes j par rapport au Topic T_i « «$tf_{j,i}$ ». Cette fréquence est pondérée par la somme des fréquences locales tf par rapport au T_i. Il est calculé par :

$$W_{T\,i,j} = \frac{tf_{j,i}}{\sum_{k \in Terms} tf_{k,i}} \qquad (3\text{-}1)$$

Notons que nous n'avons pas fait appel aux fréquences inverses « idf », car le concept « *Topic* » utilisé dans cette approche modélise l'ensemble de termes susceptible pour l'indexation. D'où nous favorisons une pondération locale au lieu d'une pondération générale. Dans ce contexte, nous avons étiqueté par rapport aux topics les fichiers de transcriptions « *.stm* » correspondant aux ressources parlées pour le calcul des fréquences locales. Notons aussi que nous pouvons avoir plusieurs topics différents dans le contenu d'un flux parlé.

En revanche, pour la représentation des résultats de transcriptions du contenu des flux parlés fournis par le module SLA-SOMI. Nous utilisons le même modèle, mais avec une pondération par rapport au document, tel que :

$$W_{D\,j,k} = \frac{tf_{k,j}}{\sum_{i \in Terms} tf_i} \qquad (3\text{-}2)$$

Cependant, ces pondérations seront calculées après l'exécution d'un processus de prétraitements lexicaux sur les résultats bruts de transcriptions. Ce processus exécute les tâches suivantes :

- La racinisation des termes « *Stemming* » par l'algorithme de « *Porter* ».
- La lemmatisation des termes « *Lemmatization* ».
- L'élimination des mots vides « Stop Words ».

Dans ce contexte, nous avons développé un environnement avec le langage de programmation Delphi. Cet environnement permet l'exécution du processus de prétraitements ainsi que le calcul des différents modèles de représentation des ressources et Topics.

3.5.2.2. Stratégie de détection des Topics

La détection des Topics du contenu des ressources parlées n'est pas une tâche de classification classique. Elle cherche de détecter tous les topics qui couvrent le contenu parlé plutôt que classer ce contenu par rapport aux topics. Entre autres, nous cherchons des degrés de couverture de chaque Topics dans le contenu d'une ressource parlé. Cette couverture sera traduite par un score ou un seuil et elle sera l'objet d'indexation du contenu de cette ressource parlée.

En revanche, pour les modèles vectoriels, les similarités sont exprimées par des distances comme la distance de *Minkowski*. Cette distance est calculée par la formule suivante :

$$d(i,j) = \sqrt[q]{\left(\left|x_{i1} - x_{j1}\right|\right)^q + \left(\left|x_{i2} - x_{j2}\right|\right)^q + \cdots + \left(\left|x_{ip} - x_{jp}\right|\right)^q} \qquad (3\text{-}3)$$

Aussi, nous trouvons la distance de « *Manhattan* » et la distance « *Euclidienne* » qui sont des cas spécifiques pour la distance de *« Minkowski* » tel que $q = 1, q = 2$ respectivement.

En effet, parmi les avantages du modèle vectoriel est l'utilisation des distances comme une métrique de détection des similarités entre les modèles de représentation des ressources. Pour cela, nous admettons leurs utilisations dans le processus de détection des topics du contenu du flux parlé transcrit.

En revanche, ces mesures ne prennent pas en charge les distances conditionnelles d'un terme par rapport aux autres termes. En effet, elles calculent le cumul des distances de chaque terme indépendamment. Elles peuvent être écrites comme :

$$d(i,j) = \sqrt[q]{\sum_{k=1}^{p} \left(\left|x_{ik} - x_{jk}\right|\right)^q} \qquad (3\text{-}4)$$

Dans ce contexte, nous inspirons du théorème de Bayes pour le calcul conditionnel. Pour la détection des topics nous calculons les distances entre les termes du document D_j avec le vecteur de définition du topic T_i en tenant compte les distances de ces termes par rapport aux autres topics. Aussi, pour le calcul de distance d'un vecteur de représentation du contenu parlé d'une ressource D_j avec le vecteur de définition d'un topic T_i, nous utilisons un mappage du vecteur D_j avec le vecteur de topic T_i pour le calcul de la valeur de pondération. Lors de cette phase, la pondération sera calculée à base des termes communs des deux vecteurs D_j et T_i seulement. Donc, la fréquence pondérée pour un terme de document D_j par rapport au topic T_i sera calculée comme suit :

$$W_{T_{i,k}D_{j,k}} = \frac{tf_{k,i}}{\sum_{s\in(Terms_i \cap Terms_j)} tf_{s,i}} \qquad (3\text{-}5)$$

Ainsi, nous utilisons la somme des distances de chaque terme du D_j par rapport aux autres topics T_p tel que $(p \neq i)$ comme un facteur de pénalisation divisée sur $W_{T_{i,k}D_{j,k}}$. En effet, cette stratégie permet de pénaliser les termes non discriminants.

La somme des distances de chaque terme de D_j par rapport au topic T_i représente le score de détection $Score_1(T_i, D_j)$. Il est calculé par :

$$Score_1(T_i, D_j) = \sum_{k \in (Terms_i \cap Terms_j)} 2 - \left| W_{D\,j,k} - W_{T_{i,k} D_{j,k}} \right| - \frac{\left| W_{D\,j,k} - \left(M_i * m - W_{T_{i,k} D_{j,k}} \right) \right|}{m-1} \quad (3\text{-}6)$$

Avec

- $W_{D\,j,k}$: la fréquence pondérée du terme k dans le document j.
- $W_{T_{i,k} D_{j,k}}$: la fréquence pondérée du terme k de document D_j par rapport au topic T_i
- $M_k = \frac{\sum_{i=1}^{m} tf_{i,k}}{m}$: la fréquence moyenne de l'apparition du mot du terme k par rapport à tous les topics.
- m : le nombre de topics.

Ensuite, nous obtenons un score de classement des topics par rapport au contenu d'un flux parlé. La formule (3-6) peut être simplifiée comme suite :

$$Score_2(T_i, D_j) = \sum_{k \in (Terms_i \cap Terms_j)} 2 - \left| W_{D\,j,k} - W_{T_{i,k} D_{j,k}} \right| - \left| W_{D\,j,k} - M_j \right| \quad (3\text{-}7)$$

Où le score représente le degré de correspondance entre un document Dj avec un topic Ti : il est exprimé en fonction de la cumule des différences de fréquence entre les termes constituant ce document et celles du vecteur de topic, avec considération des différences de fréquence d'occurrence des termes dans le document et celles de la fréquence moyenne d'occurrence dans tous les topics.

Cependant, nous utilisons une autre alternative de ce score qui possède une lecture probabiliste. Cette mesure prend en considération le cumul de la différence de fréquence de termes constituant un document et celles du vecteur de topic divisé par la fréquence moyenne d'occurrence des termes dans tous les topics.

$$Score_3(T_i, D_j) = \sum_{k \in (Terms_i \cap Terms_j)} - \frac{\left| W_{D\,j,k} - W_{T_{i,k} D_{j,k}} \right|}{M_j} \quad (3\text{-}8)$$

L'objectif de ce $Score_A$ de classement du document D_j pour le topic T_i est de stimuler l'impact des fréquences des termes de D_j par rapport aux autres topics sur les résultats.

En résumant, la figure 3-8 illustre la stratégie définie pour la modélisation et la détection des topics. Dans le quatrième chapitre, nous présentons les détails de l'environnement développé et les résultats obtenus pour les différentes étapes de ce module.

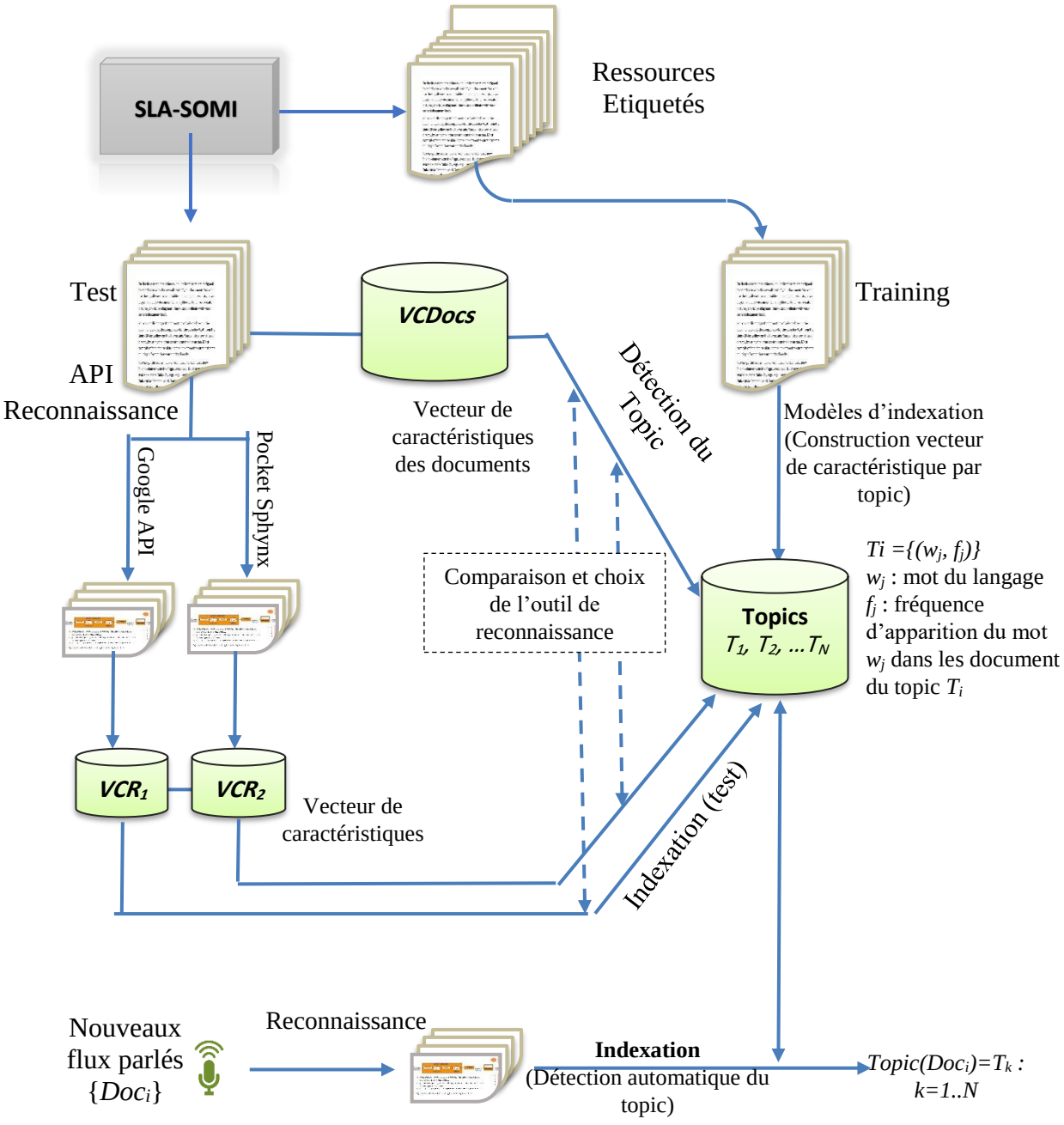

Figure 3-8 : Stratégies de détection des Topics proposée

3.5.2.3. Détection des Topics à base des transcriptions partielles

Le processus de décodage consomme beaucoup de temps de calculs et de ressources. Pour cela nous étudions dans cette section l'efficacité de la stratégie de détection de topic définie par rapport au contenu partiel. En effet, la transcription du contenu du flux parlé dans le module SLA-SOMI est effectuée à base des segments, d'où l'opportunité de tester les performances de système de détection par rapport au contenu partiel. Donc nous cherchons le passage du global vers le partiel pour surmonter les problèmes de charges de calculs avec un processus d'enrichissement sémantique qui sera présenté dans la section suivante.

Dans ce contexte, nous définissons le paramètre α qui présente le taux de partiel, par exemple $\alpha = 1$ signifie qu'on utilise la totalité de D_j pour la détection des Topics, ainsi que $\alpha = 3$ signifie qu'on utilise le tiers de D_j. On note aussi, que la pondération des termes $W_{T_{i,k}D_{j,k}}$ de D_j par rapport au T_i sera calculée sur le partiel et non sur le total. En outre, la sélection des termes du partiel s'effectue via une procédure séquentielle simple à base de « modulo ». Aussi dans le chapitre d'implémentation et d'évaluation, nous présentons l'environnement développé ainsi que les résultats obtenus.

3.5.3. L'enrichissement sémantique

La détection des topics permet de référencer les contenus des ressources parlées par rapport à leurs contextes. Les concepts fréquents de ces contextes construisent l'ensemble des termes d'indexation candidats pour le contenu de ces flux parlés. Cependant le choix des termes d'indexation discriminants à partir de cet ensemble n'est pas évident, car la notion de fréquence n'implique pas forcément l'importance et le pouvoir de discrimination. L'utilisation des mesures des similarités sémantiques entre les différents termes de cet ensemble permet la sélection des termes d'indexation valides fréquents et discriminant à la fois.

Cette détection est basée sur les transcriptions du contenu parlé obtenues par des traitements automatiques. La qualité de cette détection est liée étroitement par les performances des systèmes de reconnaissances. Ces derniers n'assurent pas une transcription du contenu intégral et ils sont incapables de détecter les termes étranges de langues et les termes techniques. Cependant, ces termes techniques sont souvent des termes d'indexation valides et discriminants.

Dans ce contexte que nous intégrons une étape d'enrichissement sémantique dans notre démarche proposer. Cette étape à deux vocations, elle améliore la capacité de système de détection des topics pour les ressources parlées dont ils englobent plusieurs topics avec des fréquences équiprobables. Ainsi, elle permet l'enrichissement de la liste des termes d'indexation des ressources parlée par de nouveaux concepts extraits à partir des ontologies. À cet effet, nous utilisons l'ontologie *WordNet* comme un support de représentation de connaissances dans ce processus d'enrichissement. Le scénario général de cette étape est illustré dans la figure 3-9.

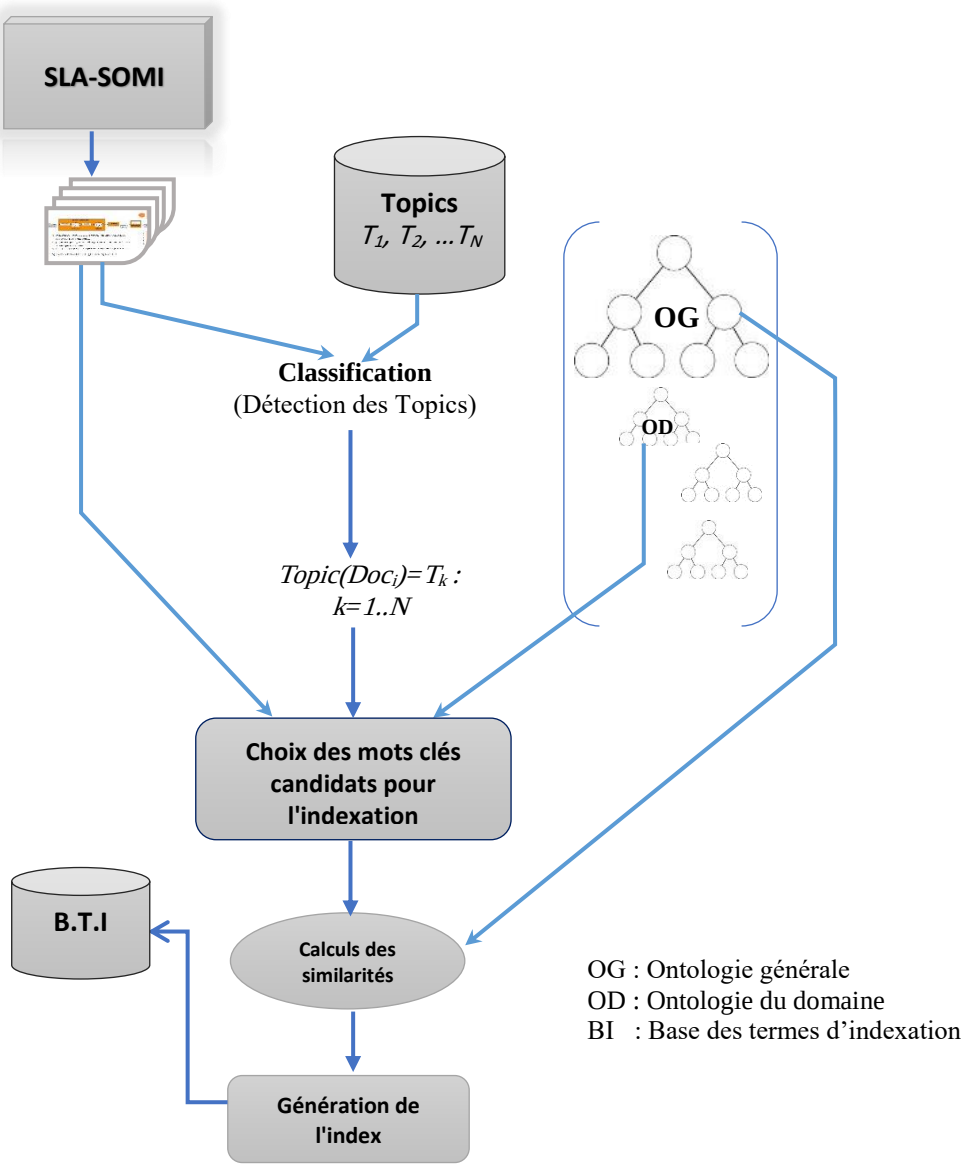

Figure 3-9 : Scenario général du processus d'enrichissement sémantique

3.5.3.1. WordNet : la ressource sémantique

Le processus d'enrichissement sémantique nécessite des modèles de représentation de connaissances. Ces modèles sont les ressources sémantiques utilisées pour les calculs des différentes métriques de similarités entre les termes susceptibles d'indexation du contenu d'un flux parlé. L'ontologie *WordNet* est une ressource largement utilisée pour le traitement du contenu des ressources en langue anglaise. Elle est structurée en trois catégories : les noms, les verbes les adjectifs et les adverbes. Avec un ensemble de relations comme : *hypernym*, *hyponym*, *has member*, *member of*, *has part*, *part of*, …etc. Ces relations et concepts sont organisés sous forme des « *Synset*s »

Afin de rendre l'ontologie *WordNet* exploitable dans l'environnement développé pour le module CSA-SOMI, nous proposons une modélisation de ces connaissances sous forme d'un modèle simplifié avec le langage UML, décrit dans la figure 3-10. Ensuite, nous avons développé une application pour la représentation de l'ontologie *WordNet* dans ce modèle relationnel. Ce modèle nous permet la visualisation des différentes relations sémantiques entre les termes et le calcul des mesures de similarités correspondantes.

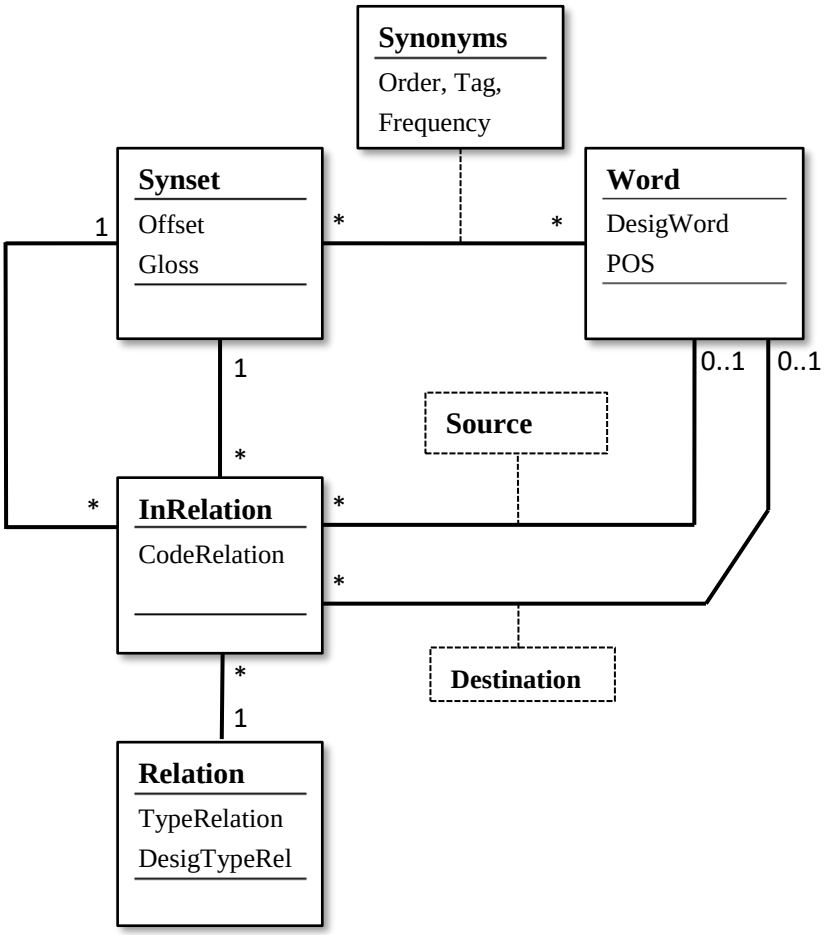

Figure 3-10 : Modèle UML simplifié proposé pour l'ontologie *WordNet*

3.5.3.2. Mesures de similarités utilisées

Les mesures de similarités permettent de capturer la généralité et le caractère concret d'un concept ou terme. Elles évaluent l'informativité et l'expressivité d'un concept en fonction de son placement dans la hiérarchie en fonction de ses ancêtres et ses descendants. Dans ce contexte, nous exploitons le riche réseau de relations entre les concepts présents *WordNet* pour évaluer les distances sémantiques dans les termes d'indexation du contenu des flux parlés. Nous avons présenté dans la troisième partie du deuxième chapitre une étude bibliographique sur les différentes mesures utilisées. Nous utilisons les mesures basées sur la hiérarchie et le nombre de nœuds et branches vu le vaste volume de connaissances disponibles dans l'ontologie WordNet.

Nous avons utilisé dans ce module d'enrichissement sémantique les mesures de similarités : *Lesk étendu*, *Rada*, *Wu & Palmer* et *Leacock & Chodorow*. Elles sont calculées par les formules suivantes :

- $Sim_{Wu\&Palmer}(t_1, t_2) = \frac{2.N_3}{N_1+N_2+2.N_3}$ (3-9)
- $Sim_{Rada}(t_1, ct_2) = N_1 + N_2$ (3-10)
- $Sim_{L\&Ch} = -log\left(\frac{N_1+N_2}{2.D}\right)$ (3-11)

Avec

- N_1 : le nombre des nœuds parcouru à partir du concept du terme t_1 jusqu'au atteindre le premier concept commun entre les termes t_1, t_2.
- N_2 : le nombre des nœuds parcouru à partir du concept du terme t_2 jusqu'au atteindre le premier concept commun entre les termes t_1, t_2
- N_3 : le nombre des nœuds parcouru à partir du concept commun entre les termes t_1, t_2 jusqu'au atteindre la racine.
- D : la profondeur de hiérarchie de la ressource sémantique.

Ces mesures sont utilisées pour la détection des similarités entre les différents termes de l'ensemble d'index candidats pour le contenu transcrit d'un flux parlé. Ces similarités permettent la sélection des termes d'indexation valide à partir de cet ensemble. Entre autres, l'intégration de plusieurs mesures dans l'environnement développé pour l'approche d'indexation proposée nous permet d'avoir des possibilités d'expérimentations et de test pour sélectionner les mesures les plus adéquates.

3.5.3.3. Stratégies d'enrichissement

La représentation par le modèle vectoriel des résultats de transcriptions nous permet la détection des fréquences d'apparition locales des termes du contenu d'un flux parlé. Cette fréquence nous permet la sélection des termes susceptibles d'indexation de ce contenu selon un seuil minimum et maximum donné. Ces termes n'ont pas le même degré de représentativité du contenu cible d'indexation. En effet, les fréquences d'apparitions ne permettent pas la distinction entre ces termes. À cet effet, nous utilisons les mesures de similarités pour détecter les termes d'indexation pertinent et efficace.

Dans ce contexte, nous calculons la matrice de similarité pour l'ensemble des termes candidats d'indexations. Ensuite, pour tout tuple $\{term_i, term_j\}$ qui dépasse le seuil minimum de similarité, nous exécutons un processus d'enrichissement sémantique de ces termes par les relations définis dans l'ontologie WordNet. La figure 3-11 présente l'algorithme de scénario d'enrichissement sémantique.

```
1. Initialisation des seuils de fréquences d'apparitions
       S_freq_min # Seuil de fréquence minimum
       Sl_freq_max # Seuil de fréquence maximum
       Seuil       # Seuil minimum de similarités
       Term_Index= {}
2. Construction de l'ensemble des termes candidats d'indexation
       Term_Index_condidats =
              If (freq_W>= S_freq_min and freq_W<= S_freq_max
                     Term_Index_condidats.insert (W)
3. Calcul des similarités sémantiques
       for each Wi in Term_Index_condidats
              for each (Wk in Term_Index_condidats) and (i<>j)
                     Mat_Sim(i,j) = Mesure_similarity(Wi,Wj)
4. Détection des termes similaires
       Similar_Terms ={}
       for i=1 : nb_Words
              for j=1 : nb_Words -1
                     if (i<>j) and (abs(Mat_Sim(i,j)- Mat_Sim(i,j+1))>= Seuil)
                     Similar_Terms.add(Wi,Wj)
                     Term_Index.insert (Wi,freq_Wi)
                     Term_Index.insert (Wj,freq_Wj)
5. Enrechissement sémantique
              for each tuple {Wi, Wj} in Similar_Terms
                     New_term = Semantic_enrechissement(Wi, Wj)
                     Term_Index.insert (New_term,Moy(freq_Wi, freq_Wj)
```

Figure 3-11 : Algorithme d'enrichissement sémantique [Bendib, 2014]

3.6. Le moteur de détection des index KDE-SOMI

3.6.1. Description du KDE-SOMI

Dans la dernière phase de cette approche, nous définissons les mécanismes nécessaires pour la recherche dans le contenu du flux parlé des ressources multimédias en utilisons les résultats des deux modules SLA/CLA SOMI. Ces modules nous permettent d'avoir des termes d'indexation pour le contenu du flux parlé. Ces termes permettent une représentation de la totalité du contenu. En revanche, nous visons dans cette approche d'identifier les segments pertinents du flux parlé par rapport aux termes d'indexation. Cette stratégie nous permet de gérer les ressources parlées volumineuses par la restitution des segments d'un flux parlé au lieu de sa totalité.

Dans le module de détection « *Keyword Detection Engine* - KDE-SOMI », nous utilisons la technique de détection des termes parlés « *Keyword Spotting* » pour identifier les intervalles des segments qui contiennent les termes d'indexation. En revanche, le traitement du contenu parlé nécessite la définition de plusieurs modèles de représentations. Ces modèles permettent la modélisation du contenu parlé pour les différentes étapes de décodage et de détection par les SGMM-HMM. La figure 3-12 présente la structure globale de ce module.

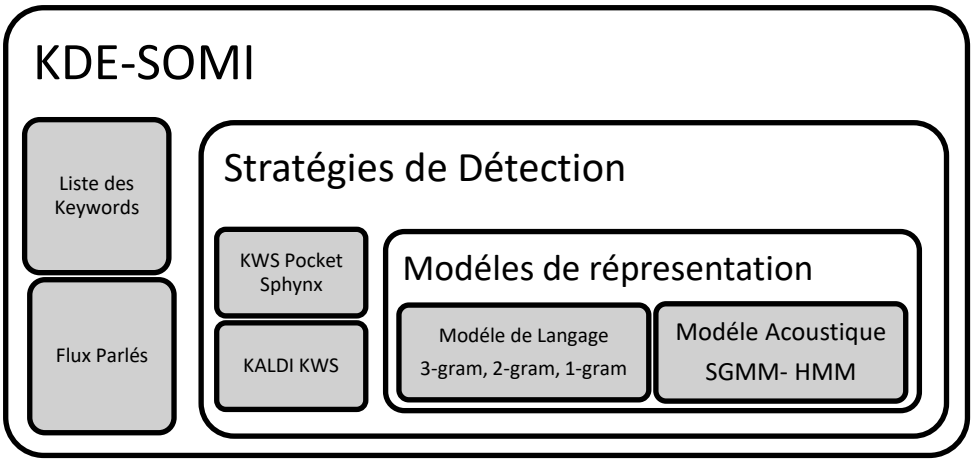

Figure 3-12 : Structure globale du KDE-SOMI [Bendib, 2018]

3.6.2. Modèles de représentations

Ces modèles sont nécessaires pour la création des systèmes de reconnaissances ou détection des termes parlés. Il existe deux possibilités, soit de définir les propres modèles nécessaires pour le décodage : modèle acoustique, modèle de langage, dictionnaire phonétique et la grammaire. Cette solution est utilisée pour une tâche de recherche dans des ressources parlées dans des domaines spécifiques. Soit d'exploiter les modèles standards disponibles dans les plateformes de développement. Cette variante permet de gérer le contenu de flux parlé hétérogène. Dans les sections suivantes, nous présentons la description de ces modèles.

3.6.2.1. Modèle de langage

Les modèles de langage statistique contiennent des probabilités des mots et des combinaisons de mots dans une langue donnée. Ces probabilités sont estimées à partir des données d'apprentissage sous différents modèles :1-*gram*, bi-*gram*, 3-*gram*, …etc. Il existe plusieurs outils de création de ces modèles comme *SRILM* « *SRI Language Modeling Toolkit* », *LMtools* « *Sphinx Knowledge Base Tool* » … etc. Les capacités de ces modèles dépendent du volume de données d'apprentissage utilisées. Le tableau 3-4 présente la description des exemples de modèles de langage créé par l'outil *LMtool* à partir des différentes sources. Cependant, nous trouvons des plateformes de développement des systèmes de reconnaisse automatique offrent des modèles standards pour quelques langues utilisées. Le tableau 3-6 présente la description de quelques modèles.

Modèle	1-gram	2-gram	3-gram
Langue Anglaise avec Vocabulaire de 70 k élagué	72354	2217781	1577579
Langue Anglaise avec Vocabulaire de 70 k non élagué	72547	9704821	12264838
Langue Allemande avec Vocabulaire de 700 k	729029	2312977	816274
Langue Espagnol avec Vocabulaire de 700 k	23500	1463799	1109989

Tableau 3-6 : description de quelques modèles de langage standards [1]

3.6.2.2. Modèle acoustique

Le modèle acoustique définit les représentations des phonèmes constituant les mots d'un modèle de langage. Il construit souvent par les techniques d'apprentissage via l'algorithme *Baum-Welch* avec les modèles *GMM -HMM*. Les performances des modèles obtenus dépendent des caractéristiques des corpus d'apprentissage utilisée comme : la fréquence d'échantillonnage, le style de parole, topologie du HMM utilisé, fonctions de distribution des observations, nombre d'itérations, … etc. le tableau 3-7 représente la description de quelques modèles acoustiques disponibles dans les plateformes de développement des LVSCR.

Modèle	Style de parole	Nb HMM utilisé	Modèle d'observation	Fréquence d'échantillonnage	Outil d'acquisition
Communicator	Dialogue	4000	64 Gaussiennes	8 KHz	Téléphone
SJ1	Dictée	4000	32 Gaussiennes	16 KHz	Microphone
SJ2	Dictée	5000	32 Gaussiennes	8 KHz	Microphone
HUB 4	Actualités	6000	8 Gaussiennes	16 KHz	Large Bande

Tableau 3-7 : description de quelques modèles Acoustiques disponibles

[1] Description réalisée sur les données extraites du site web : www.speech.cs.cmu.ed

3.6.2.3. Dictionnaire phonétique

Les dictionnaires phonétiques définissent les différentes prononciations phonétiques des termes du vocabulaire d'une langue. Il existe un certain nombre de dictionnaires libres d'accès couvrant certaines langues comme l'anglais américain, le français, l'allemand, le mandarin, …etc. Nous pouvons utilisés ces dictionnaires phonétiques dans le développement des systèmes de reconnaissance et détection Cependant, nous pouvons les adaptés via des applications tel que : *Phonetisaurus* et *Sequitur*. La figure 3-13 présente un extrait du dictionnaire phonétique en langue anglaise.

```
# CMUdict  --  Major Version: 0.07 ;
COMPUTER    K AH0 M P Y UW1 T ER0
COMPUTER'S  K AH0 M P Y UW1 T ER0 Z
PRINTED     P R IH1 N T IH0 D
PRINTER     P R IH1 N T ER0
PRINTERS's  P R IH1 N T ER0 Z
PRINTING    P R IH1 N T IH0 NG
```

Figure 3-13 : Extrai du dictionnaire phonétique *cmudict-0.7b*[1]

3.6.3. Stratégies de Détection

Il existe plusieurs plateformes et environnement qui permettent la réalisation des systèmes de détection des termes parlés comme : *Julius*, *Sphinx*, *Pocket Sphinx*, *Kaldi*, … etc. En effet, la détection des termes parlé s'effectue soit par l'utilisation des systèmes *LVCSR* avec la définition de ces termes dans une grammaire *JSGF* « *Java Speech Grammar Format* ». Soit par l'utilisation des techniques de détections « *Keyword Spotting* » avec la spécification de liste des termes à rechercher. Cette variante permette la spécification des seuils de détection pour chaque terme pour le décodage *LVSCR*. Nous présentons dans la section suivante deux systèmes qui peuvent être utilisés comme un noyau de détection pour notre approche proposée.

3.6.3.1. Pocket Sphinx KWS

L'architecture de ces systèmes est basée sur l'exploitation des modèles de représentation disponibles avec des systèmes de reconnaissance à large vocabulaire. Soit par la définition d'une grammaire de mots clés dans le cas où le système de recherche dans le flux parlé est dédié pour un contexte bien défini. Soit par l'utilisation des fichiers contenant les listes des termes parlés à détecter pour les cas de recherches dans des contenus parlés dans divers domaines. L'architecture globale de tels systèmes est présentée par la figure 3-14. En pratique, il existe plusieurs implémentations de la plateforme « *Sphinx Speech Recongnition* » dans divers environnements. Nous choisissons d'utiliser le module sous python « *pocketsphinx-python* » sous le système d'exploitation « Ubuntu 16.04».

[1] www.speech.cs.cmu.ed

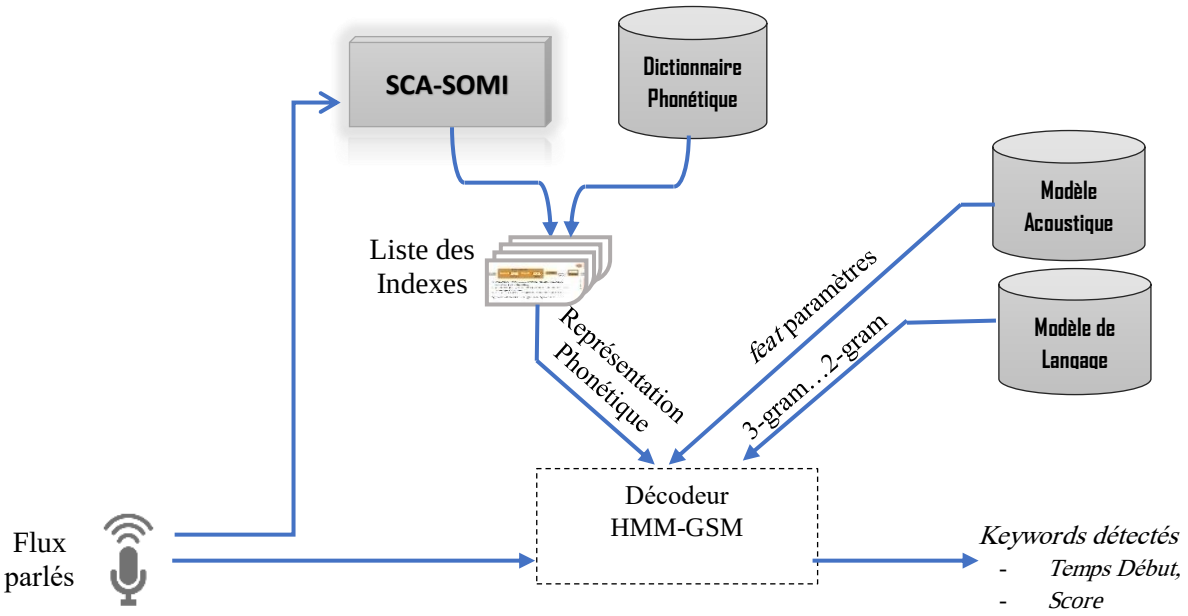

Figure 3-14 : Structure globale du KDE-SOMI

3.6.3.2. Kaldi KWS

Le « Kaldi KWS » est un module de la plateforme open source KALDI. IL utilise les transducteurs pondérés à état fini « *WFST* » pour la représentation des treillis de phonèmes de décodage LVCSR. Dans le processus de recherche, le mot-clé est converti en une simple machine à états finis avec le facteur de transducteur afin d'obtenir toutes ses occurrences dans la collection de recherche, qui conservent l'ID du segment parlé, le temps de début et de fin. Ensuite, ces occurrences seront triées en fonction de leurs probabilités a posteriori et une décision binaire est assignée à chaque instance. En pratique, les modules de la plateforme sont écrits en langage *C++* et la compilation et l'exécution des routines avec des scripts en *Perl* sous le système *Linux*. Cependant, cette plateforme la définition des modèles : langage, phonétique, et la grammaire. Tandis l'apprentissage du modèle acoustique est réalisé par les modèles HMM avec des mixtures de gaussiennes à base des sous-mots « *subword* ». Nous visons d'utiliser cette plateforme comme un module de détection de cette approche pour les ressources parlées dont leurs contextes sont spécialisés.

3.7. Conclusion

Dans ce chapitre nous avons présenté nos contributions pour améliorer les systèmes de recherches dans les flux parlés. Nous avons proposé une approche globale d'un processus de recherche basée sur un processus d'indexation sémantique du contenu des ressources parlées. Cette approche exploite les outils et les techniques de traitement automatique de la parole et celle de détection des termes parlés. Cependant nous avons exploité les similarités entre les

résultats des transcriptions automatiques pour capter l'aspect sémantique afin d'améliorer le processus d'indexation et de recherche.

Dans ce contexte, nous avons conçu le module SLA-SOMI qui utilise les APIs de reconnaissance « *Google Cloud Speech API* » et le module « *PocketSphinx* » sous *Python*. Ainsi, afin d'accroitre la qualité des résultats de ces systèmes de reconnaissance, nous avons utilisé une stratégie de segmentation automatique des flux parlés à base des durées de silence via le module « *SoX* ».

En revanche, pour le raffinement des résultats des transcriptions obtenus par les modules de reconnaissance automatique et diminuer l'impact de la stratégie de la segmentation, nous proposons le module CLA-SOMI. Il permet l'extraction des termes d'indexation les plus représentatifs du contenu par une stratégie de détection des topics du contenu du flux parlé et un processus d'enrichissement sémantique via les mesures de similarités extraites à partir de l'ontologie *WordNet*.

Finalement, pour l'accès au contenu des flux parlés nous proposons le module KDE-SOMI qui permet la détection des termes d'index fournis par CSA-SOMI via les techniques de « *Keyword Spotting* ». Dans ce contexte, nous optons pour l'utilisation le module *« PocketSphinx* » sous *Python* et nous visons d'utiliser la plateforme « *Kaldi* » comme une autre variante pour le processus de détection. Dans le chapitre suivant, nous présentons les différents outils développés pour la validation de cette approche.

Chapitre 4

Implémentation et Validation

4.1. Introduction

Dans ce chapitre, nous présentons les différentes phases de validation et d'implémentation de la contribution présentée dans le chapitre précédent. À cet effet, nous présentons l'environnement développé et les routines de programmation utilisées. Ensuite, nous étudions l'impact des différentes stratégies définies dans cette approche pour le processus de recherche dans le contenu parlé et nous discutons les résultats obtenus. Comme nous l'avons mentionné, la mise en œuvre de notre approche nécessite le traitement de deux problèmes intrinsèques : la complexité du codage et la complexité du contenu des ressources parlées par les techniques de détection et de recherche dans le domaine de traitements de la parole et la modélisation de contextes par les modèles représentation de connaissances comme *WordNet*.

En effet, la manipulation de ressources parlées de point de vue de l'indexation s'effectue à l'aide des systèmes de transcription *on-line* ou *hors-ligne*. L'efficacité de ces derniers est liée étroitement par la qualité des modèles syntaxiques et linguistiques utilisés. Tandis que, la recherche dans le contenu s'effectue par des techniques de détection des termes « mots-clés » ou indexes sous leurs formes phonétiques. Dans ce contexte, nous avons choisi d'utiliser dans notre démarche des ressources audio en langue anglaise avec des systèmes de reconnaissance on-line et hors-ligne à la fois avec une stratégie de segmentation automatique.

Entre autres, pour la manipulation efficace du contenu de ressources parlées et afin de traiter les erreurs de transcriptions, nous optons pour la modélisation et la représentation du contexte. Dans ce contexte, nous avons conçu et réalisés un environnement avec *Delphi* pour l'extraction et la détection du topic du contenu des ressources parlées à l'aide des résultats de transcriptions obtenus des segments parlés de ces ressources. Ce système utilise les techniques d'indexation et les distances de similarités sémantiques pour construire les représentations sémantiques de ces ressources avec l'utilisation des différentes relations et concepts fournis dans *WordNet*.

4.2. Choix de Ressources Parlés utilisées

4.2.1. Définition du corpus parlé

Les corpus parlés ou corpus audio sont des bases de données constituées des fichiers audios avec leurs transcriptions textuelles. Dans le domaine de la parole, ces corpus sont souvent utilisés pour la construction des modèles acoustiques. Dans le domaine du traitement et la recherche dans le contenu parlé, les corpus sont utilisés pour effectuer des recherches dans la représentation phonétique du contenu, ainsi pour la construction des modèles phonétiques d'indexation.

Généralement, il existe deux types de corpus parlé.

- *Langage lu (Read Speech) :* qui contient des séquences d'enregistrement audio pour des lectures formelles des ressources. Il comprend :
    - Des extraits de livres.
    - Des informations et discours (Broadcast news).
    - Listes de mots.
    - Des Séquences des nombres et des chiffres.
- *Discours spontané* - qui contient des séquences d'enregistrement audio pour des discours informelles. Il inclut :
    - Dialogues - entre deux personnes ou plus (comprends des réunions).
    - Narratives - une personne racontant une histoire.
    - Conversations téléphoniques.
    - Les conférences

4.2.2. Les corpus parlés existants

Dans la pratique, les recherches actuelles dans l'élaboration des corpus parlés sont orientées vers la recherche des correspondances entre les différentes sources des données audios pour une normalisation pour les mettre exploitables sous différentes plateformes et systèmes de traitement automatiques. Dans ce contexte, nous trouvons plusieurs normes de normalisation, comme le *« Linguistic Data Consortium* - LDC[1] ». Ce dernier est un consortium ouvert d'universités, de bibliothèques, de sociétés et de laboratoires de recherche gouvernementaux. Il a été formé en 1992 pour faire face à la pénurie de données critiques puis à la recherche et au développement de la technologie du langage. Ce consortium englobe plusieurs corpus parlés qui couvrent plusieurs langues ainsi que plusieurs formes d'enregistrements et modes d'utilisation comme : les appels téléphoniques, les ressources web, les émissions radio…etc. dans le tableau 4.1 nous nous présentons les corpus les plus utilisés dans le consortium LDC

[1] https://www.ldc.upenn.edu/about

Code	Description
LDC93S1	TIMIT Acoustic-Phonetic Continuous Speech Corpus
LDC2006T13	Web 1T 5-gram Version 1
LDC96L14	CELEX2
LDC2013T19	Onto Notes Release 5.0
LDC2008T19	The New York Times Annotated Corpus
LDC93S10	TIDIGITS
LDC2016T19	BOLT Chinese-English Word Alignment and Tagging -- Discussion Forum Training

Tableau 4-1 : Extrait des meilleurs corpus parlé dans LDC[1]

Dans le tableau précédent, nous trouvons que les corpus TIMIT[2] sont les plus sollicités. Ces corpus sont conçus pour le style du langage lu « *Read speech* ». Il sert à fournir des données et ressources vocales pour les études acoustiques-phonétiques et pour le développement et l'évaluation des systèmes automatiques de reconnaissance vocale. TIMIT contient des enregistrements à large bande avec dialecte d'Anglais américain. Le corpus TIMIT comprend des transcriptions orthographiques, phonétiques et de mots alignés dans le temps, ainsi qu'un fichier de flux de parole codé sur 16 bits avec une fréquence d'échantillonnage de 16 kHz.

Aussi, nous trouvons d'autres projets de coopération pour la construction des corpus parlés dans les plateformes de développement des systèmes de traitement automatique. Ces corpus sont souvent développés pour des fins des projets de recherches associés. Entre autres, la disponibilité de ces ressources pour leur utilisation est soumise sous les conditions des contrats de coopération. La majorité de ces ressources sont payantes et elles n'offrent pas des accès gratuits et complets pour leurs utilisations dans le développent des systèmes de traitement automatiques des flux parlés. Dans le tableau 4.2 nous citons quelques corpus et projets pour des systèmes de traitement automatiques de ressources parlés.

[1] Selon le classement de Top 10 sur le site officiel (https://catalog.ldc.upenn.edu/topten)

[2] Corpus réalisé avec un effort commun entre le Massachusetts Institute of Technology (MIT), SRI International (SRI) et Texas Instruments, Inc. (TI). Le discours a été enregistré à TI, transcrit au MIT et vérifié et préparé pour la production de CD-ROM par le National Institute of Standards and Technology (NIST).

Corpus	Données	Nb de Locuteurs (Speakers)	Langage	Nb de Phrases	Durée (Hours)	Accessibilité (Accebility)
AMI Meeting Corpus is a multi-modal data [Carletta, 2005]	Meeting recording	15	English	/	100	Payant
TIMIT Corpus of read speech [Garofolo, 1993]	Read speech	630	8 dialectes of English US	6300	5	Payant
The Wall Street Journal (WSJ) CSR Corpus [Paul, 1991]	Read speech mode	Speaker Independent	English	9600	80	Free
TED-LIUM corpus [Rousseau, 2012]	Real speech	698	English	/	118	Free
The Libri-Speech corpus [Panayotov, 2015]	Audio books	1166	English	/	960	Free
SWITCHBOARD CORPUS [Godfrey, 1992]	conversational speech	500	English US	2500	250	Payant
Fisher Corpus [Cieri, 2004]	conversational telephone speech	16454	variety pronunciations including English U.S.	16454	2000	Payant

Tableau 4-2 : Caractéristiques de quelques corpus audio parlés

En effet, la majorité des corpus parlés sont développés autour de la langue anglaise. Ce choix est justifié par la masse des travaux de recherches réalisés sur le développement des modèles linguistiques tel que : les modèles de langage 1-gram, bi-gram, tri-gram et même n-gram ; les modèles phonétiques et les modèles lexicaux. Néanmoins, nous trouvons aussi des travaux qui ont développé des versions de ces corpus pour d'autres langages comme les langues Mandarin, la langue française …etc. Cependant, les résultats obtenus par l'utilisation de ces ressources restent moins performants par rapport à la langue anglaise et cela dut aux richesses de ces ressources linguistiques.

Dans ce contexte, nous choisissons dans la phase de validation de notre approche proposée d'indexation sémantique pour la recherche dans le contenu parlé d'utiliser la langue anglaise. Ce choix est fondé par la disponibilité du volume important des ressources parlées pour l'expérimentation ainsi que la richesse et la diversité de ces derniers. Néanmoins, dans nos perspectives futures nous envisageons de tester d'autres langues dans notre approche SOMI.

En réalité, la communauté LDC qui a beaucoup travaillé sur l'élaboration des corpus audio standardisés riches de points de vue structure, linguistiques et annotation. Mais tous les corpus disponibles dans ses catalogues de LDC sont payants. Cependant, nous trouvons d'autres communautés qui offrent leurs corpus gratuitement tels que *LibrSpeech* et *TED-LIUM*. À cet effet, nous avons opté pour l'utilisation du corpus TED-LIUM dans la phase d'expérimentation de notre contribution.

4.2.3. Les corpus parlés utilisé – TED-LIUM

La conférence TED[1] « *Technology, Entertaitment, Design* » est un événement annuel lancé par *Richard Saul Wurman* et *Harry Marques* dans les années quatre-vingt du vingtième siècle avec un but non lucratif consacré à « *la puissance des idées qui peuvent changer le monde* ». L'objectif de la conférence TED est de diffuser les idées dans de différents domaines et disciplines. Elle met gratuitement à la disposition du public les meilleures conférences sur son site Web[2]. Les exposés couvrent un large éventail de sujets, tels que la science, les arts, la politique, les questions mondiales, l'architecture, la musique et plusieurs autres sphères de compétences. Les intervenants eux-mêmes sont d'une grande variété de disciplines. Ainsi, les conférences TED ont reçu les discours de personnalités publiques telles que l'ancien président des États-Unis *Bill Clinton*, l'inventeur du *Web Tim Berners-Lee*, le cofondateur de Wikipédia *Jimmy Wales*, les cofondateurs de Google *Sergey Brin* et *Lawrence E.* Page, … etc.

Figure 4-1 : La conférence TED [2]

En revanche, nous trouvons qu'une équipe de recherche du Laboratoire d'Informatique de l'Université du Maine « LIUM » a développé dans le cadre de projet de recherche le corpus TED-LIUM. Ce dernier a été réalisé à partir de conférences audios et de leurs transcriptions disponibles sur le site web de TED. Ce corpus diffusé sous la licence Creative Commons BY-NC-ND 3.0[3] . Ils ont préparé et filtré ces données afin de former des modèles acoustiques pour participer à l'événement « *International Workshop on Spoken Language Translation 2011* ».

1. https://fr.wikipedia.org/wiki/Conférence_TED
2. https://www.ted.com/
3. http://creativecommons.org/licenses/by-nc-nd/3.0/deed.en

Ce corpus est composé par :

- 774 fichiers audio dans la norme *NIST* avec le format sphere « *SPH* »
- 774 Transcriptions automatiques alignées sous le format *STM* format
- Dictionnaire de prononciations « 157617 *Words* ».

Ce corpus est composé des enregistrements de 774 locuteurs avec un volume d'environ 118 heures des ressources parlées.

Nombre de locuteurs	774
Nombre de segments	56800
Durée Totale	118h 4m 48s
Hommes	81h 53m 7s
Femmes	36h 11m 41s

Tableau 4-3 : Caractéristiques du corpus TED-LIUM V1[1]

Entre temps, la diversité du contexte des contenus des ressources parlé fourni dans ce corpus nous a encouragés à l'utiliser pour la validation de notre contribution. La diversité du contenu est un défi potentiel pour la validation de l'approche proposée. En effet, elle est basée sur le principe d'enrichissement sémantique du contenu du flux parlé par un processus d'indexation automatique.

Entre autres, le style utilisé dans les ressources parlées de ces conférences est le style de discours. Certes, ce style est caractérisé par sa bonne qualité phonétique ainsi que les termes de langage utilisé, mais il contient des hésitations, des émotions et même des discussions. Cependant, le contenu de cette conférence est diversifié. Il est généralement innovant et nouveau et il est souvent spécialisé et technique, ce qui augmente la tolérance d'existence des mots hors vocabulaire.

A cet effet, que nous avons opté pour utiliser ce corpus, car la philosophie et la stratégie de cette contribution est de surmonté les problèmes liés aux reconnaissances des mots étranges ; hors vocabulaires ; par le module sémantique CSA-SOMI. Dans les sections suivantes, nous présentons les différentes étapes et stratégies utilisées pour le processus de validation ainsi que la description de l'environnement développé « *MyWordNet* ».

[1] http://www-lium.univ-lemans.fr/en/content/ted-lium-corpus

4.3. Validation du module SLA-SOMI

L'objectif du ce module est le passage du contenu parlé vers une représentation textuelle à l'aide des outils de segmentation et les APIs de reconnaissance. Dans ce contexte, nous avons exécuté une série de tests pour connaitre les capacités de ce module devant les problèmes de charge des calculs, les erreurs de reconnaissance ainsi que les termes étranges et les termes techniques. À cet effet, nous exécutons dans les sections qui se suivent plusieurs scénarios pour valoriser l'efficacité de ce module.

4.3.1. Stratégies de segmentation

Le corpus TED-LUM est composé d'un ensemble d'enregistrements parlés de diverses durées. Ces durées influent étroitement sur les résultats de décodage dans les deux modes : *on-line* ou *off-line*. À cet effet, la segmentation de ces fichiers vers des segments plus petits est incontournable. Entre autres, les API des reconnaissances ne permettent pas le décodage des flux de parole de grande durée, ils sont incapables de transcrire le contenu. En effet, toutes les APIs actuelles sont conçues pour traiter des petites séquences de parole. En effet, leurs performances sont liées étroitement par la durée du flux parlé traité. Cette influence est due aux complexités de calculs et allocation des ressources dans les systèmes cloud ou dans les systèmes locaux. Dans ce qui se suit, nous essayions de valider l'impact des stratégies de segmentation automatique définies dans la section 3.4.2 de ce manuscrit dans deux scénarios différents.

4.3.1.1. Scenario 1

Dans ce scénario, nous exécutons plusieurs séries de tests sur les flux parlés basées sur les durées des silences existants dans le corpus *TEDLIUM V1*. Notre objectif est de trouver un compromis entre les valeurs optimales de silence qui peuvent être utilisées et le contenu des segments générés. Le corpus *TED-LIUM* est composé de trois ensembles. Chaque ensemble est divisé en deux répertoires : les enregistrements des discours parlés qui sont sous le format « *sph* » et leurs annotations et transcriptions sous format « *stm* ». Le tableau 4-4 décrit brièvement les caractéristiques de chaque ensemble.

Ensembles	**Nombre de fichiers**	**Taille**	**Durée Globale**
Train	774	23,7 Go	206 Heures 29 m
Dev	08	197,5 Mo	1 Heure 43 m
Test	11	353,5 Mo	3 Heures 6 m

Tableau 4-4 : Volumes et durées du documents parlé du corpus TED-LIUM V1[1]

[1] http://www-lium.univ-lemans.fr/en/content/ted-lium-corpus

Les durées de ces flux parlés se varient entre quelques minutes jusqu'aux dizaines de minutes. En effet, les durées de ces ressources parlées sont très importantes, car elles donnent une idée sur le nombre de phrases et de paragraphes de chaque discours parlé enregistrés. En plus, elles pressentent le degré de complexité de décodages et de ces modèles linguistiques respectivement. Dans ce contexte, nous présentons dans la figure 4-2 les nuages de distribution des durées des ressources parlées du corpus utilisé.

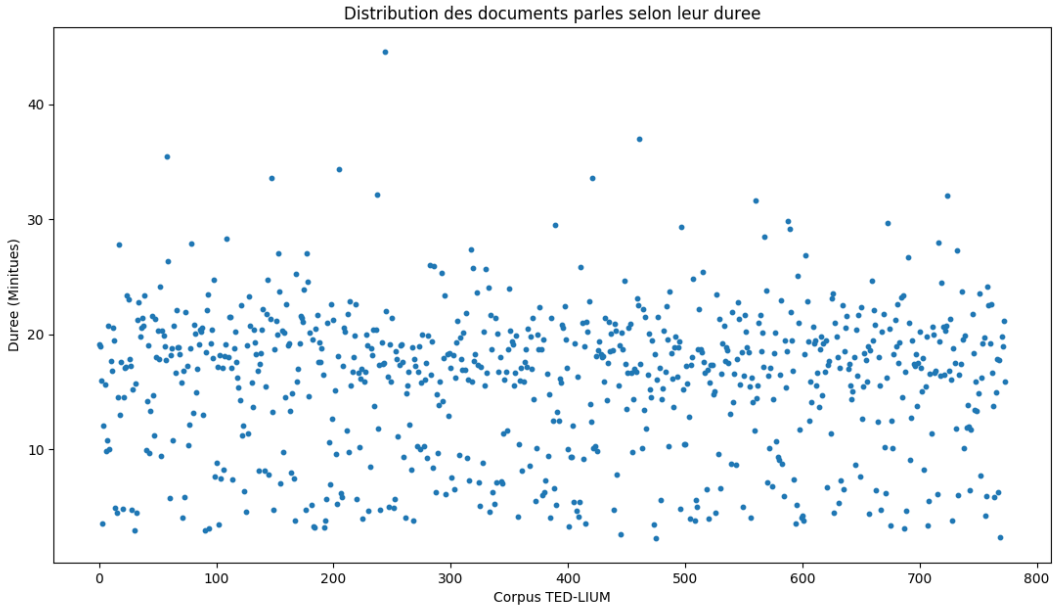

Figure 4-2 : Distribution des documents parlés selon leurs durées

Nous constatons sur cette figure que la majorité des durées de ces flux parlés sont concentrées dans l'intervalle de 10 à 25 minutes. Le décodage automatique de ces enregistrements comme une seule unité est quasiment impossible. À cet effet, la segmentation de ces ressources vers des segments plus petits est indispensable. Cependant, il faut que cette segmentation nous offre des segments valides. Des segments qui sont porteurs et utilisables. Entre autres, nous cherchons des segments qui reflètent le concept de discours dans les documents textes. Donc, nous cherchons des segments qui contiennent des phrases syntaxiquement valides, ainsi qu'ils gardent leurs contenus sémantiques.

L'objectif de ce scénario est de segmenter ces ressources à base de silence. Sachant que le silence est le paramètre fondamental pour la séparation des phrases parlé dans un discours. Il remplace le point comme signe de ponctuation dans les documents textes. Cependant, il présente aussi la virgule, les hésitations, les actes de timidités ... etc. Pour cela nous essayons avec des séries d'expérimentations sur le corpus TED-LIUM de trouver l'intervalle des durées optimales des silences qui permet la discrimination de la fin de phrase contre les autres événements des silences.

Pour cette fin, nous avons utilisé des scripts batch avec le module de traitement de signal SoX « *Sound eXchange* » dans sa version « *14.4.2* » sous le système Linux « Ubuntu 16.04 ». La figure 4-3 présente un exemple da la commande utilisée pour segmenter un seul fichier « wav ».

```
issam@issam-HP-Pavilion-15-Notebook-PC:~/Project/Test_Scenario2$ sox -V3 AimeeMullins_2009P.wav
AimeeMullins_2009P_Seg.wav silence -l 1 0.5 -40d 1 4.0 -40d : newfile : restart
sox:      SoX v14.4.2
sox INFO formats: detected file format type `wav'

Output File    : 'AimeeMullins_2009P_Seg093.wav'
Channels       : 1
Sample Rate    : 16000
Precision      : 16-bit
Sample Encoding: 16-bit Signed Integer PCM
Endian Type    : little
Reverse Nibbles: no
Reverse Bits   : no
Comment        : 'Processed by SoX'
```

Figure 4-3 : Exemple du Commande de segmentation SoX

En effet pour le traitement de la totalité des ressources parlées du corpus utilisé. Nous avons utilisé le langage python dans sa version 2.7 pour générer le script batch « .sh » pour la segmentation automatique de la totalité des ressources parlées du corpus utilisé.

Avec ces scripts, nous avons effectué plusieurs segmentations de ces ressources sur différentes valeurs de durée de silence dans un intervalle de [1.0 : 5.0]. L'objectif de cette étape est d'avoir une idée sur le nombre de segments obtenus par rapport à la durée de silence utilisée.

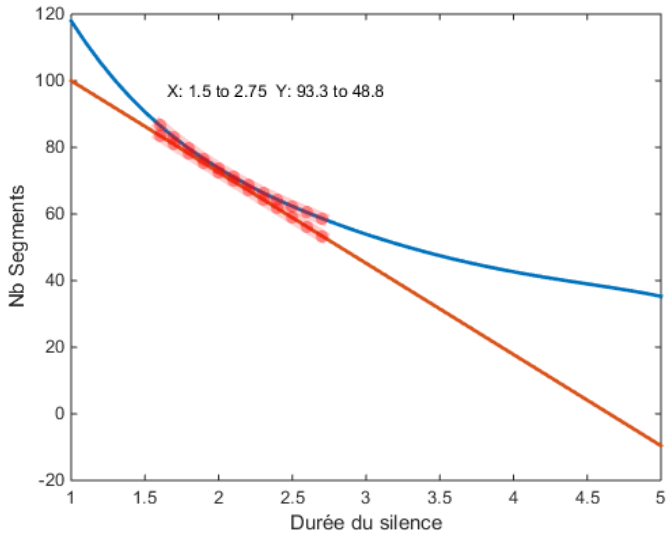

Figure 4-4 : Evolution du nombre du segments par rapport aux durées des silences

Nous trouvons que l'intervalle défini par la tangente de cette courbe représente les valeurs de silences comprises entre 1.5 et 2.75 secondes. Entre temps, nous cherchons aussi d'avoir l'impact de pas de variation de la durée du silence sur le nombre de segments obtenus. Dans la Figure 4-4, nous présentons la courbe de l'impact de cette variation de la durée par rapport au nombre moyen des segments obtenus dans le corpus TED-LIUM V1.

Entre autres, nous cherchons les durées du silence adéquates qui segmentent les documents parlés vers des segments porteurs d'informations. Nous cherchons des séquences qui assurent le compromis entre la durée et le contenu. Des segments qui contiennent des informations syntaxiquement interprétables, mais aussi décodables par les APIs de reconnaissances. La figure 4-5 présente l'impact de la variation du pas du silence sur la segmentation.

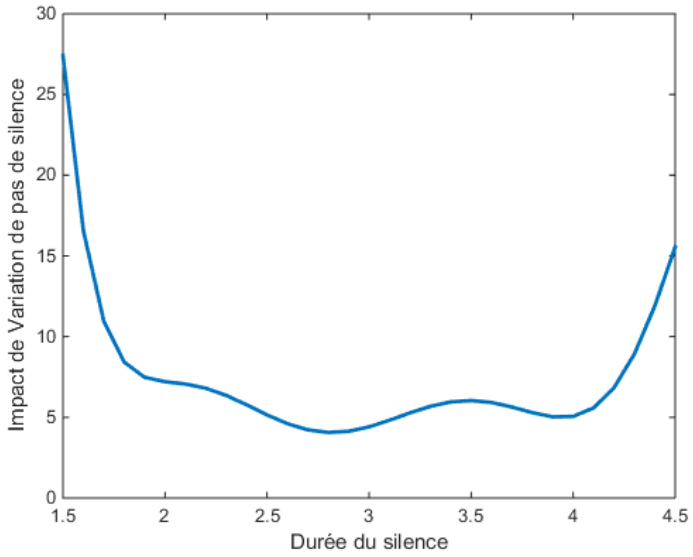

Figure 4-5 : Impact de la variation du pas du silence sur le nombre des segments

Nous constatons dans ce graphe qu'il y a une stabilité du nombre de segments obtenus dans l'intervalle de silence entre [1.75 : 4.0] secondes. Cette stabilité indique que ces durées de silences génèrent des segments « *utterances* » reflètent le concept phrase dans la structure de contenu parlé. Donc, nous pouvons constater que l'intervalle de silence [1.5,3.50] permet une bonne combinaison de point de vue variation et nombre de segments obtenus. À cet effet, que nous exécutons le deuxième scénario qui permet l'alignement de flux parlés avec leurs fichiers « *.stm* ». Cet alignement permet d'évaluer la qualité de segmentation par rapport au nombre réel des « *utterance* » des flux parlés.

4.3.1.2. Scenario 2

Dans ce deuxième scénario, nous essayons d'exploiter les annotations réalisées par l'équipe de recherche lors de l'élaboration du corpus TED-LIUM V1. L'annotation de ce corpus à était réalisé sous le format « *NIST .stm* ». Ce format est utilisé souvent pour la synchronisation du texte avec la parole dans les documents multimédias. Il est composé de plusieurs balises qui contiennent des informations concernant : le nom physique du document parlé, le nom de

locuteur, son genre, la transcription du segment avec son intervalle de temps « temps de début et le temps de fin » comme la montre la figure 4-6.

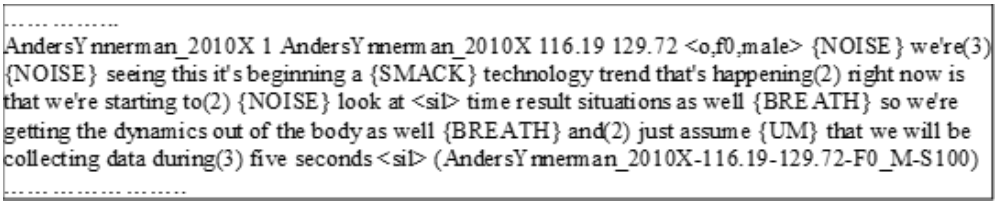

Figure 4-6 : Extrait d'une transcription sous le format *.stm*

Cependant, la création de ces fichiers d'annotation est réalisée selon des segmentations et divisions du contenu du document parlé à base de phrases. Entre autres, en respectant les structures de texte de point de vue des phrases et paragraphes. À cet effet, notre objectif dans ce scénario est de valider notre stratégie de segmentation proposée dans le premier scénario, car le modèle *« .stm »* nous donne une idée sur l'organisation de la structure du contenu du document parlé.

En pratique, le nombre de segments dépend étroitement par la taille du document parlé. Néanmoins, les longueurs des énoncés ou les phrases parlées influent sur le nombre de segments. Dans la figure 4.-6 nous présentons une courbe qui décrit le nombre de segments dans les fichiers « *.stm* » pour les documents parlés en fonction de leurs durées.

Dans ce contexte, nous avons utilisé la librairie *Python/C* « *aeneas* » dans sa version 1.7.3.0 sous Ubuntu. La fonction principale de cette librairie est d'automatiser le calcul d'un fichier de carte de synchronisation « *Synchronisation Map* » entre la ressource parlée et son fichier de transcription. En termes abstraits, une carte de synchronisation associe chaque fragment de texte à l'intervalle de temps, dans le flux parlé, lorsque ce fragment de texte est prononcé.

Le principal avantage de l'*aeneas* est d'éliminer le besoin de travail humain pour produire les timings, tout en produisant une sortie « correcte », c'est-à-dire des cartes de synchronisation indiscernables de celles qu'un opérateur humain produirait manuellement.

À cet effet, nous avons généré les cartes de synchronisation du corpus TED-LIUM sous le format « json ». La génération de ces cartes nous permet d'avoir une segmentation sur le contenu valide et complet. Ces segments sont générés selon leurs contenus lexicaux et syntaxiques. Donc, nous obtenons un modèle de structuration textuelle des documents parlés. La figure 4-7 nous présente le nombre de segments obtenus dans les cartes de synchronisations générées par « *aeneas* » du notre corpus de test.

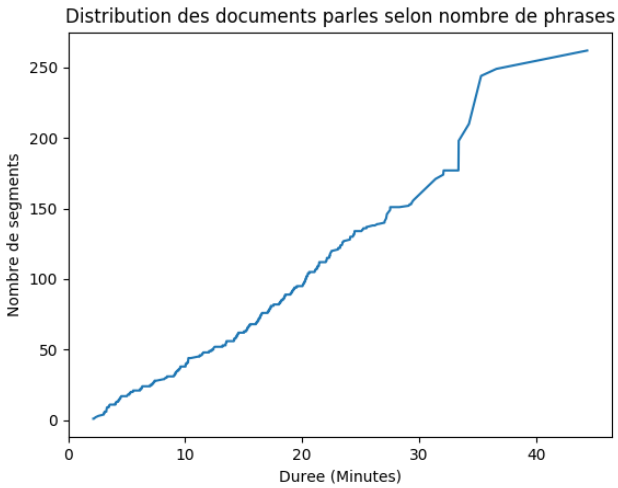

Figure 4-7 : Courbe du nombre de segments obtenus dans les carte de synchronisation du corpus TED-LIUM

L'objectif de cette phase est de connaitre la segmentation la plus proche de point de vue de contexte et structure textuelle. Car dans le premier scénario, la segmentation a été réalisée selon les durées de silence. Et comme nous avons dit précédemment que le silence présente un signe de ponctuation, mais il peut présenter aussi des signes des hésitations, repos, traques …etc.

Pour cela, cette segmentation à base de carte de synchronisation nous permet de comparer les résultats obtenus par le premier scénario avec des outils et des techniques d'analyse statistique comme les facteurs de régressions et de corrélations.

Dans le domaine de statistiques, la corrélation est une mesure statistique qui caractérise l'existence ou l'absence d'une relation entre deux échantillons de valeurs prise sur un même groupe de sujets. Le coefficient de corrélation permet de quantifier cette relation par le signe de la corrélation : positive et négative, et par la force de cette corrélation. Le degré de corrélation se mesure sur une échelle de 0 à 1. Zéro signifie une totale absence de corrélation entre les deux mesures, alors que 1 signifie une corrélation parfaite.

Dans ce contexte, nous avons calculé les coefficients de corrélations pour les différentes valeurs de silences utilisés dans le premier scénario pour connaitre la durée de silence optimale à utiliser dans le processus de segmentation automatique par rapport aux nombres de segments obtenus par la carte de synchronisation. Le tableau 4-5 décrit les valeurs obtenues après la normalisation des valeurs par la norme Z-score[1].

[1] Z-score c'est une norme de normalisation dont les valeurs d'un attribut A sont normalisés en fonction de la moyenne et l'écart-type de A.

Durée du silence	Facteur de corrélation
4.00	0.7031895
3.50	0.7513758
3.00	0.7650665
2.50	0.7804677
2.00	0.8064811
1.75	0.7925667
1.50	0.7700268
1.25	0.6866598
1.00	0.5973781

Tableau 4-5 : le facteur de corrélation entre les durées du silence et le nombre de segments.

Dans ce tableau, nous trouvons que la segmentation des ressources audio à base d'une valeur de silence dans l'intervalle [1.50s,3.5s] est la meilleure. Cependant, cette segmentation a été réalisée sans considération de contenu du segment.

4.3.1.3. Synthèse et discussions

Dans cette partie, nous avons essayé de trouver une technique automatique pour segmenter les documents parlés vers des segments qui représentent au mieux les phrases parlées. Le seul outil de segmenter le contenu de ces ressources parlées est le silence. Dans ce contexte, nous avons utilisé l'environnement « *SoX* » sous *Ubuntu* pour exécuter une série de tests sur différentes valeurs de la durée de silence pour trouver l'intervalle des durées optimales qui nous permette de générer des segments le plus présentatif du contenu. Pour la validation de cette stratégie, nous avons construits les cartes de synchronisations de ces ressources parlées avec le *Python* « *aeneas* » on utilisons les fichiers « *.stm* » correspondants.

Entre autres, la durée de silence n'influe pas seulement sur le nombre de segments obtenus, mais aussi sur le contenu des segments. Dans SOMI, cette phase de segmentation automatique a pour objectif de fournir des segments traitables pour les APIs de reconnaissance. Donc, la complexité et même les résultats de transcriptions obtenus dépendent étroitement par ce paramètre. Dans ce contexte, nous étudions dans la section suivante l'impact de cette segmentation sur la qualité de la transcription automatique du contenu obtenu.

4.3.2. Stratégies de reconnaissances

Dans cette partie, nous utilisons des moteurs de reconnaissance automatique pour la transcription des segments des ressources parlés vers des textes. En pratique, nous trouvons deux catégories de ces systèmes. La première repose sur la création des modèles essentiels du décodeur tels que : le modèle acoustique, le dictionnaire phonétique et le modèle de langage. Puis nous exécutons les routines de décodage. Cette méthode elle s'appelle souvent « décodage offline ». La deuxième variante, elle utilise les interfaces programmables « APIs ». Elle repose sur la puissance des calculs parallèles disponibles dans les systèmes de cloud, ainsi que l'utilisation des modèles de langage et les techniques de décodage standardisés. Cette technique est appelée souvent « décodage online ».

Dans ce contexte, nous avons réalisé une série d'expérimentations en variant les durées des segments audios pour mesurer la qualité des résultats des transcriptions par rapport aux transcriptions définies dans le corpus utilisé. En effet, la segmentation selon les durées du silence nous permet de gérer le nombre de mots moyen par segment. Plus le nombre de mots par segments est petit ; plus que les erreurs inter segments sont petites, mais la tolérance des erreurs intra segments s'augmente.

Pour cela, notre stratégie dans cette phase pour pallier ce problème est :

- Analyser les erreurs liées au contenu par l'analyse des exceptions par segment qui sont générées par les moteurs de reconnaissances.
- Analyser les erreurs intra segments qui présentent l'impact de segmentation sur la qualité de transcription globale générée par la concaténation des transcriptions locales.
- Analyser l'aspect temps d'exécution qui présente un facteur primordial pour l'évaluation de notre approche

À cet effet, nous exécutons deux scénarios. Le premier « online » avec *Google Cloud Speech API* et le deuxième « offline » on utilisons *PocktSphynx avec* ces modèles linguistiques de la langue anglaise.

4.3.2.1. Scenario 1

Dans ce scénario, nous utilisons la librairie Python « *SpeechRecognition* » dans sa version 3.8.1 pour effectuer la reconnaissance vocale avec le moteur de décodage online « *Google Cloud Speech API* » avec « Google API Client Library » sous Python 2.7. Cette API utilise les algorithmes de « *Deep learning* » les plus sophistiqués du marché. Cette technologie basée sur les réseaux de neurones permet une reconnaissance vocale avec une précision inégalée pour les langues bien dotées. Cependant, le degré complexité et la qualité du décodage de cette technique dépendent fortement de la durée des segments parlés. Cette API permette la configuration des modèles et les techniques de décodage utilisés comme : le langage cible, mode de décodage : *1-Best* ou *N-Best*, treillis de décodage : mots ou phrases, Type de fichier et la fréquence d'échantillonnage … etc. La figure 4-8 présente un exemple de configuration de cette interface.

```
"config": {
      "languageCode": "en-US",
      "encoding": "wav",
      "enableWordTimeOffsets": True
      "maxAlternatives": 1
      "sampleRateHertz": 16000,
      "profanityFilter": False,
      "speechContexts": [ "phrases"]:
  }
```

Figure 4-8 : Configuration du *Google API Client Library*

Dans la phase d'expérimentation, nous avons utilisé des scripts Python pour la transcription de segments générés pour chaque ressource parlée. Ainsi, la transcription globale de la ressource est le cumul des transcriptions partielles des segments. La figure 4-9 présente un extrait d'utilisation du *Google Cloud Speech API* sous python.

```
for files in h :
        f= os.path.basename(files)
        print os.path.basename(files)
        AUDIO_FILE = files
        r = sr.Recognizer()
        with sr.AudioFile(AUDIO_FILE) as source:
                audio = r.record(source)
        try:
                T_Google = r.recognize_google(audio)
                Text = Text + " \n" + T_Google
                NbValide = NbValide + 1
        except sr.UnknownValueError:
                print("Google Speech Recognition could not understand audio")
                NbErreur = NbErreur + 1
        except sr.RequestError as e:
                print("Could not request results from Google Speech Recognition service; {0}".format(e))
                NbErreur = NbErreur + 1
        i = i + 1
```

Figure 4-9 : Un extrait d'utilisation du *Google Cloud Speech API* sous python

Entre-temps, le décodage avec *Google Cloud Speech API* sous python génère des exceptions dans les cas ou le contenu du segment n'est pas interprétable ou en cas de problème de disponibilité des ressources. Pour cela, nous intéressons à connaitre le nombre de segments omis lors du décodage des segments d'une ressource parlée donnée. À cet effet, nous définissons le taux d'omissions qu'on a appelé « R_{OS} ». Ce taux représente la pondération du nombre de segments omis lors da la phase de décodage.

Pour évaluer la valeur de « R_{OS} » lors de la phase de reconnaissance. Nous avons exécuté une série d'expérimentations sur plusieurs segmentations pour chaque document parlé en variant la durée du silence du : 1.0s jusqu'au 4.0s. La figure 4-10 présente la courbe d'évolution des valeurs de R_{OS} des flux parlées du corpus utilisé.

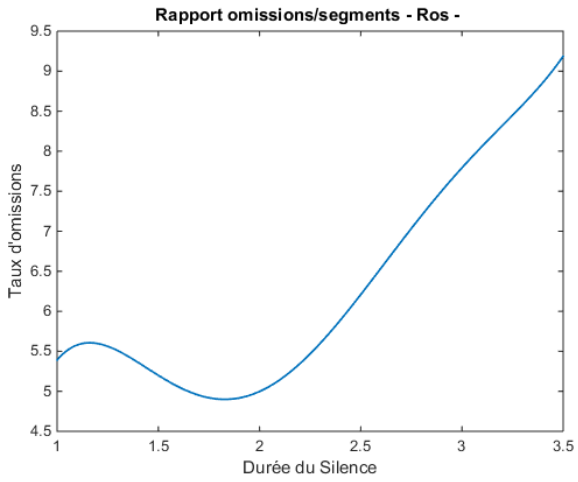

Figure 4-10 : l'évolution du R_{OS} par rapport à la durée du silence

D'après les résultats obtenus, nous constatons que la segmentation à base des valeurs de silences comprises dans l'intervalle [1.50 – 2.50] est la meilleure par rapport aux taux d'omissions. Nous constatons que l'impact de la segmentation à base de silence pour des durées comprises dans l'intervalle [1.50 :2.50] ne génère pas des segments invalides de point de vue contenu. Entre autres, la figure 4-11 présente le temps d'exécution écoulé du processus des transcriptions des segments pour les différentes valeurs de silences définies auparavant.

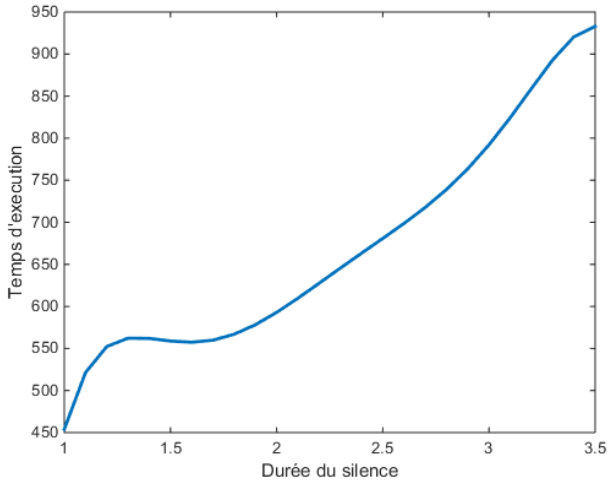

Figure 4-11 : Evolution des temps d'exécutions par rapport aux valeurs de silence

Nous constatons dans ce graphe que le temps d'exécutions consommé pour le décodage dépend fortement du nombre de segments invalides. Car dans le cas d'une situation avec un contenu invalide, l'algorithme décodage dans le graphe *Wfst* devient heuristique qui consomme beaucoup de temps et ne trouve pas généralement la décision.

Pour cela, nous supposons que la segmentation à base d'une valeur de silence comprise dans l'intervalle [1.50 :2.5] est la meilleure. Entre temps, nous analysons le contenu des

transcriptions pour évaluer les erreurs intra-segments. Ces erreurs valorisent l'impact de processus segmentation sur la qualité de transcription globale générée par la concaténation des transcriptions partielles.

Pour connaitre la qualité obtenue lors de la reconnaissance des segments générés, nous utilisons les méthodes et les techniques relatives au domaine de recherche d'informations défini dans le chapitre deux de cette thèse. À cet effet, nous utilisons la « *Similarité Cosinus* ». En effet, la similarité cosinus entre deux documents dans l'espace vectoriel est calculée par le cosinus de l'angle entre eux. Elle présente une comparaison entre des documents sur un espace normalisé par les modèles « *Tf-Idf* » de chaque document par l'angle entre les documents.

En pratique, nous avons utilisé « *scikit-learn Machine Learning* » in Python dans sa version 0.19.1 avec la librairie « *nltk - Natural Language Toolkit* » dans sa version 3.3 et les outils de détection de limite de phrase « punkt ». Ces outils permettent les tâches de traitement de texte comme : *tokenisation*, *stemming*, *lemmatization*, identification de nom propre et l'extraction des paramètres des documents.

En revanche, nous avons opté pour la division des ressources parlées du corpus utilisées dans la phase d'expérimentation en tranches selon leurs durées. Cette répartition sert à estimer la capacité de notre stratégie utilisée pour la recherche dans le contenu parlé par rapport à la durée des ressources : courte, moyenne, longue. Ensuite, nous avons calculé les similarités cosinus pour tous les résultats de transcriptions des segments obtenus pour les différentes durées de silences par rapport à la transcription cible « .stm » pour chaque ressource parlée. Premièrement, nous avons exécuté une phase de prétraitement par la lemmatisation des termes et l'élimination des mots vides « *Stop Word* ». Le tableau 4-6 présente les résultats de similarités cosinus obtenus pour les résultats du décodage par « *Google Cloud Speech API* » des flux parlés segmentés sur différentes durées de silences.

	Durée du Silence								
Tranches	**1.50**	**1.75**	**2.0**	**2.25**	**2.50**	**2.75**	**3.0**	**3.50**	**4.00**
04m – 08m	44.40%	45.10%	46.10%	44.70%	44.40%	44.50%	42.90%	42.60%	36.00%
08m – 12m	51.50%	53.60%	55.90%	56.30%	56.60%	56.70%	57.10%	55.90%	55.70%
12m – 16m	53.80%	54.00%	53.90%	53.70%	53.50%	52.50%	53.60%	51.80%	53.10%
16m – 20m	51.20%	55.00%	58.30%	63.00%	64.50%	65.60%	67.40%	65.20%	65.03%
Plus de 20m	49.00%	50.66%	52.24%	53.10%	53.41%	53.49%	53.90%	52.56%	51.18%

Tableau 4-6 : Similarités cosinus avec prétraitements des résultats de transcription automatique par Google Cloud Speech API.

Nous remarquons que les taux de similarités obtenus sont faibles parce que l'étape de lemmatisation modifie la forme syntaxique des termes reconnus. Ainsi, l'élimination de mots vides supprime un nombre important des termes. En effet, l'objectif de cette étape est la comparaison textuelle directe entre un flux parlé transcrit avec son fichier « *.stm* ». Cependant, pour connaitre l'impact de processus de lemmatisation des termes sur la qualité des transcriptions obtenues par l'API « Google Cloud Speech ». Nous calculons les similarités cosinus de ces transcriptions obtenues avec l'élimination des mots vides sans l'utilisation du processus de lemmatisation. Le tableau 4-7 présente les résultats obtenus.

	Durée du Silence								
Tranches	**1.50**	**1.75**	**2.0**	**2.25**	**2.50**	**2.75**	**3.0**	**3.50**	**4.00**
04m – 08m	46.70%	46.90%	47.00%	48.88%	49.00%	48.90%	46.00%	45.50%	44.60%
08m – 12m	46.20%	48.80%	49.60%	51.55%	51.80%	52.40%	53.10%	53.70%	54.50%
12m – 16m	50.60%	50.80%	50.90%	50.50%	50.60%	50.80%	50.90%	49.40%	49.70%
16m – 20m	44.60%	48.90%	52.40%	58.95%	59.10%	60.30%	62.30%	61.80%	61,20%
Plus de 20m	58.20%	58.70%	58.10%	59.15%	59.30%	58.10%	58.50%	57.70%	56.20%

Tableau 4-7 : Similarités cosinus sans lemmatisation des résultats de transcription automatique par Google Cloud Speech API.

Nous remarquons que les valeurs de similarités s'améliorent, car l'élimination de l'étape de lemmatisation permet de garder les termes (mots) dans leurs états bruts. Enfin, pour connaître l'impact de l'élimination des mots vides « Stop Words ». Nous ignorons aussi cette étape de prétraitements, le tableau 4-8 présente les résultats des similarités cosinus qui traitent l'intégralité des termes constituant les flux parlés transcrits.

	Durée du Silence								
Tranches	**1.50**	**1.75**	**2.0**	**2.25**	**2.50**	**2.75**	**3.0**	**3.50**	**4.00**
04m – 08m	81.10%	81.30%	81.50%	81.40%	80.70%	80.60%	80.00%	78.60%	73.70%
08m – 12m	88.10%	89.00%	89.00%	89.20%	88.80%	89.80%	89.70%	88.10%	87.30%
12m – 16m	91.40%	91.60%	91.80%	91.70%	91.70%	91.60%	91.60%	90.60%	90.40%
16m – 20m	78.20%	82.00%	87.30%	88.00%	90.50%	92.60%	94.40%	93.05%	92.90%
plus de 20m	96.50%	95.70%	96.90%	96.70%	96.70%	96.80%	96.70%	95.60%	95.10%

Tableau 4-8 : Similarités cosinus sans lemmatisation et sans StopWord des résultats de transcription automatique par Google Cloud Speech API.

L'analyse des différents résultats obtenus, montre que la qualité de la transcription automatique de la parole obtenue par Google Speech API est excellente. Cependant, le problème de reconnaissance de quelques segments accroître d'une façon significative le temps d'exécution. Entretemps, de point de vue de la stratégie de segmentation à base de silence. Nous constatons que les durées de silence dans l'intervalle [2.25 , 3.0] fournissent des bons résultats de transcription avec « Google Cloud Speech ». Le tableau 4-9 présente un récapitulatif des différents résultats de similarités cosinus obtenus.

Similarités Cosinus	Durée du Silence								
	1.50	**1.75**	**2.0**	**2.25**	**2.50**	**2.75**	**3.0**	**3.50**	**4.00**
Lemmatization & Stop Word	49.98%	51.67%	53.29%	54.16%	54.48%	54.56%	54.98%	53.61%	52.20%
Stop Word	53.26%	54.82%	55.60%	57.81%	57.96%	58.10%	58.16%	57.62%	56.25%
Sans Lemmatization & Stop Word	87.06%	88.12%	88.45%	89.40%	89.68%	90.28%	90.48%	89.19%	87.88%

Tableau 4-9 : Récapitulatif sur les différents résultats de similarités cosinus obtenus par Google Cloud Speech API.

4.3.2.2. Scenario 2

Dans ce deuxième scénario, nous utilisons le décodeur offline *Pocket Sphinx* de *CMU Sphinx*. Le modèle acoustique utilisé dans ce décodeur est basé sur les *GMM* et les *HMM*. Il contient divers outils et il est très sollicité pour la création des applications offline avec des tâches de reconnaissance vocale. *Pocket Sphinx* dans son implémentation sous Python « *SpeechRecognition 3.7.1* » contient un certain nombre de paquets pour différentes tâches et applications. C'est une bibliothèque écrite en *C*, puis en *Python* pour le développement des applications. L'utilisation de *Pocketsphinx* nécessite la configuration des modèles : acoustique, langage et le dictionnaire phonétique lors de la phase d'initialisation. Le *CMU Sphinx* mit à la disposition pour ces utilisateurs un ensemble de modèles de quelques langages en accès libre. La figure 4-12 présente les modèles utilisés lors de décodage des flux parlés en langue anglaise.

```
config = Decoder.default_config()
config.set_string('-hmm', path.join(MODELDIR, 'en-us/en-us'))
config.set_string('-lm', path.join(MODELDIR, 'en-us/en-us.lm.bin'))
config.set_string('-dict', path.join(MODELDIR, 'en-us/cmudict-en-us.dict'))
decoder = Decoder(config)
```

Figure 4-12 : Les modèles utilisés pour le de décodage des documents parlés en Anglais par Pocket Sphynx

Contrairement au premier scénario, le décodage offline génère rarement des omissions lors de décodage des segments parlés. Parce que les résultats de décodage avec l'algorithme *Viterbi* sont limités par les capacités du modèle acoustique utilisé. Elle ne cherche pas d'autres modèles acoustiques qui peuvent décoder un segment inextricable comme le cas de « *Google Cloud Speech API* » sur le Cloud. Dans ce contexte, nous présentons dans la figure 4-13 le temps d'exécutions écoulé par les deux scénarios pour le décodage des flux parlés segmentés sur différentes durées de silences.

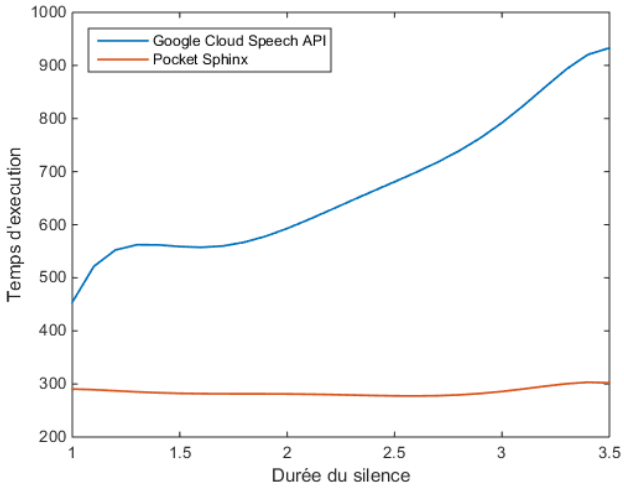

Figure 4-13 : Temps d'exécution écoulé par les deux scénarios de décodage

Nous remarquons dans la courbe obtenue que l'impact de la valeur de R_{OS} sur le temps d'exécution par le décodeur « *Google Cloud Speech API* » est colossal. Cependant, nous trouvons que le temps écoulé lors de décodage avec des ressources locales par « *Pocket Sphinx* » est généralement stable. En revanche, il faut qu'on évalue la qualité des transcriptions obtenues par les ressources locales par rapport au décodage online.

Dans ce contexte, nous gardons les mêmes métriques utilisées dans le premier scénario. Nous procédons aux calculs des similarités cosinus pour tous les résultats de transcriptions des segments obtenus pour les différentes durées de silences par rapport à la transcription cible pour chaque document parlé. Ainsi, nous utilisons la même stratégie utilisée dans le premier scénario pour la phase de prétraitement du contenu des ressources parlées. En premier lieu, nous recourons aux étapes de lemmatisation des termes et l'élimination des mots vide « *Stop Word* ». Le tableau 4-10 présente les similarités cosinus obtenues par rapport aux transcriptions effectuées avec « *Pocket Sphinx* » pour des ressources parlées du corpus « TED-LIUM-V1 » avec un processus de segmentation sur des durées de silences différentes.

	Durée du Silence								
Tranches	**1.50**	**1.75**	**2.0**	**2.25**	**2.50**	**2.75**	**3.0**	**3.50**	**4.00**
04m – 08m	23.65%	25.44%	26.92%	28.72%	29.29%	30.24%	31.07%	30.94%	30.03%
08m – 12m	27.41%	29.56%	31.39%	34.15%	34.78%	36.39%	37.90%	36.30%	35.54%
12m – 16m	33.34%	37.10%	40.14%	43.48%	44.45%	46.14%	47.56%	46.21%	45.64%
16m – 20m	33.34%	37.10%	40.14%	43.48%	44.45%	46.14%	47.56%	46.31%	45.64%
plus de 20m	34.80%	37.90%	40.07%	42.92%	43.44%	44.79%	45.95%	44.41%	43.72%

Tableau 4-10 : Similarités cosinus avec prétraitements des résultats de transcription automatique par *Pocketsphinx*.

Nous remarquons que les similarités entre la ressource parlée cible et ses transcriptions obtenues par sa segmentation sur de différentes durées de silences sont faibles et cela dut à la philosophie de la métrique utilisée. Pour cela, l'étape de prétraitements qui permette la lemmatisation des termes des documents deviennent inutile, car nous visons dans ce stade une comparaison de deux documents textes mots à mots. Le tableau 4-11 présente les similarités cosinus calculées sans passer par l'étape de lemmatisation des termes des documents.

	Durée du Silence								
Tranches	**1.50**	**1.75**	**2.0**	**2.25**	**2.50**	**2.75**	**3.0**	**3.50**	**4.00**
04m – 08m	27.00%	28.99%	30.58%	32.02%	33.19%	34.16%	34.98%	33.02%	32.50%
08m – 12m	33.74%	35.81%	37.56%	39.22%	40.74%	42.22%	43.57%	42.39%	41.89%
12m – 16m	37.29%	41.47%	44.70%	47.18%	49.26%	51.04%	52.57%	51.48%	50.76%
16m – 20m	38.25%	41.76%	44.76%	47.00%	49.18%	51.18%	52.89%	51.05%	50.64%
plus de 20m	38.97%	41.93%	44.06%	45.87%	47.30%	48.53%	49.58%	47.27%	45.21%

Tableau 4-11 : Similarités cosinus sans lemmatisation des résultats de transcription automatique par *Pocketsphinx*.

Nous remarquons que les valeurs de similarités s'améliorent légèrement, car l'élimination de l'étape de lemmatisation permet de garder les termes (mots) dans leurs états bruts. Cependant, nous avons aussi éliminé l'étape d'élimination des mots de langage appelé « Stop Word » parce qu'ils sont aussi appartenus au contenu brut des textes transcrits. Le tableau 4-12 présente les résultats des similarités cosinus qui traitent l'intégralité des termes constituant les ressources transcrites.

	Durée du Silence								
Tranches	**1.50**	**1.75**	**2.0**	**2.25**	**2.50**	**2.75**	**3.0**	**3.50**	**4.00**
04m – 08m	76.40%	77.20%	77.94%	78.45%	78.84%	79.17%	79.42%	78.15%	78.53%
08m – 12m	83.65%	84.47%	85.03%	85.47%	85.83%	86.17%	86.44%	85.33%	85.43%
12m – 16m	86.12%	86.91%	87.76%	88.30%	88.76%	89.13%	89.43%	88.10%	88.02%
16m – 20m	88.30%	88.79%	89.33%	89.72%	90.08%	90.38%	90.62%	89.83%	88.64%
plus de 20m	88.30%	88.79%	89.33%	89.72%	90.08%	90.38%	90.62%	88.33%	88.02%

Tableau 4-12 : Similarités cosinus sans lemmatisation et sans StopWord des résultats de transcription automatique par *Pocketsphinx*.

Les résultats des similarités obtenus sans le recours aux étapes de prétraitements s'améliorent considérablement. En effet, même les mots vides sont des termes constituants le contenu d'un flux parlé. Le tableau 4-13 présente un récapitulatif sur les différents résultats de similarités cosinus obtenus.

Similarités Cosinus	Durée du Silence								
	1.50	**1.75**	**2.0**	**2.25**	**2.50**	**2.75**	**3.0**	**3.50**	**4.00**
Lemmatization & Stop Word	30.51%	33.42%	35.73%	38.55%	39.28%	40.74%	42.01%	40.83%	40.12%
Stop Word	35.05%	37.99%	40.33%	42.26%	43.93%	45.43%	46.72%	45.04%	44.20%
Sans Lemmatization & Stop Word	84.56%	85.23%	85.88%	86.33%	86.72%	87.05%	87.31%	85.95%	85.73%

Tableau 4-13 : Récapitulatif sur les différents résultats de similarités cosinus obtenus par *Pocketsphinx* du corpus Test/TED-LIUM

L'analyse des différents résultats obtenus nous permet de conclure que la qualité de la transcription automatique de la parole obtenue par *« Pocketsphinx* » est inférieure à celle obtenue par « *Google Cloud Speech API* ». Cependant, ce décodage offline ne consomme pas beaucoup de temps de calcul. De point de vue de stratégie de la segmentation à base de silence, nous trouvons que la durée de silence dans l'intervalle [2.25 – 3.00] fournissent les meilleurs résultats de la transcription du contenu parlé.

4.3.2.3. Synthèse et discussions

Dans cette partie, nous avons étudié l'aspect de reconnaissance du contenu des flux parlés par les modules de reconnaissance « *Pocket Sphinx »* et « *Google Cloud Speech API* » sous *Python*. Le premier module utilise en mode local le modèle acoustique *GMM-HMM*, le modèle de langage *3-gram* et un dictionnaire phonétique de 134k termes en langue anglaise. Tandis que le deuxième module utilise ses propres modèles en cloud. Nous avons étudié les performances de ces APIs de reconnaissance sur deux aspects : temps de calcul et qualité de transcription. L'utilisation de « *Google Cloud Speech API* » fournit de bons résultats par rapport à celle du « *Pocket Sphinx »*, mais elle est sensible à la qualité de segments traités. Nos trouvons que le traitement d'un segment invalide dans le cloud consomme beaucoup de temps, car l'API cherche les meilleurs modèles afin de le décrypter. Cependant, le module « *Pocket Sphinx »* fournit la séquence la plus probable par rapport à ses modèles. En revanche, nous avons utilisé la mesure « cosinus » pour évaluer la qualité des transcriptions obtenues. Nous avons défini trois variantes : une similarité brute, similarité avec un prétraitement de lemmatisation et similarité avec un prétraitement de lemmatisation et élimination de mot vides « Stop Word ». En effet, la similarité brute présente un score global de qualité sur la totalité du flux parlé transcrit, ainsi le prétraitement « lemmatisation et mots vides » évalue la capacité de ces systèmes par rapport aux termes : spécifique, techniques et hors vocabulaire. En effet, ces systèmes n'ont aucun problème de reconnaissance pour les termes usuels. Mais leurs défis résident sur leurs capacités de détecter les autres termes qui ne sont pas souvent utilisés. Ce problème est jugé par la nature du modèle acoustique utilisé. Ces modèles sont généralement probabilistes avec les GMM-HMM et leurs performances dépendent du volume de ressource d'apprentissage utilisée ainsi que ses fréquences d'apparitions des termes.

4.3.3. Conclusion

Le module SLA-SOMI est conçu pour trouver une représentation textuelle la plus proche du contenu des ressources parlées. Nous avons utilisé une technique de segmentation acoustique avec les modules de reconnaissances « *Google Cloud Speech API* » et « *Pocket Sphinx* » avec des ressources partagées et locales respectivement. Nous constatons d'après les résultats obtenus lors de la phase d'expérimentation qu'on a une bonne qualité de transcription générale de contenu de flux parlé. Cependant, nous constatons une difficulté pour la reconnaissance des termes techniques et spécialisés. Entre autres, nous constatons aussi que le temps écoulé est très important pour la solution « *Google Cloud Speech API* », ce qui nous amène à ignorer la reconnaissance des segments qui dépassent un seuil de temps d'exécution.
Dans ce contexte, nous envisageons de chercher des stratégies pour améliorer les résultats de ces transcriptions obtenues par une analyse syntaxique et sémantique profonde de ces résultats. Dans la section suivante, nous présentons les différentes stratégies et techniques utilisées dans cet analyseur.

4.4. Validation du module CSA-SOMI

L'objectif de ce module est d'alléger les calculs et traiter les problèmes d'omissions et d'insertions dans la phase de reconnaissances par l'enrichissement des résultats du module SLA-SOMI par des ressources statistiques et sémantiques. En effet, le module CSA-SOMI utilise les résultats des transcriptions des flux parlés avec l'intégration des concepts et relations de l'ontologie *WordNet* comme ressource sémantique pour la détection des indexes discriminants du contenu des ressources parlées. Dans ce contexte, nous introduisons la notion de topics par la définition de leurs représentations vectorielles. En effet, la détection des topics emboitées dans le contenu de flux parlé permet la détection des termes candidats d'indexation valides et discriminants. Aussi, nous avons implémenté quelques mesures de similarité sémantique pour l'enrichissement de cet ensemble de termes d'indexation.

Le but de cette validation est de mesurer la capacité l'indexation sémantique à base des transcriptions partielles du contenu parlé avec un enrichissement sémantique. En effet une convergence vers les résultats de recherches sur le contenu basées sur l'utilisation de l'intégralité du contenu. Cette stratégie s'appuie sur trois axes :

- La construction des modèles de topics pour le contenu des ressources parlés. Ces modèles sont utilisés comme un support de décision pour le choix de termes d'indexations candidats pour le contenu parlé.
- L'évaluation des performances de détection par rapport au contenu partiel des ressources parlées.
- L'enrichissement des termes d'indexations candidats par les mesures de similarités à partir de l'ontologie *WordNet*.

4.4.1. Présentation de l'environnement développé

Pour la validation, nous avons développé l'environnement « *MyWordNet* » avec le langage de développement *Delphi* sous *Windows* analyseur syntaxique lexical. Cet environnement permet l'acquisition et l'adaptation des concepts et relations de l'ontologie WordNet dans sa version 3.1. Il contient un analyseur syntaxique lexical complet et il permet aussi la définition et la détection des topics, le calcul de mesures de similarité et l'enrichissement sémantique. Les différentes fonctionnalités de cet environnement seront détaillées au fur à mesure de leurs utilisations dans le processus de validation de ce module. La figure 4-14 présente l'interface de l'analyseur lexical.

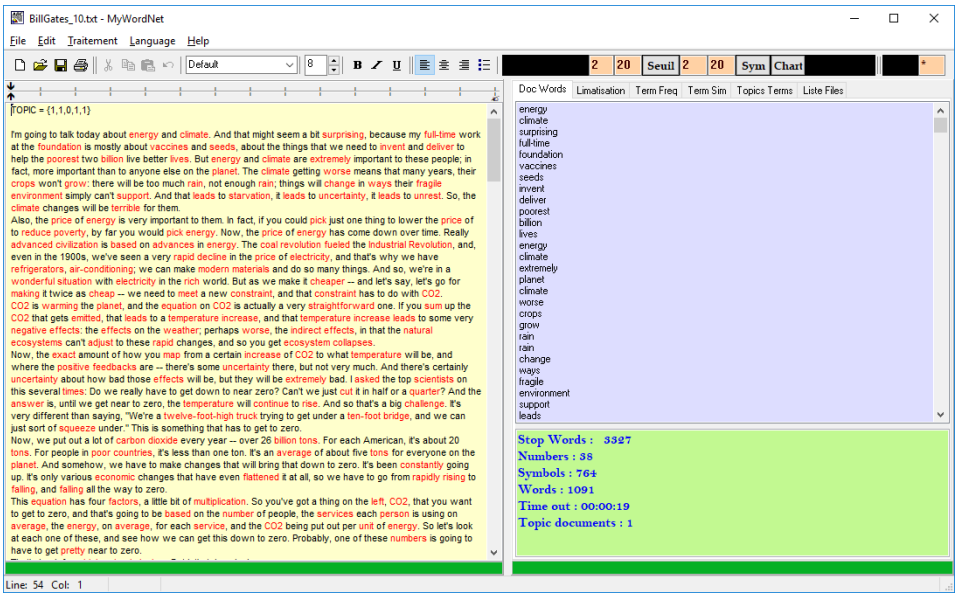

Figure 4-14 : L'analyseur lexical « *Expert Editor* »

4.4.2. Stratégie de détection des Topics

4.4.2.1. Choix des topics

La communauté de la conférence TED nous offre une large bibliothèque numérique dans différents sujets pour de différents domaines qui sont souvent en étroite liaison. Cette diversité et imbrication nous encourage à utiliser ces ressources comme un support de test la détection des topics dans le contenu des ressources parlées. En effet, le contenu parlé de ces ressources parlées fournies par TED contient plusieurs topics. Notre objectif est de détecter les topics existants dans le contenu d'une ressource parlée avec leur degré de pertinence.

À cet effet, nous avons utilisé un extrait de ces ressources pour la construction des modèles des topics. Nous avons choisi d'utiliser les topics les plus populaires par la communauté TED : *Technology, Business, Design, Science et Global issues*. Néanmoins dans nos perspectives futures nous élargissant la liste de ces topics. Dans ce contexte, nous avons construit un corpus d'apprentissage par l'annotation d'un ensemble de ressources parlées par rapport aux topics.

Le tableau 4-14 présente un extrait de la répartition de ressources parlées par rapport aux topics utilisés

	Technology	Business	Design	Science	Global issues
Technology	176	41	62	60	34
Business	41	118	32	20	45
Design	62	32	134	22	12
Science	60	20	22	148	30
Global issues	34	45	12	30	120

Tableau 4-14 : Répartition d'un extrait du ressources parlés par rapport aux topics

Nous remarquons dans cette répartition que les topics de ces ressources sont chevauchés et ils ne sont pas linéairement séparables. Cependant, ce chevauchement et cette imbrication sont indispensables pour la construction d'une représentation valide avec un pouvoir de généralisation utile et utilisable.

4.4.2.2. Définition des vecteurs de représentation des topics

Nous avons défini les caractéristiques des topics par le modèle vectoriel avec la pondération des termes « *tf/idf* ». Ce modèle est basé sur les fréquences d'apparition des termes dans le corpus utilisé pour cette fin. Dans ce contexte, nous avons créé les vecteurs de représentation de chaque topic par l'environnement « *MyWordNet* ». Ensuite, nous insérons les termes et ses fréquences d'apparition au vecteur de représentation d'un topic après un prétraitement lexical des représentations textuelles des ressources parlées. Cette procédure s'effectue par lot pour la totalité de ressources utilisées. La figure 4-15 présente un exemple de mis à jour des vecteurs de topics par une source textuelle.

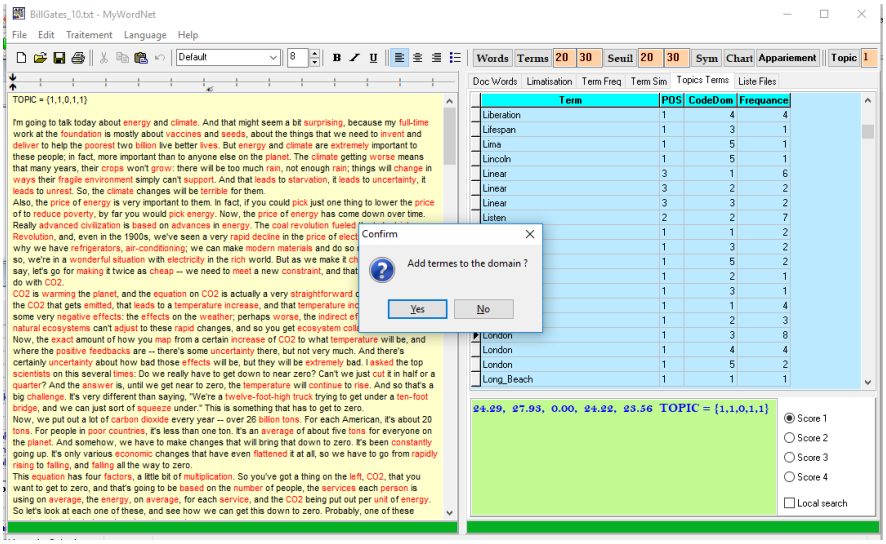

Figure 4-15 : Interface de création des vecteurs de représentation des topics

Ainsi, ces vecteurs de représentation sont soumis à une étape d'épuration. Selon la loi de *Zipf,* nous éliminions les termes rares et nous utilisons la technique de normalisation « Min-Max » par la formule :

$$V' = \frac{V - min_A}{max_A - min_A}(New\_max_A - New\_min_A) + New\_min_A \qquad (4\text{-}1)$$

Avec

- $max_A et\ min_A$: les fréquences maximum et minimum des termes d'un topic T_i.
- $New\_max_A\ et\ New\_min_A$: les bornes de l'intervalle de normalisation avec les valeurs 1 et 0 respectivement.

Les vecteurs de représentation obtenus forment ainsi une matrice de définition des topics de Dic_Topics$[N * M]$ tel que : M représente le nombre de termes utilisés et N nombre de topics. On note aussi que l'environnement permet la mise à jour de cette matrice : soit par insertion des nouveaux termes ; soit la mis à jour des pondérations des termes par topic selon les représentations textuelles des ressources utilisées.

On note aussi, nous avons procédé à l'élimination des termes qui possèdent une forte densité par rapport à tous les vecteurs de définitions de topics. Le tableau 4-15 représente une description statistique des vecteurs de représentation des topics obtenus.

	Nombre de ressources	Nombre de Termes	Total des Fréquences
Technology	176	16685	87250
Business	118	14166	62105
Design	134	11738	68345
Science	148	17612	88473
Global issues	120	13218	65592

Tableau 4-15 : Description statistique des vecteurs de représentation des topics

4.4.2.3. Détection des topics

L'utilisation du modèle vectoriel comme un support de représentation des topics et les transcriptions du contenu des ressources parlées permet l'utilisation des distances vectorielles et le calcul des scores de vraisemblances pour la détection des topics du contenu des ressources parlées. Nous avons défini dans la section 3.5.2.3 du chapitre précédent trois formules de calcul de ces scores. À cet effet, nous avons implémenté ces formules dans l'environnement « *MyWordNet* ». La figure 4-16 présente l'interface qui permet de détection des topics du contenu d'une ressources parlée.

On note aussi lors de la phase de détection, nous recourrons aux mêmes procédés utilisés lors de la définition des modèles de représentations des topics tels que :

- La sélection des termes qui représentent le contenu de la ressource parlée via les techniques de lemmatisation et élimination des termes vide « Stop Word ». Dans l'interface graphique de l'environnement, les termes sélectionnés sont coloriés en rouge.
- La pondération des termes sélectionnés par rapport à leurs fréquences dans le document source après l'élimination des termes rares.
- La normalisation des fréquences pondérées dans l'intervalle [0,1] avec la technique de normalisation « *min-max* ».

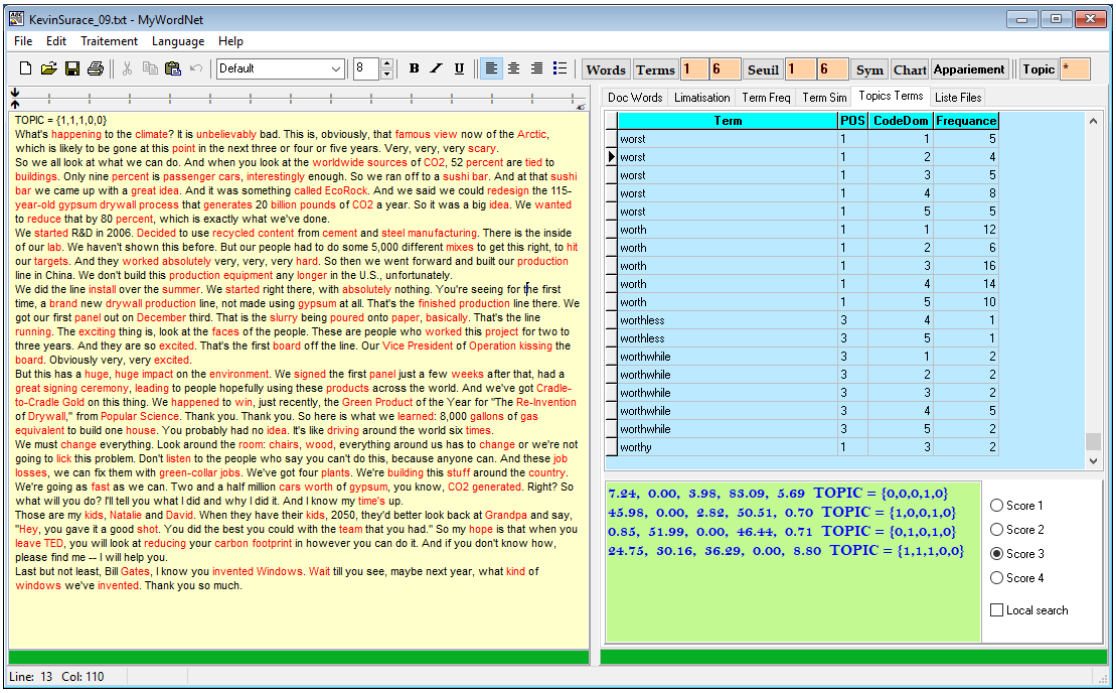

Figure 4-16 : Interface de détection des topics du contenu d'une ressource parlée

Les résultats de détection obtenus présentent des scores de vraisemblances par rapport au topics définis. Ainsi, le critère de décision utilisé est le seuil de détection « S_D ». Nous avons considéré que les topics sont équiprobables et le seuil de détection est ainsi :

$$S_D = \frac{1}{N} \qquad (4\text{-}2)$$

Avec

- N : nombre de topics.

Entre-temps, nous envisageons l'utilisation des probabilités d'existence des topics lors de la phase d'apprentissage comme un facteur pour le calcul de la valeur de S_D dans nos travaux futurs. En revanche, nous constatons que la valeur utilisée de S_D est discriminante. La figure 4-17 présente les résultats de détection du contenu de quatre ressources parlées. Notons aussi qu'on a utilisé une codification binaire pour l'annotation des topics existants dans le contenu d'une ressource parlée, par exemple :

- TOPIC = {1,0,0,1,0} : le contenu de cette ressource appartient aux premier et quatrième topic.

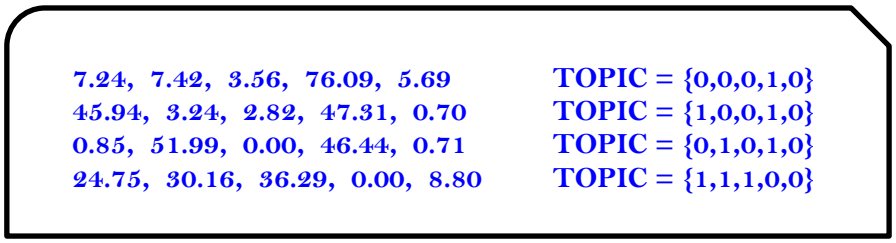

Figure 4-17 : Exemple des résultats de détection des topics pour le contenu des ressources parlées

Afin d'évaluer les performances de ces scores, nous avons annoté un corpus test par rapport aux topics utilisés dans cette étape de validation. Les ressources de corpus sont extraites à partir des transcriptions des ressources de la conférence TED. Le tableau 4-16 décrit la description du corpus test utilisé.

	Technology	Business	Design	Science	Global issues
Technology	63	15	22	21	12
Business	15	42	11	7	16
Design	22	11	48	7	4
Science	21	7	7	53	10
Global issues	12	16	4	10	43

Tableau 4-16 : Répartition des ressources parlés du corpus test

Pour l'évaluation des performances des mesures utilisées. Nous avons exécuté une série d'expérimentation pour chaque mesure de détection des topics du contenu des ressources parlées du corpus test. Ainsi, nous utilisons les métriques d'évaluation standards dans les systèmes de recherche d'information : *Précision* et *Rappel*. Les tableaux 4-17 et 4-18 présentent les valeurs de *Rappel* obtenu en utilisons la formule du $Score_1(T_i, D_j)$ et $Score_3(T_i, D_j)$ respectivement pour la détection des topics du contenu des ressources parlées du corpus *Test*.

		TP	TF	Nb de ressources	Evaluation	
					Rappel	Précision
Topic 1	**T.P**	59	6	65	0.9077	0.9077
	T.N	6	71	77	0.9221	0.9221
Topic 2	**T.P**	40	6	46	0.8696	0.8333
	T.N	8	88	96	0.9167	0.9362
Topic 3	**T.P**	44	6	50	0.8800	0.8800
	T.N	6	86	92	0.9348	0.9348
Topic 4	**T.P**	40	9	49	0.8163	0.8889
	T.N	5	88	93	0.9462	0.9072
Topic 5	**T.P**	46	6	52	0.8846	0.8846
	T.N	6	84	90	0.9333	0.9333

Tableau 4-17 : Valeurs de Rappel et Précision obtenu par le $Score_1(T_i, D_j)$ pour la détection des topics du contenu des ressources parlées du corpus Test

		TP	TF	Nb de ressources	Evaluation	
					Rappel	Précision
Topic 1	**T.P**	54	3	57	0.9474	0.9643
	T.N	2	83	85	0.9765	0.9651
Topic 2	**T.P**	39	4	43	0.9070	0.9512
	T.N	2	97	99	0.9798	0.9604
Topic 3	**T.P**	50	3	53	0.9434	0.9615
	T.N	2	87	89	0.9775	0.9667
Topic 4	**T.P**	40	3	43	0.9302	0.9524
	T.N	2	97	99	0.9798	0.9700
Topic 5	**T.P**	50	3	53	0.9434	0.9615
	T.N	2	87	89	0.9775	0.9667

Tableau 4-18 : Valeurs de Rappel et Précision obtenu par le $Score_3(T_i, D_j)$ pour la détection des topics du contenu des ressources parlées du corpus Test

Avec

- $T.P$: Topic existe et détecter.
- $T.N$: Topic non existant et non détecter.

Nous constatons que les résultats obtenus par la deuxième mesure sont les meilleurs et sont très satisfaisants. Cependant pour assurer les performances de cette mesure par rapport aux problèmes de pondération locale nous envisageons dans nos travaux futurs d'élargir le volume de corpus d'apprentissage utilisé pour la construction des vecteurs de représentations des topics.

4.4.2.4. Détection des topics à base des transcriptions partielles

Afin d'alléger le temps nécessaire pour la transcription du contenu des ressources parlées, nous étudions la capacité de notre système de détection vis-à-vis le contenu partiel. Dans ce contexte, nous recourons à tester le système de détection avec un scénario du partitionnement séquentiel du contenu. A cet effet, nous définissons le paramètre α qui présente le taux de partitionnement. Par exemple : $\alpha = 2$ signifie qu'on traite la moitié du contenu de la ressource parlée. Ensuite, la sélection des termes de vecteur de représentation s'effectue séquentiellement via une simple fonction de « *modulo* » après une étape de lemmatisation des termes et élimination des mots vides du contenu intégrale. En effet, nous envisageons un partitionnement purement statistique basé sur les fréquences d'apparitions des termes et des pondérations locales. Cette stratégie stimule typiquement la philosophie de processus d'indexation automatique utilisé dans le module SLA-SOMI. Le tableau 4-19 présente les résultats de détection obtenus pour les valeurs de α comprises dans l'intervalle [2,20] ;

Evaluation	Taux de partitionnement								
	1/2	**1/3**	**1/4**	**1/5**	**1/6**	**1/7**	**1/10**	**1/15**	**1/20**
Rappel	0.9577	0.9437	0.9296	0.8803	0.8592	0.8451	0.7746	0.7042	0.6549
Precision	0.9706	0.9590	0.9511	0.9253	0.9020	0.8899	0.8287	0.7882	0.8449

Tableau 4-19 : Evaluation de l'impact de partitionnement sur la qualité de détection des topics

Nous constatons que le système de détection des topics capable de trouver une décision pour les valeurs de $\alpha \in [2,4]$, ainsi les performances suivent une courbe descendante par rapport au taux de partitionnement. La figure 4-18 schématise l'évolution des valeurs de *Précision* et *Rappel* du système de détection de topics par rapport au paramètre de partitionnement.

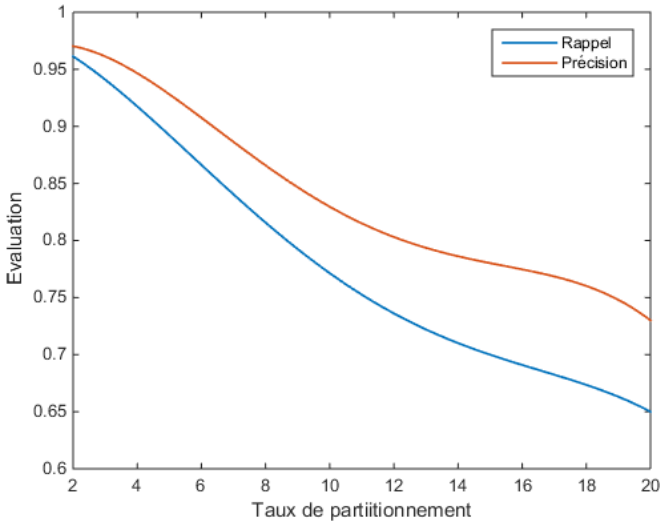

Figure 4-18 : Courbe d'évolution des valeurs de Précision et Rappel par rapport au paramètre de partitionnement

En revanche, notre approche proposée est sur l'utilisation du contenu partiel des ressources parlées comme un support d'indexation. À cet effet, recourrons à un processus d'enrichissement sémantique de ce partiel pour augmenter le taux de détection.

4.4.3. L'enrichissement sémantique

L'objectif de cette étape est l'enrichissement de l'ensemble des termes fréquents obtenu à partir de transcription partielle du contenu des ressources parlées par de nouveaux concepts communs. Entre autres, nous cherchons des rapprochements sémantiques des termes candidats d'indexation par les mesures de similarités entre eux. Dans ce contexte, nous avons intégré les différents concepts et relations de l'ontologie WordNet dans l'environnement « *MyWordNet* ». Ainsi nous pouvons trouver plusieurs types de relations comme *synonyms*, *hypernyms* et *hyponyms* pour un concept donné. En effet, ces relations nous permettent l'enrichissement des termes d'indexation par des termes similaires. Cependant, nous exploitons les autres relations comme : *member of*, *part of* pour l'enrichissement avec des termes d'indexation plus générale. La figure 4-19 présente un exemple des termes liés au concept « *Similarity* » par la relation *hyponyms*.

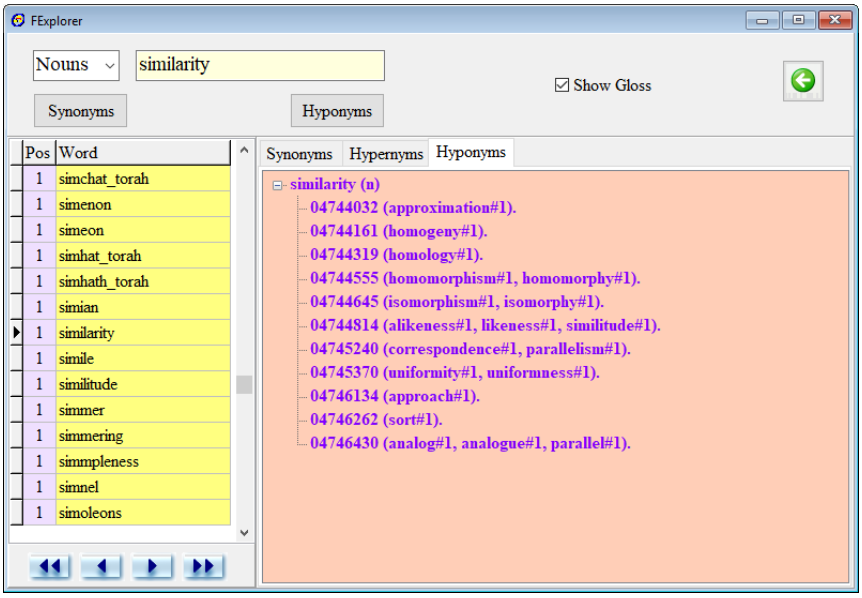

Figure 4-19 : Exemple des termes liée au concept « *Similarity* » par la relation *hyponyms*.

4.4.3.1. Les mesures de similarités

L'enrichissement par des relations telles que : *synonyms*, *hypernyms* et *hyponyms* s'effectue par la liste termes de *Synset* qui couvre le concept lié au terme. Cependant pour la recherche des concepts plus généraux ou plus détaillés d'un concept, nous utilisons les mesures de similarité à base de traits lexicaux et distances taxonomiques. Dans ce contexte, nous avons implémenté dans l'environnement « *MyWordNet* » quatre mesures de similarités qui sont : *Lesk étendu*, *Rada*, *Wu&P*, *Wu&P2* et *Leacock & Chodorow*. La figure 4-20 présente un exemple de calcul de similarité à base des distances taxonomiques.

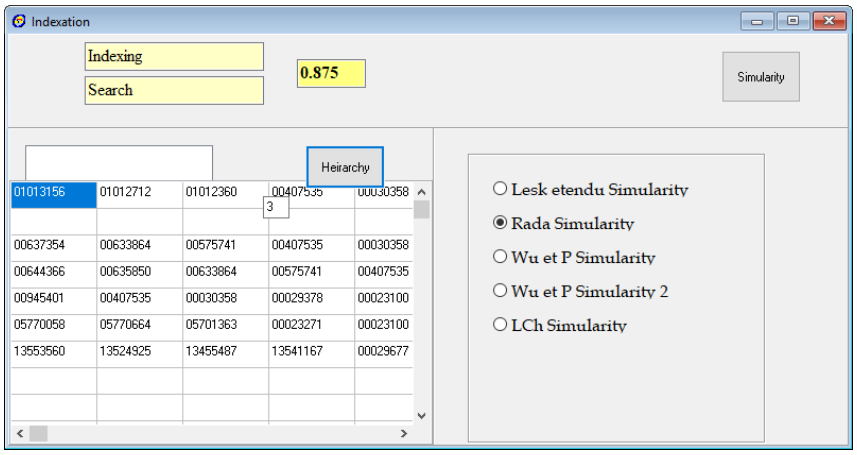

Figure 4-20 : la mesure de similarité *Rada* entre les termes « *Indexing* » et « *Search* »

4.4.3.2. Stratégie d'enrichissement

La phase d'enrichissement permet de couvrir les erreurs de reconnaissances liées aux capacités du modèle acoustique et du modèle de langage utilisés. Elle permet aussi d'augmenter la représentativité informationnelle du vecteur de représentation du contenu partiel de la ressource parlée. Dans ce contexte, nous recourons au calcul des similarités entre les termes fréquents du contenu partiel d'une ressource parlée. La figure 4-21 présente un exemple de calcul de similarités entre les termes fréquents du contenu d'une ressource parlée.

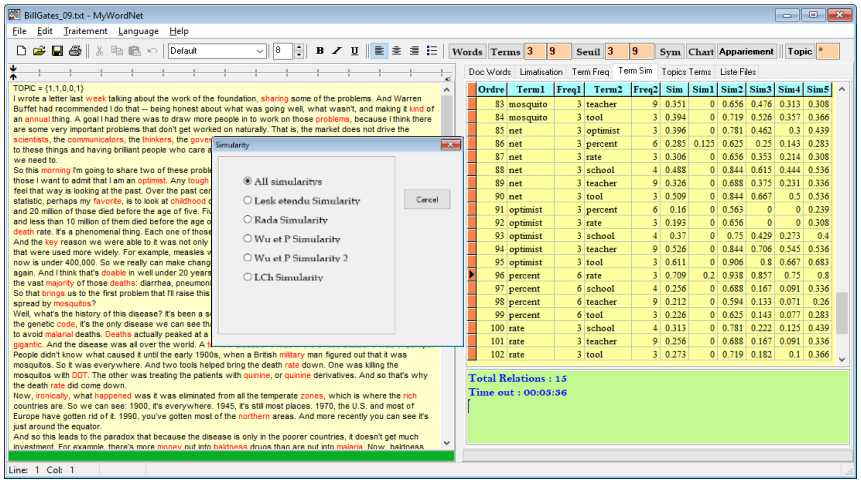

Figure 4-21 : Exemple des valeurs de similarités obtenues pour les termes fréquents de contenu d'une ressource parlée

Notons aussi que nous avons implémenté un ensemble de mesure de similarités. Les pouvoirs de représentation de ces mesures dépendent aux caractéristiques de la mesure ainsi que le contexte utilisé. Nous remarquons dans la figure 4-22 que la mesure de *Lesk* est non discriminante pour ce contexte.

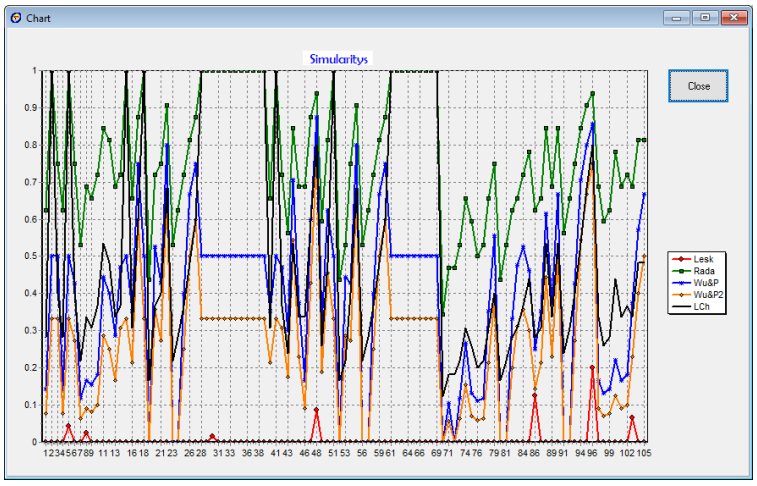

Figure 4-22 : Représentation graphique des valeurs de similarités par rapport aux mesures utilisées

Enfin, l'enrichissement de cet ensemble s'effectue via les relations : *member of, part of* pour les termes qui ont des fortes relations sémantiques.

4.4.3.3. Impact d'enrichissement sur le system de détection des topics

Afin de mesurer l'impact du processus d'enrichissement sur la qualité des termes d'indexation obtenue. Nous réexécutons le même scénario de détection de topic utilisé dans la section 4.4.2.4 mais avec l'ajout des termes d'enrichissements. Les figures 4-23 et 4-24 représentent respectivement l'impact du processus d'enrichissement sur les valeurs de Rappel et Précision du système de détection des topics.

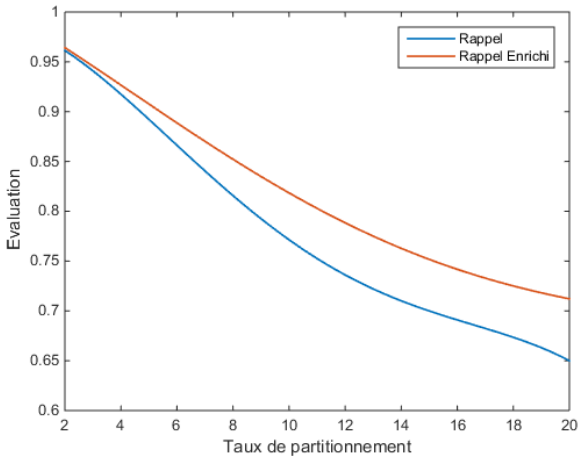

Figure 4-23 : Impact du processus d'enrichissement sur les valeurs de *Rappel* de système de détection partiel des topics

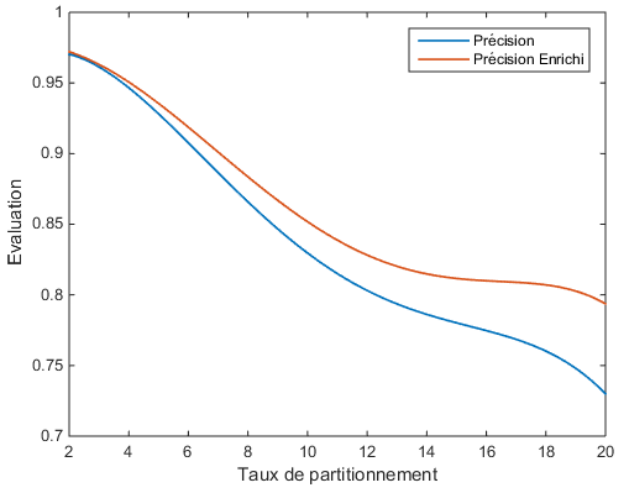

Figure 4-24 : Impact du processus d'enrichissement sur les valeurs de *Précision* de système de détection partiel des topics

Les résultats obtenus montrent qu'il y a une amélioration des performances de système de détection par rapport au contenu partiel. Cependant, ces performances sont liées aux tailles des ressources parlées. En effet, la sélection des termes susceptibles d'indexation et qui seront l'objet d'enrichissement est effectuée à base des fréquences d'apparitions. Les valeurs de ces fréquences s'affaiblissent lors de processus de partitionnement, ce qui mit le choix de termes fréquents n'est pas explicite. En revanche, nous envisageons dans nos objectifs futurs d'utiliser un processus de segmentation hiérarchique des ressources parlées afin de simuler les structures de la hiérarchie des phrases et paragraphes.

4.4.4. Conclusion

Le module SLA-SOMI est conçu pour la détection des termes d'indexation valides et discriminant pour le contenu des ressources parlées. Dans ce contexte, nous avons introduit le concept « Topic » pour catégoriser les termes fréquents du contexte de la ressource parlée. À cet effet, nous avons conçu un système de détection des topics par les modèles de représentations vectoriels avec les fréquences pondérées. Ainsi, nous avons étudié la capacité de ce système par rapport au contenu partiel de ces ressources parlées. En effet, cet objectif est primordial pour surmonter les problèmes liés aux charges de calculs. Enfin, nous recourons aux mesures de similarités sémantiques pour enrichir la transcription du contenu partiel de ces ressources parlées afin d'accroitre les performances du système de détection. Entre temps, ce processus d'enrichissement permet d'améliorer la qualité des transcriptions obtenue par le module SLA-SOMI par rapport aux termes techniques, termes étranges te les termes hors vocabulaire. Pour la validation et les tests d'expérimentations de ce module, nous avons développé l'environnement « *MyWordNet* » avec le langage *Delphi* qui assure toutes les fonctionnalités du module CSA-SOMI.

4.5. Validation du module KDE-SOMI

Ce module permet de chercher la possibilité de correspondances entre les termes d'indexation du contenu d'une ressource parlée avec sa structure physique. En effet, les termes d'indexation obtenus à partir des modules SLA-SOMI et CSA-SOMI sont très suffisants pour le fonctionnement d'un système de recherches dans les flux parlés. Cependant, nous envisageons dans notre approche d'améliorer le mode de restitution des résultats de recherches par la possibilité de restituer les segments pertinents de la ressource parlée détectée. Dans ce contexte, nous avons utilisé les techniques de reconnaissance et détection dans les flux parlés tels que : « Keyword Spotting ». Ainsi, nous envisageons d'utiliser les techniques de détection à base phonétique avec les systèmes LVCSR. Dans les sections suivantes, nous prestons les deux stratégies de détection avec quelques exemples des tests d'expérimentations.

4.5.1. Scenario de détection basée « Keyword Spotting »

Afin d'accéder au contenu des flux parlé à l'aide des index discriminants obtenus lors des phases précédentes, nous avons utilisé dans ce scénario la technique « *Keyword Spotting* » pour la détection de probabilité d'existence des termes dans une séquence parlée. Nous avons utilisé les modèles de représentation en mode local avec le module « *sphinxbase* » dans sa version « 0.8 » et le module « *pocketsphinx* » sous python.

En effet, la technique de « *Keyword Spotting* » est utilisée essentiellement pour les systèmes de commandes vocales et elle traite des segments parlés de petite taille. Dans ce contexte, nous recourrons à la même stratégie utilisée dans le module SLA-SOMI pour augmenter l'efficacité de cette technique est de réduire le temps de détection des termes.

Dans la phase des tests, nous avons utilisé le modèle acoustique « *hub4wsj_sc_8k* ». C'est un modèle markovien construit à partir de 330 heures des infos diffusées de *Wall Street Journal* avec une fréquence d'échantillonnage de 8 kHz. Ainsi, nous avons utilisé le dictionnaire d'anglais phonétique « *cmudict-en-us.dict* » et le modèle de langage « *cmu-sphinx* ». La figure 4-25 présente l'extrait de configuration du module de détection de *Pocketsphinx*.

```
config = Decoder.default_config()
config.set_string('-hmm', "/home/issam/Project/pocketsphinx-0.8/model/hmm/en_US/hub4wsj_sc_8k")
config.set_string('-lm', "/home/issam/Project/pocketsphinx-0.8/model/lm/en_US/en-us.lm.bin")
config.set_string('-dict',"/home/issam/Project/pocketsphinx-0.8/model/lm/en_US/cmudict-en-us.dict")
config.set_string('-kws', 'keyword.list')
decoder = Decoder(config)
```

Figure 4-25 : Les modèles utilisés pour le module *KWS Pocket Sphynx*

Entretemps, pour la tâche de « *KWS* », il faut « *keyword .list* » qu'on définît la liste des termes à détecter avec ses critère de détection par les valeurs de « *kws_threshold* » pour chaque termes. Sa valeur est souvent entre [1 e-10 , 1 5-50] et elle est calculée par rapport aux représentations phonétiques des termes. Dans les figues 4-26 et 4-27 présentent respectivement un exemple de représentation des termes fréquents du contenu d'une ressource parlée obtenu par le module CSA-SOMI dans un modèle.

```
DEATH   D EH TH
KID     K IH D
MALARIA M AH L EH R IY AH
MOSQUITO        M AH S K IY T OW
```

Figure 4-26 : Les modèle de représentation phonétique des termes à détecter

```
DEATH /1e-10/
KID /1e-10/
MALARIA /1e-10/
MOSQUITO /1e-10/
```

Figure 4-27 : exemple d'un fichier *keyword .list*

Ensuite, nous avons effectué une série des tests sur les valeurs de « *kws_threshold* » sur un ensemble de ressources parlées. Dans le premier scénario, nous avons utilisé la même valeur de « *kws_threshold* » pour tous les termes recherchés. Le tableau 4-20 présente les taux moyens de détection obtenus pour chaque valeur de « *kws_threshold* ».

Evaluation	Taux de partitionnement								
	1 e-10	**1 e-15**	**1 e-20**	**1 e-25**	**1 e-30**	**1 e-35**	**1 e-40**	**1 e-45**	**1 e-50**
Taux de détection moyen	0.8523	0.8179	0.7987	0.8194	0.8438	0.8632	0.8591	0.8233	0. 8045

Tableau 4-20 : Taux de détection moyen par rapport a la valeur de *kws_threshold*

Dans le deuxième scénario, la valeur de « *kws_threshold* » est calculée en fonction de nombre de phonèmes pour chaque terme par la formule suivante :

$$kws_{threshold} = 1e - (5 * (n - 1)) \quad (4\text{-}3)$$

Avec

- n : nombre de phonèmes du terme
- $si\ n = 2,\ kws_{threshold} = 1e - 10$
- $si\ n > 11,\ kws_{threshold} = 1e - 50$

La figure 4-27 présente l gain obtenu après l'application de cette stratégie.

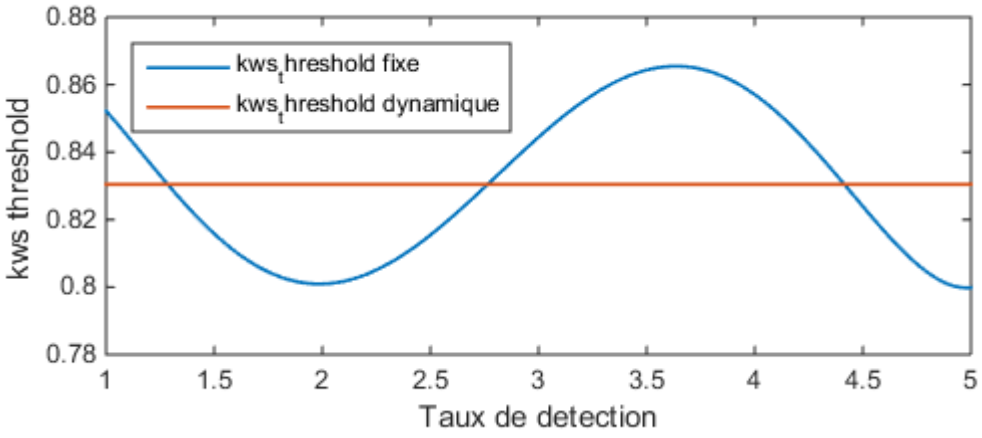

Figure 4-28 : Impact des valeurs de *kws_threshold* sur le taux de détection

En revanche, l'utilisation de la technique « keyword spotting » avec des modèles de représentation standards pour les systèmes de recherches dans les flux parlés dans des contextes généraux. Cependant, elle ne permet pas de pallier effectivement le problème de certains termes techniques ou nouveaux dans les systèmes de recherches de flux parlés spécialisés. En effet, le processus de détection utilise les probabilités d'émission et d'observations définies dans le modèle acoustique markovienne. Dans ce contexte, nous trouvons l'autre alternative pour la détection des termes parlés qui sera détaillée dans la section suivante.

4.5.2. Scenario de détection basée LVCSR

Dans ce scénario, nous envisageons la construction des modèles de représentation propres pour le système de recherche dans les flux parlés. Dans ce contexte, nous trouvons la plateforme *« Kaldi ASR »* qui intègre un ensemble d'outils de traitements tel que :

- Le F4DE « Framework for Detection Evaluation **» :** c'est un module en langage Perl qui contient un ensemble d'outils d'évaluation pour les systèmes de détection.
- SRILM « SRI Language Modeling Toolkit » : c'est un outil pour la création des modèles de langages et les autres ressources linguistiques : 2-gram, 3-gram…etc.
- WFST « Open Source WFST-based Decoder Toolkit » : c'est un outil de décodage vocal basé sur WFST léger et portable écrit en C ++.

Cette plateforme permet la construction des modèles acoustiques markovienne SGMM et les modèles neuronaux : *DNN*. Cependant l'apprentissage des modèles DNN nécessite des ressources matérielles énormes et les calculs sont exécutés généralement dans les Grid Computing. Entretemps, les corpus nécessaires pour l'apprentissage du modèle acoustique sont payants. A cet effet, nous n'avons pas pu de tester ce scénario et il est parmi nos objectifs futurs.

4.5.3. Conclusion

L'objectif de ce module est d'appliquer les techniques de détection dans les systèmes de recherche dans le contenu parlé. Certes, que l'indexation est l'étape primordiale pour annoter et restituer les ressources parlées dans les systèmes de recherche, mais nous envisageons dans notre approche proposée que le système de recherche puisse fournit des segments de la ressource parlée au lieu de sa totalité. En effet, ce choix est jugé par le volume important des ressources parlées traitées. Pratiquement, nous avons utilisé les techniques de « *keyword spotting* » avec des modèles de représentation linguistique générale. Les résultats obtenus sur les tests obtenus sont acceptables, mais elles ne sont pas généralisées à l'échelle. Cependant, le développement des propres modèles de représentations dédié aux systèmes de recherche dans le contenu parlé par la plateforme *Kaldi-Asr* révèle aussi important afin d'accroitre les performances de détection vis-à-vis les termes étranges de langues et les termes techniques. Cette solution est utile pour les systèmes de recherches dédiés aux contextes spécifiques tels que la recherche scientifique, les sciences biomédicales et les bibliothèques numériques spécialisées. À cet effet, nous envisageons dans nos perspectives futures de construire des systèmes de reconnaissance et détection personnalisés et pour les autres langages.

Conclusion Générale

L'émergence des nouvelles technologies dans notre vie quotidienne et professionnelle a engendré la création d'un large volume des ressources multimédias. L'accès et la recherche dans ce volume nécessitent l'amélioration et l'adaptation des techniques et méthodes utilisées pour la gestion de leurs contenus. Les performances de ces systèmes de recherches dépendent étroitement des techniques utilisées pour la modélisation du contenu ainsi que les stratégies d'indexations utilisées.

Actuellement, l'accès au contenu de ces ressources multimédias s'effectue par le biais des métadonnées. Ces métadonnées représentent les caractéristiques physiques comme le titre, l'auteur, la taille, le format, le codec utilisé … etc. Elles représentent aussi dans certaines ressources multimédias les annotations et les références. Ces annotations et références sont souvent manuelles et dépendent des capacités cognitives humaines. En effet, avec la croissance colossale du volume de ces ressources, la recherche des techniques et approches pour l'amélioration de ce processus par des procédés automatiques est largement sollicitée.

Dans ce contexte, nous avons présenté dans ce manuscrit des contributions pour l'amélioration des systèmes de recherche dans le contenu parlé par la proposition d'une approche d'indexation sémantique pour le contenu parlé des ressources multimédias. Cette approche qu'on a appelée SOMI définit les mécanismes d'accès au contenu parlé en utilisant les techniques de recherche d'informations et celles du contenu, les techniques de reconnaissance automatique de la parole et les techniques de représentations et modélisation des connaissances par les ontologies et les distances de similarités sémantiques. L'approche proposée est composée de trois modules : SLA-SOMI, CSA-SOMI et KDE-SOMI reflètent impérativement les disciplines étudiées. Pour cela, ce travail a été réalisé et développé en trois axes qui représentent les trois problématiques essentielles : recherche sur le contenu parlé, modélisation de la structure de flux parlé et les techniques et modalités d'accès.

Dans le premier module, nous avons travaillé sur l'aspect structurel du contenu. A cet effet, nous avons exploité les APIs de reconnaissances automatiques de la parole : « *Google Cloud Speech API* » et « *PocketSphinx* » pour accéder au contenu des ressources parlées. Ces systèmes ne permettent pas la transcription des longues durées de flux parlés. Ce qui nous amène à proposer une stratégie d'indexation automatique des flux parlés vers des segments valides afin de manipuler des séquences audios de grandes tailles. Nous avons effectué une série d'expérimentations pour évaluer l'impact de cette stratégie sur la qualité des systèmes en reconnaissance.

Tandis que, nous avons travaillé dans le deuxième module sur l'amélioration des résultats de transcription du contenu obtenu par le module SLA -SOMI et la recherche des termes d'indexation pertinents et discriminants. En effet, les erreurs engendrées par les APIs de reconnaissance comme les omissions et les insertions affectent le processus d'indexation. Dans ce cas, nous avons défini le module CSA-SOMI. Il utilise une stratégie des détections des topics qui couvrent le contenu des flux parlés. Nous avons utilisé les techniques d'indexation *tf/idf* avec des pondérations locales par topic pour la détection des termes candidats d'indexation du contenu des flux parlés.

Cette stratégie permet de situer le contenu des flux parlés avec leurs contextes. Ensuite, nous avons procédé à l'utilisation d'un processus d'enrichissement sémantique via l'ontologie *WordNet* avec les mesures de similarités : *Lesk étendu, Rada, Wu & Palmer* et *Leacock & Chodorow*. Ce processus permet l'enrichissement de l'ensemble d'index candidats par de nouveaux concepts via les relations de l'ontologie WordNet comme : *hypernym, hyponym, member of, has part*.

En revanche, dans le troisième module KDE-SOMI, nous avons étudié l'utilisation des systèmes de détection phonétique des termes parlés STD avec la technique : *Keyword spotting*. Dans ce contexte, nous avons opté en premier lieu d'utiliser les modèles de représentations standard : le modèle acoustique, le modèle de langage et le dictionnaire phonétique de plateforme *Sphinx Speech Recognition* pour la détection des termes d'indexation dans le contenu des flux parlés. Ainsi, nous avons étudié l'utilisation de la plateforme *KALDI* avec l'utilisation des propres modèles de représentation.

Dans la phase d'implémentation et de validation, nous avons utilisé le corpus open source *TED-LIUM V1* pour la validation des différents modules de l'approche SOMI. Pour le premier module nous avons utilisé les modules Google *Cloud Speech API* et *PocketSphinx* sous *Python* et l'environnement *SoX* et la bibliothèque *Python/aeneas* dans le processus de segmentation automatique. Tandis que pour la validation du CSA-SOMI, nous avons développé tout un environnement sous *Delphi* qui intègre la totalité des concepts et relations de l'ontologie WordNet dans un schéma relationnel.

Cet environnement contient un analyseur lexical qui permet la représentation des transcriptions des flux parlés par le modèle vectoriel avec des fréquences pondérées. Aussi, il permet la définition des topics et la détection des topics. Il permet aussi le calcul de plusieurs mesures de similarités comme : *Wu & Palmer* et *Leacock & Chodorow*, …etc. Ainsi, il permet l'enrichissement via les relations de WordNet modélisées dans le schéma

relationnel. Pour le module KDE-SOMI, nous avons utilisé le module PocketSphinx KWS avec les modèles : acoustiques, langage et phonétique de la plateforme *Sphynx Speech Recognition.* Les résultats obtenus lors de l'étape de validation sont satisfaisants. Ceci nous encourage à tester cette approche avec d'autres ressources et contextes.

Cependant, pour nos perspectives futures, nous espérons de travailler sur les aspects suivants :

- L'amélioration de la stratégie de segmentation automatique par un processus de détection dynamique des intervalles de silences.
- L'utilisation d'autres modèles de représentation de connaissance entre les termes d'indexation comme « Word2Vect ».
- La construction d'un système LVSCR avec la plateforme *KALDI* avec des modèles de langage, acoustiques et phonétique appropriés.
- L'utilisation de cette approche comme un support d'indexation des ressources bibliographiques multimédias dans les bibliothèques numériques spécialisées « *E-librairies* ».
- L'utilisation de cette approche dans des contextes spécifiques comme la médecine, recherche scientifique … etc. En effet, l'exploitation de la stratégie « détection des topics » dans un contexte fermé permet l'amélioration des performances des systèmes de recherche dans les flux parlés.
- L'exploitation des ressources parlées pour d'autres langages et celles moins dotées comme les ressources parlées en arabe

Bibliographie

[Adami, 2010] A. G. Adami, *Automatic speech recognition: from the beginning to portuguese language*, In Proceedings of PROPOR, 2010.

[Adamson, 1974] G. Adamson et J.Boreham. *The use of an association measure based on character structure to identify semantically related pairs of words and document titles*. In Information Storage and Retrieval, 10, p. 253–60, 1974.

[Allauzen, 2004] C. Allauzen, M. Mohri and M. Saraclar, *General indexation of weighted automata: application to spoken utterance retrieval*, In Proceedings of the Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval at HLT-NAACL'04, Association for Computational Linguistics, pp. 33—40, 2004.

[Beazley, 2009] D.M. Beazley, *Python Essential Reference*, 4th edi, New Riders Publishing, Thousand Oaks, CA, USA, 2009

[Bellanger, 2008] M. Bellanger, *Traitement numérique du signal Théorie et pratique*, 2e édition, Dunod, Collection : Sciences Sup, ISBN : 2-10-050162-3, 2008

[Benayed, 2003] Benayed Y, Fohr D, Haton JP, Chollet G, *Confidence measure for keyword spotting using support vector machines*. In: Proceedings of ICASSP. pp 588–591, 2003

[Bendib, 2018] Bendib I, laouar M.R., *A semantic indexing approach of multimedia documents content based partial transcription*. In Proceding of 2nd International Conference on Natural Language and Speech Processing ICNLSP'18 – IEEE, 2018

[Bendib, 2014] Bendib I, Laouar MR, Hacken R and Miles M, *Semantic ontologies for multimedia indexing (SOMI): Application In the e-library domain*, In Library Hi Tech, V: 32, N°2, pp=206--218, 2014

[Bendib, 2012] Bendib I, laouar M.R., *Approaches for the Detection of the Keywords in Spoken Documents Application for the Field of E-Libraries*. In: Huang T., Zeng Z., Li C., Leung C.S. (eds) Neural Information Processing. ICONIP'12. Lecture Notes in Computer Science, vol 7666. Springer, Berlin, Heidelberg, 2012

[Bertoldi, 2003] N. Bertoldi et M.Federico, *Cross-Language Spoken Document Retrieval on the TREC SDR Collection*. In: Peters C., Braschler M., Gonzalo J., Kluck M. (eds) Advances in Cross-Language Information Retrieval. CLEF 2002. Lecture Notes in Computer Science, vol 2785. Springer, Berlin, Heidelberg, 2003

[Bishop, 1995] C. Bishop, *Neural Networks for Pattern Recognition*, Oxford: Clarendon Press, 1995.

[Bourlard, 1994] H. Bourlard and N. Morgan, *Connectionist Speech Recognition - A Hybrid Approach*, Boston: Kluwer Academic Publishers, 1994.

[Boves, 2009] L. Boves, R. Carlson, E. W. Hinrichs, D. House, S. Krauwer, L. Lemnitzer, M. Vainio and P. Wittenburg, *Resources for speech research: present and future infrastructure needs*, In INTERSPEECH, pp. 1803—1806, 2009.

[Carletta, 2005] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec,V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I.McCowan, W. Post, D. Reidsma, P. Wellner, *The AMI meeting corpus: a preannouncement*, In Proc of the Workshop on Machine Learning for Multimodal Interaction (MLMI), Edinburgh, Jul. 2005.

[Can, 2011] D. Can and M. Saraclar, *Lattice indexing for spoken term detection*, Audio, Speech, and Language Processing, IEEE Transactions on, pp. 2338--2347, 2011.

[Cieri, 2004] C. Cieri David, D. Miller et K. Walker, *The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text*, In Proceedings of 4th International Conference on Language Resources and Evaluation, pp: 69—71, 2004

[Champclaux, 2010] Y.Champclaux, *Un modèle de recherche d'information basé sur les graphes et les similarités structurelles pour l'amélioration du processus de recherche d'information*. Thèse de Doctorat, Université Toulouse III - France, 2010

[Chelba, 2005] C. Chelba and A. Acero, *Position specific posterior lattices for indexing speech*," In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 443-450, 2005

[Chou, 2003] W. Chou and F. Juang, *Pattern Recognition in Speech and Language Processing*, CRC Press, 2003.

[Dahlbäck, 1997] N. Dahlbäck, *Towards a Dialogue Taxonomy*. In Dialogue Processing in Spoken Language Systems, 1997.

[Dudognon, 2010] D. Dudognon, G. Hubert, and B.J Victorino Ralalason, *Proxigénéa: Une mesure de similarité conceptuelle*. In Proceedings of the Colloque Veille Stratégique Scientifique et Technologique, 2010

[Fankam, 2009] C. Fankam, L. Bellatreche, D. Hondjack, Y. A. Ameur and G. Pierra, *Conception de Bases de Données à partir d'Ontologies de Domaine*, Technique et Science Informatiques, p. 1233–1261, 2009.

[Fiscus, 2007 J. Fiscus, J. Ajot, J. Garofolo, G. Doddington, Results of the 2006 Spoken Term Detection Evaluation, In Procedding SIGIR'07 Workshop on Searching Spontaneous Conversational Speech, July 2007.

[Frakes, 1992] W.B Frakes, *Information Retrieval Data Structures and Algorithms- Chap 8*, Ricardo Baeza-Yates Prentice Hall, 1992

[Fucks, 2017] Fuchs. T, Keshet. J, *Spoken Term Detection Automatically Adjusted for a Given Threshold*, IEEE Journal of Selected Topics in Signal Processing, 11, 1310-1317, 2017.

[Fukunaga, 1990] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, vol. II, Academic Press, 1990.

[Gaida, 2014] C. Gaida, P. Lange, R. Petrick, P. Proba, A. Malatawy, and D. Suendermann-Oeft, *Comparing open-source speech recognition toolkits*, Tech. rep, Technical report, Project OASIS, 2014.

[Garcia, 2006] A. Garcia and H. Gish, *Keyword spotting of arbitrary words using minimal speech resources*, In Procceedings of Acoustics, Speech and Signal'06 .IEEE, pp. I—I, 2006

[Garofolo, 2000] J.S. Garofolo, . C. G. P. Auzanne and E. M. Voorhees, *The trec spoken document retrieval task: A success story*, In Proceedings of RIAO: Content Based Multimedia Information Access Conference, Paris, France, 2000.

[Garofolo, 1993] J.S. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallet, N. Dahlgren. *Darpa, TIMIT, Acoustic-phonetic continuous speech corpus*, NISTIR Publication No 4930), Washington, DC: US Department of Commerce, 1993

[Godfrey, 1992] J.J. Godfrey, E.C. Holliman et J. McDaniel, *SWITCHBOARD: telephone speech corpus for research and development*. In Proceedings of the International Conference on Acoustics, speech and Signal Processing, volume I, pp. 517–520. March 1992.

[Gold, 2011] B. Gold, N. Morgan and D. Ellis, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*, vol. II, Hoboken, New Jersey: John Wiley & Sons, Inc., 2011.

[Gomez Perez, 2004] A. Gómez-Pérez, M. Fernández-López and Ó. Corcho, *Ontological Engineering: With Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web*, In Advanced Information and Knowledge Processing, 2004.

[Gomez Perez, 1999] A. Gomez Perez, *Ontological Engineering : A State of the Art*, Expert Update 2, , pp. 33-44, 1999.

[Grangier, 2009] D. Grangier, J. Keshet and S. Bengio, *Chapter on discriminative keyword spotting*, In Automatic speech and speaker recognition: large margin and kernel methods., NewYork:Wiley, 2009.

[Gruber, 2009] Gruber, T, Ontology in Encyclopedia of Database Systems, Ling Liu and M. Tamer Özsu (Eds.), Springer-Verlag, 2009.

[Gruber, 1993] T. Gruber, "A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition," Current issues in knowledge modeling, pp. 199-220, 1993.

[Gruhn, 2011] R.E. Gruhn et al., Statistical Pronunciation Modeling for Non-Native Speech Processing, Signals and Communication Technology, Springer-Verlag Berlin Heidelberg 2011

[Guarino, 1999] N. Guarino N., C. Masolo, and G. Vetere, OntoSeek : *Using Large Linguistic Ontologies for Accessing On-Line Yellow Pages and Product Catalogs*. National Research Council, LADSEBCNR : Padavo, Italy, 1999

[Guarino, 1995] N. Guarino, *Formal Ontology, Conceptual Analysis and Knowledge Representation*, International Journal of Human-Computer Studies Volume 43, Issues 5–6, p. 625–640., 1995.

[Hrtmam, 2016] Hartmann W, Le Z, and Kerri B, Hsiao R, Stavros Tand Schwartz , *Comparison of Multiple System Combination Techniques for Keyword Spotting*, In Interspeech'16, pp=1913—1917, 2016.

[Hain, 2004] T. Hain, P. C. Woodland, G. Evermann, M. J. F. Gales, X. Liu, G. L. Moore, D. Povey and L. Wang, *Automatic Transcription of conversational telephone speech*, IEEE Transaction, 2004.

[Hakkani-Tür, 2003] D. Hakkani-Tür and G. Riccardi, *A general algorithm for word graph matrix decomposition*, In Proceedings Acoustics, Speech, and Signal Processing (ICASSP'03).pp. I—596, 2003

[Hirst, 1998] G. Hirst & D.D. St-onge, *Lexical chains as representations of context for the detection and correction of malapropisms. WordNet : An electronic Lexical Database*. C. Fellbaum. Ed. MIT Press. Cambridge. MA, pp. 305–332. Ed. MIT Press, 1998.

[Holmes, 2001] J. Holmes and W. Holmes, *Speech Synthesis and Recognition*, vol. 2nd edition, Taylor & Francis, 2001.

[Huang, 2001] X. Huang, A. Acero and H. W. Hon, *Spoken language processing : a guide to theory, algorithm, and system development*, Upper Saddle River, N.J.: Prentice Hall PTR, 2001.

[Hubert, 2009] G. Hubert, J. Mothe and B. Ralalason, *Modèle d'indexation dynamique à base d'ontologies*, In Conférence en Recherche d'Information et Application., 2009.

[Huggins-Daines, 2006] D. Huggins-**Daines**, M. Kumar, A. Chan and A. Black, *Pocketsphinx: A Free, Real-Time Continuous Speech Recognition System for Hand-Held Devices*, in ICASSP, Toulouse, France, 2006.

[James, 1994] D. A. James and S. J. Young, *A fast lattice-based approach to vocabulary independent wordspotting*, In Acoustics, Speech, and Signal Processing. ICASSP-94, IEEE International Conference, pp. I—377, 1994

[Tejedor, 2017] Tejedor J, Toledano D, Lopez-Otero P, Serrano L, ALBAYZIN 2016 spoken term detection evaluation: an international open competitive evaluation in Spanish, In Journal on Audio, Speech, and Music Processing, V: 2017-02, 2017

[Kai, 2006] Y. Kai, *Adaptive Training for Large Vocabulary Continuous Speech Recognition*, Thèse de Doctorat , Cambridge University Engineering Department, 2006

[Keshet, 2009] J. Keshet, D. Grangier, S. Bengio, *Discriminative keyword spotting*, Speech Communication, Volume 51, Issue 4,Pages 317-329,2009

[Kneser, 1995] R. Kneser and H. Ney, "Improved backing-off for M-gram language modeling," In Proceedings of ICASSP, 1995.

[Kowalski, 2010] Gerald Kowalski, *Information Retrieval Architecture and Algorithms*, Springer-Verlag New York, Inc., New ,York, NY, 2010

[Lamere, 2003] P. Lamere, P. Kwok, E. Gouvea, B. Raj , . R. Singh, W. Walker, M. Warmuth and P. Wolf, *The CMU SPHINX-4 Speech Recognition System*, in ICASSP'03, Hong Kong, China, 2003.

[Lassila, 2001] O. Lassila and D. McGuinness, *The Role of Frame-Based Representation on the Semantic Web*, Knowledge Systems Laboratory Report KSL-01-02, Stanford University, 2001.

[Leacock, 1998] C. Leacock & M. Chodrow, *Combining local context and WordNet similarity for word sense identication*. In Fellbaum, WordNet: An Electronic Lexical Database, MIT Press, 1998

[Lecouteux, 2008] B. Lecouteux. *Reconnaissance automatique de la parole guidée par des transcriptions a priori. Informatique et langage*, Thèse de Doctorat, Université d'Avignon et des Pays de Vaucluse, 2008.

[Lefevre, 2000] F. Lefevre, *Estimation de probabilité non-paramétrique pour la reconnaissance Markovienne de la parole*, Thèse de Doctorat, Pierre and Marie Curie University, 2000

[Lee, 2001] A. Lee, T. Kawahara and K. Shikano, Julius – an open source realtime large vocabulary recognition engine, In Proceedding of EUROSPEECH, p. 1691–1694, 2001

[Lesk, 1986] M. Lesk M., *Automatic sense disambiguation using machine readable dictionaries : how to tell a pine cone from an ice cream cone*. In Proceedings of the 5th annual international conference on Systems documentation, SIGDOC'86, pp. 24–26, New York, USA, 1986

[Lin-shan, 2015] Lin-shan L, Glass J, Hung-yi L, Chun-an C. 2015. Spoken content retrieval: beyond cascading speech recognition with text retrieval. IEEE/ACM Trans. Audio, Speech and Lang. Proc. 23, 9, 1389-1420, 2015

[Lin, 1998] D. Lin, *An information-theoric definition of similarity*, In Proceedings of the 15th international conference on Machine Learning, p. 296-304, 1998.

[Luhn, 1959] H.P Luhn. *Keyword-in-Context Index for Technical Literature (KWIC Index)*. International Business Machines Corp. Yorktown Heights, NY, 1959

[Mamou, 2007] J. Mamou, B. Ramabhadran and O. Siohan, *Vocabulary independent spoken term detection*, In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, New York USA, pp. 615—622, 2007

[Mandal, 2014] Mandal, A., Prasanna Kumar, K.R. & Mitra, P, *Recent developments in spoken term detection: a survey*, In Int J Speech Technol, 2014

[Mangu, 2000] L. Mangu, E. Brill and A. Stolcke, *Finding consensus in speech recognition: word error minimization and other applications of confusion networks*, Computer Speech & Language, vol. 14, no. 4, pp. 373--400, 2000.

[Marlow, 2006] C. Marlow, N. Naaman, D. Boyd and M. Davis, *HT06, Tagging Paper, Taxonomy, Flickr, Academic Article*, in Proceedings of the Seventeenth Conference on Hypertext and Hypermedia, 2006.

[Mohri, 2008] M. Mohri, F. Pereira, and M. Riley, *Speech recognition with weighted nite-state transducers.* In Springer Handbook of Speech Processing, pp. 559-584, Springer, 2008.

[Ng, 2000] K. Ng, *Subword-based Approaches for Spoken Document Retrieval*, PHD thesis, Massachusetts Institute of Technology, Massachusetts, 2000.

[Novotney, 2009] S. Novotney, R. Schwartz and J. Ma, *Unsupervised acoustic and language model training with small amounts of labelled data*, In Acoustics, Speech and Signal Processing,. ICASSP 2009. IEEE International Conference on, IEEE, pp. 4297—4300, 2009

[Paice, 1996] C.D Paice, *Method for evaluation of stemming algorithms based on error counting*. Journal of the American Society for Information Science 47 (8) p. 632-49, 1996.

[Pan, 2010] Y.-C. Pan and L.-s. Lee, *Performance analysis for lattice-based speech indexing approaches using words and subword units*, Audio, Speech, and Language Processing, IEEE Transactions on vol 18, no 6, pp 1562--1574, 2010.

[Paul, 1991] D.B. Paul et J.M. Baker, *The design for the wall street journal-based CSR corpus*. In Proceedings of the workshop on Speech and Natural Language (HLT '91). Association for Computational Linguistics, Stroudsburg, PA, USA, 357-362, 1991

[Panayotov, 2015] V. Panayotov, G. Chien, D. Povey et S. Khudanpur, *Librispeech: an asr corpus based on public domain audio books*. In Acoustics, Speech and Signal Processing (ICASSP'15),IEEE International Conference, pp. 5206–5210, 2015

[Pellegrini, 2008] T. Pellegrini. Transcription automatique de langues peu dotées. Informatique, Thèse de Doctorat ,Université Paris Sud - Paris XI, 2008.

[Prasad, 2005] R. Prasad, S. Matsoukas, C. L. Kao, J. Ma, . D.-X. Xu, T. Colthurst, O. Kimball and R. Schwartz, *The 2004 BBN/LIMSI 20xRT English Conversational Telephone Speech Recognition*, in Interspeech 2005.

[Porter, 1980] M. F. Porter, *An algorithm for suffix stripping Program*, Vols. 14-3, Program, 1980.

[Povey, 2011] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer and K. Vesely, *The Kaldi speech recognition toolkit*, In ASRU, Hawaii, USA, 2011.

[Ostendorf, 2008] M. Ostendorf, B. Favre, R. *Grishman, Speech segmentation and spoken document processing*, IEEE Signal Processing Magazine, vol. 25, no. 3, p. 59–69, 2008.

[Rabiner, 1993] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, NJ: Prentice-Hall, 1993.

[Rada, 1989] R. Rada, H. Mili H., E. Bicknell et M. Blettner, *Development and application of a metric on semantic nets*. Systems Man and Cybernetics, IEEE Transactions on, 19(1): pp. 17-30, 1989.

[Rensik, 1995] P. Resnik, *Using information content to evaluate semantic similarity in a taxonomy*, In Proceedings of the 14th International Joint Conference on Artificial intelligence - V1, IJCAI'95, pages 448–453, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc, 1995

[Robertson et al., 1997] S. E. Robertson and S, Walker. *On relevance weights with little relevance information*. In Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval, pages 16–24. ACM Press, 1997.

[Rohlicek, 1995] J.R, Rohlicek, *Word Spotting*. In: Ramachandran R.P., Mammone R.J. (eds) Modern Methods of Speech Processing. The Springer International Series in Engineering and Computer Science (VLSI, Computer Architecture and Digital Signal Processing), vol 327. Springer, Boston, MA, 1995

[Rose, 1996] R. C. Rose, *Word spotting from continuous speech utterances*, In Automatic speech and speaker recognition, Springer , pp. 303—329, 1996.

[Rousseau, 2012] A.Rousseau, P. Deleglise et Y. Esteve, *TED-LIUM: an Automatic Speech Recognition dedicated corpus*, LREC, pp. 125-129, 2012.

[Roussey, 2011] C. Roussey, F. Pinet and M. Ah Kang, *An Introduction to Ontologies and Ontology Engineering*, In Ontologies in Urban Development Projects, New York, Springer London Dordrecht Heidelberg, 2011.

[Rybach, 2009] D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Loof, R. Schluter and H. Ney, The RWTH Aachen University Open Source Speech Recognition System, In Proceeding of INTERSPEECH, p. 2111–2114, 2009.

[Salton, 1993] G. Salton and M. J. McGill. *Introduction to Moderne Information Retrieval*, New York, 1983.

[Sandness, 2000] E. D. Sandness and I. L. Hetherington, Keyword-based discriminative training of acoustic models, In Proceedings of INTERSPEECH'2000, pp. 135—138, 2000.

[Saraclar, 2004] M. Saraclar and R. Sproat, *Lattice based search for spoken utterance retrieval*, Urbana, vol. 51, p. 61801, 2004.

[Sarikaya, 2005] R. Sarikaya, Y. Gao, M. Picheny and H. Erdogan, *Semantic Confidence Measurement for Spoken Dialog Systems*, IEEE Transactions On Speech And Audio Processing, vol. 13, no. 4, 2005

[Savoy, 2005] J. SAVOY. *Indexation manuelle et automatique : une évaluation comparative basée sur un corpus en langue française.* In Actes de Coria, pages 9–23, Grenoble, 2005.

[Savoy, 1993] J. Savoy. *Stemming of French words based on grammatical categories*. Journal of the American Society for Information Science, 44(1), p. 1-9, 1993.

[Schmandt, 1994] C. Schmandt, *Voice Communication with Computers (Conversational Systems)*, New York, 1994.

[Scuturici, 2002] M. Scuturici. *Contribution aux Techniques Orientées Objet de Gestion des Séquences Vidéo pour les Serveurs Web*, Thèse de Doctorat, Institut National des Sciences Appliquées, Lyon, 2002.

[Sean, 2003] B. Sean, . M. Ralf and C. Peter, *The DIG description logic interface: DIG/1.1*, in Proceedings of the 2003 International Workshop on Description Logics, Rome Italy, 2003.

[Seco, 2004] N. Seco, T. Veale & J. Hayes, *An intrinsic information content metric for semantic similarity in Wordnet*, In Proceedings of ECAI'2004, the 16th European Conference on Artificial Intelligence, 2004

[Singhal et al., 1996] A. Singhal, C. Buckley, M. Mitra. *Pivoted document length normalization*. In Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval . Zurich, Switzerland .Pages: 21 - 29 . 1996

[Sukkar, 1996] R. A. Sukkar, A. R. Setlur, M. G. Rahim and C.-H. Lee, *Utterance verification of keyword strings using word-based minimum verification error (WB-MVE) training*, In Acoustics, Speech, and Signal Processing, ICASSP-96, pp. 518—521, 1996

[Sy, 2012] M.-F. Sy, *Utilisation d'ontologies comme support à la recherche et à la navigation dans une collection de documents*, Thèse de Doctorat, Université Montpellier II, 2012.

[Szoke, 2008] I. Szoke, L. Burget, J. Cernocky and M. Fapso, *Sub-word modeling of out of vocabulary words in spoken term detection, I*n Spoken Language Technology Workshop, Goa, India, IEEE, pp. 273—276, 2008.

[Szoke, 2005] I. Szoke, P. Schwarz, P. Patejka, L. Burget, M. Karafiat, M. Fapso and J. Cernocky, Comparison of keyword spotting approaches for informal continuous speech., In Interspeech, Lisbon, Portugal, pp. 633—636, 2005

[Tambellini, 2007] C. Tambellini, *Un système de recherche d'information adapté aux données incertaines : adaptation du modèle de langue*. Thèse de Doctorat, Université de Grenoble I - France, 2007.

[Thambiratnam, 2005] K. Thambiratnam and S. Sridharan, *Dynamic Match Phone-Lattice Searches For Very Fast And Accurate Unrestricted Vocabulary Keyword Spotting*, In ICASSP, pp. 465—468, 2005

[Tchechmedjiev, 2012] A. Tchechmedjiev, *État de l'art : mesures de similarité sémantique locales et algorithmes globaux pour la désambiguïsation lexicale à base de connaissances*. In Proceedings of the Joint Conference JEP-TALN-RECITAL'12, volume 3, 295-308, 2012

[Young, 2008] S. Young, *HMMs and Related Speech Recognition Technologies*, In Springer Handbook of Speech Processing, Heidelberg, Berlin, Springer-Verlag, 2008, pp. 539-583.

[Young, 2015] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, "The HTK Book Version 3.5". University of Cambridge, 2015

[Young, 1994] S. J. Young and P. C. Woodland, *State clustering in hidden Markov Model-based continuous speech recognition*, In Computer Speech and Language, 1994.

[Young, 1993] S. J. Young and P. C. Woodland, *The use of state tying in continuous speech recognition*, In EuroSpeech, 1993.

[Youssoufou, 2009] S. Youssoufou. *Un système pour l'annotation semi-automatique des vidéos et application à l'indexation.* Thèse de Doctorat, Université du Québec à Trois-Rivières, 2009.

[Yu, 2004] K. Yu and M. J. F. Gales, *Adaptive training using structured transforms*, In Proceedings of ICASSP, 2004.

[Vergyri, 2007] D. Vergyri, I. Shafran, A. Stolcke, V. R. R. Gadde, M. Akbacak, B. Roark and W. Wang, *The SRI/OGI 2006 spoken term detection system*, In INTERSPEECH, Citeseer, pp. 2393—2396, 2007,

[Walker, 2004] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, J. Woelfel, *Sphinx-4: A flexible open source framework for speech recognition*, Tech. Rep., Sun Microsystems Inc., 2004.

[Wallace, 2010] R. Wallace, R. Vogt, B. Baker, and S. Sridharan, *Optimising figure of merit for phonetic spoken term detection*, In Procedding of ICASSP , pp. 5298—5301, 2010

[Wang, 2009] D. Wang, *Out-of-Vocabulary Spoken Term Detection*, Thèse de Doctorat, University of Edinburgh, Decembre 2009

[Weintraub, 1997] M. Weintraub, F. Beaufays, Z. Rivlin, Y. Konig and A. Stolcke, *Neural-network based measures of confidence for word recognition*, In ICASSP, IEEE, p. 887, 1997.

[Wer, 2012] L. v. d. Wer, *Evaluation of Noisy Transcripts for Spoken Document Retrieval*, Thèse de Doctorat , Center for Telematics and Information Technology, Enschede, The Netherlands, 2012.

[Woodland, 2002] P. C. Woodland and D. Povey, *Large scale discriminative training of hidden Markov models for speech recognition*, Computer Speech & Language, no. 16, pp. 25-47., 2002.

Productions Scientifiques

- ✓ Bendib I, Laouar M.R., A semantic indexing approach of multimedia documents content based partial transcription. In Proceeding of 2nd International Conference on Natural Language and Speech Processing ICNLSP'18 – IEEE, Directorate General for Scientific Research and Technological Development, Algeria, 2018,
- ✓ Bendib I, Khelifa, B Laouar M.R., Ontology based semantic content indexing for spoken documents. In Proceeding of International Conference on Pattern Analysis and Intelligent Systems PAIS'15, Tebessa, Algeria, 2015.
- ✓ Bendib I, Laouar MR, Hacken R and Miles M, *Semantic ontologies for multimedia indexing (SOMI): Application In the e-library domain*, In Library Hi Tech, V: 32, N°2, pp=206--218, 2014.
- ✓ Bendib I, Laouar M.R., *Hybrid Method with Confusion Network for Indexing Spoken Documents*. In Proceeding of the Third International Conference on Digital Information Processing and Communications ICDIPC'13, Islamic Azad University (IAU), Dubai, UAE, 2013
- ✓ Bendib I, Laouar M.R., *Approaches for the Detection of the Keywords in Spoken Documents Application for the Field of E-Libraries*. In: Huang T., Zeng Z., Li C., Leung C.S. (eds) Neural Information Processing. ICONIP'12. Qatar, Doha, Lecture Notes in Computer Science, vol 7666. Springer, Berlin, Heidelberg, 2012.