HAL
open science

Approche par patrons linguistiques pour la détection automatique du locuteur : application à l'indexation par le contenu des journaux télévisés

Mbarek Charhad, Georges Quénot

To cite this version:

Mbarek Charhad, Georges Quénot. Approche par patrons linguistiques pour la détection automatique du locuteur : application à l'indexation par le contenu des journaux télévisés. Compression et Représentation des Signaux Audiovisuels (CORESA'05), 2005, Rennes. hal-00954057

HAL Id: hal-00954057

https://inria.hal.science/hal-00954057

Submitted on 3 Mar 2014

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

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

Approche par patrons linguistiques pour la Détection Automatique de l'Identité du Locuteur : Application à l'indexation par le contenu des journaux télévisés

Mbarek Charhad[1] Georges Quénot[1]

Laboratoire CLIPS-IMAG

385, rue de la Bibliothèque, B.P. 53,
38041 Grenoble Cedex

{Mabrek.Charhad, Georges.Quenot}@imag.fr

Mots clés : indexation et recherche vidéo, transcription automatique de la parole, segmentation par locuteur, patrons linguistiques.
Concours jeune chercheur : Non

Approche par patrons linguistiques pour la Détection Automatique de l'Identité du Locuteur : Application à l'indexation par le contenu des journaux télévisés

Résumé

L'identité des personnes dans les documents audiovisuels représente une information sémantique importante pour un processus d'indexation et de recherche par le contenu. La tâche de détection de l'identité des locuteurs peut être réalisée en exploitant des éléments d'informations issues de différentes modalités (texte, image et son).

Dans cet article, nous proposons une approche pour l'indexation de l'identité des locuteurs dans les journaux télévisés en exploitant le contenu audio. Après une phase de segmentation en locuteurs, une identité est attribuée à des segments de parole par l'intermédiaire de patrons linguistiques appliqués à leur transcription produite par reconnaissance vocale. Trois types de patrons sont utilisés pour prédire l'identité du locuteur dans les segments précédents, courants ou suivants. Ces prédictions sont ensuite propagées à d'autres segments par similarité au niveau acoustique. Des évaluations ont été menées sur une partie du corpus TREC 2003 : une identité de locuteur a pu être attribuée à 53% du corpus annoté avec une précision de 82%.

Mots clefs

Indexation et recherche vidéo, transcription automatique de la parole, segmentation par locuteur, patrons linguistiques.

1 Introduction

La manipulation des documents audiovisuels dans le contexte de la recherche d'information nécessite des nouveaux outils pour exploiter le contenu. Parmi ceux-ci on trouve par exemple la détection et l'identification des locuteurs. Il s'agit de déterminer qui parle afin d'associer à chaque segment audiovisuel l'identité du locuteur approprié.

La variété des données contenues dans un document audiovisuel engendre généralement une multitude possibilités de clés d'indexation et de recherche. Une clé d'indexation peut être par exemple un simple mot clé prononcé dans le discours, comme elle peut être un thème (sport, commerce, etc.)[5]. Dans ce cas tous les mots relatifs à ce thème seront recherchés. Une autre clé d'indexation qui nous paraît très utile consiste à retrouver les segments audiovisuels où un locuteur particulier parle. Dans le cas d'un journal télévisé, il est plus facile d'appliquer les techniques de suivi et de détection des locuteurs. En effet, ce genre de document audiovisuel possède une structure bien définie. Ce qui permet de bien exploiter les informations contenues dans le document dans des applications variées telles que l'indexation et la recherche d'information.

Les approches basées sur l'utilisation de la transcription et de la segmentation de l'audio sont particulièrement intéressantes car elles permettent d'indexer les locuteurs sans avoir à construire de modèles acoustiques de ceux-ci et donc de construire des systèmes ne nécessitant pas d'apprentissage spécifique. Elles peuvent en outre être appliquées à des documents purement audio. A l'inverse, un système basé sur l'utilisation de la piste image des vidéos nécessite un apprentissage des visages à reconnaître et la capacité de détecter une activité vocale (par analyse du mouvement des lèvres par exemple). Il est limité aux personnes apprises et pourrait manquer les passages dans lesquels une personne parle mais n'est pas vue ou bien elle est vue mais sous un angle ou à une échelle inappropriée pour la méthode.

Alternativement, encore la reconnaissance de texte par OCR *(Optical Character Recognition)* dans la piste image des vidéos pourrait donner une information utile mais la performance actuelle des systèmes apparaît encore insuffisante et il n'est pas non plus facile d'associer un nom affiché à l'image à un locuteur.

L'approche d'indexation du locuteur proposée ici est basée sur l'utilisation de patrons linguistiques. Elle est similaire à celle proposée dans [10]. La recherche d'expressions régulières paramétrables dans la transcription permet de proposer une identité de locuteur pour les segments dans lesquels elle apparaît, pour les segments précédents ou pour les segments suivants. Cette approche a été complétée par la propagation des identités trouvées aux autres segments non encore indexés et qui sont identifiés au niveau acoustique comme provenant d'un même locuteur. Des évaluations ont été faites en utilisant une partie du corpus TREC vidéo 2003.

La suite de cet article est structurée comme suit : dans la section 2, nous présentons quelques travaux de recherche sur l'exploitation de l'audio dans un document audiovisuel pour des applications variées. Dans la section 3, nous détaillons notre approche de détection de l'identité du locuteur par patrons linguistiques. Nous présentons nos résultats d'évaluations dans la section 4. La section 5 de cet article porte sur une discussion et l'application de cette approche. Nous évoquons en conclusion quelques perspectives que nous souhaitions développer ultérieurement.

2 Contexte du travail

L'audio, constitue une composante du document vidéo ayant un contenu très riche et qui représente en même temps une multitude clés d'indexation au niveau signal et aussi au niveau sémantique. L'information auditive constitue une base d'un grand nombre d'approches proposées dans la littérature. Ces approches se distinguent par la spécificité du cadre applicatif généralement très restreint à des applications données.

L'audio possède suffisamment des caractéristiques de manière qu'il constitue en lui-même le fondement de plusieurs travaux de recherche [1]. D'un point de vue contenu sémantique, la parole est plus sollicitée que d'autres types d'informations, tels que par exemple la musique ou le bruit [7]. Dans [11], les auteurs proposent une méthode de suivi et de détection de locuteurs différents dans le cas des conversations en se basant sur le modèle de Markov caché. Ils exploitent les résultats de détection pour pouvoir segmenter le contenu de documents audiovisuels.

Un autre aspect d'exploitation de l'audio consiste à la reconnaissance automatique de la parole [3], [6]. Il s'agit d'une transcription textuelle de la parole contenue dans le document qui permet d'avoir énormément d'information sur le contenu. Par ailleurs ceci permet d'effectuer des recherches efficaces sur de simple information textuelle mais s'adressant à des bases d'informations audiovisuelles. La technique de transcription automatique a été mise en place dans le cadre de plusieurs travaux de recherche. Citons par exemple le système ANTS[1] [2]. Ceci rend plus facile l'exploitation du contenu audio. En effet, les résultats de la transcription engendrent un document audiovisuel segmenté automatiquement par locuteur. Chaque segment contient la parole d'un seul locuteur non identifié pour le moment. Ce que nous proposons dans cet article, consiste à exploiter ces résultats pour pouvoir identifier le nom du locuteur.

2.1 Structure d'un journal télévisé

Un journal télévisé possède une structure bien définie. Cette structure peut aider à la génération des clés d'indexation. Le concept locuteur constitue l'acteur principal dans ce genre de document. Nous voulons exploiter cet élément d'information pour étiqueter les différents segments du journal. D'une manière générale, il existe deux façons pour segmenter le contenu d'un journal télévisé : La première se base sur la découpe du contenu thématique. Un journal sera vu comme étant une suite des thèmes (politique, sport, météo) séparés souvent par des jungles, publicités etc. Par contre la deuxième possibilité exploite la structure plateau / reportage pour segmenter le contenu du journal. C'est souvent au cours de cette deuxième forme de segmentation qu'il y a transition et changement de locuteur. La figure suivante (fig. 1) montre un exemple d'organisation du contenu d'un journal télévisé.

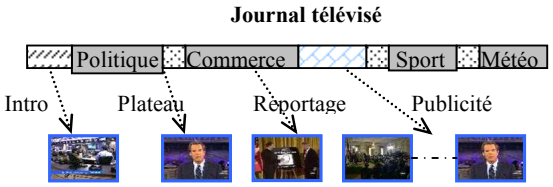

Figure 1- Structure d'un journal télévisé

En partant de cette structure, nous désirons identifier pour chaque segment le nom de locuteur. Nous exploitons pour cela les résultats de la transcription automatique de la parole fournis par le LIMSI[2]. Chaque transcription est segmentée suivant les locuteurs.

2.2 Segmentation

L'objectif de la segmentation du contenu audio est d'obtenir des segments plus cohérents. Chaque segment a un contenu spécifique : musique, bruit ou parole. Dans la littérature, plusieurs travaux sur la segmentation automatique de la parole ont été proposés. Kwon et Narayanan par exemple ont proposé dans [12] une technique pour la segmentation du document en se basant sur le calcul de distance pondéré des points de changement de locuteur dans le flux audio. L'importance de ce type d'approche apparaît surtout lorsqu'il s'agit de détecter plusieurs locuteurs différents. D'autres techniques plus génériques de segmentation exploitent les caractéristiques du contenu audio pour segmenter le contenu telles que par exemple détecter les passages audios contenant un silence significatif ou bien aussi la détection des jungles qui peuvent aussi inférer des transitions.

Nous nous intéressons à la première catégorie de segmentation (segmentation selon le locuteur) pour la mise en œuvre de notre approche. Cette catégorie de segmentation est aussi considérée comme une tâche relevant du domaine de traitement automatique de la parole.

2.3 Les patrons linguistiques

En plus de la structure physique (plateau / reportage), un journal télévisé possède aussi une structure linguistique lorsqu'il y a un changement du locuteur. En effet, les journalistes utilisent souvent des expressions verbales très marquées dans leurs discours pour par exemple passer la

ANTS [1] (Automatic News Transcription System) développé dans le cadre du projet ESTER

[2] LISMI Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur

parole d'un locuteur à un autre. Le cas où un passage entre deux locuteurs n'est pas marqué est rare. Nous voulons exploiter cette structure dans l'objectif de mettre en place une technique de détection et de reconnaissance automatique des identités des locuteurs. L'ensemble des expressions verbales sera considéré comme étant des patrons linguistiques.

L'avantage d'utiliser ce type de patrons c'est qu'ils permettent de distinguer facilement les identités des personnes citées dans le discours de celles des locuteurs. Pour le moment, nous appliquons l'approche de détection sur la transcription automatique de la parole contenue dans le journal. Cependant, il reste un grand nombre de cas ambigus qui résultent des erreurs générées lors de la phase de reconnaissance automatique de la parole (ASR). Par exemple, s'il y a une erreur sur l'identité d'une personne ou bien sur un patron linguistique, nos résultats le seront aussi.

Nous proposons de classer les patrons en trois catégories : la première concerne les patrons pour la détection d'identité du locuteur du segment précédent, la deuxième catégorie contient les patrons qui permettent de détecter l'identité du locuteur dans le segment courant et enfin la troisième catégorie des patrons permet d'identifier l'identité du locuteur dans le segment suivant. Nous détaillerons ces trois catégories dans la section suivante.

3 Détection de l'identité du locuteur par patrons linguistique

La détection de l'identité du locuteur se fait à partir de la transcription de ce qui est dit et à partir d'une segmentation en locuteurs de la reconnaissance de la parole [9] et/ou par un système de segmentation en locuteurs [4]. Outre la transcription et la segmentation en locuteur, il est nécessaire de disposer de connaissances générales permettant d'identifier la présence d'entités nommées. Des listes de noms, de prénoms, de lieux et d'organisations peuvent être utilisées. Nous avons récupéré la plupart d'entre elles depuis des sources publiques disponibles sur le Web (par exemple, Wikipedia).

La figure 2 montre la structure d'un fichier intégrant les informations issues de la reconnaissance de la parole et de la segmentation en locuteurs. Il est à noter que les résultats de segmentation ne contiennent pas d'information sur l'identité des locuteurs (aucun modèle acoustique de locuteur n'est utilisé) mais seulement le résultat d'une détection des transitions d'un locuteur à un autre suivie de la fusion des segments ainsi produits par similarité acoustique.

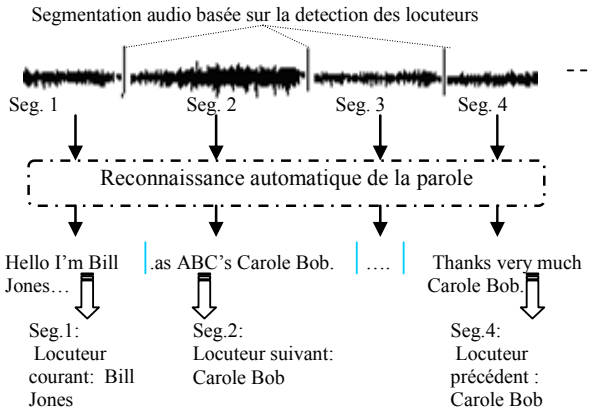

Figure 2- Cas de segments audios contenant des patrons

3.1 Affectation directe

Les patrons linguistiques sont appliqués à des morceaux de texte qui correspondent à la transcription du document audiovisuel restreinte à un segment issu de la segmentation en locuteurs. Ils correspondent à des expressions régulières, en général paramétrables, censées préciser l'identité de la personne qui vient de parler, qui parle ou qui va parler. Ils sont appliqués à chaque segment et, lorsqu'ils sont détectés, ils permettent de faire une prédiction de locuteur pour le segment précédent, le même segment ou le segment suivant. Une catégorie de patrons est définie pour chacun de ces trois cas :

- La première catégorie permet de détecter l'identité du locuteur qui est entrain de parler. Par exemple, le locuteur se présente : « …this is c.n.n news i'm [nom] …. » ou bien, lorsqu'il s'agit d'un reportage, généralement à la fin, la personne qui parle mentionne son identité.
- La deuxième catégorie contient les patrons pour détecter l'identité de locuteur qui vient de parler juste avant le locuteur actuel, par exemple « thank you very much [nom] … ».
- La troisième catégorie permet de détecter l'identité de la personne qui va parler (locuteur du segment suivant). Ce passage est souvent exploité lors passage plateau / reportage. Par exemple à la fin du discours du présentateur on trouve une expression de type « …..[nom] has the latest … ».

À chaque catégorie correspond une liste de patrons de détection. L'utilisation des patrons permet de distinguer entre l'identité d'une personne mentionnée dans le discours de celle d'un locuteur. Le tableau 1 décrit des exemples de patrons de détection, que nous avons utilisés dans notre approche. Ces patrons sont en partie spécifiques au corpus sur lequel nous travaillons (TREC vidéo 2003 et 2004, journaux télévisés de CNN et ABC).

Patron précédent (SP)	Patron courant (SC)	Patron suivant (SS)
thank you … (nom) thanks …(nom) (nom) reporting good morning (nom)	(nom) for a.b.c. news I'm (nom) (nom) c.n.n (nom) a.b.c.	tonight with (nom) a.b.c.'s (nom) (nom) reports good morning (nom)

Tableau 1- Exemple de catégories de patrons

Le patron linguistique « good morning » est souvent exploité pour signaler soit un locuteur du segment précédent soit un locuteur du segment suivant. La position de ce patron dans le segment est donc capitale. En effet, si ce terme apparaît à la fin du segment, il s'agit alors d'un patron pour détecter un locuteur dans le segment suivant.

Par contre dans le cas où ce patron apparaîtrait au début du segment, il s'agit alors d'un patron pour détecter un locuteur dans le segment précédent.

Les patrons sont tous paramétrés. Ils ont tous au moins un paramètre correspondant à un nom, un prénom ou un couple prénom+nom. Ils peuvent aussi contenir des paramètres correspondant à des lieux géographiques, à des organisations (CNN ou ABC par exemple) ou à des formules comme « good morning », « good evening », « thanks ».

Pour chaque segment, nous analysons le contenu afin de détecter des patrons déduisant l'apparition du nom du locuteur. Les patrons sont généralement situés à la fin du segment audio. Par exemple, le patron suivant infère un passage de la parole entre le présentateur du journal et un reporter : « …a.b.c's Sheila Macvicar has the latest ».

Pour la reconnaissance de prénom de la personne dans chaque segment, notre approche se base sur une liste de mots contenant environ 12400 prénoms classés en prénom femme / prénom homme. Par contre, pour la reconnaissance de nom de famille, nous exploitons une liste de noms communs pour filtrer les termes correspondants à des noms de famille, généralement. Le nom apparaît toujours après les prénoms.

3.2 Affectation par propagation

Dans le cas d'un journal télévisé, le locuteur se présente une seule fois lors de sa première intervention. Pour cela, exploiter uniquement des patrons linguistiques pour détecter l'identité des locuteurs, parait insuffisant. Pour cela, nous avons spécifié en complément de l'affectation directe, un cas d'étude qui consiste à propager les résultats obtenus par application des patrons sur le reste des segments.

Nous exploitons pour cela des informations générées dans l'étape de segmentation et détection de changement du locuteur. Ces informations sur les locuteurs qui sont de type spkr # (où spkr désigne le locuteur et le symbole '#' indique l'indice de locuteur). Rappelons que lors d'un processus de transcription automatique de la parole, quelques informations sont automatiquement générées par le système. Parmi ces informations, on trouve par exemple la référence du locuteur (homme ou femme) et aussi un indice qui nous va nous permettre d'identifier les locuteurs qui apparaissent plus qu'une seule fois dans le document. La balise suivante montre un exemple d'entête de segment audio transcrit. Dans cette balise, l'information *"FS3"* indique que le locuteur est une femme (*FS : female speech*). Le chiffre trois indique l'indice de locuteur.

```
<SpeechSegment spkr="FS3" stime=".." etime="..">
…
<Word stime=".." dur=".." conf=".."> in </Word>
<Word stime=".." dur=".." conf=".."> the </Word>
<Word stime=".." dur=".." conf=".."> fog </Word>
…

 </SpeechSegment>
```

4 Évaluation

Les évaluations ont été effectuées sur une partie de la collection TREC vidéo 2003 (quatre journaux télévisés : deux de CNN et deux d'ABC d'une demi-heure environ chacun). Le tableau 2 résume les caractéristiques du corpus effectivement utilisé.

Durée totale	7009.0 s
Durée totale de parole	5249.1 s
Durée totale des journaux	4250.3 s
Parole dans les journaux	3767.1 s
Locuteurs annotés dans les journaux	3677.5 s

Tableau 2- Caractéristiques du corpus

Durée totale représente la durée totale du document vidéo y compris les génériques (musique) et les annonces publicitaires etc. la durée totale des journaux est la durée sans compter les passages musicaux par contre la ligne parole dans les journaux représente uniquement la durée la parole dans le journal télévisé c'est à dire sans la parole dans les passages publicitaires.

Le tableau ci-dessous (tableau 3) montre les résultats obtenus pour l'indexation de l'identité du locuteur par patrons linguistiques. Les types de patrons « locuteur précédent », « locuteur courant » et « locuteur suivant » permettent de faire une prédiction dans 1.0 %, 6.8 % et 7.0 % des cas (en durée de parole) respectivement, soit une

prédiction « directe » dans 14.8 % des cas. La propagation de ces prédictions aux autres segments attribués à un même locuteur au niveau acoustique permet de faire une prédiction dans 52.7 % des cas. Les précisions obtenues sont respectivement de 100 %, 90.2 %, 74.1 %, 83.3 % et 82.4 %. Il est à noter que la propagation a introduit assez peu d'erreurs tout en augmentant de manière très significative la proportion de prédiction. Les erreurs sont dues à une mauvaise détermination de la position des extrémités des segments de parole, à une mauvaise association de segments à un même locuteur ou à une erreur dans la reconnaissance du nom d'une personne. Trois confusions ont été faites par le système de reconnaissance : « Dean Reynolds » pour « Tim Reynolds », « Jim Wooten » pour « Jim Wutton » et « Shimbun Darrow » pour « Siobhan Darrow ». Pour mesurer les effets respectifs des différentes sources d'erreurs, nous avons refait l'évaluation en corrigeant manuellement ces erreurs sur les noms (toutes les autres erreurs de transcriptions étant laissées telles quelles). La précision avant propagation est très nettement améliorée mais la propagation introduit alors plus d'erreurs. Diverses sources d'informations peuvent permettre de filtrer ou de corriger ces erreurs sur les noms (par exemple, les noms proposés ne correspondent pas à ceux de personnes connues).

Prédiction	**Durée prédite**		**Durée correcte**	
Précédent	37.2	1.0 %	37.2	100 %
Courant	250.3	6.8 %	225.9	90.2 %
Suivant	258.2	7.0 %	191.4	74.1 %
Directe	545.8	**14.8** %	454.6	**83.3** %
Propagation	1936.8	**52.7** %	1595.9	**82.4** %

Tableau 3- Résultats pour l'indexation de l'identité des locuteurs. Le pourcentage de durée prédite est relatif à la durée totale ayant pu être annotée manuellement. Le pourcentage de la durée correcte est relatif à la durée prédite.

5 Discussion et travaux futurs

La transcription de la parole est une source d'information sémantiquement très riche. Dans le cas de documents audiovisuels, il existe généralement une relation entre l'information auditive (ce qui est dit) et l'information visuelle (ce qui apparaît à l'écran).

Dans l'approche décrite tout au long de cet article, nous avons exploité les résultats de la transcription de la parole. La détection et la reconnaissance d'identité de locuteur s'intègrent dans une problématique de modélisation sémantique du contenu audiovisuel pour l'indexation et la recherche par le contenu.

Nous envisagerons de poursuivre cette démarche afin d'extraire plus d'information sur le locuteur tel que par exemple détecter le lieu (nom de ville, nom de pays) ou bien détecter le thème (de quoi il parle), la fonction, etc. D'autre part, il est possible d'exploiter les résultats obtenus par notre système pour pouvoir détecter les changements de sujet (segmentation en histoires).

Pour exploiter cette approche dans le cadre d'un système d'indexation par le contenu des documents audiovisuels. Il est intéressant d'intégrer l'information auditive et visuelle de manière similaire à ce qui a été proposé dans [8]. Cependant, dans notre approche, à chaque nom d'une personne identifiée ne correspond pas forcement une image dans le flux visuel. Grâce à la structure prédéfinie de journal télévisé (plateau / reportage), il est possible d'estimer, pour les identités détectées, les conditions dans lesquelles on peut voir et entendre la personne en même temps. C'est le cas par exemple pour le présentateur du journal. Cette information est nécessaire surtout pour évaluer la pertinence des réponses du point de vue utilisateur et aussi l'exactitude des résultats de recherche surtout lorsqu'il s'agit des locuteurs peu connus par le public.

L'alignement temporel de la transcription audio facilite le processus de détection et permet surtout de savoir à quel moment il y a eu changement de locuteur. Par similarité avec le flux visuel, cette étape est semblable à la détection des transitions entre les plans. D'autre part, pour le contenu visuel, l'unité de repérage classique est le plan vidéo. Un plan vidéo est de durée généralement courte par rapport à la segmentation de l'audio. Dans notre approche, nous proposons d'exploiter l'information temporelle permettant d'associer à chaque segment audio l'image clé qui lui correspond dans le flux visuel. Ceci permet de faciliter le jugement des résultats.

6 Conclusion

Nous avons proposé et implémenté une approche pour la détection d'identité de locuteur appliquée à la recherche par le contenu des documents audiovisuels.

Nous avons défini une liste de patrons linguistiques pour la détection d'identité de locuteur. Ces patrons sont classés dans trois catégories selon qu'ils permettent de trouver l'identité de locuteur en train de parler, qui vient de parler ou qui va parler.

La partie principale de ce travail consiste à analyser des informations textuelles générées par la transcription automatique de la parole. Plusieurs informations d'ordre sémantique sont exploitables dans la transcription telle que les lieux, les organisations ou les personnes.

Les évaluations menées sur une partie du corpus TREC vidéo 2003 ont montré que la méthode proposée permet l'affectation d'une identité de locuteur pour 53 % du temps dans les journaux télévisés avec une précision de 82 % (47 % reste non affectée, 43 % est affectée correctement et 10 % est affectée de manière incorrecte).

La reconnaissance des locuteurs représente à la fois, un apport pour l'indexation par le contenu vidéo et aussi pour la modélisation relationnelle. En effet, la reconnaissance du nom locuteur infère des relations sémantiques de type « parler » ou bien dans un contexte thématique la relation « parler de ».

Dans le cas de journaux télévisés, il est parfois très difficile de détecter l'identité de la personne en exploitant uniquement l'information auditive par exemple lorsqu'il s'agit de la parole d'un intervenant. L'identité de ce dernier apparaît souvent directement à l'écran (sous-titre) et rarement citée dans le discours. Pour cela, nous envisageons d'améliorer notre approche de détection en combinant d'autres sources d'informations telles que par exemple les résultats de reconnaissance automatique du texte dans la vidéo.

Références

[1] A. Albiol, L. Torres, E. J. Delp: Video Preprocessing for Audiovisual Indexing. In 5th IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI) April 2002.

[2] A. Brun, C. Cerisara, D. Fohr, I. Illina, D. Langlois, O. Mella, K. Smaïli : "ANTS : le système de transcription automatiquedu LORIA" Journées d'Etude sur la Parole (JEP'04) se tiendra, du 19 au 22 avril 2004 Fes, Maroc.

[3] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman: "Transcriber: development and use of a tool for assisting speech corpora production", Speech Communication special issue on Speech Annotation and Corpus Tools, Vol 33, No 1-2, January 2000.

[4] D. Moraru, S. Meignier, C. Fredouille, L. Besacier, J-F Bonastre, Segmentation selon le locuteur : les activités du Consortium ELISA dans le cadre de Nist RT03", JEP 2004, Fès, Maroc, 19-22 avril 2004.

[5] Fiscus, J., Doddington, G., Garofolo, J., and Martin, A.: Topic Detection and Tracking, Evaluation (TDT), Fifth European Conf, On Speech Comm. and Tech., Vol. 4, pp. 247-250, 1998.

[6] Garofolo John S., Ellen M. Voorhees, Cedric G. P. Anzanne, and Vincent M. Stanford: "Spoken document retrieval: 1998 evaluation and investigation of new metrics". In Proceedings of the ESCA Workshop: Accessing Information in Spoken Audio, pages 1-7, Cambridge, UK, April, 1999.

[7] J. Pinquier, C. Sénac and R. André-Obrecht: "Speech and music classification in audio documents", in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2002) Orlando, Floride, Mai 2002.

[8] J. Yang, A. Hauptmann, M-Y. Chen: "Finding Person X: Correlating Names with Visual Appearances", International Conference on Image and Video Retrieval (CIVR'04), Dublin City University, Ireland, July 21-23, 2004.

[9] J.L. Gauvain, L. Lamel, and G. Adda: The LIMSI Broadcast News transcription system, in Speech Communication 37, 2002, pp. 89-108.

[10] L. Canseco-Rodriguez L., L. Lamel, and J.L. Gauvain: "Speaker Diarization from Speech Transcripts". In International Conference on Speech and Language Processing, pages 1272-1275, Jeju Island, October 2004.

[11] M. K. Sönmez, L. Heck, M. Weintraub: ``Speaker Tracking and Detection with Multiple Speakers," in *Proc. EUROSPEECH 99*, Volume 5, Page 2219-2222 Budapest, Hungary, September 1999.

[12] S. Kwon and S. Narayanan: "Speaker Change Detection Using a New Weighted Distance Measure", In Proceedings of International Conference Spoken Language Processing. Denver, Colorado, U.S.A., September 16-20, 2002.

T. Kemp and A. Waibel: Reducing the Oov Rate In Broadcast News Speech Recognition, In Proceedings of the ICSLP, Sydney, Australia, 1998.