Identification non-supervisée de personnes dans les flux télévisés

Johann Poignant

UNIVERSITÉ DE GRENOBLE

Thèse

Pour obtenir le grade de

Docteur de l'Université de Grenoble

Spécialité : **Informatique**

Arrêté ministériel : 7 août 2006

Présentée par

Johann Poignant

Thèse dirigée par **Laurent Besacier**
et codirigée par **Georges Quénot**

préparée au sein **Laboratiore d'informatique de Grenoble**
et de **l'École Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique (MSTII)**

Identification non-supervisée de personnes dans les flux télévisés.

Thèse soutenue publiquement le **18 octobre 2013**,
devant le jury composé de :

M. Georges Linares
Professeur, Laboratoire Informatique d'Avignon, Président
M. Frédéric Béchet
Professeur, Laboratoire d'Informatique Fondamentale de Marseille, Rapporteur
M. Bernard Mérialdo
Professeur, Eurecom, Rapporteur
M. Philippe Joly
Professeur, Institut de Recherche en Informatique de Toulouse, Examinateur
M. Laurent Besacier...
Professeur, Université Joseph Fourrier, Directeur de thèse
M. Georges Quénot
Directeur de recherche CNRS, CNRS, (Membre), Co-Directeur de thèse

Remerciements

Écrire cette partie est un exercice plaisant bien sûr mais aussi délicat : comment remercier en quelques lignes pour tout ce qu'on a reçu pendant un doctorat ! ? De la simple attention aux échanges les plus riches, merci à tous.

Je tiens aussi à remercier plus particulièrement mes deux encadrants, formidable duo aux talents complémentaires, deux personnalités, deux visions, qui ont su prêter attention à toutes les idées qui ont pu germer dans la tête d'un apprenti chercheur.
Je remercie ainsi Laurent Besacier qui m'a guidé dans mes pérégrinations, pour la richesse de nos échanges et ses bons conseils. J'ai reçu de toi un soutien sans faille, merci d'avoir cru en moi. Tu m'as poussé toujours plus en avant pour que je donne le meilleur de moi.
Et Georges Quénot, qui a apporté le meilleur environnement possible pour ce doctorat, que ce soit matériel, technique ou par ses nombreuses connaissances. Grâce aux nombreuses présentations et rédactions que tu m'as incité à réaliser j'ai pu développer des compétences importantes pour ce métier.
Vous m'avez offert une liberté et une autonomie d'action très appréciable tout en m'apprenant à prioriser mes efforts en fonction des objectifs spécifiques de la recherche. Merci à vous deux, j'aurai plaisir à retravailler avec vous.

Je tiens ensuite à remercier Frédéric Béchet, Bernard Mérialdo, Georges Linares et Philippe Joly pour le temps consacré à la lecture et à l'évaluation de mon travail. Merci pour l'attention que vous y avez porté, en espérant échanger à nouveau bientôt.
Une pensée pour Sylvain Meignier qui n'a pu être présent dans ce jury. J'espère que nos chemins se recroiseront.

Comment faire de la recherche sans une équipe ? Pour moi, il y en eu deux : MRIM et GETALP. A l'image de mes directeurs, elles ont chacunes leurs personnalités mais c'est leurs complémentarités qui m'ont permis de mener à bien mon travail. A vous tous, merci pour nos échanges, votre soutien, votre aide et tout ce qui m'a permis de me construire en tant que chercheur.
Franck, tu es parti vers de nouveaux horizons, mais je n'oublie pas la sympathie que tu m'as témoigné et tes précieux conseils qui m'ont permis de bien démarrer cette aventure.
Décidément, le bureau B215 me réussit. Il a été le terrain de bien des discussions sérieuses (ou pas ...). Que ce soit avec toi Marion, mon amie de master et de thèse.

Merci pour nos déjeuners, ta bonne humeur et tes saperlipopettes. Ou avec toi, David, le G.O. du labo, que de bons moments passés et futurs (je compte sur toi) à se divertir mais aussi à travailler.
Et comment ne pas citer les Qcomperiens : Hervé, Guillaume, Viet Bac, Makarand, Sophie, Claude, Jakob et bien d'autres encore, qui ont fournit une quantité et une qualité de travail considérable pour que le projet dans lequel s'est déroulé ce doctorat se passe pour le mieux.

Et bien sur les amis, ceux qui sont à mes cotés quotidiennement, qui m'ont diverti et encouragé. Mais aussi ceux qui sont loin, pour qui j'ai toujours une pensée. Merci pour votre présence et votre soutien.
Enfin, j'ai une attention particulière à adresser à mon papa et ma maman qui ont su croire en moi envers et contre tout. Et bien sûr pour ma sœur et mon frère qui, même si vous êtes loin, je vous garde dans mon cœur. Je vous témoigne ici toute mon affection.
Voici sans doute les remerciements qui ont le plus d'importance à mon cœur, ceux pour ma femme, Audrey. Je ne pourrais jamais exprimer en quelques mots tout ce que tu mérites. Merci pour ta patience, tes précieux conseils, d'avoir été là pour m'écouter, me soutenir dans les moments de doutes, mais aussi pour tous les moments de joie que tu me procures. Avec tout mon amour, merci.

Il est très étonnant comme les dessins d'un jeune reporter, Tintin, se prête à la représentation des difficultés d'un doctorat. A Hergé pour les quelques illustrations que je lui ai emprunté.

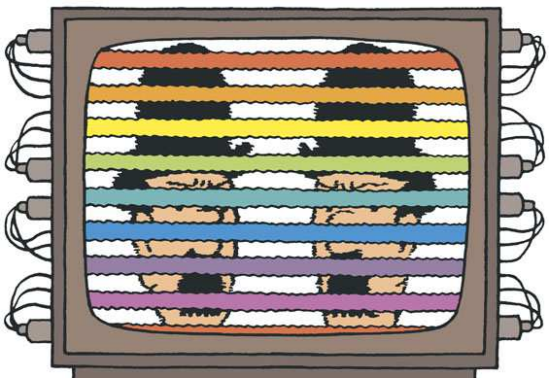
A celui qui lit ces lignes, tu ne sais pas encore ce qui t'attend...

Table des matières

Introduction

Savoir « qui a dit quoi », « qui parle à qui », « qui était où » dans de larges collections de vidéos est très utile pour fournir un accès efficace à l'information dans toutes sortes de documents audio-visuels tels que les vidéos du web, les fictions ou encore les émissions de télévisions. Par conséquent, l'identification des personnes présentes dans ces vidéos est incontournable pour la recherche et la navigation dans ce type de contenu. Cette tâche peut répondre à plusieurs besoins :

- **L'annotation automatique ou semi-automatique** identifie les visages-voix pour remplacer-aider l'annotation manuelle. Elle peut être très utile, par exemple, pour l'information aux malentendants, malvoyants, etc. Dans ce cas, elle va permettre de minimiser l'effort de post-annotation manuelle.
- **L'extraction de segments audio ou d'images de visages automatiquement** peut remplacer les annotations manuelles. Elle est très intéressante pour la construction de modèles biométriques. Cependant, il est nécessaire d'identifier les visages-voix automatiquement avec une très forte précision, tout en ayant une durée de signal extraite suffisamment longue pour construire ces modèles.
- **L'indexation automatique** peut être intéressante pour proposer plusieurs segments vidéos en réponse à une requête. La qualité de la liste d'extraits retournée doit être un équilibre entre précision et exhaustivité. Dans cette tâche, on ne va pas identifier les visages-voix mais leur présence à l'écran ou dans la bande son (par exemple, on n'a pas besoin de dire que le visage de droite est « Dupont » et celui de gauche « Dupond », on doit juste savoir qu'ils apparaissent à l'écran).



Nous nous sommes intéressés plus particulièrement aux flux télévisés, parce qu'il y a une variété de personnes à identifier importante (du présentateur à l'inconnu interviewé dans la rue en passant par l'invité du jour). Ces personnes peuvent apparaitre dans une ou plusieurs vidéos dans des contextes variés (journaux télévisés, émissions politiques ou de débats, etc).

Pour répondre aux trois besoins mentionnés précédemment, la solution la plus naturelle est de construire des modèles biométriques et d'utiliser des systèmes qui comparent chaque modèle avec la personne à reconnaître. Deux orientations sont possibles :

- **Être exhaustif** : il faut collecter suffisamment de données (c'est-à-dire de temps de signal audio pour les modèles de voix et d'images de visages pour les modèles faciaux) pour être sûr que la personne qu'on veut identifier est présente dans ces données et qu'on pourra donc construire un modèle biométrique lui correspondant. Cette solution est peu viable pour deux raisons. La première est le coût des annotations manuelles nécessaires à la construction de tous ces modèles. La deuxième est qu'un système utilisant tous ces modèles aura forcément des difficultés à choisir lequel correspond le mieux à la personne à identifier.
- **Être spécifique** : il faut réduire le nombre de modèles à ceux nécessaires. Toutefois, dans ce cas, comment avoir une connaissance a priori des personnes présentes dans les émissions de télévisions ? Pour les présentateurs, chroniqueurs, envoyés spéciaux, c'est envisageable. Ils font partie de la distribution de l'émission. Par contre, pour les personnes les plus intéressantes du point de vue de l'information (celles qui sont invitées sur le plateau de télévision, celles interviewées dans la rue, celles qui s'expriment devant une assemblée de journalistes, celles dont on veut savoir ce qu'elles ont dit, fait, etc), il est difficile d'avoir connaissance de leur présence dans une vidéo sans que quelqu'un le spécifie.

L'utilisation de modèles biométriques est donc à restreindre à quelques personnes spécifiques. Il n'est en effet pas viable, à l'heure actuelle, à cause du coût d'annotations manuelles, de construire des modèles pour toutes les personnes à identifier. D'autant plus que ces modèles sont très sensibles aux conditions d'enregistrement, aux variations dans le temps des personnes, etc. Il faudrait donc les réadapter trop souvent.

D'autres sources d'informations peuvent nous renseigner sur les personnes présentes dans une vidéo. En premier lieu, on peut utiliser les **méta-données** liées à une vidéo, comme les tags pour une vidéo du web ou les noms dans le programme télévisé de l'émission. Toutefois, ces sources sont souvent incomplètes et ne spécifient pas à quel moment une personne a dit quelque chose dans une vidéo.

Le **script** ou les **sous-titres** d'une vidéo de fiction contiennent beaucoup plus d'informations mais ces sources font appel à des annotations humaines, ce que nous voulons éviter.

D'autres modalités intrinsèques à une vidéo peuvent nous fournir cette information de manière automatique.

En premier lieu, les **noms prononcés** : le présentateur se charge d'introduire les invités en citant leur nom, la voix-off d'un journaliste va citer le nom des personnes apparaissant à l'écran, etc. Néanmoins, cette source d'information n'est pas très précise. Ce n'est pas parce qu'un nom est cité que la personne correspondante est présente. Une deuxième difficulté s'ajoute, il s'agit de la qualité d'extraction des noms prononcés. De plus, cette extraction est très dépendante de la langue utilisée.

Une autre source d'informations est utilisée par les émissions de télévision pour introduire une personne : les **noms écrits** à l'écran dans un cartouche. Un cartouche correspond à l'emplacement utilisé par une émission pour écrire un nom, en vue d'introduire la personne correspondante.

FIG. 1 – Exemples d'images du corpus *REPERE*

Si on regarde les exemples de la figure 1, on s'aperçoit que les personnages principaux de chaque extrait sont introduits à l'image (et probablement dans la bande son) par leur nom écrit en bas de chaque capture d'image.

Donc, au regard de tous ces éléments, les **axes de travail** à envisager dépendent de la définition :

- De la tâche : Qui parle et/ou apparaît
- Du média à cibler : émission télévisée, vidéo de fiction, etc.
- Des ressources à utiliser : modèles biométriques, script, noms prononcés, noms écrits, données issues du web, etc.

Deux projets ont donné un cadre à ce travail de thèse : le projet *QUAERO* avec ses sous-tâches sur l'identification des personnes dans les vidéos ; et le projet *QCompere* qui participe à la compagne d'évaluation *REPERE* (REconnaissance de PERsonnes dans les Emissions audiovisuelles).

Dans cette thèse, nous avons orienté notre travail sur l'identification des personnes invitées, interviewées, etc. parlant et/ou apparaissant dans les émissions télévisées, parce qu'elles constituent les personnes les plus intéressantes à identifier dans ce type de média. Les présentateurs et journalistes étant très bien reconnus par les systèmes à base de modèles biométriques.

Pour reconnaître ces personnes, nous avons voulu utiliser le moins d'annotations manuelles possible, et donc proposer des méthodes non supervisées. Pour la suite de ce document, « non supervisé » sous-entend la non utilisation de modèles biométriques, appris a priori à partir de données annotées manuellement pour identifier les personnes présentes. Nous nous sommes limités aux ressources contenues dans les vidéos donc nous n'avons pas utilisé de ressources externes spécifiques aux vidéos (pas de programme TV, ...) ou aux personnes à reconnaître (pas d'image de personnes issues d'une recherche web, etc).

Il y a encore quelques années, la mauvaise qualité des vidéos empêchait d'extraire correctement les noms écrits à l'écran, du à un d'un trop grand nombre d'erreurs de transcription. Donc, assez peu d'articles de l'état de l'art ont travaillé sur des méthodes de nommage à partir de cette modalité spécifiquement. Cependant, ces dernières années ont vu une augmentation de la qualité des vidéos disponibles. C'est pourquoi, il nous parait intéressant de développer des méthodes de nommage de personnes à partir des noms écrits.

Tous ces éléments nous ont donc conduit à nous poser la question suivante :

Comment nommer correctement les personnes présentes dans les émissions de télévision de façon non supervisée ?

Pour tenter d'y répondre, nous allons d'abord proposer un état de l'art sur les méthodes de nommage des personnes dans les documents audio-visuels : émissions de télévision mais aussi radios, vidéos de fictions, etc. Ensuite, nous présenterons le matériau expérimental (corpus, métriques, briques de base) sur lequel sont basés nos travaux. Le chapitre suivant présentera LOOV, un système de reconnaissance des caractères dans les vidéos que nous avons développé pour extraire les noms écrits à l'écran. Lesquels seront comparés aux noms prononcés, dans le chapitre 4, pour leur capacité à proposer le nom des personnes présentes dans les vidéos. Le chapitre 5 décrira les différentes méthodes de nommage de clusters mono-modaux (voix ou visage) que nous proposons. Dans le chapitre 6, nous présenterons l'extension d'une de nos méthodes de nommage à des clusters multimodaux (voix et visage). Le manuscrit se terminera par une conclusion et les perspectives à venir dans ce domaine.


Si vous n'avez rien de mieux à faire, venez chez nous. Nous vous montrerons les documents que nous avons déjà rassemblés pour notre enquête.
Volontiers.

Chapitre 1

État de l'art : Nommage des personnes dans les documents audio-visuels

Dans le chapitre précédent, nous avons introduit trois critères qui définissent l'identification de personnes dans les vidéos : la tâche, le média ciblé et les ressources utilisées. Bien que notre travail soit ciblé plus précisément sur certains de ces critères, notre état de l'art est un peu plus large. Il va couvrir les deux tâches (qui parle, qui apparaît) quelles que soient les personnes à identifier, bien évidement dans les émissions de télévisions mais aussi dans les vidéos de fiction (films et séries télévisées) et les émissions radiophoniques. Pour les ressources utilisées, nous avons limité l'état de l'art aux méthodes n'utilisant pas de modèle biométrique appris sur des données annotées manuellement (avec quelques exceptions lorsqu'un article proposait une utilisation intéressante de ces modèles).

Une vue globale des articles présentés est donnée par la figure 1.2. Ce graphe représente les articles (nœuds du graphe) classés par année de bas en haut avec les citations entre ces articles (liens du graphe). Un code de couleurs et de formes permet de définir la tâche cible et les modalités utilisées pour extraire les noms.

La couleur va définir la tâche à laquelle l'article essaye de répondre :

- Rouge : nommage des locuteurs
- Bleu : nommage des visages
- Violet : nommage des locuteurs et des visages

La forme va représenter la modalité utilisée pour extraire les noms :

- Rond : noms extraits de la parole
- Triangle : noms extraits des textes écrits à l'écran
- Losange : noms extraits de la parole et des textes écrits à l'écran
- Rectangle : noms extraits de données externes à la vidéo (le script d'un film, programme TV, ...)

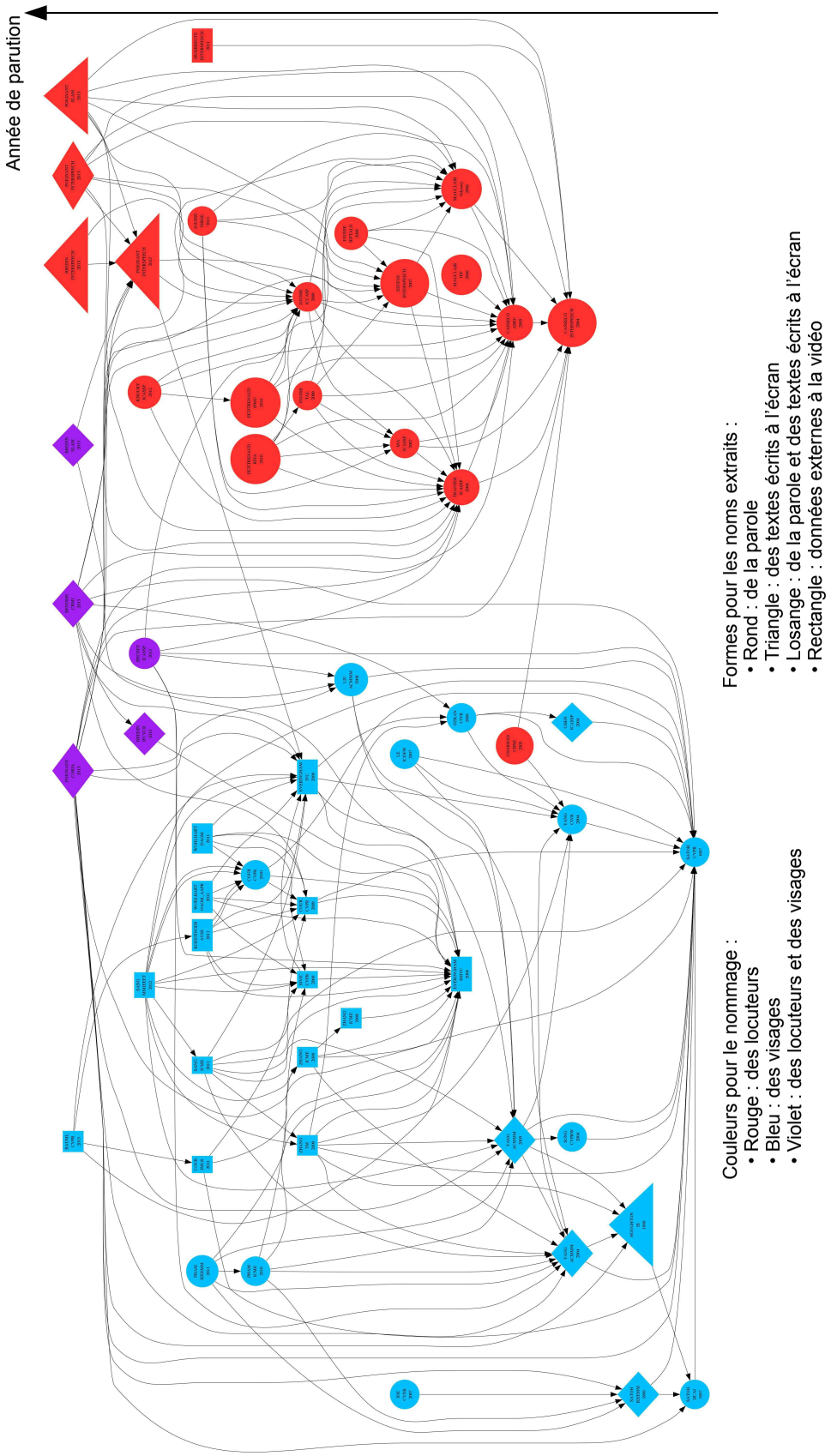


FIG. 1.1 – Connexions entre les articles de la littérature traitants du nommage des personnes dans les documents audio-visuels. Ce graphe est visible en haute définition à l'adresse suivante http://mrim.imag.fr/johann.poignant/graph_etat_de_l_art.pdf

On peut voir que ce graphe[1] est divisé entre deux communautés. La première s'est intéressée à la reconnaissance des visages (en bleu), la seconde s'est concentrée sur celle des locuteurs (en rouge). Ces dernières années plusieurs travaux ont fait des liens entre ces deux communautés. On observe aussi que les noms cités à l'oral (ronds et losanges) et les ressources externes (rectangles) sont utilisés majoritairement pour extraire les noms des personnes. C'est principalement la mauvaise qualité des images des vidéos utilisées dans l'état de l'art, qui engendre une mauvaise qualité de transcription des noms écrits à l'écran (triangles et losanges).

Cadre général

Pour nommer les personnes dans les vidéos sans modèles biométriques, un cadre général en trois étapes est généralement utilisé :

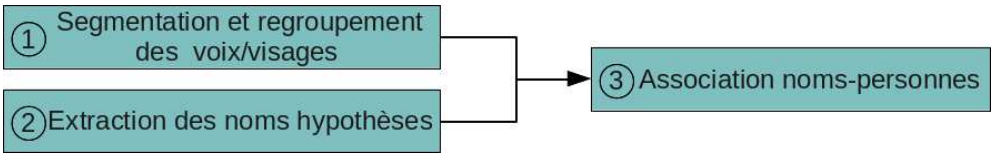


FIG. 1.2 – Cadre général en trois étapes pour le nommage non supervisé des personnes

Les deux premières étapes sont indépendantes l'une de l'autre et peuvent donc être effectuées en parallèle. La segmentation et le regroupement des voix/visages consistent à créer des clusters de personnes. Idéalement un cluster correspond à une personne et inversement. L'extraction des noms hypothèses peut être effectuée sur plusieurs flux de la vidéo (noms prononcés à l'oral, ou écrits à l'écran) ou sur des données externes (script, guide des programmes télévisés).

La dernière étape, l'association noms-personnes, est très dépendante de la modalité utilisée pour extraire les noms. Par exemple, un nom prononcé peut faire référence à une personne visible au moment de la citation ou qui le sera dans le plan suivant, précédent ou pas du tout. Un nom écrit dans le programme de l'émission de télévision fait référence à une personne présente dans la vidéo sans spécifier le moment où elle parle/apparaît. C'est donc cette étape qui va choisir si un nom fait référence à une personne et si oui à laquelle.

Ce schéma a largement été repris dans l'état de l'art avec quelques adaptations inhérentes aux sources d'informations utilisées.

Nous allons détailler ces différentes étapes à partir de l'exemple visible dans les pages suivantes. Il montre un extrait de journal télévisé avec le présentateur, un journaliste et une personne interviewée. Seulement deux modalités sont utilisées pour les noms (prononcés à l'oral et écrits à l'écran).

[1] Ce graphe a été généré automatiquement à l'aide de l'outil Graphviz [Gra]. Cet outil minimise le nombre de liens se croisant.

Sur la première capture d'image, on voit le présentateur *Jean-Rémi Baudot* parler à son envoyé spécial, *Philippe Salvador*. Ensuite, l'envoyé spécial présente un reportage. L'image suivante montre une personne, *Roland Vierne*, interviewé par des journalistes. Et enfin le présentateur clôt le reportage.

Dans cet extrait, il y a trois personnes principales. L'étape ① du processus décrit précédemment consiste à extraire les différentes images de visages/tours de parole d'une personne et à les regrouper en clusters. Ici trois clusters apparaissent avec trois couleurs différentes.

L'étape ② consiste à extraire les noms des personnes. Dans cet exemple, ces noms sont extraits de la parole mais aussi des textes écrits dans un cartouche (en bas des images). On retrouve deux des trois noms (*Philippe Salvador* et *Roland Vierne*) de nos intervenants mais aussi deux autres (*Nathalie Koscuisko-Morizet* et *Romain Garrigues*) ne correspondant pas à des personnes présentes dans cet extrait.

L'étape ③ doit sélectionner les clusters candidats pour chaque nom. Nous nous sommes limités aux visages/tours de parole co-occurrents pour les noms écrits et à ceux présents dans les plans/tours de parole contigus pour les noms prononcés. L'image ci-dessous donne le détail pour *Philippe Salvador*.

On retrouve tous les liens possibles entre les noms et les visages/tours de paroles dans l'image suivante. La première difficulté réside dans le choix des clusters candidats pour chaque nom hypothèse. Cette étape est donc très dépendante de la modalité utilisée pour extraire les noms.

Ensuite, on doit tenter de résoudre l'association entre noms et clusters de personnes en sélectionnant les liens (flèches) pertinents. L'augmentation du nombre de clusters candidats permet d'augmenter les chances de sélectionner le bon cluster mais complexifie aussi le choix à faire parmi tous ces candidats. Durant cette étape, on doit aussi choisir d'utiliser ou non un nom hypothèse.

Cette illustration, assez représentative de ce qui se passe sur l'ensemble des données que nous ciblons, nous permet de faire deux remarques :

- Le présentateur n'a pas pu être nommé à partir de cet extrait. Effectivement une personne n'est identifiable qu'à partir du moment où son nom a été proposé. L'utilisation de la vidéo complète ou encore d'informations issues d'autres modalités peuvent peut-être palier à ce manque.
- Deux noms prononcés (*Nathalie Koscuisko-Morizet* et *Romain Garrigues*) n'introduisaient pas les personnes correspondantes dans cet extrait. Ces noms peuvent gêner l'étape d'association noms-personnes et ce n'est pas parce qu'un nom est proposé comme hypothèse qu'il faut forcement nommer quelqu'un avec.

Malgré ces deux remarques, la personne la plus intéressante est identifiée, c'est *Roland Vierne*. C'est lui qui est le centre du sujet et qui apporte une information importante. C'est donc le troisième plan de l'exemple que l'on veut retrouver dans une grande collection de vidéos et non pas celui où le présentateur remercie le journaliste.

Cet exemple nous permet d'illustrer les points importants d'un système basé sur une telle approche :

- Le premier est la qualité du regroupement en clusters. Plus il est correct, c'est-à-dire avec une plus grande pureté (un cluster ne doit correspondre qu'à une seule personne), et couvrant (un cluster doit couvrir la totalité des tours de parole/images de visage de la personne), plus ce sera facile de nommer toutes les apparitions/prises de parole d'une personne. Ce clustering est basé sur l'extraction de caractéristiques uniques à chaque personne. L'amélioration de l'extraction de ces caractéristiques ne fait pas partie des objectifs de ce manuscrit. Nous ne détaillerons donc pas cet aspect dans ce chapitre.
- Le deuxième est le choix de la modalité utilisée pour extraire les noms. Outre la qualité de cette extraction, le choix des clusters candidats pour chaque nom dépend de la modalité utilisée. Nous allons donc présenter les modalités utilisées pour extraire ces noms. Et comment ont été choisis les clusters candidats pour chaque nom dans les travaux de l'état de l'art.
- Le troisième est l'étape d'association. Cette étape aussi est dépendante de la modalité utilisée pour les noms. Chacune de ces modalités a ses avantages et ses inconvénients. Nous allons donc aussi détailler comment les précédents travaux ont profité de ces avantages et contourné ces inconvénients.

Cet état de l'art essaye de faire un tour d'horizon de toutes les propositions faites afin de tirer profit des modalités utilisées ; pour surmonter les difficultés inhérentes aux types de données utilisées et aux tâches ciblées.

Comme les deux communautés (voix/visage) ont fait des avancées parallèles mais sans beaucoup de liens entre elles, nous allons d'abord détailler les travaux sur l'identification des visages dans la prochaine section, puis ceux sur l'identification des locuteurs.

1.1 Nommage des visages

On peut différencier deux « courants » dans la littérature. Le premier est apparu à la fin des années 1990-début 2000. Il propose d'identifier les visages dans les émissions télévisées (surtout les journaux télévisés) à partir des **noms cités** à l'oral. Quelques tentatives d'utilisation des **noms écrits** à l'écran sont tout de même à signaler.

Le deuxième « courant » a émergé un peu plus tard (2006), il s'est intéressé à l'identification des personnages dans les **vidéos de fictions**. Les sources principales des noms sont le **script** et les **sous-titres** liés à ce type de vidéos.

1.1.1 Premiers travaux sur les émissions télévisées

Utilisation de la redondance des co-occurrences entre noms et visages

Satoh et al. avec le système `Name-It` [SK97] ont été les premiers à introduire dans la littérature le principe d'associer un nom et un visage apparaissant à l'écran en se basant sur leurs co-occurrences. L'hypothèse est que lorsqu'un nom est prononcé (les noms sont détectés à l'aide d'un dictionnaire de noms dans les transcriptions **manuelles** de la parole), il est probable que la personne correspondant à ce nom apparaisse à l'écran dans une fenêtre temporelle centrée sur le moment de citation.

C'est la redondance des co-occurrences entre noms et visages qui permet de nommer un visage. En effet, si à chaque fois que le même nom est prononcé on voit le visage de la même personne apparaître, il y a de fortes chances pour que le nom et la personne soient reliés.

Les auteurs calculent un score de correspondance entre chaque nom et chaque visage basé sur le nombre de co-occurrences entre-eux. Ces scores sont pondérés (à l'aide d'une gaussienne) en fonction de la distance temporelle entre le moment de citation du nom et l'apparition du visage. Dans cet article, les auteurs attribuent à chaque cluster de visages une liste de trois noms classés en fonction de leurs scores et inversement (un nom pour trois clusters de visages). Ce système a été utilisé sur 9 vidéos de journaux télévisés (CNN) de 30 minutes chacune, sans réelle évaluation de la qualité.

Dans [SNK97], ces mêmes auteurs ont étendu le système `Name-It` avec l'utilisation d'informations lexicales et grammaticales permettant l'extraction des entités nommées à partir des sous-titres. Cette méthode d'extraction permet d'obtenir un score de confiance pour chaque entité nommée. Ces scores sont intégrés dans le processus d'association entre les noms et les visages.

Dans [SNK99], ce système est encore étendu avec l'extraction automatique des noms écrits en surimpression à l'écran. En raison d'un fort taux d'erreur en mots (52%), ces noms sont détectés et corrigés à partir d'un dictionnaire de noms. L'utilisation des noms écrits est limitée à la liste fermée des personnes présentes dans le dictionnaire.

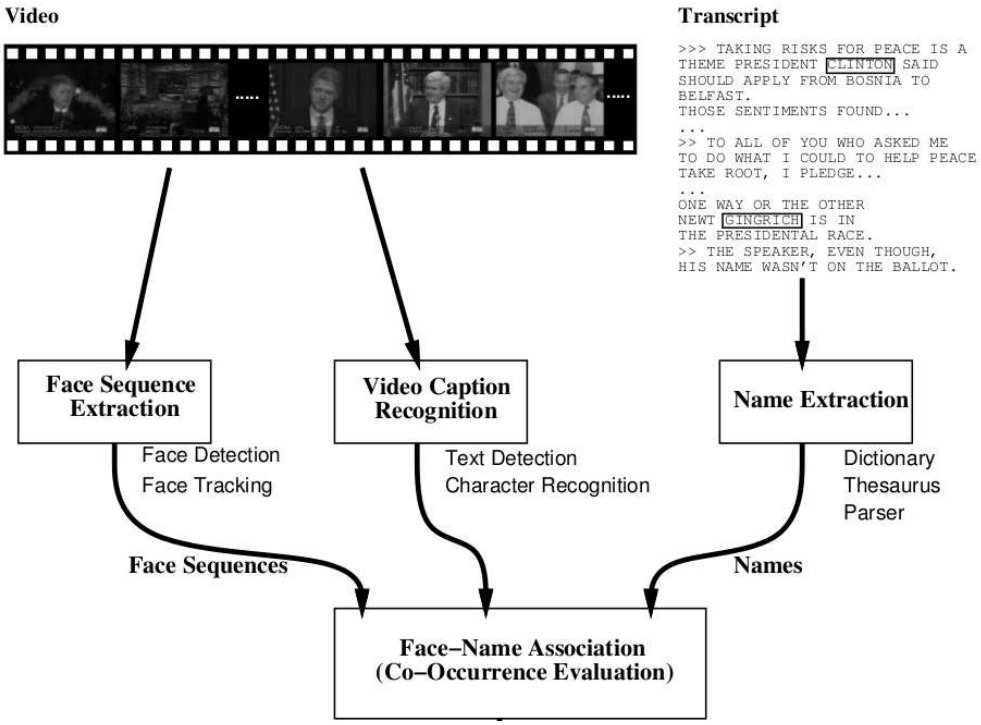


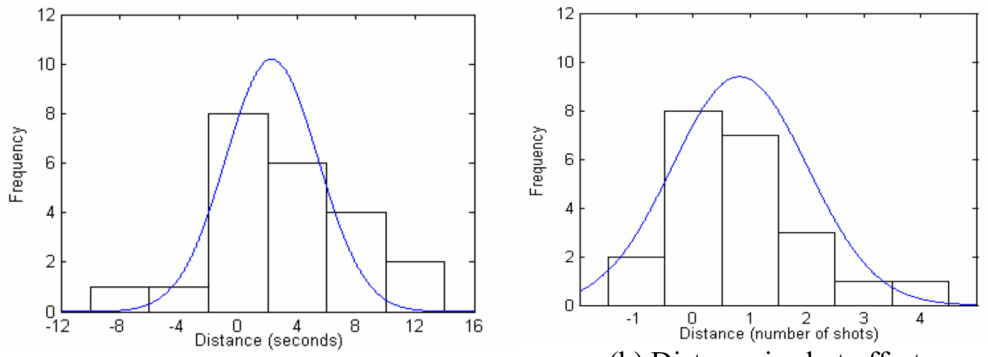
FIG. 1.3 – Vue globale du système `Name-It`. Image extraite de [SNK99]

Dans l'image 1.3, on retrouve la vue globale du système `Name-It`. L'approche suit le schéma que nous avons décrit au début de ce chapitre ; avec d'une part l'extraction des éléments d'identités à reconnaître (ici les images de visage) et d'autre part l'extraction des noms hypothèses (ici les noms écrits à l'écran et cités à l'oral). La dernière étape, symbolisée par le rectangle en bas, se charge de résoudre l'association entre les noms et les visages. Ce modèle sera largement repris par les travaux suivants dans la littérature.

En 1999, *Houghton* [Hou99] a proposé de construire une base de données de visages nommés. L'idée étant de pouvoir interroger cette base de données soit avec un nom pour obtenir une image du visage correspondant, soit avec une image de visage pour avoir le nom correspondant. Cette base de données est remplie d'images de visages associées automatiquement aux noms correspondants. Ces associations sont issues d'articles de journaux de sites web et de vidéos de journaux télévisés. Pour ces derniers, l'association entre un nom et un visage est effectuée à partir des noms écrits à l'écran (dans les sous-titres) extraits par un système de reconnaissance des caractères. Cependant, le taux d'erreur de mots de 65% sur les noms transcrits a obligé l'auteur à avoir recours à un dictionnaire de noms pour les corriger (réduction du taux d'erreur à 45%). Ce dictionnaire est construit à partir de noms détectés dans des articles de presse du web. Malgré ces corrections, le taux d'erreur reste très important et limite beaucoup l'utilisation d'informations issues des vidéos.

Quelques années plus tard, *Yang et al.* dans [YCH04] ont utilisé la même méthode que dans [SNK97] : le nommage des visages à partir des noms prononcés (extraction manuelle). Pour améliorer l'association, ils ont d'abord étudié le moment d'apparition (distribution selon une séquence de 4 secondes ou selon les plans vidéos) de 20 personnes par rapport au moment de citation de leur nom.

Sur la figure 1.4, on peut voir le décalage temporel (en secondes et en plans) entre le moment de citation du nom « *Bill Gates* » et l'apparition de son visage à l'écran. En effet, plus le plan est loin temporellement du moment de citation du nom, moins la personne correspondante a de chances d'apparaître à l'écran. Le score d'affiliation, d'un nom aux visages apparaissant dans les plans adjacents, est calculé à partir d'une courbe gaussienne. Ils ont utilisé différentes gaussiennes pour augmenter les chances d'affilier un nom au bon visage tout en réduisant l'impact des mauvaises affiliations (nom affilié à un visage autre que celui de la personne).



(a) Distance in seconds (b) Distance in shot offset

FIG. 1.4 – Décalage temporel en secondes et en plans de l'apparition de « *Bill Gates* » par rapport au moment de citation de son nom. Image extraite de [YCH04]

Ce système a été évalué sur les données de TRECVID 2003 (journaux télévisés de ABC, CNN et C-SPAN). Seulement 20 personnes cibles devaient être nommées, ce qui simplifie la tâche d'extraction des noms hypothèses à 20 noms différents possibles. Il est en effet plus facile de corriger les transcriptions si l'on sait qui va apparaître dans les vidéos.

Faciliter l'association noms-visages à l'aide de modèles biométriques

Au regard des articles présentés jusqu'à présent, une première conclusion s'impose : la simple redondance des co-occurrences entre noms prononcés et visages n'est pas une information suffisante pour nommer avec certitude un visage. Il faut donc trouver d'autres indices ou sources d'informations pour augmenter cette certitude.

C'est pourquoi *Chen et al.* [CH04] proposent de confirmer le choix fait par ce type de méthodes par le score d'un système supervisé utilisant des modèles biométriques. La première étape consiste à extraire tous les plans où est susceptible

d'apparaître une personne cible. La citation d'un nom (à l'oral ou écrit à l'écran) donne une fenêtre temporelle de la possible présence d'une personne.

Dans cette fenêtre, les plans où le présentateur a été détecté sont supprimés automatiquement. Cette détection automatique est basée sur un histogramme de couleur. Leur hypothèse est que la couleur en arrière plan du présentateur est toujours identique. Il est en effet important de supprimer le visage du présentateur comme hypothèse pour chaque nom. Le présentateur est la personne la plus visible et celle qui cite le plus de noms alors que son propre nom est très peu cité. Son visage a donc un très fort nombre de co-occurrences avec tous les noms cités à l'oral.

Sur les plans ainsi sélectionnés, *Chen et al.* ont appliqué un système supervisé de reconnaissance des visages basé sur des modèles biométriques. Enfin, pour construire la prédiction finale, ils ont utilisé une combinaison linéaire des résultats du système supervisé mais aussi des résultats du système de nommage non-supervisé. Cette méthode et les apports de chacunes des informations utilisées ont été évalués sur 65h de journaux télévisés du corpus TRECVID 2003. Toutefois, ici encore, seulement 5 personnes étaient à reconnaître, ce qui facile grandement la tâche.

1.1.2 Les méthodes à base d'apprentissage

Yang et al. [YH04] ont choisi une autre voie pour améliorer la confiance d'un système sans modèle biométrique. Ils proposent d'utiliser d'autres informations pour construire un modèle d'apprentissage SVM[2]. Ces informations supplémentaires sont à classer dans deux catégories : des caractéristiques et des contraintes.

Dans ces caractéristiques, on compte par exemple les noms écrits à l'écran, les noms cités à l'oral, des indices dans les transcriptions pour reconnaître les présentateurs et les journalistes, des informations à partir du clustering des locuteurs, la position temporelle d'un nom par rapport au plan ou encore la structure temporelle des journaux télévisés.

Les contraintes reposent sur la similarité des tours de parole et l'apparence similaire des plans (et non sur celui des visages, l'hypothèse étant qu'une personne apparaissant plusieurs fois dans la même vidéo apparaît toujours avec le même décor, les mêmes vêtements, etc.). Le SVM, entrainé sur des données annotées manuellement, fournit en sortie un score pour chacun des couples plan-nom. L'agrégation de tous les scores calculés sur la vidéo permet de choisir le meilleur nom pour un plan. Les auteurs obtiennent de biens meilleurs résultats avec cette méthode d'apprentissage qu'avec une méthode basée sur de simples règles.

Le nommage des visages dans les émissions de télévision comme un problème d'apprentissage d'instances multiples

Les méthodes d'apprentissage d'instances multiples (MIL[3]) ([DLLP97]) sont une catégorie de méthode d'apprentissage basées sur des sacs positifs (avec au

[2]SVM : Support Vector Machine
[3]MIL : multiple instance learning

moins un élément positif mais aussi des éléments négatifs) et de sacs négatifs (ne contenant que des éléments négatifs). Appliquons cette méthode à notre problème : lorsqu'un nom est cité, plusieurs images de visages (celles apparaissant autour du moment de citation) sont candidates. Or, seulement certaines sont positives (les images de visages correspondant bien au nom) alors que les autres sont négatives (les images de visages des interlocuteurs par exemple). Cette formulation permet d'avoir des données annotées automatiquement.

Une fois les sacs constitués, un algorithme (« diverse density » [MLP98]) va chercher un vecteur proche des sacs positifs et loin des sacs négatifs. Il va donc chercher les éléments dans l'intersection des sacs positifs moins l'union des sacs négatifs pour trouver les images de visages positives.

En 2004, *Song et al.* [SLS04] ont été les premiers à proposer un système basé sur MIL. Ils ont adapté cette méthode à la problématique visée en introduisant la notion de « sac quasi-positif » (un sac positif peut ne pas contenir d'instances positives). Cette formulation est basée sur l'hypothèse que lorsqu'un nom est cité, le visage correspondant au nom n'est pas forcement visible. Ils ont aussi étendu l'algorithme « diverse density » [MLP98] avec un algorithme appelé « extended diverse density » en supprimant l'influence des sacs faux-positifs (sacs positifs sans instances positives). Cet algorithme permet d'extraire des images positives de la personne à reconnaître.

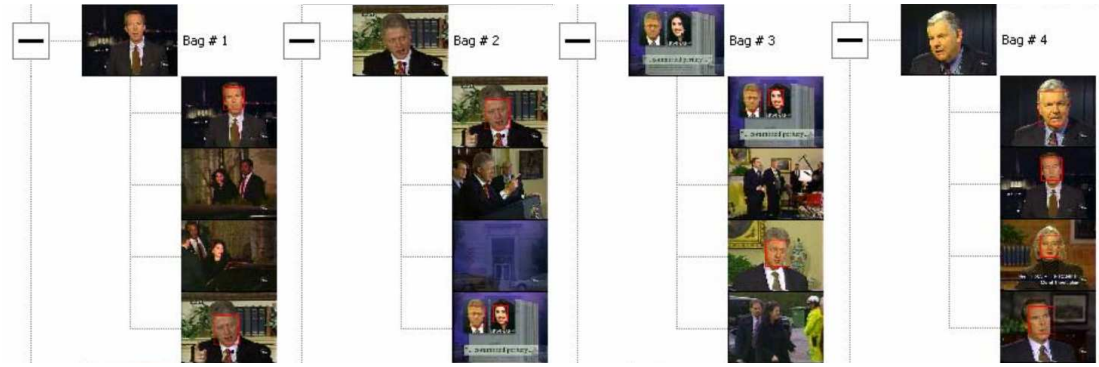


FIG. 1.5 – Exemples de « quasi positive bag » pour « *Bill Clinton* ». Les trois premiers sacs contiennent des images positives alors que le dernier non. Image extraite de [SLS04]

FIG. 1.6 – Images de visages extraites à l'aide de l'algorithme « extended diverse density » à partir des sacs de l'image 1.5. Image extraite de [SLS04]

Par exemple, dans la figure 1.5, on peut voir les « quasi positive bags » pour « *Bill Clinton* ». Le quatrième sac ne contient pas d'images de « *Bill Clinton* »

et l'algorithme ne retourne que les trois images de visages correctes (figure 1.6). Cette méthode a été testée sur les données TRECVID 2003 pour identifier « *Madeleine Albright* ». Parmi les images extraites, les 8 meilleures ont été sélectionnées (6 correctes, 2 erronées). Ces images ont permis de construire un modèle biométrique testé et comparé à un modèle construit à partir d'images de visages annotées manuellement. Toutefois, les résultats sur seulement une personne cible ne permettent pas de tirer de conclusions objectives sur la qualité d'une telle méthode.

L'année suivante, *Yang et al.* [YYH05] ont remplacé le classifieur SVM utilisé dans leurs précédents travaux [YH04] par un système MIL. Ils ont proposé trois modifications de la méthode MIL. La première, nommée « exclusive MIL », s'intéresse à la constitution des sacs. Ils proposent de n'avoir qu'un et un seul exemple positif par sac positif. Pour chacun de ces exemples, ils proposent d'utiliser plusieurs caractéristiques pour les définir (distance temporelle entre le visage et le nom, texte présent ou non à l'image où on voit le visage, ...).

Les deux autres propositions qu'ils ont faites s'intéressent à la résolution du problème MIL avec ce type de sac. La méthode, nommée « exclusive density », essaye de trouver un vecteur, dans l'espace des caractéristiques, qui pour chaque sac positif a une seule instance proche alors que chaque instance des sacs négatifs est plus loin dans l'espace. La dernière, nommée « iterative exclusive dentsity » est une modification de leur deuxième proposition pour une implémentation plus efficace.

Ces propositions ont été évaluées sur 20 journaux télévisés de 30 minutes de la chaîne ABC. La tâche étant d'identifier les personnes faisant un monologue à l'écran. Parmi tous les visages détectés, les visages des présentateurs et des journalistes n'ont pas été considérés. Dans ces vidéos, 234 personnes était identifiables avec une moyenne de 4.7 noms candidats et 242 personnes n'ont pas été nommées dans les sous-titres. Ce qui correspond à 476 sacs (234 positifs, 242 négatifs) avec 2236 instances. Les auteurs ont obtenu des résultats similaires ($\approx$ 60% de précision) à leur système à base de SVM. Par contre, ce dernier nécessitait des annotations manuelles pour pouvoir être entraîné.

1.1.3 Quelques cas d'études intéressants

C'est à partir de l'année 2006 que les propositions ont moins suivi une tendance et sont devenues plus hétéroclites. Nous allons passer en revue ces propositions dans cette sous-section.

Association noms-visages comme un graphe

Ozkan et Duygulu [OD06] ont transformé le problème de nommage des personnes en un problème de recherche du sous-graphe le plus dense, dans un graphe de similarité entre visages. Pour chaque citation d'un nom à l'oral, un graphe local est construit. Les nœuds correspondent aux visages qui apparaissent dans les plans autour de la citation de ce nom et le poids des liens correspond à la similarité entre visages.

La multiplication des citations de ce nom dans la vidéo permet de constituer un graphe global à partir des graphes locaux. Les nœuds du sous-graphe le plus dense dans ce graphe global correspondent aux images de visage de la personne correspondant au nom. Les visages des présentateurs sont aussi détectés et supprimés du graphe global.

L'évaluation a été faite sur 229 journaux de 30 minutes issus du corpus TRECVID 2004. Cependant, encore une fois, cette évaluation ne porte que sur 5 personnes cibles. Cette liste très restreinte ne nous permet pas de nous faire une idée de la qualité d'un tel système sur une liste de personnes plus large ou une liste de personnes ouvertes.

Le nommage des personnes vu comme une tâche de traduction

Pour identifier les visages des personnes importantes dans les vidéos d'une collection large (130h de journaux télévisés de ABC et CNN, TRECVID 2003), *Le et al.* [LSHN07] ont modélisé le problème en faisant l'analogie avec la traduction statistique, et plus précisément comme un problème d'alignement de « mots » source-cible en traduction. L'hypothèse est que les visages apparaissant dans une fenêtre temporelle peuvent former une phrase dans une langue source ; alors que les noms prononcés dans cette même fenêtre peuvent former une phrase dans une langue cible. A partir de phrases alignées entre les deux langues, on peut aligner les noms et les visages.

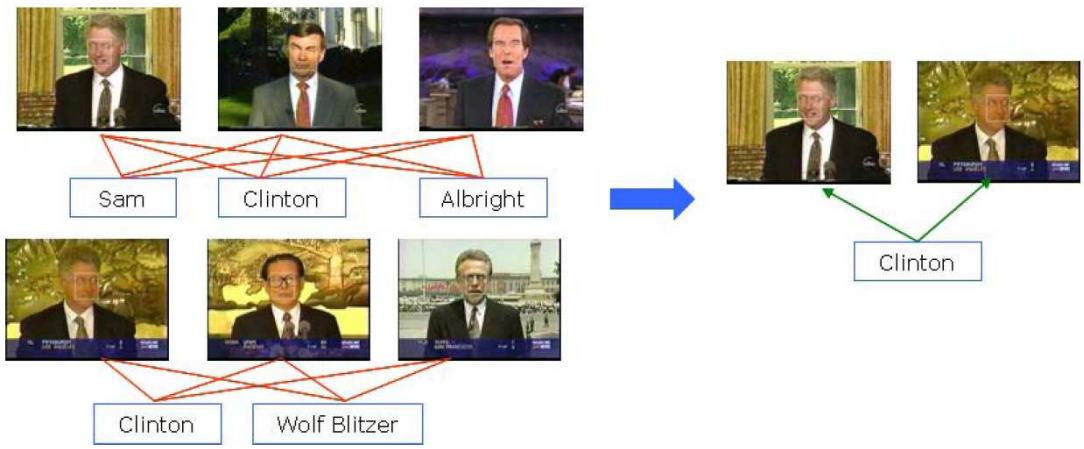


FIG. 1.7 – Image extraite de [LSHN07]

Dans la figure 1.7, on peut voir deux exemples d'alignement entre les phrases de visages et les phrases de noms. Une segmentation en histoires a permis de sélectionner 73 clusters correspondants aux personnes cibles. Ces clusters ont été générés automatiquement à l'aide du modèle RSC[4] [Hou06]. Les clusters de présentateurs (détectés automatiquement) et les clusters n'apparaissant que dans une seule histoire du corpus ont été filtrés. Les noms sont extraits automatiquement de la parole (transcription fournie avec le corpus TRECVID 2003) à l'aide

[4]RSC : Relevant-Set Correlation

de l'outil LingPipe [lin]. L'outil GIZA++ [ON03], très populaire en traduction, a été utilisé pour associer les visages et les noms.

Sur les 73 clusters, 37 ont été jugés comme importants (à partir d'un jugement humain). Seulement 11% des images n'ont pas été regroupées dans le bon cluster. 28 des 37 clusters ont bien été nommés par au moins un des trois premiers noms retournés par cette méthode.

Utilisation d'images du web

Dans [LJH08], les auteurs transforment le problème d'affiliation en utilisant une autre source d'informations pour connaître le nom d'un visage. La figure 1.8 montre le schéma général de leur proposition.

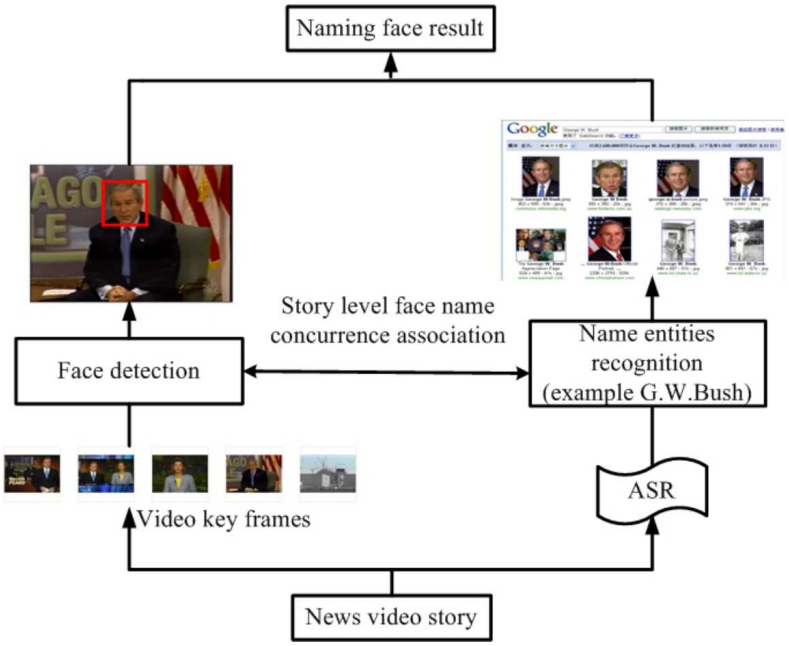


FIG. 1.8 – Image extraite de [LJH08]

Dans une première étape, ils extraient les noms de la transcription automatique de la parole ou des sous-titres. Une requête sur internet leur permet d'extraire 20 images de visages (1 seul visage de face par image) correspondant aux noms hypothèses. Après un reclassement des images selon leurs similarités, ils sélectionnent les 5 premières pour construire des modèles biométriques. Enfin, ils comparent les visages extraits de la vidéo aux modèles construits.

Ce système a été testé sur 5 journaux télévisés (3 CNN, 2 MSNBC de TRECVID 2005). L'analyse s'est concentrée sur les affaires internationales (les nouvelles locales et de divertissements ont été rejetées). 139 visages et 86 noms ont été extraits, 91 visages ont été correctement nommés et 16 mal identifiés, 32 n'ont pas été nommés.

Cette méthode d'utilisation d'annotations que l'on peut qualifier de semi-supervisée ne peut fonctionner que pour les visages de personnes connues. Les requêtes pour les personnes peu ou pas connues peuvent retourner beaucoup d'images ne correspondant pas.

Utilisation de quelques annotations manuelles

Pham et al. [PMT10, PTM11] proposent d'utiliser quelques annotations manuelles pour bien démarrer l'étape d'association noms-visages. Ils reprennent le schéma classique (utilisation des noms issus de la parole / extraction des visages, voir figure 1.9) avec en plus une étape de détection des présentateurs (anchor detection) permettant de les traiter à part.

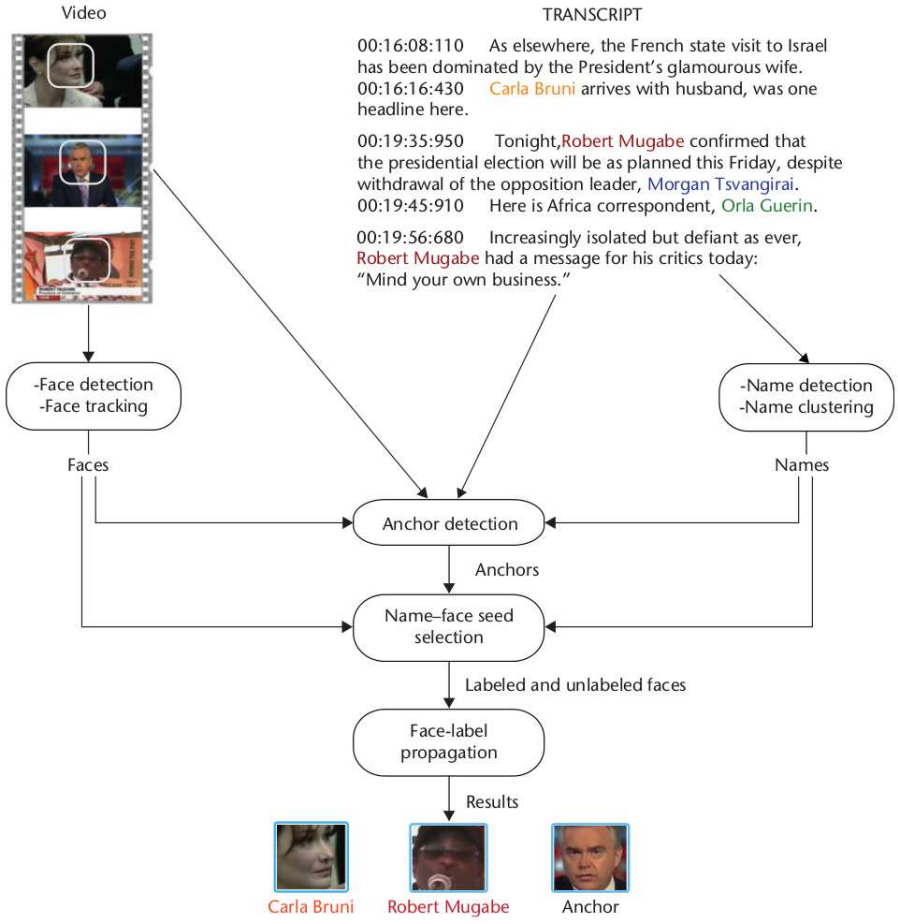


FIG. 1.9 – Image extraite de [PTM11]

L'idée principale est de sélectionner manuellement, pour chaque nom prononcé, quelques images de visages lui correspondant. Cette annotation manuelle légère nomme certains visages alors que d'autres restent inconnus.

A partir de la similarité entre tous les visages, un graphe peut être construit (nœuds : visages, liens : similarité entre visages). A ce graphe on peut ajouter deux contraintes liées aux vidéos :

- Toutes les images d'un visage qui a été suivi dans un plan doivent avoir le même nom.
- Deux visages sur la même image doivent avoir des noms différents.

Dans ce graphe, autour de chaque visage nommé se dessinent des régions de haute densité (forte similarité entre visages). Les noms sont propagés itérativement d'abord sur les visages de ces zones de haute densité puis sur les visages des zones de plus faible densité jusqu'à un certain seuil.

Cette méthode permet de profiter de l'influence des visages qui étaient anonymes au début du processus. Ces visages anonymes, une fois nommés, peuvent permettre de nommer d'autres visages anonymes.

Ce système est évalué sur 9 journaux télévisés (BBC) de 30 minutes, avec une moyenne de 120 visages et 21 noms uniques par émission. Cette méthode de propagation des noms a obtenu de meilleurs résultats que deux classifieurs (KNN et SVM) entraînés à partir des mêmes données annotées.

Cas de l'utilisation de plusieurs variantes pour un même nom

Ide et al. [IOTM07] proposent de traiter un sous-problème du nommage des personnes : l'utilisation de différents noms pour une même personne. Dans les émissions de télévision, il arrive souvent que les personnes s'interpellent par plusieurs nominations différentes, par exemple par le nom complet, juste le prénom, le titre honorifique,.... Un changement de fonction peut aussi se produire (ministre, ex-ministre).

Après une étape de nommage des visages à partir des noms dans les sous-titres, ils proposent simplement d'utiliser la similarité des visages nommés pour vérifier s'ils ne correspondent pas à la même personne.

Utilisation de l'activité des lèvres pour propager le nom d'un locuteur vers un visage

En 2013, *Bendris et al.* [BFC+13], pour aider à l'identification non supervisée des visages apparaissant dans les émissions de télévision, ont utilisé l'activité des lèvres pour propager les noms des locuteurs vers les visages. Leur hypothèse est que, comme le clustering des locuteurs obtient de meilleurs résultats que celui des visages (parce que ces derniers apparaissent avec de multiples variations), il peut être intéressant d'utiliser l'identité du locuteur pour nommer le visage détecté comme parlant.

Avant d'effectuer cette propagation, ils ont d'abord extrait les noms écrits et les noms prononcés et aussi calculé les distances entre les visages et entre les tours de parole.

Ensuite, ils ont d'abord identifiés directement les visages et les tours de parole co-occurrents avec les noms :

- Nom écrit → visage : les visages sont nommés directement lorsqu'il n'y a pas d'ambiguïté (un seul nom co-occurre avec un seul visage et inversement). Ensuite les auteurs utilisent une décision globale pour les cas ambigus, avec un graphe bi-partite basé sur les co-occurrences.
- Nom écrit → tour de parole : les tours de parole sont nommés avec les noms écrits les plus co-occurrents. Lorsqu'aucun nom ne co-occurre, le tour de parole reste inconnu.
- Nom prononcé → visage : Cette stratégie n'est utilisée que pour identifier les images de photographies (détectées automatiquement).
- Nom prononcé → locuteur : les auteurs ont utilisé la méthode proposée par *Jousse et al.* dans [JPRM+09] décrite dans la suite de ce chapitre.

Et enfin, les visages sont identifiés indirectement avec une propagation des noms :

- Visages nommés → visages anonymes : deux solutions sont envisagées :
  - Après un clustering des visages, les clusters qui contiennent un visage déjà identifié sont nommés par le même nom.
  - A partir des visages déjà identifiés, les auteurs construisent un modèle biométrique pour chaque nom. Ensuite, la similarité entre les visages anonymes et les modèles est utilisée pour propager les noms.
- Locuteurs nommés → locuteurs anonymes : les noms des tours de parole identifiés sont propagés vers les clusters de locuteurs.
- Locuteurs nommés → visages anonymes : les noms des tours de parole identifiés sont propagés vers les visages parlant ayant la plus forte probabilité de correspondre au locuteur. Cette correspondance est basée sur le seul descripteur de l'activité des lèvres.

Cette méthode de nommage non-supervisée a été validée sur le corpus *REPERE* (voir tableau 1.1) avec cinq stratégies différentes faisant intervenir différentes règles de nommage dans un certain ordre :

- `DirectFace (DF)` : identification directe des visages par les noms écrits puis par les noms prononcés.
- `DF+clu` : `DF` puis visages nommés → visages anonymes via le clustering des visages.
- `DF+sim` : `DF` puis visages nommés → visages anonymes via la similarité aux modèles.
- `DF+clu+Lip` : `DF+clu` puis voix → visage.
- `DF+sim+Lip` : `DF+sim` puis voix → visage.

Système	Précision	Rappel	F-mesure	EGER
`DirectFace (DF)`	**93.5**	21.4	34.8	78.9
`DF+clu`	71.3	47.3	56.9	55.9
`DF+sim`	84.2	32.8	47.2	68.4
`DF+clu+Lip`	65.1	**52.8**	58.3	**52.1**
`DF+sim+Lip`	68.7	52.0	**59.2**	**52.1**

TAB. 1.1 – Reproduction des résultats issus de [BFC+13]

On remarque que la première stratégie obtient une très bonne précision (93.5%) mais un faible rappel. Cela s'explique par le fait que les noms hypothèses ne couvrent pas toutes les occurrences d'apparition d'une personne. L'ajout d'une propagation basée sur le clustering obtient une meilleure F-mesure (56.9% pour `DF+clu`) que celle basée sur la similarité aux modèles (47.2% pour `DF+sim`). En revanche, cette dernière permet de conserver une précision plus importante (84.2% pour `DF+sim`), ce qui laisse plus de visages inconnus pour la propagation des noms des locuteurs vers les visages. Cette propagation à l'aide des lèvres augmente le rappel mais abaisse beaucoup la précision. Les erreurs ajoutées viennent des erreurs d'identification des locuteurs et des erreurs d'association voix → visage par le seul descripteur de l'activité des lèvres.

1.1.4 Nommage des personnages dans les vidéos de fictions

Dans cette sous-section, nous allons nous éloigner un peu des objectifs de ce manuscrit avec des travaux sur un cas particulier : le nommage des personnages dans les vidéos de fiction. Malgré tout, ce domaine nous permet de tirer des enseignements intéressants.

Les vidéos de fiction (films, séries télévisées, . . .) ont l'avantage de proposer le script et les sous-titres comme sources d'informations supplémentaires par rapport aux émissions de télévisions (seul les sous-titres y sont parfois disponibles).

- Le **script** contient la transcription des dialogues mais aussi l'identité des locuteurs et d'autres informations liées à la vidéo (mouvement des personnages, bruits, ...).
- Les **sous-titres** ne contiennent que la transcription de la parole (sans l'identité du locuteur) mais sont eux alignés temporellement à l'image, ce qui n'est pas le cas du script.

Premiers travaux avec l'utilisation de classifieurs

Les premiers travaux sur les vidéos de fictions sont apparus en 2006 dans [ESZ06]. Les auteurs ont aligné temporellement le script et les sous-titres à l'aide de la mesure « dynamic time warping ». Cet alignement permet de connaître l'identité du locuteur à chaque instant de la vidéo.

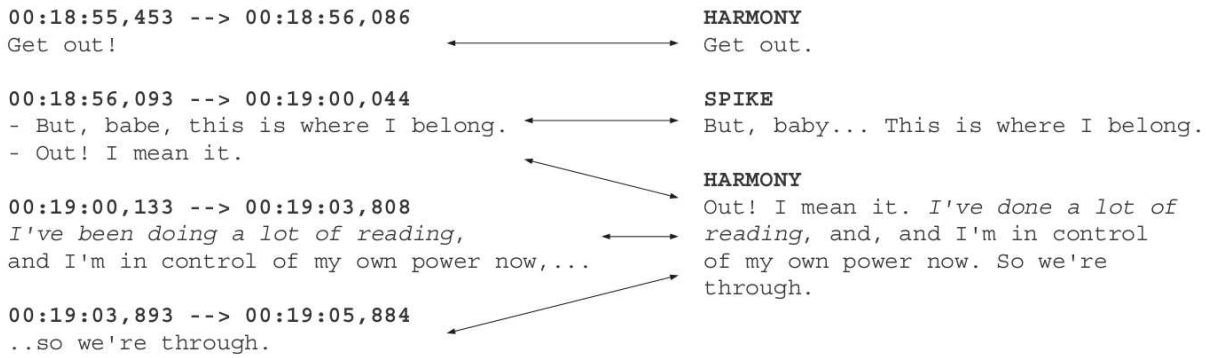


FIG. 1.10 – Image extraite de [ESZ06]

Un exemple d'alignement est montré dans la figure 1.10. Sur la partie gauche, on voit les sous-titres avec l'information temporelle. Sur la partie de droite, le script avec l'identité des locuteurs.

Dans ces travaux, les auteurs sont intéressés par l'identification des visages puisque les locuteurs sont déjà nommés par le script. Ils ont donc préalablement détecté, suivi et regroupé les visages apparaissant dans la vidéo en clusters. Ensuite, ces visages sont identifiés en deux étapes.

La première utilise la détection du mouvement des lèvres qui permet de nommer les visages (détectés comme parlant) avec le nom du locuteur courant. Dans la deuxième étape, pour chaque cluster nommé, ils construisent un modèle biométrique basé sur l'apparence visuelle (visage, vêtements). Ensuite, chaque visage encore inconnu est comparé à ces modèles à l'aide d'un classifieur des k plus proches voisins.

Un premier problème apparaît : toutes les erreurs faites durant la première étape sont propagées sur la seconde. Ces travaux ont été évalués sur le corpus de la série télévisée « Buffy the Vampires Slayer ». Deux épisodes ont été utilisés avec une moyenne de 11 personnages à reconnaître par épisode. Ils ont correctement nommé 80% des visages.

En 2009, une extension de ces travaux est proposée dans [ESZ09, SEZ09] en utilisant un classifieur SVM à la place du KNN. Le SVM est présenté comme moins influencé par les valeurs extrêmes. Il est donc moins sensible aux erreurs faites durant la première étape décrite ci-dessus. Cette méthode SVM est comparée à celle utilisant un KNN issue des premiers travaux. Elle est aussi comparée à un KNN basé sur des annotations manuelles (identification manuelle des visages, donc on évite les erreurs de la première étape).

Le classifieur SVM a une meilleur précision (aux alentours de 90%) que le KNN lorsque le système nomme moins de 65% (environ) des visages (voir figure 1.11). Au dessus de 65% de rappel, le système basé sur le KNN est plus précis. Un classifieur SVM peut s'affranchir des points aberrants (erreurs de la première étape) mais selon les auteurs ce n'est pas une solution complète de correction de ces erreurs.

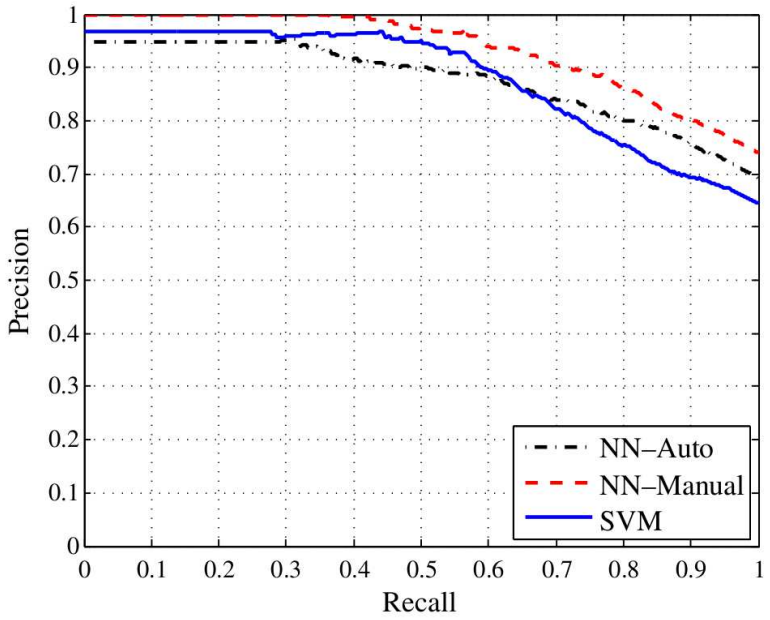


FIG. 1.11 – Évolution de la précision en fonction du rappel pour l'épisode 05-05. Image extraite de [ESZ09]

Très récemment, dans [BTS13], les auteurs ont utilisé le même schéma que [ESZ06] mais avec un classifieur de type MLR [5] [HTF01]. Les auteurs proposent deux apports supplémentaires : L_u qui intègre l'influence des données non annotées et L_c qui ajoute des contraintes du type « deux visages apparaissant en même temps appartiennent à des personnes différentes ».

Dans le tableau 1.2, on montre la précision de l'identification obtenue lorsque l'on nomme tous les personnages nommables (c'est-à-dire dont le nom est dans le script). Ces résultats, pour un système avec le classifieur MLR et pour les deux évolutions proposées sont comparés avec trois autres classifieurs (un KNN comparable à celui de [ESZ06], un SVM et une régression logistique (LR). Ce tableau est un condensé de celui présent dans l'article, avec seulement les résultats moyens par séries télévisées (BBT : *Big Bang Theory*, BF : *Buffy the vampire slayer*) et non avec le détail par épisode comme proposé dans la version originale.

system	BBT	BF
KNN [ESZ06]	64.24	56.50
one-vs-all LR	76.19	64.77
one-vs-all SVM [TBS12]	76.34	65.30
MLR	77.40	65.82
MLR + L_u	77.46	66.31
MLR + L_u + L_c	79.44	66.37

TAB. 1.2 – Condensé des résultats issus de [BTS13] sur deux séries télévisées différentes, BBT : *Big Bang Theory*, BF : *Buffy the vampire slayer*

Outre la qualité des résultats obtenus avec les différents classifieurs, l'enseignement qui nous semble le plus intéressant à mettre en avant est la différence d'évolution des résultats entre les deux séries télévisées. On voit que L_c (ligne 7 *vs* 8) apporte un gain de précision (+2%) pour BBT alors qu'il n'apporte rien pour BF. Les auteurs expliquent cette divergence par des différences de taille de casting (plus important pour BF) et de choix de cadrage (dans BBT beaucoup de plans contiennent plusieurs personnes alors que BF favorise les gros plans sur les visages). Cette divergence nous montre que le choix du corpus, à type de média identique, a une incidence sur les résultats des méthodes développées. Il nous paraît donc nécessaire d'utiliser un corpus avec une variété importante de type d'émission et de type de montage ; pour proposer des méthodes moins dépendantes des choix de conception du producteur des vidéos.

Minimisation d'une fonction de coût

Cour et al. dans [CSJT09, CST11] se placent dans le même contexte de séries télévisées (séries « Lost » et « CSI ») avec les noms des locuteurs extraits du script et des sous-titres alignés (comme dans [ESZ06]). Dans leurs travaux, ils se proposent de traiter, par la minimisation d'une fonction de coût, deux difficultés :

- La non fiabilité : erreur de nommage des visages correspondant aux erreurs de la première étape décrite ci-dessus pour les travaux de *Everingham et al.* [ESZ06]). Elles correspondent à environ 1% des visages nommés.

[5] MLR : Multinomial logistic regression - régression logistique Multinomiale

- L'ambiguïté : un nom peut correspondre à plusieurs visages. Sur l'image 1.12, les noms en vert correspondent aux noms possibles, ceux en rouge correspondent à la vérité terrain de chaque plan.

FIG. 1.12 – Affiliation de plusieurs noms hypothèses (issus du script et des sous-titres) à chaque visage. Image extraite de [CST11]

Cette fonction de coût est basée sur plusieurs caractéristiques : proximité des visages, mouvement des lèvres, genre féminin/masculin. *Cour et al.* font aussi l'hypothèse que deux visages sur deux plans consécutifs sont différents.

Ces mêmes auteurs, dans [CNT10], ont supprimé les informations disponibles dans le script. Ils ne connaissent donc plus l'identité des locuteurs. Pour retrouver cette information, ils vont utiliser à la fois les sous-titres et la liste du casting comme sources. La liste des noms issus du casting permet juste de corriger les noms prononcés extraits des sous-titres et de connaître le genre (féminin/masculin) d'un nom. Pour remplacer la connaissance du nom du locuteur, *Cour et al.* cherchent des indices dans les dialogues afin de trouver à qui fait référence un nom : à la première, deuxième ou troisième personne ?
A partir de ces indices, ils établissent des contraintes :

- Si le nom fait référence à la première personne (« Je suis Jack ») ou à la deuxième personne (« Hey, Jack »), le visage de cette personne est visible dans un fenêtre de 10 secondes autour de la citation.
- Si le nom fait référence à la troisième personne (« Jack est parti ») le visage correspondant au nom n'est pas présent dans cette fenêtre temporelle.

A partir de ces contraintes et de celles utilisées dans [CST11] (genre, distance entre visages), les auteurs résolvent l'association noms-visages en minimisant (ici aussi) une fonction de coût basée sur une combinaison linaire de toutes ces contraintes. Ils obtiennent de moins bons résultats que dans [CST11] mais avec une source en moins (pas de script).

Association à l'aide d'une vision globale et non locale

Une autre solution pour s'affranchir de cette incertitude est de changer de point de vue. Ne plus regarder un environnement local autour du moment de citation d'un nom pour sélectionner les visages candidats mais avoir un point de vue global sur la vidéo.

L'hypothèse originale de *Zhang et al.* [ZXL08] est que le nombre de phrases prononcées par une personne (les locuteurs sont identifiés dans le script) est corrélé au nombre d'apparitions à l'écran de cette personne. Après une étape de créa-

tion de clusters correspondant aux visages des personnages d'un film, ils classent les clusters en fonction du nombre d'images qu'ils contiennent et les noms des locuteurs par le nombre de phrases prononcées. Ensuite, ils associent les noms et les clusters selon le rang qu'ils occupent.

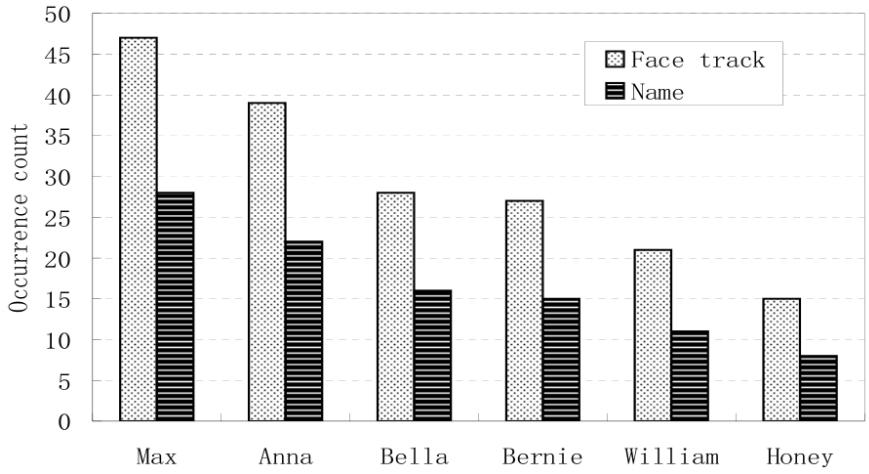


FIG. 1.13 – Alignement des noms et des clusters de personnes en fonction de leurs rangs (classement en fonction de la taille des clusters et du nombre de tours de parole d'un nom). Image extraite de [ZXL08]

On peut voir, dans la figure 1.13, la taille des clusters de noms et des clusters de personnages du film « Notting Hill ». Un problème survient avec cette méthode lorsque deux d'entre eux ont des clusters de taille proche (dans l'exemple « Bella » et « Bernie »). Pour éviter ce problème les auteurs proposent de ne pas nommer immédiatement les clusters où il y a une ambiguïté. Pour identifier un cluster non-nommé, les auteurs utilisent la proximité sociale entre un visage nommé et un visage anonyme (par exemple, deux visages apparaissant souvent dans des plans contigus). Leur hypothèse est que cette proximité se retrouve dans le script. Ils affectent aux clusters encore anonymes le nom (pas encore attribué) le plus proche de celui du visage nommé (par exemple le nom de l'interlocuteur le plus régulier du cluster déjà nommé). Avec cette méthode, 90% des visages sont nommés avec une précision proche de 95% dans deux films.

Dans leurs travaux suivants, *Zhang et al.* [ZXCL09, ZXLH09] ont poursuivi cette vision globale mais avec cette fois la construction de graphes de réseaux sociaux des personnages de fiction. Leur méthode est présentée dans la figure 1.14.

Cette méthode utilise, d'une part, un graphe de proximité des relations sociales des clusters de visages des personnages. Une relation est d'autant plus forte que deux clusters de visages apparaissent souvent dans les mêmes scènes. D'autre part, un deuxième graphe est construit à partir des relations sociales des noms des locuteurs. Là aussi, une relation est d'autant plus forte que deux personnages sont des interlocuteurs réguliers. Pour effectuer l'association, les auteurs cherchent les points communs entre les deux graphes. Un nom et un visage sont associés s'ils ont le même profil. Cette technique obtient une meilleure précision

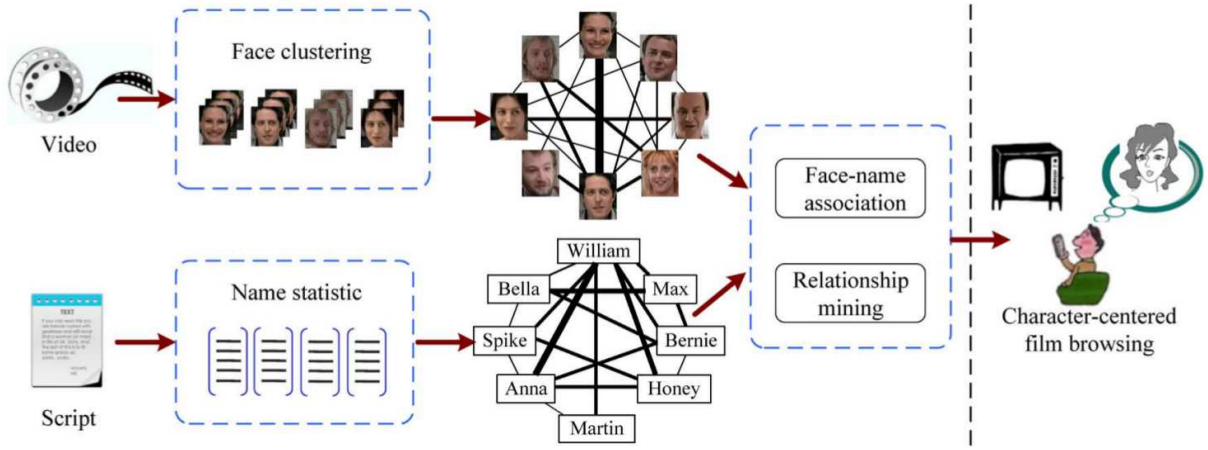


FIG. 1.14 – Vue d'ensemble de l'association noms-visages proposée par *Zhang et al.*. Image extraite de [ZXLH09]

que la méthode locale de *Everingham et al.* [ESZ06] quel que soit le nombre de visages nommés.

Dans le graphe de proximité des visages, du bruit peut apparaître consécutivement aux erreurs de suivi et de clustering des visages. *Sang et al.* dans [SLXC11, SX12] proposent de construire un graphe moins sensible au bruit, basé non plus sur le nombre de co-occurrences dans les plans mais sur un classement des co-occurrences. En d'autres termes, pour chacun des personnages, ils classent les autres personnages en fonction du nombre de co-occurrences. C'est le rang dans ce classement qui donne le poids du lien dans le graphe.

Dans la figure 1.15, on retrouve deux exemples de matrices de scores transformées en matrices de rangs.

(a) Name Affinity Matrix

	WIL	SPI	ANN	MAX	BEL
WIL	0.173	0.024	0.129	0.009	0.013
SPI	0.024	0.017	0.007	0.001	0.002
ANN	0.129	0.007	0.144	0	0
MAX	0.009	0.001	0	0.009	0.006
BEL	0.013	0.002	0	0.006	0.011

(b) Face Affinity Matrix

	Face1	Face2	Face3	Face4	Face5
Face1	0.186	0.041	0.147	0.008	0.021
Face2	0.041	0.012	0.005	0.002	0.004
Face3	0.147	0.005	0.157	0	0.003
Face4	0.008	0.002	0	0.005	0.007
Face5	0.021	0.004	0.003	0.007	0.009

(a) Matrice basée sur les scores

(a) Name Rank Ordinal Matrix

	WIL	SPI	ANN	MAX	BEL
WIL	5	3	4	1	2
SPI	4	3	3	1	2
ANN	2	1	4	0	0
MAX	3	1	0	1	2
BEL	3	1	0	2	2

(b) Face Rank Ordinal Matrix

	Face1	Face2	Face3	Face4	Face5
Face1	5	3	4	1	2
Face2	4	3	3	1	2
Face3	***3***	***2***	4	0	***1***
Face4	3	1	0	1	2
Face5	***4***	***2***	***1***	***3***	2

(b) Matrice basée sur les rangs

FIG. 1.15 – Image extraite de [SLXC11]

Pour associer les graphes de noms et de visages comme dans [ZXCL09, ZXLH09]

présenté précédemment, *Sang et al.* ont adapté un algorithme d'ECGM[6] [Bun97] à leur problème. Cet algorithme va trouver le nombre d'opérations nécessaires (suppression, ajout, substitution de nœuds et de liens) pour transformer un graphe en un autre.

Pour vérifier la sensibilité de ce graphe au bruit, deux types de bruits simulés sont ajoutés : du bruit de couverture (création, suppression de lien) et du bruit d'intensité (changement du poids des liens). Ils montrent ainsi que la méthode proposée de construction et de comparaison des graphes permet de s'affranchir des problèmes de bruits d'intensité. Par contre, les bruits de couverture, qui changent la topologie du graphe, dégradent les résultats.

Apprentissage d'Instances Multiples dans les fictions

Trois articles similaires [KWRB11, WKRB11b, WKRB11a] ont repris l'idée de l'apprentissage d'instances multiples (MIL) pour identifier les personnages de la série télévisée « Buffy the Vampires Slayer ». Ils proposent un algorithme nommé « Semi Supervised Multiple Instance Learning (SSMIL) ». Les noms sont issus des sous-titres et du script aligné temporellement comme dans [ESZ06]. Un détecteur d'activité des lèvres leur permet de savoir si le locuteur est visible ou non. A partir de ces données annotées de manière semi-supervisée, ils sont en mesure de construire des exemples positifs (visages parlant) et des exemples négatifs (visages ne parlant pas). Deux visages parlant en même temps sont considérés tous les deux comme négatifs. Un sac est constitué à partir de tous les exemples d'un plan. Les résultats obtenus sont meilleurs qu'avec une classification basée sur un KNN [ESZ06] ou un SVM [ESZ09].

[6]ECGM : Error Correcting Graph Matching - Correction des erreurs de correspondance entre graphes

1.2 Nommage des locuteurs

Jusqu'à très récemment, les travaux de l'état de l'art sur l'identification du locuteur sans modèle biométrique ont utilisé la modalité la plus proche pour extraire les noms des locuteurs : les noms prononcés issus de la transcription de la parole. Dans cette section, nous allons passer en revue les différentes propositions faites dans la littérature.

1.2.1 Utilisation de patrons linguistiques

Les premiers travaux ont été proposés par *Canseco et al.* dans [CRLG04]. Ils ont utilisé des patrons linguistiques définis manuellement pour déterminer à qui fait référence un nom cité : au locuteur courant, suivant ou précédent. La figure 1.16 nous montre quelques exemples de patrons.

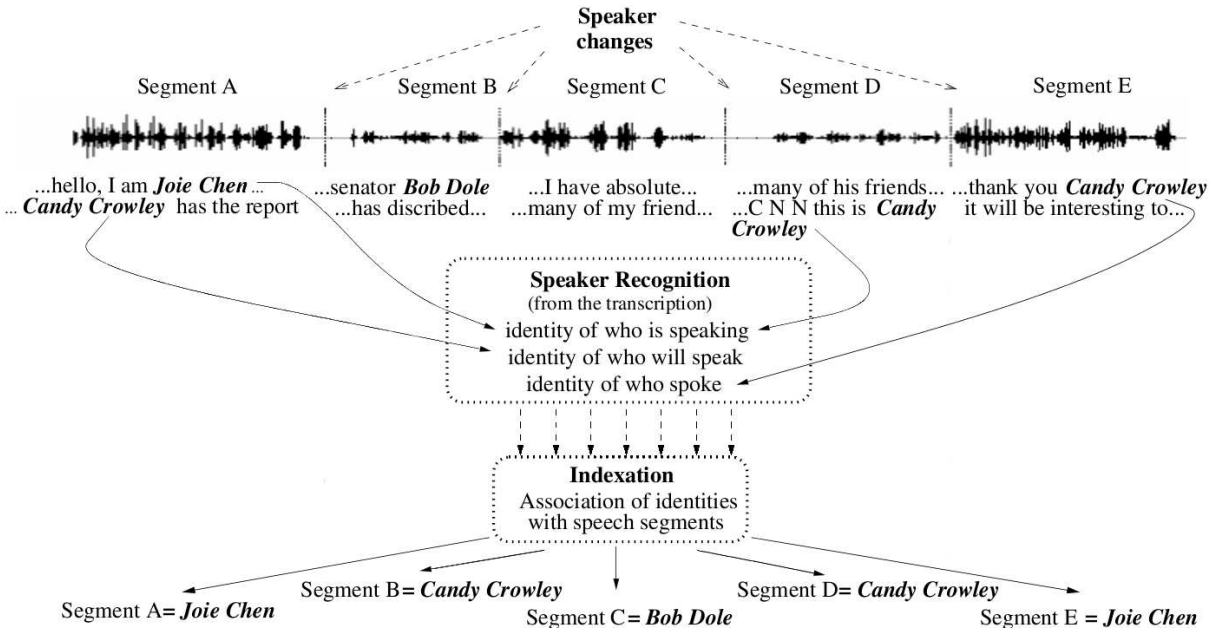


FIG. 1.16 – Exemple de patrons linguistiques utilisés pour le nommage des locuteurs. Image extraite de [CRLG04]

Dans cette figure, en partie haute, on voit 5 segments de locuteurs de A à E non identifiés. Grâce aux noms prononcés dans la transcription de la parole associée à ces segments, on peut les nommer :

- Le premier segment « I am Joie Chen » défini clairement que le locuteur courant s'appelle « Joie Chen ».
- De même « This is Candy Crowley » défini que le segment D est prononcé par « Candy Crowley ». Cela est renforcé par la phrase « thank you Candy Crowley » prononcée dans le segment E.
- Le segment B correspond à « Candy Crowley » aussi puisqu'il a été introduit par le locuteur du segment A avec la phrase « Candy Crowley has the report ».

- Le nom du locuteur du segment C « Bob Dole » est introduit par « Candy Crowley » dans le segment B.
- Et enfin le segment E peut-être nommé grâce au résultat de la diarization qui l'associe au segment A (« Joie Chen »).

Cet exemple nous montre que des informations pertinentes sont incluses dans la parole pour identifier les locuteurs d'une émission de télévision. Les données utilisées sont issues du corpus Hub4-E. Ces données sont variées : 4 chaînes de radios et télévisions avec plusieurs émissions différentes par chaîne. Les auteurs n'ont pas utilisé de liste fermée de personnes à identifier mais une diarization et une transcription de la parole **manuelle**.

Pour la détection des noms dans les transcriptions, ils ont employé des dictionnaires de noms. Ils n'ont pas pris en compte le contexte linguistique autour des noms, ce qui peut produire des erreurs de détection (par exemple, un lieu peut avoir le même nom qu'une personne).

Dans leurs travaux suivant [CLG05], *Canseco et al.* ont remplacé la transcription de la parole manuelle par une transcription issue d'un système automatique. Pour les segments de parole qui n'ont pas pu être nommés, ils ont aussi proposé de détecter le rôle du locuteur pour lui assigner un nom. Par exemple, si un tour de parole a été détecté comme provenant du présentateur, il suffit de lui donner le nom du présentateur.

Dans [CMAQ05] les auteurs ont utilisé la même méthode de patrons linguistiques mais avec des systèmes automatiques (diarization automatique basée sur la distance BIC [7] et transcription automatique de la parole issue du LIMSI [GLA02]). Ils ont pu identifier 53% du temps de parole, avec une précision de 82%, sur 2 heures de journaux télévisés (CNN et ABC, TRECVID 2003).

En 2006, *Tranter* [Tra06] va remplacer les règles définies manuellement par une phase d'apprentissage de séquences de n-grammes avec des probabilités associées. Sur le corpus Hub-4, elle a montré que moins de locuteurs sont nommables avec les systèmes automatiques (47.3%) qu'avec les annotations manuelles (76.8%), sur l'ensemble de test (4,2 heures, 138 personnes). Cette réduction abaisse le temps de parole correctement identifié de 38% à 26% avec une précision réglée à 95%.

Dans [MNM07], *Ma et al.* ont remplacé le système de calcul des règles de [Tra06] par l'utilisation d'un modèle à maximum d'entropie. Ils ont, en plus, enrichi la méthode avec les informations de position du nom dans la phrase et utilisé la correspondance du genre (masculin/féminin) entre le nom et le locuteur.

Ils ont aussi ajouté des informations issues de modèles de locuteurs obtenues à partir de l'ensemble d'entrainement. Les données utilisées sont les mêmes que dans [Tra06]. Dans la figure 1.17, on peut voir l'évolution de la précision en fonction du rappel selon les informations utilisées.

Dans cette figure, on peut voir que l'apport des modèles biométriques n'est

[7] BIC : Bayesian Information Criterion

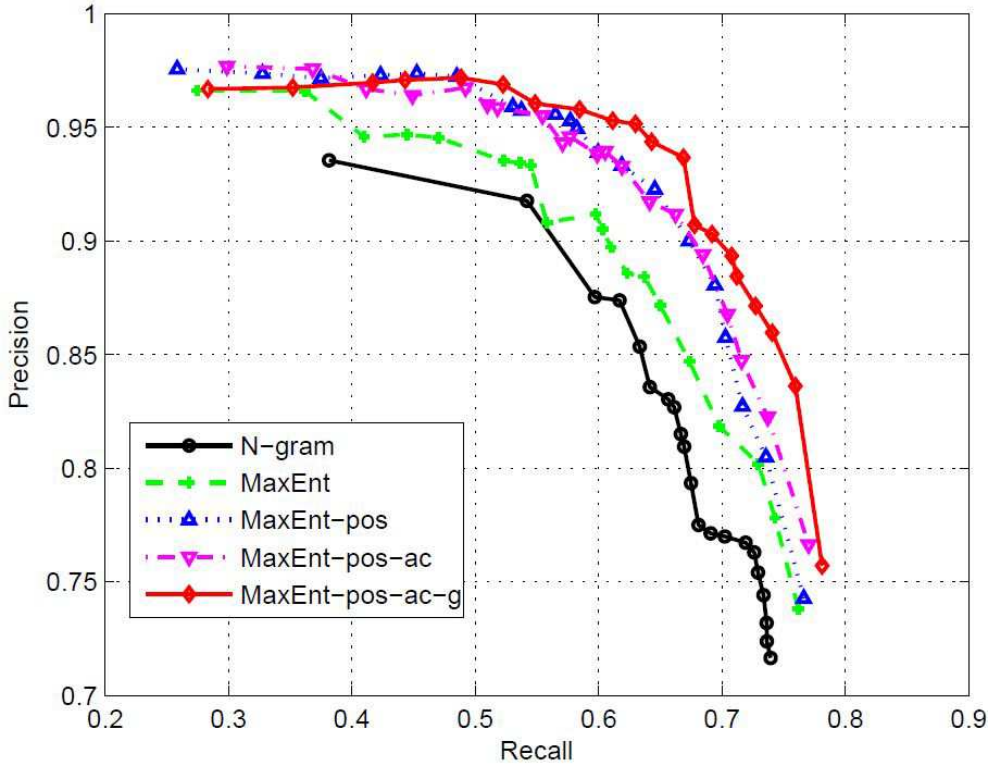


FIG. 1.17 – Comparaison méthode n-grammes *vs* maxent en fonction des sources d'informations utilisées. pos : position temporelle du nom dans la phrase, ac : modèles acoustiques, g : correspondance entre le genre du locuteur et celui du nom, Corpus Hub-4. Image extraite de [MNM07]

pas très convaincant. C'est en grande partie dû au peu de locuteurs présents à la fois dans l'ensemble d'entrainement et de test (10% de locuteurs en commun qui correspondent à 30% du temps de parole du test).

1.2.2 Arbre de classification sémantique

Mauclair et al. dans [MME06b, MME06a] ont changé la méthode d'association entre un nom et une personne grâce à l'utilisation d'arbres de classification sémantique (SCT[8]). Ces arbres sont construits localement pour affilier un nom au locuteur précédent, courant, suivant ou à un autre locuteur.

Dans l'exemple, en figure 1.18, un nom est prononcé dans une phrase. Ce nom peut faire référence à différents locuteurs (symbolisés par les flèches). A partir de cette phrase, il est possible de construire un arbre de classification sémantique. Ainsi, en fonction des termes qui entourent un nom il est possible de calculer les probabilités (apprises sur une base annotée) qu'il fasse référence au locuteur courant, précédent, suivant ou à un autre. Seul le lien avec la plus grande probabilité est pris en compte.

Dans la figure 1.19, on peut voir que deux segments de parole (en bleu) pro-

[8]SCT : semantic classification tree

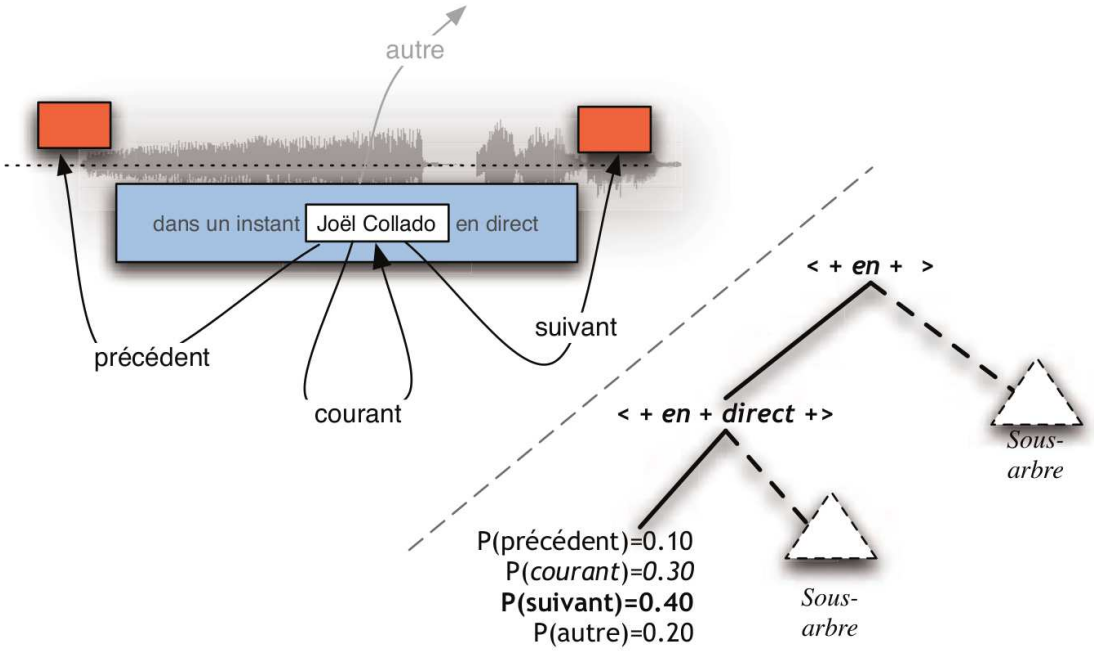


FIG. 1.18 – Exemple d'arbre de classification sémantique. Image extraite de [Jou11]

posent le même nom pour le segment orange avec des probabilités différentes. Leurs probabilités sont donc additionnées.

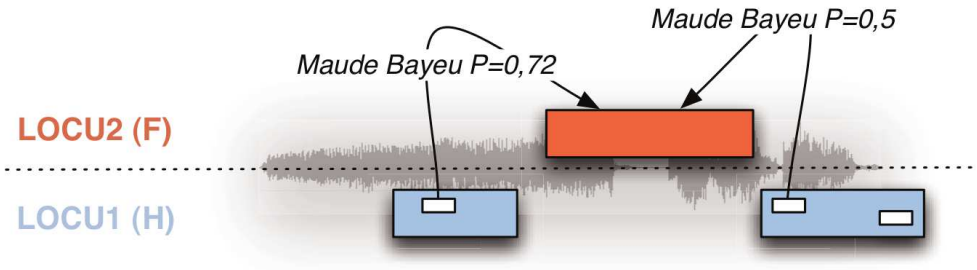


FIG. 1.19 – Exemple d'attributions locales. Image extraite de [Jou11]

Lors du processus d'association noms-personnes au niveau global, chaque locuteur est nommé par le nom avec le plus grand score. Comme chaque locuteur est traité indépendamment, il est possible d'associer le même nom à deux locuteurs détectés comme différents lors de la diarization. Cette méthode a été testée sur le corpus ESTER (émissions radiophoniques françaises) à l'aide de transcriptions et d'une diarization manuelles. 70% (environ) de la durée totale des émissions a été correctement identifiée (18% d'erreur et 12% de non identifiés).

Estève et al. [EMDM07] ont fait une comparaison des SCT par rapport à une méthode utilisant des n-grammes (voir figure 1.20).

Sur la partie gauche de l'image, on voit les résultats de la précision en fonction

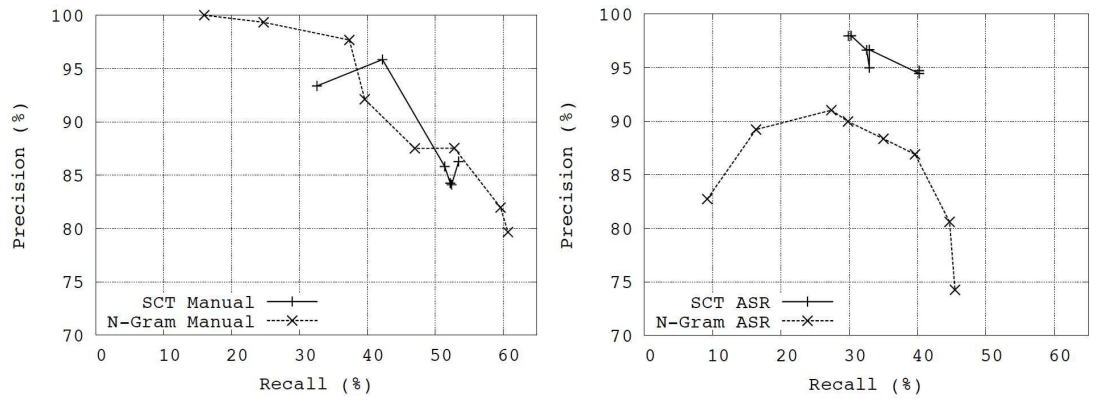


FIG. 1.20 – Méthode SCT *vs* n-grammes, transcription de la parole manuelle (a) ou automatique (b), diarization manuelle. Image extraite de [EMDM07]

du rappel, avec l'utilisation d'une transcription de la parole manuelle. Les deux systèmes sont à peu près équivalents. Par contre, sur la partie de droite, on peut observer les résultats lorsque la transcription est issue d'un système automatique. Les SCT sont moins sensibles que les séquences de n-grammes à l'utilisation de transcriptions automatiques de la parole sur le corpus ESTER.

Jousse et al. ont d'abord étendu la méthode des SCT avec l'utilisation d'un système complètement automatique [JJM+08] (la diarization et la transcription de la parole [DEMM05], la détection des entités nommées à l'aide de l'outil *Nemesis* [Fou11]). Dans cet article, ils présentent une étude des différentes erreurs générées par l'utilisation d'un système automatique et en déduisent la nécessité d'utiliser un système d'extraction des entités nommées adapté aux SCT. Ces même auteurs [JPRM+09, JMJ+09, Jou11] ont amélioré l'utilisation des arbres de classification sémantique avec une décision locale (affiliations des noms aux tours de parole proches) puis globale (propagation aux clusters de locuteurs). Ils ont aussi montré une augmentation du taux d'erreurs d'identification en durée de 16.66% (annotations manuelles) à 75.15% (système tout automatique) sur les données ESTER 1 phase 2, 10h de test, 213 locuteurs dont 24 inconnus.

1.2.3 Fonctions de croyance

En 2010, *Petitrenaud et al.*, dans [PJME10, PRJME10], ont utilisé le même contexte que dans les travaux de *Jousse*. Cependant, ils ont remplacé le système de décision basé sur les SCT par un système à base de fonctions de croyance. Elles ont la particularité de prendre en compte la cohérence des informations au sein de tours de parole contigus. Les auteurs ont observé que le taux d'erreur d'identification passe de 16.6% avec les SCT à 13.7% avec l'utilisation de fonctions de croyance. Ceci est observé seulement dans le cadre de l'utilisation de données annotées manuellement.

Khoury et al. [EKLMP12] ont étendu ces fonctions de croyance avec d'une part l'utilisation de briques de base automatiques ou manuelles (diarization et transcription). La détection des entités nommées étant, elle, automatique. Et d'autre part avec l'ajout d'informations issues des scores d'un système de reconnaissance du locuteur à base de modèles biométriques GMM-UBM. Ces scores ont été transformés en fonction de croyance pour s'intégrer aux précédents travaux. La métrique utilisée pour l'évaluation est différente de celle utilisée dans [PJME10, PRJME10].

system	Sub	Del	Ins	Err
Using reference segmentations and transcripts				
Transcript-based system	3.57	6.29	0.20	10.06
GMM-based system	1.82	58.05	2.20	62.07
Transcript+GMM system	1.31	0.94	2.40	4.65
Using automatic segmentations and transcripts				
Transcript-based system	14.25	25.69	1.24	41.17
GMM-based system	2.67	57.15	3.13	62.95
Transcript+GMM system	13.85	14.69	4.17	32.72

TAB. 1.3 – Reproduction des résultats issus de [EKLMP12] (*Transcript-based system* : système de nommage non-supervisé, *GMM-based system* : système à base de modèles biométriques)

Dans le tableau 1.3, on peut voir que le taux d'erreur du système de décision à base de fonctions de croyance (« Transcript-based system ») sans l'aide des modèles biométriques augmente de 10% (briques de base manuelles) à 41.1% avec l'utilisation de briques de base automatiques. L'ajout des informations issues des modèles GMM-UBM (« Transcript+GMM system ») permet de réduire cette erreur à respectivement 4.6% et 32.7%. En sachant que le taux d'erreur issu des modèles biométriques était de 62% (segmentation manuelle) et 63% (segmentation automatique). Cette combinaison montre donc que les modèles biométriques et les systèmes de nommage non-supervisés ont tendance à ne pas identifier les mêmes personnes. Les systèmes biométriques ne pourront reconnaître que les locuteurs ayant un modèle comme les présentateurs (taux de suppression de 58% et 57%) avec une très bonne précision (taux de substitution de 1.8% et 2.7%). L'utilisation de différentes sources d'informations en fonction du rôle des personnes est donc une orientation très intéressante.

Huijbregts et al. [HvL11] proposent d'utiliser une toute autre source pour le nom des locuteurs : le guide des programmes télévisés. En effet, les noms des intervenants peuvent être contenus dans le programme télévisé de l'émission. Ceci peut permettre d'identifier certains locuteurs d'une vidéo, a priori les plus importants. Cela ne peut marcher qu'avec certains types d'émissions télévisées comme celles où un invité est attendu. Cette information sûre (un nom correspond forcément à un locuteur) mais incomplète (tous les locuteurs n'ont pas leur nom dans cette source) ne spécifie pas à quel moment de la vidéo un nom a le plus de chances de correspondre à un locuteur, comme dans le cadre des noms prononcés.

Comme cette information est au niveau de la vidéo et plus au niveau de quelques segments, les auteurs ont dû constituer une très large collection de vidéos pour utiliser la redondance des noms avec une diarization sur la collection complète. Avant de compter les co-occurrences entre les noms et les clusters de locuteurs, ils ont classé les locuteurs en trois rôles : présentateur, chroniqueur et invité. Cette classification leur a permis d'éviter de mauvaises associations lorsque le présentateur n'était pas cité dans le programme télévisé. Enfin, la redondance des co-occurrences entre noms et clusters leur permet de nommer les clusters avec le nom le plus co-occurrent.

1.2.4 Utilisation des noms écrits comme seule source de noms

Pour obtenir de bons résultats, les travaux proposés dans l'état de l'art ont jusqu'à présent employé, pour extraire des noms hypothèses, soit une source sûre (script+sous-titres, guide de programme télévisé) mais issue d'une supervision ; soit une source peu sûre (noms prononcés, noms écrits) en essayant de pallier aux défauts de celle-ci.

Ces dernières années, l'augmentation de la qualité des vidéos permet d'envisager une extraction des noms écrits avec très peu d'erreurs. Nous avons été les premiers à proposer un système d'identification des locuteurs basé uniquement sur les noms écrits à l'écran (extraits à l'aide de l'outil LOOV [Poi11, PTQB11, PBQT12]) dans les émissions de télévision [BPT+12, PBL+12, BPF+13, PBB+13]. Ces méthodes seront décrites dans le chapitre 5.

Programmation Linéaire en Nombre Entier (PLNE)

Une collaboration entre le laboratoire LIMSI et le LIG a permis d'étendre ce clustering avec l'ajout d'informations issues des noms écrits pour identifier les personnes parlant dans les émissions de télévisions. Ces travaux ne seront détaillés que dans ce chapitre, plus de détails pourront être trouvés dans [BP13].

L'idée principale est de remplacer un clustering BIC classique par un clustering par PLNE. Précédemment, *Finkel et Manning* [FM08] ont proposé de voir le clustering comme un problème d'optimisation. En fonction de la similarité entre éléments, un clustering selon la méthode PLNE va chercher la solution optimale de regroupement en maximisant la similarité des éléments intra-classe et en minimisant la similarité des éléments inter-classes. En 2012, *Rouvier et al.* dans [RM12, DRMY12] ont testé cette approche pour le clustering des locuteurs.

Ces travaux ont été étendus dans [BP13] par l'ajout d'informations issues des noms écrits. Ils nous permettent d'identifier les clusters de locuteurs. La première étape du processus est la construction d'un graphe de similarité multimodale, présenté dans la figure 1.21

Dans ce graphe, on peut voir apparaître les liens intra-modalité qui correspondent à la similarité des éléments mono-modaux (tours de parole, noms

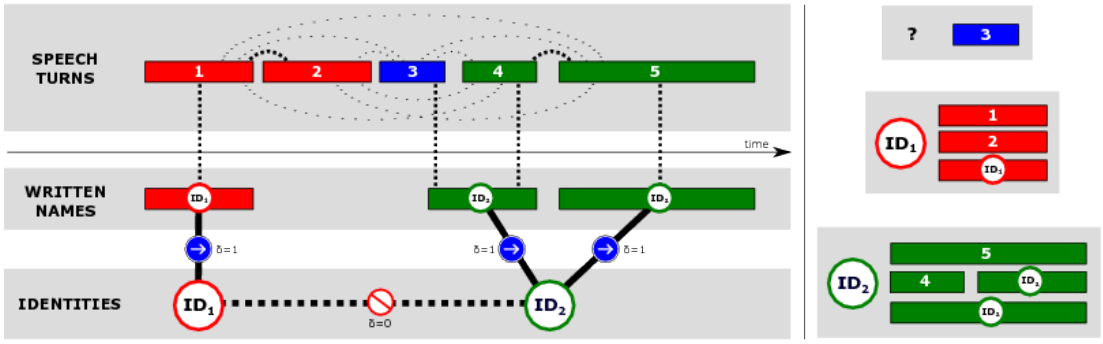


FIG. 1.21 – Graphe multimodal pour un clustering ILP. Image extraite de [BP13]

écrits). Certaines interdictions apparaissent, par exemple, entre les noms différents ($ID_1 \neq ID_2$). On retrouve aussi les liens de co-occurrences multi-modaux entre les noms écrits et les tours de parole. Comme un nom écrit a de fortes probabilités de correspondre au locuteur courant, l'identification des tours de parole par la résolution PLNE revient donc à trouver le chemin le moins coûteux (dans les lignes de pointillés fin ou gras) qui relie un nom et un tour de parole.

Les résultats de la diarization obtenus sont comparables à un clustering classique basé sur BIC. Par contre, nous obtenons des clusters plus purs, ce qui peut permettre d'intégrer cette solution comme pré-clustering dans une solution en plusieurs étapes (BIC+CLR/*i-vector* [DKD+11]).

Approche	**Précision**	**Rappel**	**F-Mesure**
Oracle	100.0%	62.1%	76.6%
Fusion tardive [PBL+12]	79.8%	59.4%	68.1%
Fusion précoce [PBB+13]	83.9%	**61.2%**	70.7%
PLNE [BP13]	**90.6%**	58.2%	**70.9%**

TAB. 1.4 – Résultats de l'identification comparés à nos autres travaux. Reproduction issue de [BP13])

Dans le tableau 1.4, nous retrouvons la performance d'identification comparée à celles que nous avons obtenue précédemment dans [PBL+12, PBB+13] (décrites dans le chapitre 5) et à celle d'un oracle. Cet oracle identifie correctement les locuteurs si leur nom a été extrait des noms écrits à l'écran au moins une fois dans la vidéo. Les résultats sont sensiblement identiques en termes d'IER[9] et de F-mesure. Cependant, la méthode PLNE obtient une très bonne précision alors que la fusion précoce obtient un rappel très proche de celui de l'oracle.

[9]IER : Identifcation Error Rate

1.3 Étude sur la capacité de nommage des noms cités à l'oral ou écrits à l'écran

Une étude intéressante, proposée par *Bechet et al.* [BFD12], analyse la capacité des noms prononcés (extraits manuellement) à nommer les personnes parlant et/ou apparaissant dans les vidéos. Dans les données utilisées (corpus *REPERE*, phase 0), 72% des personnes ont leur nom cité à l'oral. Parmi les noms cités, 447 des 717 noms cités correspondent à une personne présente (précision de 62%).

Pour intégrer cette source d'informations dans un processus de décision multimodale plus complet, les auteurs ont proposé de diminuer le nombre de noms utilisés pour augmenter la précision. Un classifieur adaboost [SS99] va sélectionner les noms en fonction de deux types d'informations :

- A partir de caractéristiques linguistiques, ce qui correspond à la façon dont un nom est mentionné par un locuteur. Par exemple, si le verbe est après le nom ("John Doe *reports* about..."), si le nom est en fin de phrase, ...
- A partir de caractéristiques structurelles, ce qui correspond au contexte d'occurrences de citation d'un nom dans une émission (combien de fois a-t-il été répété ? Par qui ? Dans quel genre de discours ? ...).

Ces deux types d'informations ont permis au classifieur de sélectionner les noms ayant la plus grande probabilité de correspondre à une personne présente dans la vidéo. Ce qui a permis de réduire le nombre de noms hypothèses inutiles (augmentation de la précision de 62% à 68%) mais qui a aussi entrainé une réduction du nombre de personnes nommables.

Pour vérifier la difficulté de la tâche, les auteurs ont demandé à deux juges s'il était possible de prédire la présence des personnes à partir des transcriptions de la parole seulement. Les juges avaient le choix entre trois possibilités : la personne correspondant au nom n'est pas présente, est présente, ne sait pas. Ces jugements ont été comparés aux annotations manuelles qui certifient la présence ou non d'une personne à l'image ou dans la bande son.

Le résultat de cette étude montre qu'il est difficile de prédire qu'une personne sera absente avec certitude. Une autre information intéressante relevée par les auteurs est qu'il est difficile de déterminer si une personne est présente seulement à l'image et qu'il est beaucoup plus facile de déterminer si un nom prononcé fait référence à un locuteur ou non.

1.4 Conclusion

Cet état de l'art a parcouru, sur un large spectre, les propositions de nommage des personnes dans les documents audio-visuels qui n'ont pas ou peu utilisé de modèles biométriques appris sur des bases annotées.

Le premier enseignement que l'on peut en tirer est que la qualité de l'affiliation noms-clusters candidats dépend de la source utilisée pour les noms. Moins cette affiliation est sûre, plus la résolution de l'association noms-personnes devra s'affranchir de cette incertitude. Il est donc important de travailler avec une source pour les noms fiable et non ambigüe.
Repassons en revue les différentes sources de noms identifiées dans la littérature :

Le guide des programmes télévisés :

- Chaque nom correspond forcément à une personne visible et/ou parlant dans la vidéo.
- Seuls les quelques noms les plus importants sont écrits dans cette source.
- Il n'y a pas d'indication temporelle du moment de présence des personnes.
- Cette source oblige à utiliser de grandes collections de vidéos pour avoir une redondance suffisante de l'information.
- Elle n'est pas forcement disponible pour tous les types d'émissions de télévision.

Le script :

- Il est seulement disponible pour les vidéos de fiction.
- Pour être créé, il a besoin d'annotations manuelles.
- Il n'est pas aligné temporellement avec la vidéo mais cette contrainte est facilement contournable avec les sous-titres.

Les sous-titres :

- Ils sont disponibles pour les vidéos de fiction, pas toujours pour les émissions de télévision.
- Pour être créés, ils ont besoin d'annotations manuelles.
- Ils apportent une information temporelle sur la possible apparition d'une personne.
- L'extraction automatique des noms complets (nom de famille + prénom) prononcés peut être sujette à des erreurs et est dépendante de la langue.
- Tous les noms ne correspondent pas forcément à une personne présente.

Transcription automatique de la parole :

- Elle apporte une information temporelle sur la possible apparition de la personne.
- Tous les noms ne correspondent pas forcément à une personne présente.
- Elle doit être extraite automatiquement :
    1. Les systèmes de reconnaissance de la parole peuvent faire des erreurs dans la transcription des noms proches (« Dupont » et « Dupond »).

2. La détection d'entités nommées peut-être erronée (« Dupont », « du pont »).
3. Chaque langue a besoin de modèles spécifiques.

- Il n'est pas toujours évident de savoir à qui fait référence un nom (au tour de parole courant ? suivant ? précédent ? à un autre ?) ou encore à quel visage dans les plans adjacents. Le nom peut aussi ne pas faire référence à quelqu'un présent dans la vidéo.

Transcription automatique des textes écrits à l'écran :

- Elle apporte une information temporelle sur l'apparition d'une personne.
- Les textes écrits à l'écran ne sont pas disponibles pour toutes les vidéos.
- Le même type de difficultés que pour les noms prononcés apparait pour leur extraction, mais :
  1. L'augmentation de la qualité des vidéos permet de réduire les erreurs de transcription.
  2. Chaque émission utilise un gabarit avec des emplacements spécifiques pour écrire les textes. La difficulté de la détection des noms écrits pour introduire la personne correspondante réside donc dans la détection des positions spatiales des cartouches.
  3. Chaque système d'écriture a besoin d'un modèle de caractères spécifiques, mais il est facile d'utiliser un modèle générique.
- Généralement, un nom est écrit à l'écran pendant que la personne est présente/parle, donc l'affiliation est beaucoup plus simple que pour les noms prononcés. Néanmoins, il reste une difficulté pour l'affiliation des visages : si plusieurs visages sont présents, à quel visage affilier le nom ?

Nous sommes intéressés par l'identification des personnes dans les émissions de télévision. Trois sources sont à écarter d'office. Le **script**, parce qu'il n'existe pas pour ce type de média. Le **guide des programmes télévisés** parce qu'il est trop peu complet : les journaux télévisés font partie des émissions visées et pour ce type de vidéo le guide des programmes télévisés ne spécifie pas le nom des personnes présentes. Les **sous-titres** car ils demandent une annotation manuelle importante et c'est justement cette annotation que nous voulons éviter.

Pour réaliser correctement l'association noms cités à l'oral - personnes, beaucoup de travaux ont cherché à prendre en compte l'incertitude liée à cette modalité (est ce qu'un nom correspond bien à une personne et à laquelle ?). Même si ces travaux semblent avoir surmonté cette difficulté à l'aide de méthodes d'apprentissage (SCT, MIL) pour les noms extraits de données annotées manuellement, il faut encore une forte amélioration de l'extraction automatique de ces noms pour espérer les utiliser dans un système complètement automatique. Et même dans ce cas, *Bechet et al.* [BFD12] ont montré qu'un jugement humain a du mal à prédire si un nom prononcé correspond bien à une personne présente dans la vidéo. Il restera donc toujours une incertitude liée à cette modalité

Deux raisons nous amènent à préférer les noms écrits à l'écran comme source d'information principale pour connaitre l'identité des personnes présentes dans les émissions de télévision. La première est que les vidéos que nous ciblons (journaux télévisés, débats, émissions politiques) introduisent souvent les personnes dans les vidéos par leurs noms écrits à l'écran dans un cartouche. L'utilisation de cette modalité réduit l'incertitude liée à l'affiliation d'un nom (un nom écrit correspond au locuteur courant et à un des visages visibles au moment de l'écriture du nom). La deuxième est que l'augmentation de la qualité des vidéos permet une extraction de ces noms avec une très bonne qualité (ceci sera détaillé dans le chapitre 3).

Comme aucune comparaison systématique n'a été faite entre les capacités des noms écrits et des noms cités à proposer le nom des personnes présentes dans les vidéos, nous avons fait une étude comparative de ces deux modalités. Cette étude a été publiée dans [PBQ13, PBL+13], nous la développerons plus en détails dans le chapitre 4 de ce manuscrit.

Maintenant que la source principale pour les noms est choisie, regardons les autres enseignements que l'on peut tirer de l'état de l'art :

Les premiers travaux utilisaient les **co-occurrences** entre noms et clusters ; mais les résultats n'étaient pas satisfaisant à cause de l'incertitude liée à la source des noms utilisés. Toutefois, nous avons choisi une source avec une très faible incertitude, il faut donc ré-envisager l'utilisation de méthodes basées sur les co-occurrences et, dans ce cadre, il n'est pas nécessaire d'utiliser des méthodes qui essayent de palier l'incertitude des noms hypothèses (comme MIL, les patrons linguistiques ou les SCT).

L'utilisation de **ressources externes**, comme les images web issues d'une requête à partir d'un nom, peut apporter une information intéressante pour construire des modèles biométriques de manière semi-supervisée. Toutefois, elle ne peut pas fonctionner pour toutes les personnes dans les vidéos. Les personnes peu ou pas connues, comme dans l'exemple présenté au début de ce chapitre, ne retournent que très peu ou pas d'image leur correspondant. Ce qui ajoute donc des images parasites dans la construction des modèles biométriques. Un deuxième point important à noter est que le changement d'apparence des personnes dans le temps ou les conditions de prise de vue peuvent avoir une incidence sur la qualité des résultats. Pour les modèles de voix, il est évidemment encore plus difficile de collecter de bonnes données permettant de les construire.

Une proposition qui revient souvent dans les articles de l'état de l'art est la détection des présentateurs. Cette détection permet de les traiter à part mais elle est dépendante des choix éditoriaux. Par exemple, si on considère que le présentateur a toujours le même fond d'image fixe lorsqu'il apparaît, il est facile de le détecter. Or, ce n'est pas forcement une hypothèse juste pour toutes les émissions.

Une autre solution proposée est la détection du rôle des personnes dans la parole. Toutefois, elle est dépendante de données annotées spécifiques à chaque émission (détection du type discours, ...). Quitte à utiliser des données annotées, autant utiliser l'effort humain dans le sens de la tâche souhaitée, c'est-à-dire la construction de quelques modèles biométriques. Pour chaque émission, on peut en effet avoir une connaissance a priori des présentateurs, chroniqueurs et journalistes réguliers et donc avoir une liste restreinte de modèles biométriques à utiliser. Par ailleurs, les systèmes qui utilisent des modèles identifient les personnes avec une bonne précision lorsqu'ils ont le modèle adéquat.

Les méthodes plus globales où aucun lien local n'est fait entre noms et personnes (comme la correspondance d'un graphe biométrique et d'un graphe de lien entre noms) ne peuvent pas s'appliquer à notre cas. En effet, pour utiliser ce type de solution, il faut que les personnes interviennent plusieurs fois et que leurs noms puissent être extraits plusieurs fois. Ce n'est pas le cas dans les journaux télévisés par exemple, où de nombreuses personnes n'interviennent qu'une fois.

Ce qui pose la question de la granularité (vidéo *vs* émission *vs* collection). Il est intéressant de développer des méthodes qui puissent à la fois fonctionner à la granularité la plus fine (vidéo ou extrait vidéo), qui profite de la répétition d'apparition des présentateurs, journalistes sur plusieurs vidéos d'une même émission et qui ne soit pas ou peu dépendante des choix éditoriaux (qui généralise bien et donc qui soit peu dépendante de réglages spécifiques à une émission). On note donc l'importance d'avoir un corpus varié.

Sang et la. ont soulevé un dernier point important : le développement de méthodes d'association peu dépendantes des erreurs de regroupement voire des méthodes qui pourraient remettre en cause la constitution des clusters. Les noms écrits à l'écran apportent une information assez précise sur les personnes à identifier pour pouvoir corriger les erreurs issues de la détection ou du regroupement des tours de parole/visages en clusters.

Chapitre 2

Matériau expérimental

Nous avons vu dans le chapitre précédent de nombreux travaux sur l'identification des personnes dans les vidéos. Tous ces travaux sont basés sur un matériau expérimental pré-existant. Ce second chapitre est l'occasion de présenter celui sur lequel sont basés les travaux des chapitres suivants. En premier lieu, nous décrirons les deux corpus *JT France 2* et *REPERE* (section 2.1). Le premier a été créé au cours du projet *QUAERO*, le second pour la campagne d'évaluation *REPERE*.

Le défi *REPERE* s'intéresse à la reconnaissance des personnes dans les émissions de télévision. Pour ce défi, le Laboratoire d'Informatique de Grenoble (dans lequel cette thèse a été effectuée) et ses 6 autres partenaires ont constitué le consortium *QCompere*. Parmi les partenaires, on compte 4 équipes de recherche en plus du LIG (le LIMSI, l'équipe LEAR de l'INRIA, le laboratoire GREYC et le laboratoire KIT) ainsi que 2 entreprises (Vocapia et Yacast).

La section 2.2 décrit les différentes métriques utilisées dans ce manuscrit. Enfin, dans la section 2.3, nous décrirons les différentes briques de base utilisées au cours de cette thèse, issues de ces autres partenaires, à part LOOV (l'outil d'extraction des textes écrits) qui sera décrit dans le chapitre 3.

2.1 Corpus

Pour l'identification des personnes dans les émissions de télévision, nous avons besoin d'un corpus qui propose à la fois des annotations concernant le son (transcription de la parole, identités des locuteurs, ...) et des annotations concernant l'image (identités des visages, transcriptions des textes écrits, ...). Plusieurs corpus répondent en partie à ces besoins. On peut citer :

- *Hub-4*[1], *ESTER, ESTER2 et ETAPE*[2] : composés d'une grande variété d'enregistrements d'émissions radiophoniques d'actualités. Toutefois, ces corpus ne proposent pas l'aspect visuel que l'on recherche.
- *Canal9*[3] : 72 émissions de débats politiques. Ce corpus ne repose que sur un seul programme, ce qui ne permet pas d'être sûr que les méthodes

[1] www.ldc.upenn.edu
[2] www.afcp-parole.org
[3] http://canal9-db.sspnet.eu

développées soient généralisables. De plus il n'y a pas de transcription de la parole, ni des textes écrits à l'écran.
- *On n'a pas tout dit* [BEN11] : 5 vidéos de 50 minutes avec l'identité des personnes présentes et la transcription de la parole.
- *TRECVID*[4] : très vaste collection de tous types de vidéos, mais l'annotation sur les personnes n'est effectuée que pour quelques individus.

Ces corpus, de par leur manque de variabilité dans les types d'émissions et l'insuffisance des annotations, ne permettent pas un étude complète sur une identification multimodale des personnes dans les émissions de télévision.

Nous avons donc basé nos études sur deux autres corpus : le corpus *JT France 2* et le corpus *REPERE*. Le premier, déjà disponible en début de thèse, a été utilisé pour valider la faisabilité de l'utilisation des textes écrits à l'écran, pour l'identification des personnes présentes dans les journaux télévisés.

Au début de la deuxième année de cette thèse, les premières données du corpus *REPERE* ont été disponibles. Ce corpus est composé d'une plus grande variabilité dans le type d'émissions (journaux télévisés, émissions de débats, questions à l'Assemblée mais aussi une émission courte de 2 minutes sur l'actualité people). Ce corpus comporte aussi une annotation assez complète de plusieurs modalités.

2.1.1 JT France 2

Le corpus *JT France 2* a été créé pour le projet *QUAERO* et plus particulièrement pour la tâche d'identification des personnes dans les vidéos. Il est composé de 59 vidéos du journal télévisé de 20 heures de la chaîne France 2. Ces vidéos ont été enregistrées entre le 1 février 2007 et le 31 mars 2007. La durée moyenne est de 36 minutes ; ce qui représente un peu plus de 35 heures de vidéos. La qualité d'enregistrement est assez faible (MPEG1, 352x288, 25 frames/sec).

Le signal audio a été complètement annoté manuellement, avec l'identité des locuteurs quand elle est connue, ainsi que la transcription de la parole. Alors que pour l'image, la présence de 24 personnes (4 présentateurs, 18 politiciens, et 2 sportifs) a été annotée sur 1 image par seconde. Les autres personnes apparaissant dans ces images ont été annotées comme inconnues.

Pour évaluer la capacité des noms écrits à l'écran à identifier les personnes présentes il nous a d'abord fallu évaluer la qualité d'extraction des textes écrits à l'écran. Nous avons donc effectué deux types d'annotations :
- La détection des textes : nous avons annoté la présence de textes en surimpression sur 2 heures de vidéos (3 journaux télévisés). 512 boîtes de texte ont été annotées temporellement et spatialement. Ces annotations ont pris environ 4 heures.
- La transcription des textes : nous avons annoté 29166 images extraites automatiquement [QMB03]. 4414 comportaient du texte écrit dans 9256 boîtes de texte. Le texte de ces boîtes a été corrigé manuellement après une transcription automatique. Lorsque ce texte correspondait au nom ou à la

[4] http://trecvid.nist.gov

fonction d'une personne, une étiquette a été ajoutée. Une autre étiquette a été utilisée pour indiquer si la personne correspondante était visible. Ces annotations ont pris environ 100 heures.

Nous pouvons voir dans les images 2.1 quelques captures d'écran issues de ce corpus. Sur les trois images du bas (d, e, f), on peut voir que les noms écrits dans le cartouche introduisent les personnes correspondant à l'image et probablement dans la bande son. Sur les images (a, d), du texte de scène est visible en fond d'image. Ce type de texte n'apporte pas d'information pour la reconnaissance des personnes dans les vidéos, il n'a donc pas été annoté. Certains textes sont directement sur-imprimés sans fond (a, b), alors que, lorsqu'ils sont écrits dans un cartouche, un fond gris transparent est ajouté avant (d, e, f).

(a) (b) (c)

(d) (e) (f)

FIG. 2.1 – Exemples d'images du corpus *JT France 2*

La qualité d'écriture de ces textes est montrée sur la figure 2.2. Sur les images (a, c), malgré un fond uni, la compression et la basse résolution des images rendent difficile leurs lectures. L'image (b) montre que le fond peut gêner la lecture alors que le fond gris transparent (d) permet de bien différencier les caractères du fond.

(a) (b)

(c) (d)

FIG. 2.2 – Exemples de textes sur-imprimés dans le corpus *JT France 2*

2.1.2 *REPERE*

Le corpus *REPERE* (REconnaissance de PERsonnes dans des Emissions audiovisuelles) a été constitué pour le challenge du même nom organisé par l'ANR[5]. Ce challenge propose de réaliser un système de reconnaissance de personnes dans des émissions audiovisuelles, en s'appuyant sur les différentes sources d'informations présentes dans ces émissions.
Cette reconnaissance peut se traduire par la réponse à quatre questions :

- Qui parle ?
- Qui apparaît ?
- Quel est le nom écrit à l'écran ?
- Quel est le nom prononcé ?

En plus de ces quatre questions principales, de multiples sous-tâches sont évaluées au cours du challenge *REPERE* ; par exemple, la segmentation des têtes incrustées, la transcription de la parole, etc. Chacune de ces sous-tâches correspond à des briques de base pour les 4 tâches principales.

Ce corpus est composé de 7 émissions différentes enregistrées sur deux chaînes de télévision française, BFMTV et LCP. Ce sont des émissions de journaux télévisés, des émissions de débats, les questions à l'Assemblée mais aussi une émission courte de 2 minutes sur l'actualité people. Les enregistrements ont été effectués au cours des années 2011 et 2012, au format 720*576 en mpeg2. La très bonne qualité de ces enregistrements nous permettra d'utiliser les textes écrits à l'écran. Le corpus *REPERE* a été constitué en 3 phases. Chaque phase est accompagnée d'un nouveau jeu de données. On peut voir dans le tableau 2.1 le détail de la répartition (effectuée par les organisateurs du défi *REPERE*) du nombre d'heures de vidéo sur la première et la deuxième phase ; la troisième phase étant postérieure à l'écriture de ce manuscrit.

Phase	Segments	Apprentissage	Développement	Évaluation
Dryrun	Vidéos complètes	X	14h	13h
	Segments UEM	X	3h	3h
Phase1	Vidéos complètes	58h	13h	15h
	Segments UEM	24h	3h	3h

Tab. 2.1 – Répartition du nombre d'heures des vidéos du corpus *REPERE* sur les deux premières phases

Annotations

Pour le défi *REPERE*, ces vidéos ont été partiellement annotées, un ou plusieurs segments UEM[6] ont été sélectionnés sur chacune d'elles. Sur ces segments UEM, la modalité audio a été complètement annotée manuellement alors que pour la modalité image, seulement une image par plan et au moins une image

[5]ANR : Agence nationale de la recherche
[6]Segments UEM : segments à traiter pour l'évaluation (Unpartitioned Evaluation Map)

toutes les dix secondes a été annotée. Pour chaque plan, l'image annotée a été choisie en essayant de maximiser la quantité d'informations contenues (nombre de visages et bonne orientation de ceux ci, présence de texte...).

Pour la modalité audio, une transcription manuelle de la parole a été effectuée. Les noms de personnes ont été étiquetés et les locuteurs ont été identifiés lorsqu'ils étaient connus.

Pour la modalité image, les visages ont été détourés et identifiés lorsque la personne n'était pas inconnue. Même les visages partiellement visibles ou au second plan, si leur taille n'était pas trop petite (supérieure à 2000 pixels carré), ont été identifiés. Le texte en surimpression a été, lui aussi, détouré, transcrit et les noms de personnes ont été étiquetés. Si le nom était écrit dans un cartouche, un marquage supplémentaire a été ajouté.

Des publicités ainsi que des émissions en dehors de celles ciblées peuvent être contenues dans les vidéos brutes, mais aussi des segments non annotés des émissions ciblées. Ces segments non annotés peuvent permettre d'extraire plus de noms, d'avoir plus d'occurrences d'un nom ou encore d'avoir des noms peu souvent cités (les présentateurs peuvent avoir leurs noms cités seulement en début de journal). L'annexe A montre la répartition du nombre de vidéos et de la durée de ces vidéos (vidéos brutes et segments UEM) des émissions sur ce corpus. On trouvera plus de détails sur le corpus *REPERE* et les annotations manuelles dans [GCM+12].

(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

FIG. 2.3 – Exemples d'images du corpus *REPERE*

La figure 2.3 présente quelques exemples d'images extraites du corpus *REPERE*. Dans ces exemples, nous pouvons voir que les conditions d'enregistrement peuvent être variables : studio (a,d,e,g,f), extérieur (b), salle de meeting (f), assemblée nationale (i), etc. Les visages peuvent être de face ou de profil (b, h, i), de grande ou de petite taille (f). Il est à noter que, dans l'image (i), trois personnes ont été identifiées alors qu'une seule est le sujet principal de l'image.

La qualité des images rend beaucoup plus lisibles les textes écrits (voir figure 2.4) et donc leur transcription sera plus facile.

(a)

(b)

FIG. 2.4 – Exemples de textes sur-imprimés dans le corpus *REPERE*

Les personnes présentes

Dans ce corpus, la grande majorité des personnes parlant/apparaissant ont été identifiées. La figure 2.5 nous montre la répartition entre connues et inconnues. Seulement 1.5% du temps de parole de l'ensemble d'apprentissage n'a pu être identifié.

Pour les visages, cette proportion est plus importante (environ 30%). Certains visages au second plan peuvent avoir une surface supérieure à 2000 pixels carrés (limite de l'annotation) mais ne pas faire partie du sujet de la vidéo et donc être difficilement identifiables lors de l'annotation.

Les personnes par rôle

Pour le corpus *REPERE*, cinq types de catégories différentes ont été définies pour classer les personnes présentes (présentateur, chroniqueur, reporter, invité, autre). Dans le tableau 2.2, on retrouve la répartition du temps de présence et du nombre de personnes (sur l'ensemble de test de la phase 1) en fonction du rôle qu'elles occupent dans les vidéos. Il est à noter que la somme du nombre de personnes à reconnaître par rôle ne correspond pas au nombre de personnes à reconnaître dans la première ligne. Les personnes pouvant avoir des rôles différents selon l'émission.

Sur l'ensemble de test de la phase 1, un peu moins de 3 heures ont été sélectionnées pour l'annotation. Pour le signal audio, 10615 secondes de signal audio ont été annotées (10148 correspondaient à du temps de parole). A cela, il faut ajouter 199 secondes de parole superposée. Durant ces 3 heures de vidéos annotées, 1252 images ont été sélectionnées. Sur ces images, 1151 contenaient au

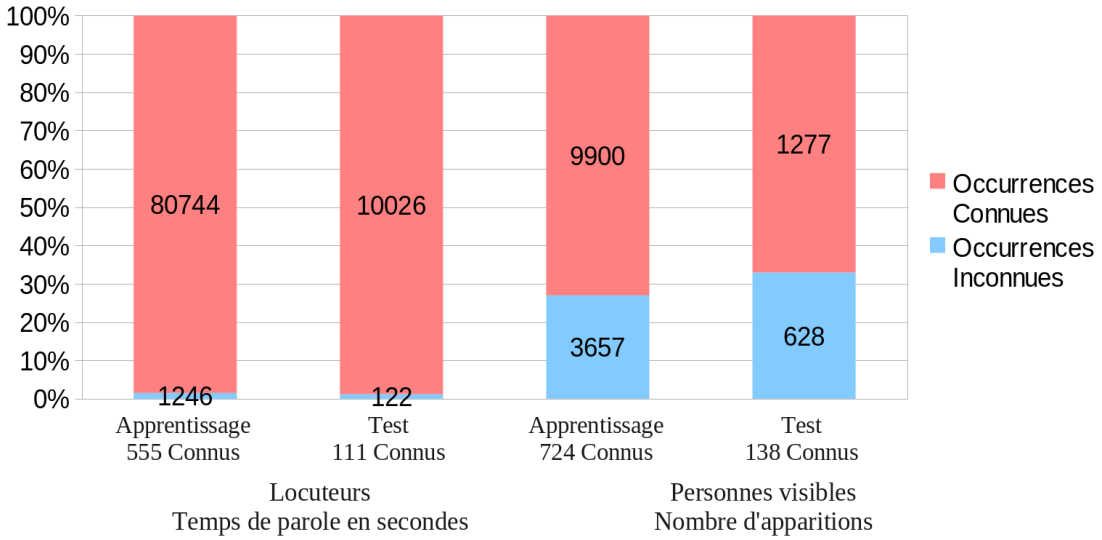


FIG. 2.5 – Répartition des personnes connues et inconnues sur l'ensemble d'apprentissage et de test.

Type	# locuteur	Temps de parole en secondes	# personnes apparaissant	# apparition à l'écran
all	111	10347	138	1449
R1	11 (10%)	2204 (21%)	11 (8%)	244 (19%)
R2	7 (6%)	1069 (10%)	8 (6%)	232 (14%)
R3	15 (13%)	1052 (10%)	5 (4%)	26 (2%)
R4	30 (27%)	3751 (36%)	36 (26%)	492 (35%)
R5	50 (44%)	2471 (23%)	81 (57%)	455 (29%)
R123	33 (29%)	4325 (41%)	23 (17%)	502 (35%)
R45	80 (71%)	6222 (59%)	115 (83%)	947 (65%)

TAB. 2.2 – Répartition du temps de présence en fonction du rôle des personnes dans les vidéos, ensemble de test de la phase 1 du corpus *REPERE*

moins un visage de plus de 2000 pixels carrés avec un total de 1277 visages connus.

Ces 5 rôles peuvent être regroupés en deux catégories :

- R123 : pour les personnes des rôles R1, R2 et R3 (présentateur, chroniqueur, journaliste), il est facile d'avoir un a priori sur leur présence dans les émissions.
- R45 : pour les personnes des rôles R4, R5 (invité, autre), il est difficile de connaître leur présence au préalable.

Entre ces deux catégories, on voit apparaître (dans le tableau 2.2) une légère disproportion du temps de présence par rapport au nombre de personnes dans ces deux catégories. Les personnes des rôles R123 représentent proportionnellement un peu plus du temps de parole/nombre d'apparition que la proportion du nombre de personnes dans ces rôles.

Plus le corpus est grand, plus cette disproportion est importante. Si on compare le temps de présence entre l'ensemble de test (3 heures annotées) et l'ensemble d'entraînement qui est 8 fois plus gros (24 heures annotées), le nombre de personnes différentes de R123 change peu (2 à 2.5 fois plus, 33 à 84 pour les locuteurs, 24 à 48 pour les visages) alors que le nombre de R45 a été multiplié par 6 (80 à 475 pour les locuteurs, 117 à 680 pour les visages) (voir tableau 2.3).

Rôle	# locuteur	Temps de parole en secondes	# personnes apparaissant	#apparition à l'écran
R123	84 (15%)	37920 (45%)	48 (7%)	2935 (30%)
R45	475 (85%)	46980 (55%)	680 (93%)	6861 (70%)

TAB. 2.3 – Répartition de la présence des personnes en fonction de leurs rôles, phase 1 du corpus *REPERE*, partie apprentissage. **R1,2,3** : Présentateur/chroniqueur /reporter, **R4,5** : Invité/autre.

Cette différence entraîne une augmentation de la disproportion du temps de présence des personnes en fonction du rôle qu'elles occupent. Les personnes de R123 occupent 45% du temps de parole alors qu'elles ne correspondent qu'à 15% des locuteurs. Elles représentent aussi 30% des visages visibles alors qu'elles n'appartiennent qu'à 7% des personnes visibles. Il est à noter que seulement 48 des 84 personnes de R123 sont visibles (voix-off des journalistes). Et inversement 475 personnes de R45 parlent alors que 680 sont visibles (le résultat est principalement dû aux personnes annotées comme visibles mais au second plan comme sur l'image de l'Assemblée Nationale 2.3.i).

2.2 Métriques

Avant de décrire les briques de bases que nous avons utilisé, nous allons détailler les métriques utilisées pour évaluer l'ensemble de ces briques ainsi que les méthodes que nous proposons pour évaluer certains aspects de nos recherches.

2.2.1 Transcription de la parole : WER et CER

Les métriques qui vont évaluer la qualité de transcription de la parole et des textes écrits à l'écran sont le WER (Word Error Rate) qui calcule le taux d'erreur de mots et le CER (Character Error Rate) pour le taux d'erreur de caractères.

$$\text{WER ou CER} = \frac{I + D + S}{R} \tag{2.1}$$

Ce taux d'erreur est calculé en alignant l'hypothèse et la référence. Pour la parole, l'alignement est effectué à l'aide de trois opérations : insertions (I), suppressions (D) et substitutions (S). Le taux d'erreur compte le nombre minimum d'opérations, effectuées au niveau des mots ou au niveau des caractères, pour transformer l'hypothèse en la référence divisée par le nombre (R) total de mots ou de caractères de la référence.

Pour la parole une normalisation est effectuée sur la référence et l'hypothèse avant évaluation :

- Dé-capitalisation.
- Suppression des ponctuations.
- Remplacement des tirets par des espaces.
- Tokénisation : séparation des mots et des apostrophes (l'autre → l' autre) sauf pour un nombre limité d'exceptions (aujourd'hui).

Pour les textes écrits à l'écran, avant de pouvoir calculer le taux d'erreur par les mêmes opérations, il faut relier les boîtes de textes de la référence à celles de l'hypothèse. Le corpus *JT France 2* ne contenant pas les positions spatiales des textes à évaluer, les boîtes de textes de la référence et de l'hypothèse sont reliées lorsque la similarité (basée sur la distance de Levenshtein) est supérieure à 50%. Le calcul du taux d'erreur est le même que pour la parole, avec en plus les erreurs de toutes les boîtes de textes de la référence auxquelles aucune boite de textes de l'hypothèse n'a pu correspondre.

Pour les textes écrits aussi, une normalisation est effectuée sur la référence et l'hypothèse avant évaluation :

- Dé-capitalisation.
- Dé-accentuation.

2.2.2 Slot Error Rate : SER

Pour évaluer la qualité des systèmes d'extraction des entités nommées de type nom de personne à partir des transcriptions de la parole, nous avons utilisé le SER (Slot Error Rate).

$$\text{SER} = \frac{I + D + 0.5 * (T + F)}{R} \tag{2.2}$$

Cette métrique réalise un alignement forcé entre les noms de la référence et ceux de l'hypothèse. Cet alignement, avec une tolérance de 250 millisecondes, nous permet d'obtenir 4 types d'erreurs :

- Erreurs d'insertion (I) : l'hypothèse propose un nom alors que la référence n'en propose pas dans l'intervalle temporel autour du nom de l'hypothèse.
- Erreurs de suppression (D) : la référence propose un nom alors que l'hypothèse n'en propose pas dans l'intervalle temporel autour du nom de la référence.
- Erreurs de type (T) : la référence et l'hypothèse proposent des noms différents dans le même intervalle de temps.
- Erreurs de frontière (F) : la référence et l'hypothèse proposent le même nom mais l'intervalle de temps entre les deux est supérieur à la tolérance de 250 millisecondes.

Le SER est alors calculé en cumulant les erreurs d'insertion et de suppression avec un poids de 1, les erreurs de type et de frontière avec un poids de 0,5. Le tout est divisé par le nombre d'entrées de la référence (R).

2.2.3 Diarization Error Rate : DER

Pour évaluer la qualité de la diarization ou regroupement, nous utilisons la métrique classique du domaine, la DER. Cette métrique compte la durée des erreurs par rapport au temps total de parole.

$$\text{DER} = \frac{\#\text{fa} + \#\text{miss} + \#\text{conf}}{\#\text{total}} \tag{2.3}$$

Où

- #fa : durée de parole où l'hypothèse indique un locuteur alors que personne ne parle.
- #miss : durée de parole où la référence indique un locuteur alors que l'hypothèse non.
- #conf : durée de parole où l'hypothèse et la référence sont en désaccord sur le locuteur.
- #total : durée de parole de la référence.

Pour cette métrique, l'identité des locuteurs est inconnue, donc la durée de la confusion n'est pas connue immédiatement. Elle est calculée après la sélection de l'alignement 1-à-1 entre les locuteurs de la référence et ceux de l'hypothèse qui minimisent cette confusion. Les frontières des interventions des locuteurs étant difficiles à poser, une tolérance de 250 millisecondes est utilisée autour des frontières de la référence pour minimiser leur impact sur la DER.

2.2.4 Estimate Global Error Rate : EGER

La métrique officielle du défi *REPERE*, l'EGER, estime le taux d'erreur d'identification des personnes. Cette métrique comptabilise 3 types d'erreurs sur des instantanés de vidéos (images annotées manuellement). Sur chacune de ces images, deux listes de personnes sont constituées, une pour la référence et l'autre pour l'hypothèse. Ces deux listes sont comparées en associant les personnes une à une. Chaque personne ne pouvant être associée au plus qu'une fois.
3 types d'erreurs peuvent apparaître :

- Confusion (#conf) : nombre d'associations entre deux personnes avec des noms différents.
- Fausse alarme (#fa) : nombre de personnes de l'hypothèse non associées.
- Oubli (#miss) : nombre de personnes de la référence non associées.

De toutes les associations possibles, est choisie celle qui donne le nombre d'erreurs le plus faible. La somme de tous ces comptes d'erreurs par image permet d'obtenir le nombre d'erreurs global. Le nombre global d'entrées attendues (#total) est lui aussi comptabilisé en cumulant le nombre de personnes présentes dans la référence à chaque image. Le taux d'erreur est alors le nombre d'erreurs global divisé par le nombre global d'entrées attendues :

$$\text{EGER} = \frac{\#\text{fa} + \#\text{miss} + \#\text{conf}}{\#\text{total}} \tag{2.4}$$

Ne sont pas considérés les visages inférieurs à 2000 pixels carrés et les personnes qui n'ont pu être identifiées pendant l'annotation

2.2.5 Précision, Rappel, F-mesure

La précision (P), le rappel (R) et la F-mesure (F) mesurent la capacité d'un système à trouver l'information pertinente (le nom des locuteurs, les noms écrits, etc). Ils sont définis à partir du nombre d'éléments correctement annotés comme appartenant à la classe positive (vp : vrai positifs), du nombre d'éléments annotés positifs alors qu'ils ne l'étaient pas (fp : faux positifs) et du nombre d'éléments annotés négatifs alors qu'ils étaient positifs (fn : faux négatifs).

La **précision** correspond à la fraction d'éléments correctement retrouvés par le système par rapport au nombre d'éléments retournés par le système :

$$\text{Précision} = \frac{vp}{vp + fp} \tag{2.5}$$

Le **rappel** correspond à la fraction d'éléments correctement retrouvés par le système par rapport au nombre d'éléments à retrouver :

$$\text{Rappel} = \frac{vp}{vp + fn} \tag{2.6}$$

La **F-mesure** est la combinaison pondérée de la précision et du rappel de façon égale :

$$F = \frac{2PR}{P + R} \tag{2.7}$$

2.3 Technologies externes utilisées

Dans cette section, nous allons décrire les systèmes utilisés comme briques de base pour les travaux effectués au cours de cette thèse. Ces systèmes ont été produits par les autres partenaires du consortium QCompere. Le tableau 2.4 récapitule les différentes compétences apportées par nos partenaires.

Partenaire	Compétence
KIT	Détection et suivi des visages
INRIA - LEAR	Création d'une matrice de distance entre séquences de visages Identification des visages à partir des modèles biométriques
LIMSI	Segmentation et regroupement en locuteurs Identification des locuteurs à partir de modèles biométriques Transcription de la parole Détection des entités nommées

Tab. 2.4 – Récapitulatif des compétences apportées par nos partenaires.

Nous allons commencer par ceux travaillant sur le traitement de la vidéo et plus particulièrement sur les visages de personnes (détection, matrice de similarité et identification) puis poursuivre par ceux travaillant sur l'audio (diarization et identification des locuteurs, transcription de la parole et détection des noms dans les transcriptions).

2.3.1 Traitements vidéo

Pour l'extraction des visages des personnes visibles, plusieurs niveaux de granularité doivent être considérés (voir image 2.6) :

- **Niveau image** : image d'un visage, encadré en vert.
- **Niveau plan** : séquence de visages, elle est constituée de plusieurs images de visage de face ou non dans le même plan (cadre rouge).
- **Niveau vidéo/collection** : cluster de personnes, il est constitué de plusieurs séquences de visages apparaissant dans la même vidéo ou dans plusieurs vidéos (ovale bleu).

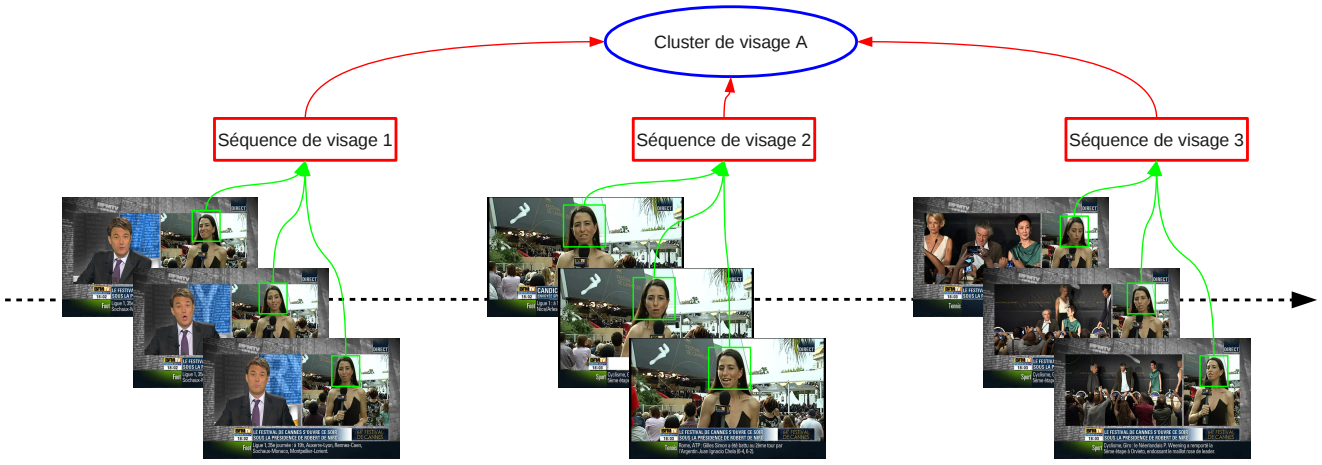


FIG. 2.6 – Trois niveaux de granularité : image de visage (vert), séquence de visages (rouge), cluster de visages (bleu)

Pour cette modalité, nous avons utilisé le travail de deux équipes de recherche du consortium *QCompere*. Le laboratoire KIT[7] nous a fourni la détection et le suivi des visages pour constituer des séquences de visages (les deux premiers niveaux). L'équipe LEAR du laboratoire INRIA nous a fourni les matrices de distance entre séquences de visages et l'identification des séquences de visages à partir de modèles biométriques (le dernier niveau).

2.3.1.1 Détection et suivi des visages, KIT

La première tâche, détection et suivi des visages, s'intéresse à la constitution des séquences de visages à partir des images d'une vidéo. Chaque vidéo est d'abord divisée en plans, ensuite la détection des visages est effectuée en utilisant un système à base de filtres particulaires [BBF+10]. Les visages sont initialisés par un balayage de la première image de chaque plan et toutes les 5 trames suivantes. Ce balayage utilise un détecteur de visages de face, de trois quart et de profil [FE04], ce qui rend la détection indépendante de la pose initiale. Le détecteur de visages atteint déjà un taux de faux positifs faible, qui est encore réduit par le suivi des visages sur les trames successives.

[7]KIT : Karlsruhe Institute of Technilogy

Ce suivi est effectué de manière linéaire en utilisant l'état de la trame précédente pour déduire la position, la taille et l'orientation des têtes dans la trame courante. 11 détecteurs d'orientation de tête sont combinés pour mesurer chaque particule d'une séquence de visage. Ce traqueur est proche du temps réel ($\sim$ 8 fps[8]). L'image 2.7 montre un exemple de détection de visages effectuée par ce système.

FIG. 2.7 – Exemples de détection de visage

2.3.1.2 Création d'une matrice de distance entre séquences de visages et identification des séquences de visages, INRIA

A partir des séquences de visages détectées par le système décrit ci-dessus, il est intéressant d'exploiter deux types d'information :

- La distance entre une séquence de visages et un modèle biométrique. Cette distance permet de faire de l'identification de visages.
- La distance entre deux séquences de visages d'une même vidéo. A partir de la distance entre tous les séquences, un regroupement en clusters de visages peut être effectué.

Pour calculer ces distances, il faut d'abord extraire un descripteur caractérisant chaque séquence de visages.

Extraction d'un descripteur par séquence - modèle

Pour chacune des images de visage d'une séquence, un maillage de neuf points [ESZ06] (2 par œil, 3 pour le nez, 2 pour le coin des lèvres) a été automatiquement appliqué. Ce maillage doit être utilisé seulement sur les visages de face. Un score de confiance, lié à la position des 9 points, indique si un visage est exploitable ou non. L'image 2.8.a. présente une détection des points correctement alignés. Les

[8]fps : frames per seconds - trames par secondes

points sur l'image 2.8.b. ont obtenu un score de confiance sous le seuil fixé donc cette image n'a pas été exploitée.

Ce maillage de neuf points permet de calculer un descripteur global pour chaque image de visage lorsque cette image est exploitable. Ce descripteur global correspond à la concaténation des 9 descripteurs (HOG à 490 dimensions) des points du visages [DT05]. Le descripteur HOG est utilisé avec une taille de bloc rectangulaire de 7×7 cellules de 7×7 pixels chacune et 10 bins d'orientation.

(a) (b) (c)

FIG. 2.8 – Exemples d'alignement du maillage de 9 points : (a) bien alignés (vert), (b) et (c) mal alignés (rouge)

Ces vecteurs permettent de construire :

- Un descripteur moyen pour chaque séquence de visages : ce descripteur moyen correspond à la moyenne des vecteurs de chaque image d'une séquence. Ces descripteurs moyens ont été calculés, pour les visages détectés lorsqu'au moins une image de visage est exploitable (confiance dans les 9 points alignés supérieure à un seuil), sur les ensembles d'apprentissage, de développement et de test du corpus *REPERE*.
- Un modèle biométrique pour chaque personne de l'ensemble d'apprentissage : chaque personne visible dans l'ensemble d'apprentissage a été annotée comme correspondant à une ou plusieurs séquences de visages. Les modèles biométriques ont été construits à partir de la moyenne des vecteurs de chaque image de chaque séquence correspondant à ces personnes.

Identification des visages

A partir de ces descripteurs moyens, un classifieur des k plus proches voisins (KNN) est utilisé pour calculer des matrices de distance entre séquences à identifier et modèles biométriques. Le classifieur n'utilise pas directement les descripteurs moyens, ils sont d'abord projetés dans un espace à 200 dimensions à l'aide de l'approche LDML[9] [GMVS12].

A partir de ces distances projetées, le classifieur peut calculer les scores de proximité de type séquence-séquence ou séquence-modèle. Avec ces scores deux

[9]LDML : Logistic Discriminant Metric Learning

matrices sont construites. La première, la matrice « séquence versus séquence », nous donne la distance entre les séquences de visages d'une même vidéo. La seconde, la matrice « séquence versus modèle », nous donne la distance entre les séquences de visages de l'ensemble de test et les modèles biométriques constitués à partir de la base d'apprentissage.

Ces deux matrices sont utilisées consécutivement pour identifier les visages des collections :

- D'abord, on effectue une identification à l'aide de la matrice « séquence versus modèle » avec une forte précision : un visage sera nommé avec le nom du modèle le plus proche s'il y a une faible distance entre le visage et le modèle le plus proche et une forte distance entre le visage et le modèle arrivé second.
- Ensuite, on propage les noms des séquences nommées aux autres séquences encore anonymes à l'aide de la matrice « séquence versus séquence ».

Pour ces deux étapes, un seuil est appris pour arrêter le processus.

2.3.1.3 Identification à partir des modèles biométriques

Pour connaitre la capacité des modèles biométriques à bien identifier les visages présents, nous avons mis en place trois oracles. Ces oracles prennent en compte deux paramètres : est-ce que le visage a bien été détecté et est-ce que le visage a un modèle correspondant.
Nous avons donc défini les oracles comme suit :

- L'oracle **modèle** nomme correctement les visages annotés dans la référence et qui ont un modèle construit sur la base d'apprentissage.
- L'oracle **visage** nomme correctement les visages qui ont été détectés et qui ont pu être exploités par le module d'identification des visages (confiance dans l'alignement des 9 points au-dessus d'un seuil).
- L'oracle **visage-modèle** est une combinaison des contraintes des deux oracles ci-dessus.

Ces oracles et le système automatique vont utiliser les 557 modèles construits à partir du corpus d'apprentissage *REPERE*. On peut voir l'ensemble des résultats dans le tableau 2.5 (protocole d'évaluation du défi *REPERE*).

Système	# ref	# hyp	# correct	Précision	Rappel
Oracle-modèle	1449	934	934	100.0	64.5
Oracle-visage	1449	894	869	97.2	60.0
Oracle-visage-modèle	1449	575	575	100.0	39.7
KNN	1449	779	467	59.9	32.2

Tab. 2.5 – Identification des visages sur les images annotées du corpus *REPERE*, ensemble de test de la phase 1

Seulement 78 des 557 modèles correspondaient à une personne dans l'ensemble de test alors qu'il y avait 138 personnes visibles différentes à nommer. Sur les

images annotées, on obtient 64.5% des visages nommables (rappel oracle-modèle). Cependant, certaines séquences de visages n'ont pas pu être exploitées par le système automatique d'identification, ce qui abaisse à 60.0% (869) le nombre de visages nommables (rappel oracle-visage). Parmi eux, seulement 39.7% ont un modèle biométrique leur correspondant (rappel oracle visage-modèle). Le système automatique d'identification essaye de nommer 779 visages des 894 détectés/exploitables avec un précision de 59.9%, il obtient un rappel de 32.2%.

Malgré les 557 modèles de visage, ce système n'obtient pas de très bons résultats à cause du manque de couverture des personnes à reconnaître. Notre hypothèse est que les modèles biométriques peuvent être utilisés avec efficacité seulement pour les personnes des rôles R123 (présentateur, chroniqueur et reporter). C'est-à-dire que nous n'allons utiliser que les modèles des personnes annotées comme R123 visibles sur l'ensemble d'apprentissage d'un type d'émission pour détecter les personnes visibles dans le même type d'émission sur l'ensemble de test. Par exemple, pour *BFM Story* le système n'utilisera que les 22 modèles qui ont pu être construits (voir tableau en annexe A) sur les vidéos de *BFM Story* de l'ensemble d'entrainement.

L'oracle-modèle couvre 88% des visages de R123 à reconnaître sur l'ensemble de test (voir tableau 2.6), ce qui correspond à près du tiers des visages à identifier (442 des 1449) en dépit du peu de modèles utilisés. Si on enlève les séquences inexploitables, on obtient 57.0% de visages nommables (rappel de l'oracle-visage-modèle R123). Le système automatique, n'utilisant que cette liste restreinte de modèles, nomme correctement 50.8% des visages de R123 avec une précision de 96.6%. Donc, ce système automatique obtient de très bons résultats seulement si il essaye d'identifier les personnes dont il a le modèle correspondant.

Système	# ref	# hyp	# correct	Précision	Rappel
Oracle-modèle R123	502	442	442	100.0	88.0
Oracle-visage-modèle R123	502	286	286	100.0	57.0
KNN R123	502	264	255	96.6	50.8

Tab. 2.6 – Identification des visages de R123 sur les images annotées du corpus *REPERE*, ensemble de test de la phase 1

2.3.2 Traitements audio

Pour cette modalité, toutes les briques ont été fournies par le LIMSI, avec : la diarization, deux systèmes d'identification des locuteurs ainsi que l'extraction des noms prononcés à partir de la transcription de la parole.

2.3.2.1 Segmentation et regroupement en locuteurs

La segmentation et le regroupement en locuteurs (ou diarization) correspond au processus de découpage du signal audio en clusters homogènes, sans connaissance des voix des locuteurs (sans modèles biométriques). A l'issue de ce proces-

sus, chaque cluster est labellisé avec un label inconnu. Pour que le résultat soit parfait, chaque cluster doit correspondre à un seul locuteur et inversement.

Le système du LIMSI, issu de [BZMG06], utilise une méthode agglomérative (bottom-up) en plusieurs étapes :

1. **Segmentation en tours de parole** : avant de segmenter le signal en tours de parole, il faut détecter les zones de parole des zones de non-parole. Une fois cette étape réalisée, les zones de parole sont segmentées par la recherche des points de rupture dans le flux audio. Ces points de rupture sont détectés à l'aide de deux fenêtres glissantes adjacentes se déplaçant sur le signal audio. A chaque pas d'avancement, les données contenues dans chaque fenêtre sont comparées à l'aide du critère d'information bayésien (BIC[10]) [CG98]. Si le signal est mieux modélisé par deux distributions au lieu d'une seule, un point de rupture est détecté et un nouveau tour de parole commence. Chaque tour de parole ou cluster est labellisé avec un identifiant unique.
2. **Regroupement hiérarchique BIC** : le regroupement hiérarchique va d'abord calculer une matrice de similarité entre tous les clusters de locuteurs (via BIC) puis fusionner les deux clusters les plus proches. De manière itérative, il va recommencer le calcul de la matrice (mise à jour seulement) puis la fusion des clusters les plus proches jusqu'à un critère d'arrêt.
3. **Re-décodage du signal** L'ensemble des clusters est ensuite modélisé à l'aide d'un modèle HMM et le signal audio est complètement re-décodé en affectant aux segments le label du modèle le plus proche. Cette étape permet de corriger certains regroupements faits dans les deux étapes précédentes.
4. **Regroupement hiérarchique CLR** : La dernière étape va encore regrouper quelques clusters à l'aide de la métrique CLR[11] [RSC+98]. Cette métrique plus précise que BIC a tout de même besoin de plus de temps de signal pour modéliser correctement le signal audio, c'est pour cette raison qu'elle est effectuée après un regroupement à l'aide de BIC.

Nous retrouvons les résultats de cette diarization, évaluée à l'aide de la métrique DER, dans le tableau 2.7. Le système BIC correspond à la sortie du système après le regroupement hiérarchique BIC (étape 3).

Système	Ensemble d'apprentissage	Ensemble de test
BIC	22	18.4
CLR	15.7	13.3

TAB. 2.7 – Résultats de la DER sur le corpus *REPERE*, ensemble de test de la phase 1.

[10]BIC : Bayésien Information Criterion
[11]CLR : Cross Likelihood Ratio

2.3.2.2 Identification des locuteurs

Les deux systèmes d'identification supervisée des locuteurs [LBF0] vont utiliser une sortie de diarization comme celle présentée ci-dessus. Le premier est un système classique basé sur le modèle GMM-UBM[12] et l'autre sur un modèle GSV-SVM. Ce dernier utilise un classifieur SVM pour chaque locuteur entrainé avec un super vecteur (construit à partir de la concaténation de la moyenne des modèles GMM-UBM adaptés). Pour les deux systèmes, chaque cluster est comparé à tous les modèles de locuteurs du même genre (féminin-masculin). Le modèle ayant obtenu le meilleur score est choisi si son score est plus élevé que le seuil de décision.

Pour construire les modèles de locuteurs, 3 sources ont été utilisées : les données d'apprentissage du corpus *REPERE*, les données d'apprentissage et développement du corpus *ETAPE*[13], et des données issues de la radio (pour les politiciens français). Ces trois sources ont permis de construire 626 modèles de locuteurs.

Sur l'ensemble de test, 111 locuteurs différents ont parlé sur les images annotées, mais seulement 65 d'entre eux avaient un modèle construit à partir des bases d'apprentissage. Pour comparer les systèmes automatiques, nous avons mis une première ligne « Oracle modèle » dans le tableau 2.8. Cet oracle nous permet de voir quel serait le résultat obtenu par un système parfait mais limité aux modèles appris à partir des 3 sources citées précédemment.

Système	# ref	# hyp	# correct	Précision	Rappel
Oracle modèle	1229	860	860	100.0	70.0
GMM-UBM	1229	1156	612	52.9	49.8
GSV-SVM	1229	1101	666	60.5	54.2

TAB. 2.8 – Identification des locuteurs sur les images annotées du corpus *REPERE*, ensemble de test de la phase 1

On peut voir que, malgré presque deux fois moins de modèles utilisables que de personnes intervenants dans l'ensemble de test, le rappel sur les images annotées est de 61.8% pour cet oracle. Le système de base GMM-UBM obtient de moins bons résultats que le système GSV-SVM. Toutefois, comme pour le visage, le manque de couverture des locuteurs ne permet pas d'obtenir de bons résultats.

Nous avons posé la même hypothèse que pour les visages : utiliser les modèles biométriques, seulement pour les locuteurs de R123 de chaque émission, appris sur la base d'apprentissage (voir l'annexe A pour le détail du nombre de modèles construits par type d'émission).

Dans le tableau 2.9, nous retrouvons l'oracle-modèle utilisant cette liste restreinte de modèles avec un très bon rappel. Ce qui démontre que l'on peut avoir une connaissance a priori des présentateurs / chroniqueurs / journalistes suscep-

[12]GMM-UBM : Gaussian Mixture Model-Universal Background Model
[13]http ://www.afcp-parole.org/etape.html

tibles d'intervenir dans chaque émission. Les deux systèmes automatiques obtiennent aussi de bons résultats pour les personnes des rôles R123.

Système	# ref	# hyp	# correct	Précision	Rappel
Oracle-modèle R123	510	443	443	100.0	86.9
GMM-UBM R123	510	398	364	91.5	71.4
GSV-SVM R123	510	408	371	90.9	72.7

TAB. 2.9 – Identification des locuteurs de R123 sur les images annotées du corpus *REPERE*, ensemble de test de la phase 1

2.3.2.3 Reconnaissance automatique de la parole et détection des entités nommées

Transcription de la parole

Pour extraire les noms prononcés, il faut tout d'abord transcrire la parole. Le LIMSI nous a fourni les sorties de leur système Speech-To-Text [LCD+11] pour le corpus *REPERE*. Ce système a été développé pour la transcription du français. Il n'utilise pas d'adaptation spécifique pour les corpus cibles, c'est à dire qu'aucun modèle acoustique ou modèle de langage n'a été re-entraîné sur les données du corpus *REPERE*.

Ce système de reconnaissance de la parole utilise la même technique statistique de modélisation et de décodage que le système « LIMSI English BN » [GLA02]. Avant la transcription à proprement parler, la première étape du traitement est de segmenter et de partitionner les données, puis d'identifier les parties contenant des données vocales à transcrire [GLA98]. Ensuite, un regroupement des segments en clusters est effectué, où idéalement un cluster représente une personne. Ce regroupement est similaire à celui qui a été décrit dans un précédent paragraphe.

Enfin, le décodage de la parole est effectué en deux passes. Chaque passe de décodage produit un treillis de mots. L'hypothèse finale est obtenue lors d'une dernière passe, avec un modèle de langage 4-grammes et des probabilités de prononciation.

Ce système a obtenu un taux d'erreurs de mots de 16.87% (pour environ 36000 mots) pendant la première campagne d'évaluation du défi *REPERE* (ensemble de test de la phase 1).

Détection des entités nommées NER (Name Entities Recognition)

Le LIMSI nous a fourni les sorties de deux systèmes basés sur leurs expériences au sein du projet QUAERO [MR11]. Le premier utilise des modèles de CRF spécifiques à l'aide de Wapiti [LCY10] entrainés sur les données de *QUAERO* :

- Un modèle pour détecter la mention d'une personne avec au moins son prénom ou son nom de famille
- Un modèle pour détecter les parties d'un nom (prénom, nom de famille)

Ces modèles utilisent les mêmes caractéristiques que celles présentées dans [MR11] :

- Un ensemble de fonctions standards comme les mots préfixes et suffixes de longueur de 1 à 5, ainsi que certains paramètres tels que : *Est ce que le mot commence par une majuscule ? Est ce que le mot ne contient pas de caractère alphanumérique ?* ...
- Des caractéristiques morpho-syntaxiques (outil *tagger* [ABM08]).
- Des caractéristiques extraites à partir de la sortie d'un analyseur multi-niveaux, utilisées dans un système de questions-réponses [BRG+09], qui contiennent des informations morpho-syntaxiques détaillées ainsi que des informations sémantiques au même niveau que les entités nommées.

Dans le tableau 2.10, on retrouve le score de SER de ce système à partir de transcriptions automatiques sur l'ensemble de test de la phase 1 du défi *REPERE*. La référence ne contenant que les noms complets (prénom + nom). Les noms prononcés comme incomplets, par exemple juste le prénom, ont été complétés par l'annotateur si cela était possible. Un pourcentage non négligeable d'erreurs vient des noms incomplets proposés par le système.

Système	# hyp	# C	# I	# D	# T	# F	# T et F	SER
Sortie du NER	407	204	61	128	19	79	44	60.0%
Noms complets	278	204	16	210	19	27	12	55.5%
+ Noms complétés	320	224	21	173	24	33	18	51.2%

TAB. 2.10 – Performance (SER) du système de détections d'entités nommées, 470 noms dans la référence sur le corpus *REPERE*, ensemble de test de la phase 1.

Dans la seconde ligne, nous avons supprimé tous les noms incomplets de l'hypothèse, ce qui réduit le nombre d'erreurs d'insertion (I), de type (T) et de frontière (F) mais augmente le nombre d'erreurs de suppression (D). Pour réduire ce dernier, nous avons essayé de complèter les noms incomplets par le nom complet le plus proche (même prénom ou nom de famille) dans la même vidéo. Cette complétion nous a permis de réduire le SER à 51.2%.

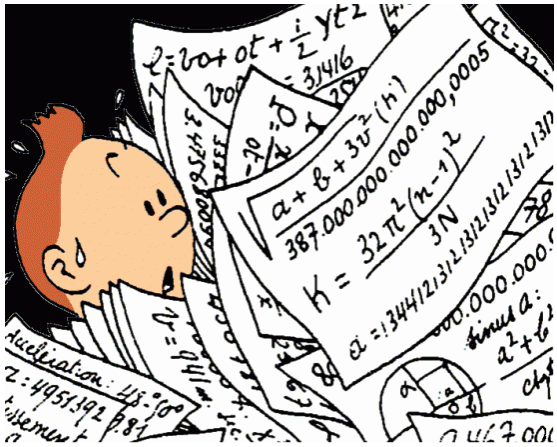
Le LIMSI a également développé un deuxième système fondé sur des règles simples afin de détecter qui désigne le nom prononcé : est ce qu'il correspond à la personne qui parle ? Qui va parler ? Est-ce quelqu'un dont le locuteur est en train de parler ? Pour ce système, la précision a été privilégiée.

Pour les modèles CRF, le score fourni par les CRF a été utilisé. Pour le système à base de règles, la précision obtenue par chaque règle sur les données de développement du corpus *REPERE* de la phase dryrun est utilisée. Ce système annote chaque mot de la transcription automatique avec un marqueur (prénom, nom de famille, ...), ce qui nous permet de créer une liste des noms prononcés dans les émissions.

2.4 Conclusion

La présentation de notre matériau expérimental nous permet de tirer plusieurs conclusions :

- Le corpus *JT France 2* va nous permettre d'évaluer la capacité de notre système (présentée dans le prochain chapitre) à extraire les textes écrits à l'écran dans des vidéos de qualité moyenne.
- Le corpus *REPERE* est suffisamment varié dans le type d'émissions et dans les personnes à reconnaître (liste ouverte) pour nous permettre d'être sûrs que les méthodes de nommage non-supervisées que nous proposons se généralisent bien. La bonne qualité d'enregistrement de ce corpus va nous permettre d'extraire les textes avec peu d'erreurs. L'annotation des rôles des personnes va nous permettre d'évaluer une méthode mixte (avec modèles biométriques pour R123, sans pour R45).
- Nos partenaires du consortium QCompere nous ont fourni deux briques de base de qualité : la matrice de similarité des visages et la diarization des locuteurs. Elles sont incontournables pour identifier les clusters de personnes à l'aide des noms écrits.
- L'identification des visages et des locuteurs pour les personnes de R123 à l'aide de modèles biométriques est de qualité correcte.
- Une comparaison entre l'usage des noms cités et des noms écrits peut être faite grâce à l'extraction des noms prononcés fournie par le LIMSI.



Chapitre 3

Extraction des noms écrits dans les vidéos

Dans le chapitre précédent, nous avons présenté presque tous les éléments de base nécessaires au nommage non supervisé des personnes. Toutefois, il manque le plus important : l'extraction des noms écrits à l'écran (ces noms introduisent les personnes correspondantes dans la vidéo). En effet, il sera plus facile de nommer les personnes présentes dans les émissions de télévision si les noms écrits à l'écran sont extraits avec une très bonne qualité.

C'est pourquoi, nous avons développé LOOV (Lig Overlaid OCR in Video). C'est un outil de vidéo OCR qui effectue l'extraction des textes écrits à l'écran dans les vidéos. Il a été développé dans le cadre du projet *QUAERO* et du défi *REPERE* (un binaire pour linux est téléchargeable à l'adresse `http://mrim.imag.fr/johann.poignant/`).

Avant de décrire cet outil, nous allons d'abord présenter les travaux de l'état de l'art qui nous ont permis de développer LOOV. Ensuite, une description globale de notre système permettra d'en découvrir les différentes composantes, avant de les détailler. La troisième section sera dédiée aux résultats obtenus par ce système sur le corpus *JT france 2*. Enfin, dans la dernière partie, nous proposerons une méthode simple et non-supervisée permettant d'extraire, de ces transcriptions, les noms de personnes écrits dans un cartouche.

Le travail décrit dans ce chapitre a fait l'objet d'un publication à la conférence IEEE ICME 2012 [PBQT12] (présentation orale, taux d'acceptation 13%)

3.1 Travaux de l'état de l'art

Avant de décrire les méthodes d'extraction des textes écrits dans les vidéos, il faut souligner que, dans les vidéos d'émissions de télévision, deux types de textes peuvent apparaître :

- Les **textes de scène** qui sont les textes filmés par la caméra. Par exemple, un texte écrit sur un panneau de signalisation routière ou encore sur des affiches, comme sur l'image 3.1. L'extraction de ce type de texte n'est pas traitée dans cette thèse, parce qu'il n'apporte pas d'information pertinente

sur la présence des personnes dans les émissions télévisées.

- Les **textes superposés** tels que définis par *Lienhart et al.* [LE98]. Ces auteurs ont relevé les caractéristiques suivantes (certaines sont plus spécifiques aux caractères de l'alphabet latin) :
  - Les caractères sont au premier plan, ils ne sont donc jamais occultés, même partiellement.
  - Les caractères sont monochromes.
  - Les caractères sont « rigides ». Ils ne changent pas de forme, de taille ou d'orientation d'une trame à l'autre de la vidéo.
  - Les caractères ont des restrictions de taille. Une lettre n'est pas aussi grande que la totalité de l'écran. Les caractères ne sont pas plus petits qu'un certain nombre de pixels afin qu'ils soient lisibles pour les téléspectateurs.
  - Les caractères sont, pour la plupart, disposés à la verticale.
  - Les caractères sont stationnaires ou en mouvement linéaire. Les caractères mobiles ont aussi un sens de mouvement dominant : horizontalement de droite à gauche ou verticalement de bas en haut.
  - Les caractères contrastent avec le fond, le texte artificiel est conçu pour être lu facilement.
  - Les mêmes caractères apparaissent dans plusieurs trames consécutives, pour avoir une durée d'apparition suffisante à leur lecture.
  - Les caractères apparaissent en groupes, avec une faible distance entre eux et alignés sur une ligne horizontale. C'est la méthode naturelle pour écrire les mots et groupes de mots.

FIG. 3.1 – Texte de scène et texte superposé dans un journal télévisé, corpus *JT France 2*

Dans la figure 3.1, on retrouve ces deux types de texte. Dans cet exemple, les caractéristiques décrites par *Lienhart et al.* sont respectées pour les textes superposés. Seuls ces textes apportent les informations que nous désirons.

Or, pour utiliser ces textes, il faut les extraire. La plupart des méthodes de l'état de l'art découpent le travail en deux étapes : d'abord la détection des zones de texte et ensuite la transcription du texte contenu dans ces zones. Les deux sous-sections suivantes seront donc dédiées à ces étapes.

3.1.1 La détection des boîtes de texte

En général, on effectue la détection de zones spatiales souvent rectangulaires que l'on peut appeler boîtes de texte. Ces boîtes contiennent plusieurs mots contigus sur la même ligne et thématiquement homogènes (mots de la même phrase, par exemple). Nous ne sommes intéressés que par les textes non défilants, écrits horizontalement en alphabet latin (pour la détection des textes écrits avec différents angles d'orientation, le lecteur pourra se référer aux travaux de *Lua et al.* [HCWZ01]). Ces textes peuvent être détectés par une méthode utilisant une des particularités de l'alphabet latin : une texture de barres verticales reliées par des barres horizontales [LE98, WJC02].

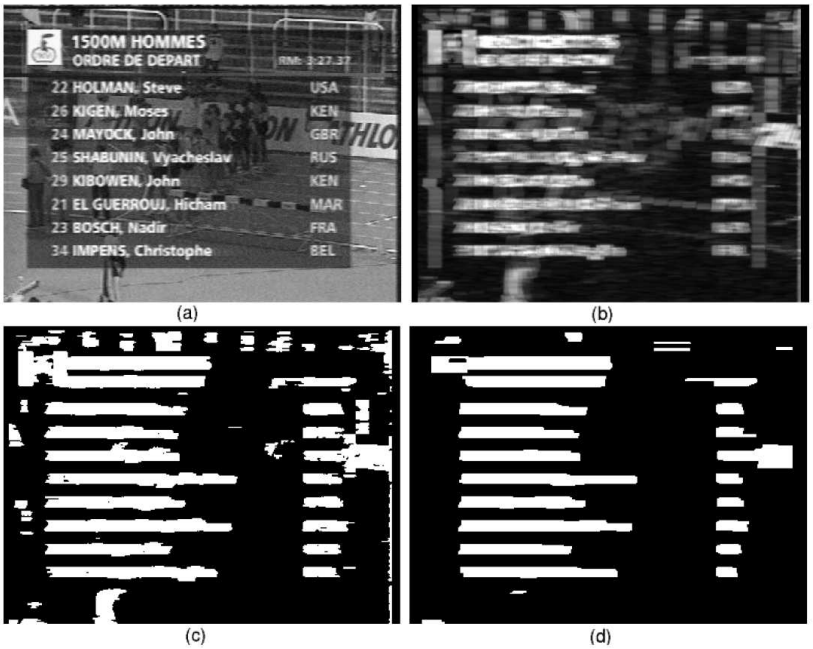
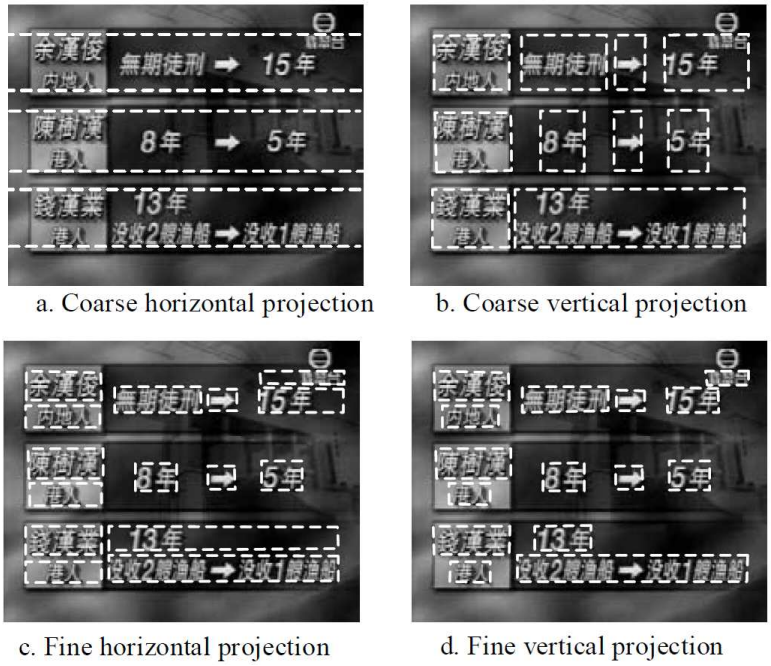


FIG. 3.2 – Détection des textes basée sur la texture, la couleur, le contraste et la géométrie. Image extraite de [WJC02]

Dans la figure 3.2, on peut voir les différentes étapes de l'extraction des zones de texte proposées par *Wolf et al.* [WJC02]. En (a), on retrouve l'image originale. L'image (b) montre la détection de la texture révélée par une mesure du gradient cumulé (filtre Sobel[1] horizontal). L'image est binarisée (c). Des opérateurs morphologiques (dilatation, érosion) font apparaître les boîtes de texte (d). Ensuite, ces boîtes sont sélectionnées en fonction de contraintes géométriques (rapport hauteur/largeur, nombre de pixels par boîtes, etc).

[1]Un filtre Sobel est un opérateur utilisé en traitement d'image pour la détection de contours

Pour améliorer la précision de la détection des coordonnées spatiales de ces boîtes, *Cai et al.* [CSL02] présentent une méthode en deux phases. D'abord une détection grossière puis ensuite un affinage des coordonnées (voir figure 3.3).



a. Coarse horizontal projection b. Coarse vertical projection

c. Fine horizontal projection d. Fine vertical projection

FIG. 3.3 – Détection grossière puis fine proposée par *Cai et al.*. Image extraite de [CSL02]

La détection grossière utilise une projection horizontale puis verticale. Ensuite, l'affinage des coordonnées est lui aussi effectué à l'aide de ces deux projections. Seule la projection grossière horizontale est effectuée sur l'image entière, les autres sont appliquées localement. Les auteurs vont aussi utiliser la dernière projection pour filtrer les boîtes de texte qui ont une trop faible densité de pixels.

Par la suite, de nombreux travaux de la littérature [YH05, JLK09] ont utilisé cette stratégie : trouver le maximum de boîtes de textes candidates, puis affiner les coordonnées et enfin supprimer les boîtes non pertinentes. De même, *Anthimopoulos et al.* [AGP10] appliquent à la fois un filtre Sobel vertical et horizontal afin de détecter les bords des caractères. S'en suit une dilatation de quelques itérations permettant de connecter les caractères ensemble. Une première opération d'ouverture isole les composants connexes. Puis, les lignes de texte sont détectées avec un procédé basé sur des projections horizontales et verticales. Cela génère beaucoup de faux positifs mais une seconde détection locale avec de l'apprentissage automatique est utilisée pour filtrer ces fausses alarmes.

Toutes ces méthodes sont basées sur l'image uniquement. L'avantage des vidéos est que les boîtes de textes sont affichées sur des trames consécutives, ce qui permet de filtrer les fausses alarmes (boîtes inexistantes) [WJC02] et de récupérer les détection manquées (boîtes existantes) sur certaines images.

3.1.2 Reconnaissance du texte

Pour la reconnaissance du texte, deux stratégies sont possibles. On peut développer un outil d'OCR dédié, choix que nous n'avons pas fait. Toutefois, nous renvoyons le lecteur intéressé à [EIH07].

L'autre solution consiste à utiliser un logiciel d'OCR classique, auquel on adapte les images à transcrire. Avant d'envoyer l'image au logiciel d'OCR, *Wolf et al.* [WJC02] proposent d'augmenter la résolution de l'image par une interpolation bi-linéaire.

Il est aussi intéressant d'utiliser la redondance de l'apparition du texte sur les trames consécutives pour améliorer la qualité de l'image du texte (image moyenne sur plusieurs trames) [XHC+01, WJC02]. Ainsi sur la figure 3.4, on note une très légère amélioration de la lisibilité du texte.

(a) Image sur une seule trame

(b) Image moyenne sur plusieurs trames consécutives

FIG. 3.4 – Image extraite de [XHC+01]

L'une des principales difficultés est de fournir au logiciel d'OCR une image très bien binarisée (en noir et blanc uniquement, pas de niveau de gris). Plusieurs solutions existent dans la littérature. Nous pouvons mentionner les travaux de *Sauvola et Pietikäinen* [SP00] et ceux de *Wolf et al.* [WJC02] parmi d'autres études. Ces derniers proposent une comparaison de leurs deux méthodes. On peut en voir un exemple dans la figure 3.5.

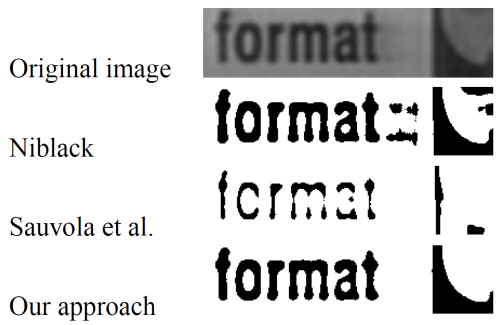


FIG. 3.5 – Binarisation du texte. Image extraite de [WJC02]

Le texte à reconnaitre est écrit en surimpression sur les images et donc, selon le fond (ajout d'un cadre uniforme ou superposition directe sur l'image de la vidéo), l'image peut être plus ou moins difficile à binariser.

Une idée proposée par *Prasad et al.* [PSM+08], confirmée l'année suivante par *Liu et al.* [LFF+09], utilise la combinaison de plusieurs transcriptions pour un même texte. Ces transcriptions sont issues de textes reconnus à partir d'images temporellement espacées. Ces hypothèses sont ensuite combinées, soit à l'aide de l'algorithme NIST ROVER dans [PSM+08] ou à partir d'un réseau de confusion où le meilleur chemin est trouvé à l'aide d'un modèle de langage [LFF+09].

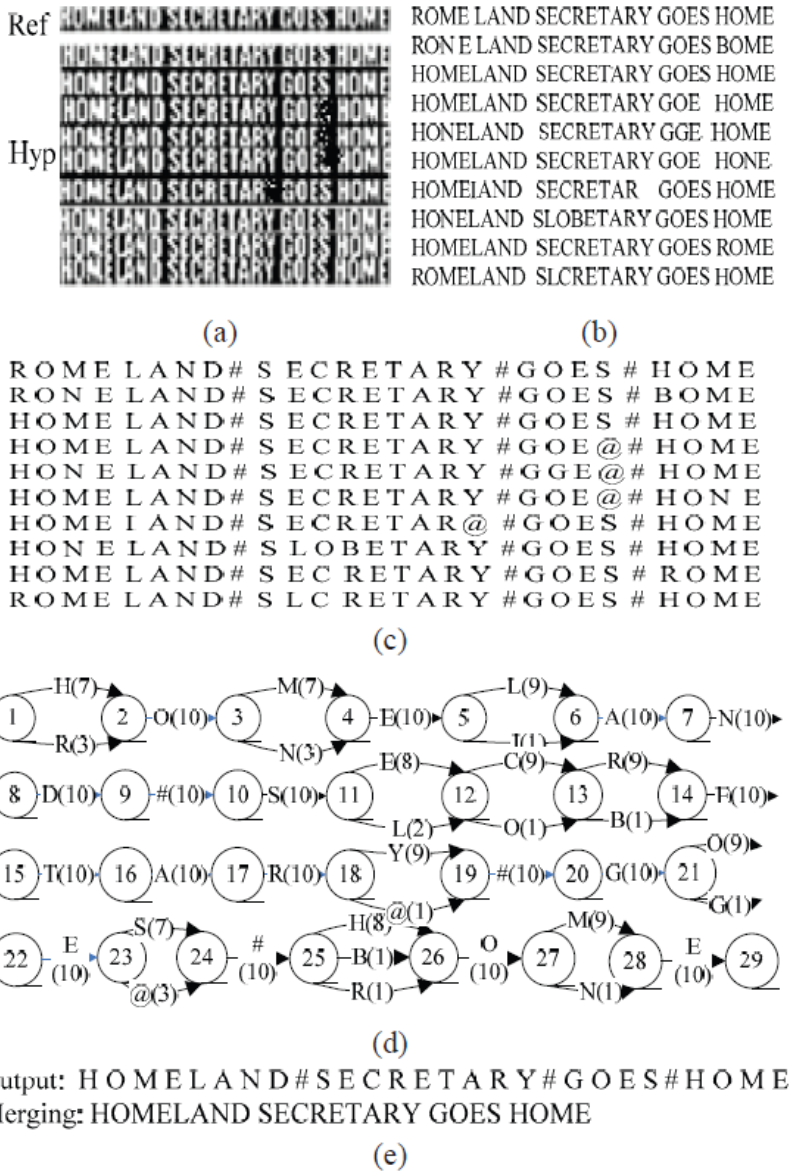


FIG. 3.6 – Combinaison de plusieurs transcriptions. Image extraite de [LFF+09]

Dans la figure 3.6, on retrouve la méthode proposée par *Liu et al.*. En (a) les images de textes à transcrire. En (b), les transcriptions obtenues par l'OCR avec quelques variations. En (c), un post-traitement permettant d'aligner les transcriptions. En (d), le réseau de confusion. Et en (e), la sortie après décodage du réseau.

3.2 LOOV : Lig Overlaid OCR in Video

3.2.1 Vue globale du système

Pour développer ce système, nous nous sommes intéressés aux vidéos de qualité moyenne (Corpus *JT France 2* , 288*352, mpeg1). Nous avons considéré seulement les textes écrits en alphabet latin horizontalement et non défilants. Notre détection se fait aussi en deux étapes (voir figure 3.7) où la détection grossière est très proche de celle de [WJC02]. C'est-à-dire qu'elle est obtenue grâce à un filtre Sobel, suivi d'une étape de dilatation/érosion. La deuxième étape de la détection effectue un affinage des coordonnées à l'aide d'opérateurs morphologiques appliqués localement. Cette deuxième étape permet de filtrer quelques faux positifs. Ensuite, notre traitement temporel inclus celui de [WJC02] ; mais aussi une étape de récupération qui permet de corriger les moments d'apparition/disparition d'une boîte de texte. Un réglage des paramètres semi-supervisé peut-être utilisé pour adapter le système à un corpus particulier.

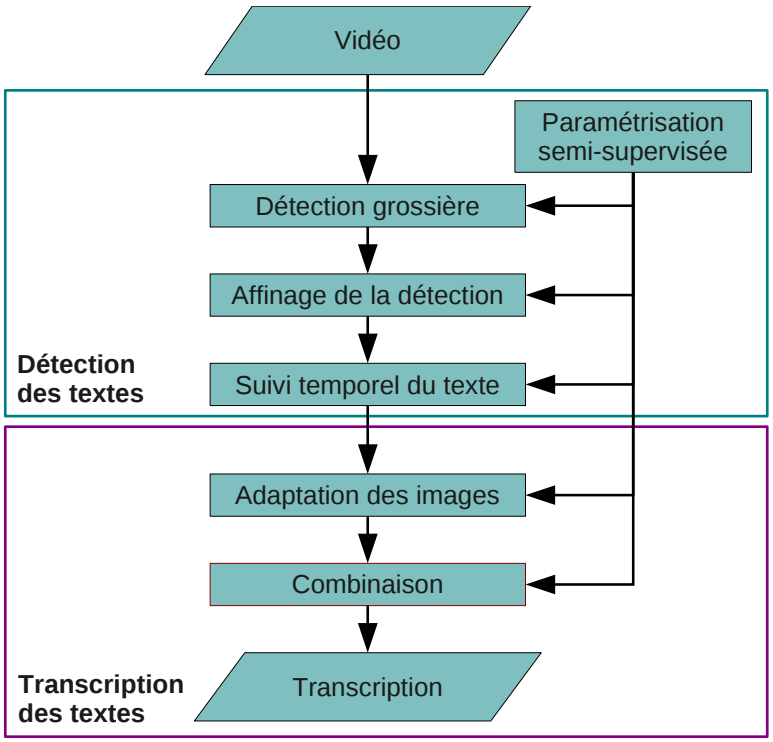


FIG. 3.7 – Vue globale du système LOOV.

Pour la reconnaissance des textes, nous nous appuyons sur le logiciel OCR `Tesseract` de Google[2]. Comme dans [PSM+08, LFF+09], nous utilisons un réseau de confusion à partir des différentes transcriptions d'un même texte. Par contre, nous n'intégrons pas de modèles de langage pour éviter de corriger de manière erronée les noms proposés.

[2] `http://code.google.com/p/tesseract-ocr/`

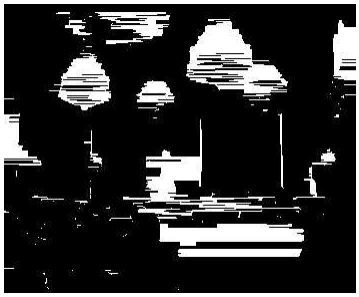
Nous allons maintenant détailler les étapes de notre système de reconnaissance des textes. Tout d'abord, la détection des boîtes de texte et ensuite la transcription du texte contenu dans ces boîtes.

3.2.2 Détection des boîtes de texte

Comme mentionné ci-dessus, la détection spatiale s'effectue en deux parties. La première, la détection grossière, sélectionne sur la globalité de l'image les composantes connexes correspondant à du texte, avec un fort taux de rappel. La seconde, l'affinage des coordonnées, filtre les fausses détections et trouve les coordonnées précises des boîtes de texte (voir figure 3.8.a, l'image originale de notre exemple)

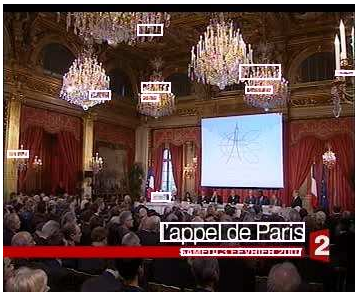
(a) Image originale de notre exemple.

(b) Filtre Sobel.



(c) Étape de dilatation et d'érosion.

(d) Étape de filtrage.



(e) Détection grossière.

(f) Affinage de la détection.

(g) Résultat final de la détection spatio-temporelle.

FIG. 3.8 – Les différentes étapes de la détection spatio-temporelle.

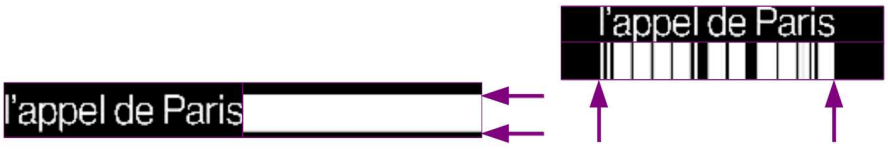
3.2.2.1 Détection grossière

La détection du texte est effectuée sur toutes les images de la vidéo. Elle commence par un filtre Sobel horizontal, qui met en évidence l'une des principales caractéristiques de l'alphabet latin : une texture de barres verticales reliées par des barres horizontales. Nous pouvons voir le résultat dans la figure 3.8.b. Après binarisation, une opération de dilatation et d'érosion horizontale connecte les caractères d'une même chaîne (figure 3.8.c). Ensuite, un filtrage (érosion verticale et horizontale suivies d'une dilatation) nettoie l'image (figure 3.8.d). Dans l'image qui en résulte, nous sélectionnons les zones qui respectent des contraintes géométriques. On obtient ainsi une détection grossière des coordonnées des boîtes de textes (figure 3.8.e).

Il est à noter, à cette étape du processus, que d'autres algorithmes comme celui de [AGP10] auraient proposé de nombreuses boîtes de texte candidates. Ce n'est pas un problème en termes de qualité finale (ces boîtes seraient probablement filtrées durant l'étape d'affinement à base d'apprentissage), mais cela peut l'être d'un point de vue computationnel.

3.2.2.2 Affinage des coordonnées

Une seconde détection est effectuée sur chacune des zones de texte séparément. Après binarisation de l'image originale en niveau de gris, en utilisant l'algorithme de Sauvola [SP00], nous détectons si le texte est écrit en noir sur blanc ou vice versa, dans l'image binarisée à partir du nombre et de la variance de l'aire des composantes connexes blanches.



(a) coordonnées en Y (b) coordonnées en X

FIG. 3.9 – Affinage de la détection.

Cette détection est effectuée en deux étapes afin de trouver les coordonnées horizontales et verticales. Une opération morphologique de dilatation horizontale montre les coordonnées verticales des textes (flèche sur la figure 3.9 a). On procède de la même manière pour les coordonnées horizontales (figure 3.9 b) avec une dilatation dans l'autre sens.

Cet affinage des coordonnées nous permet de filtrer les boîtes qui ne respectent pas les contraintes géométriques du texte. Dans notre exemple (figure 3.8.f), la détection grossière a trouvé onze boîtes, cinq d'entre elles ont été supprimées par la détection locale. Bien que le fond soit assez complexe, seuls quatre faux positifs restent, et les vrais positifs n'ont pas été filtrés.

3.2.2.3 Suivi temporel du texte

En plus de la détection spatiale en deux étapes nous profitons du fait qu'un texte s'affiche sur plusieurs trames successives. L'information temporelle est donc utilisée pour filtrer quelques faux positifs, mais aussi pour récupérer les boîtes pour lesquelles la détection a échoué sur quelques trames. Nous pouvons voir dans notre exemple (figure 3.8.g) que quatre boîtes ne sont pas détectées systématiquement sur des trames consécutives.

3.2.3 Transcription des boîtes de texte détectées

3.2.3.1 Adaptation des images pour l'outil d'OCR

Après cette étape de détection, il faut adapter les images issues des boîtes de texte détectées pour le logiciel d'OCR standard `Tesseract` de Google. Après une augmentation de la résolution des images à 200 pixels de haut à l'aide d'une interpolation, les images sont binarisées avec l'algorithme de Sauvola [SP00].

Les images de la figure 3.10 montrent que l'interpolation avant la binarisation permet de lisser les caractères. On atténue donc l'importance d'un pixel, choisi comme noir ou blanc par la binarisation, sur la forme générale d'un caractère.

(a) Image originale (32 pixels de haut) (b) Image binarisée sans interpolation

(c) Image après interpolation (200 pixels de haut) (d) Image binarisée après interpolation

FIG. 3.10 – Augmentation de la résolution des images avant/après binarisation

3.2.3.2 Combinaison de plusieurs transcriptions pour une même boîte

Pour améliorer la qualité de la transcription du texte, nous appliquons l'OCR, pour une même boîte de texte, sur plusieurs groupes d'images consécutives d'une vidéo. (figure 3.11).

Notons $\bar{I}_{i,j}$ l'image construite sur la moyenne des images de la série $[i, j]$. Nous avons choisi 10 images à partir d'expériences préliminaires. Deux types d'images moyennes sont calculées : l'image moyenne globale $\bar{I}_{1,M}$ et un ensemble d'images moyennes locales calculées sur un sous-ensemble de taille n : $\bar{I}_{k,l}$, où $l-k = n-1$. Les transcriptions sont calculées à la fois pour l'image moyenne globale et pour les images locales.

Dans la figure 3.11, nous présentons les transcriptions pour notre exemple : la première transcription correspond à l'image moyenne globale ($\bar{I}_{1,M}$), les autres correspondent aux transcriptions temporellement décalées ($\bar{I}_{1,n}, \bar{I}_{n+1,2n}, ...$).

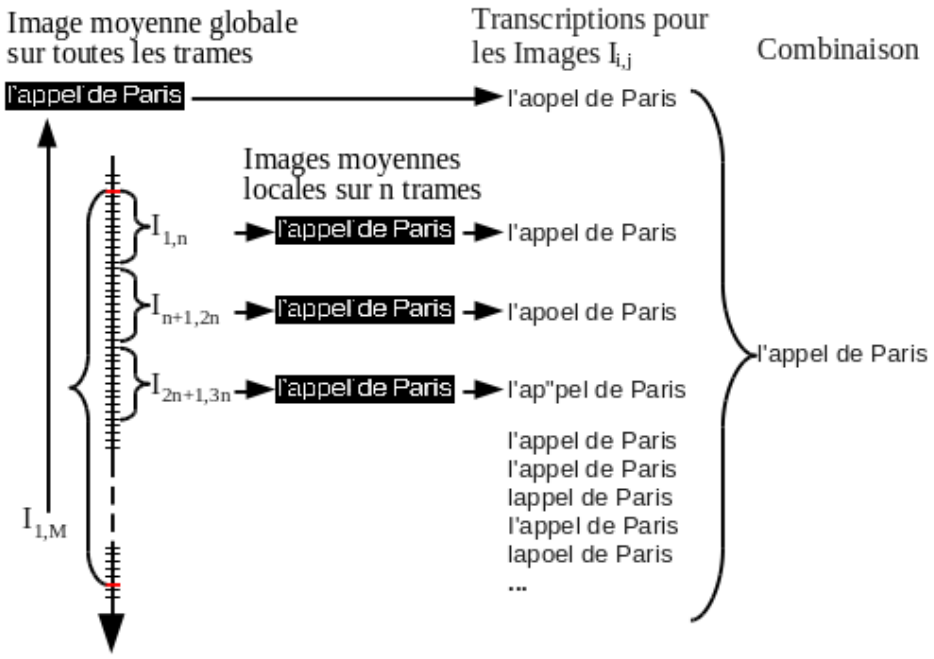


FIG. 3.11 – Multiples transcriptions pour une boîte de texte.

La transcription obtenue à partir de l'image moyenne globale (à savoir *l'aopel de Paris*) produit une erreur de caractère. La transcription de la première plage locale ($\bar{I}_{1,n}$) ne produit pas d'erreur alors que les deux suivantes ($\bar{I}_{n,2n}$, $\bar{I}_{2n+1,3n}$) en produisent deux différentes (*l'apoel de Paris*, *l'ap"pel de Paris*).

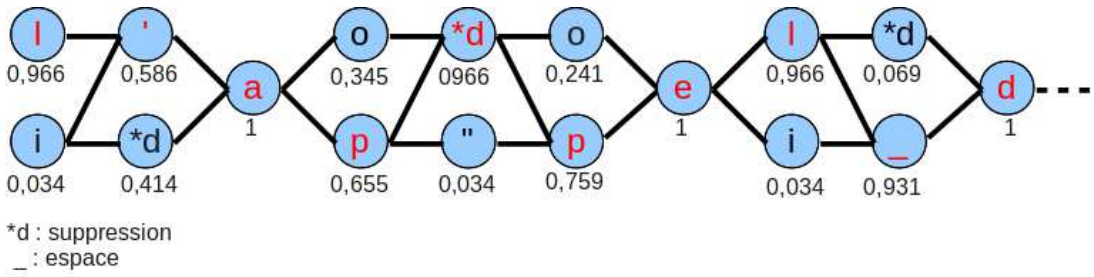


FIG. 3.12 – Début du réseau de confusion pour notre exemple

Afin de combiner toutes ces transcriptions, nous construisons un réseau de confusion (voir figure 3.12), avec comme hypothèse de base, la transcription de l'image moyenne globale. Il est obtenu en utilisant un alignement dynamique (implémenté dans la boite à outils SRILM[3]). L'algorithme de Viterbi est ensuite utilisé pour sélectionner le chemin le plus probable. Grâce à cette combinaison, l'erreur de transcription faite sur l'image moyenne globale est corrigée.

[3] www.speech.sri.com/projects/srilm/

3.2.4 Réglage semi-supervisé des paramètres

Dans cette section, nous allons voir quelle quantité de boîtes de textes, annotées manuellement, est nécessaire pour le bon paramétrage de LOOV. En effet, nous avons développé cet outil pour qu'il soit entièrement paramétrable et qu'il puisse donc s'adapter à un changement de corpus. Les paramètres par défaut peuvent être utilisés ; mais en les adaptant à un corpus particulier, on peut améliorer la qualité de la transcription.

Pour régler notre outil, nous avons besoin des positions spatio-temporelles des boîtes de textes annotées manuellement. Elles permettront de sélectionner les meilleurs paramètres à utiliser. Nous avons découpé ce paramétrage en plusieurs étapes permettant de fixer un groupe de paramètres lors de chacune d'elles (le découpage de ce paramétrage suit le schéma d'ordonnancement général de notre outil, voir figure 3.7) :

- Détection grossière :
  - Seuil pour la binarisation de l'image après application du filtre Sobel (figure 3.8.b).
  - Nombre d'itérations des opérateurs morphologiques pour la connexion des caractères et le filtrage du bruit (figure 3.8.c et d).
  - Seuils (hauteur, largeur, ratio) sur les contraintes géométriques des boîtes de texte (figure 3.8.e).
- Affinage de la détection (figure 3.9) :
  - Critères de sélection (nombre et variance de l'aire des composantes connexes) pour connaitre la couleur du texte (noir sur blanc ou inversement).
  - Marge d'élargissement des boîtes de texte après la détection grossière.
- Suivi temporel (figure 3.8.g) :
  - Tolérance sur le recouvrement pour considérer que deux boîtes de texte sur des trames consécutives correspondent au même texte.
  - Durée minimum d'apparition.
  - Durée maximum entre deux détections à la même position pour considérer qu'elles correspondent à la même boîte.
  - Critère de vérification du changement du texte pour des boîtes consécutives se situant à la même position spatiale. Ce critère est basé sur la différence de couleur entre deux images de deux trames successives d'une boîte de texte. Si cette différence est supérieure au critère, une nouvelle boîte est créée.
- Adaptation des images :
  - Hauteur de l'image après interpolation (figure 3.10).
- Combinaison :
  - Durée de la fenêtre pour les images moyennes locales (figure 3.11).

A chacune des étapes nous choisissons les réglages qui donnent la détection automatique des boîtes de texte la plus proche de l'annotation manuelle. Ils seront ensuite utilisés dans les étapes suivantes.

Pour évaluer la quantité d'annotations nécessaires pour le réglage des paramètres, nous avons annoté 2 heures de notre corpus *JT France 2*, ce qui correspond à trois vidéos. 512 boîtes de texte ont été temporellement et spatialement annotées. Cette annotation manuelle a pris 4 heures.

Nous avons voulu savoir comment variait la qualité de la détection en fonction du nombre de boîtes annotées utilisées pour paramétrer. Nous avons par conséquent réglé notre système sur une vidéo (correspondant à 212 boîtes). Nous avons choisi au hasard 10 ensembles avec respectivement 1, 25, 50, 75, 100, 150 et 200 boîtes annotées. La performance du système est alors évaluée sur les deux autres vidéos.

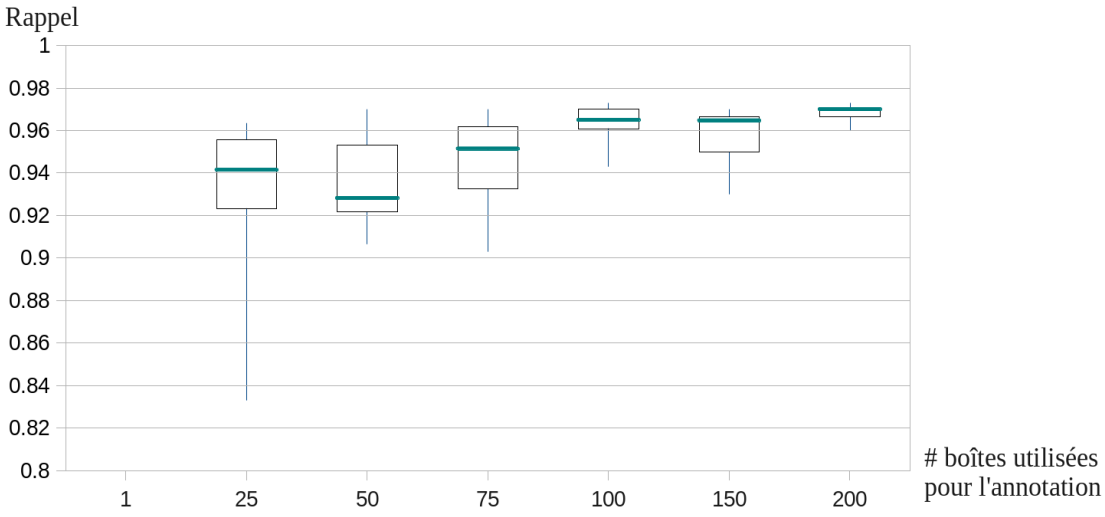


Fig. 3.13 – Évolution du rappel du nombre de boîtes détectées en fonction du nombre de boîtes annotées utilisées pour le paramétrage. Ces boxplots sont calculés sur dix ensembles de boîtes de textes tirés aléatoirement, corpus *JT France 2*.

La figure 3.13 montre un graphique de boxplots du rappel en nombre de boîtes de texte correctement détectées en fonction du nombre de boîtes annotées utilisées pour la paramétrisation. Une boîte est considérée comme étant correctement reconnue si la F-mesure calculée sur les pixels entre la zone de référence et la zone hypothèse est plus grande que 0.5.

L'annotation d'une seule boîte donne un taux de détection très faible (percentiles : min, 1/4, 1/2, 3/4, max : 0.144, 0.533, 0.572, 0.622, 0.763). Par contre, on peut voir que lorsque l'on utilise 25 boîtes, le taux de rappel atteint une médiane de 94%. Il est stabilisé entre 100 et 200 boîtes avec une médiane à plus de 96%.

Au vu des résultats, on peut donc considérer que 100 à 150 annotations sont suffisantes pour régler le système pour ce corpus, réduisant ainsi l'exigence en annotation manuelle à environ une heure.

3.3 Évaluation du système de reconnaissance de texte sur le corpus *JT France 2*

Nous évaluons maintenant l'ensemble du système LOOV sur la base de ses performances au niveau de la reconnaissance des caractères. Pour rester cohérent avec la section précédente, nous avons utilisé la paramétrisation pour la détection des textes obtenue avec 200 boîtes annotées.

Nous avons évalué notre système sur le corpus *JT France 2* qui a une qualité moyenne d'images (352*288, mpeg1) avec des textes écrits allant de 5 à 20 pixels de haut. Nous n'avons pas évalué la totalité des textes écrits dans ces vidéos mais ceux apparaissant dans 4414 images. Dans ces images, il y avait 9256 boîtes de texte (30905 mots et à 154904 caractères). Un simple post-traitement ad hoc a été appliqué pour supprimer les erreurs les plus courantes. Par exemple « ii » est systématiquement remplacé par « M » à la sortie du logiciel d'OCR.

Comme métrique, nous avons utilisé le taux d'erreur de mots (WER) et de caractères (CER) défini dans le chapitre 2. Les boîtes de textes de la référence et de l'hypothèse ont été alignées si la similarité, basée sur la distance de Levenshtein, était supérieure à 50%. Toutes les boîtes de la référence qui n'ont pas pu être alignées avec une boîte de l'hypothèse comptent comme des erreurs. C'est-à-dire que le taux d'erreur prend en compte à la fois les erreurs de transcription et les erreurs de détection. Les résultats sont présentés dans le tableau 3.1.

	Nombre de		Sauvola [SP00]	
Type	Mots	caractères	tx err de mots (%)	tx d'err de caractères (%)
Tous types de textes	30905	154904	19.2 (16.5)	8.6 (6.2)
Noms⋆	3230	19248	9.6 (6.4)	3.4 (1.7)
Fonctions	5794	32472	9.3 (9.1)	2.8 (2.6)
			Sauvola [SP00]+combinaison	
Tous types de textes	30905	154904	**11.6** (9.3)	**4.6** (2.7)
Noms⋆	3230	19248	**7.1** (4.0)	**2.6** (0.9)
Fonctions	5794	32472	**6.3** (6.0)	**1.7** (1.5)

⋆ Noms qui apparaissent seuls et pas dans la liste des personnes créditées en fin de reportage.

TAB. 3.1 – Taux d'erreur de caractères et de mots avec et sans combinaison. Le score entre parenthèses donne la performance pour l'OCR seulement sur les boîtes de texte détectées, corpus *JT France 2*.

Pour montrer l'apport de la combinaison, nous avons utilisé comme système contrastif le texte reconnu sur les images moyennes globales (binarisées avec l'algorithme de Sauvola [SP00]). Par souci d'exhaustivité, nous fournissons le taux d'erreur de mots, même s'il n'y a pas de post-traitement basé sur un modèle de langue. L'utilisation d'un modèle de langue pourrait diminuer ce taux d'erreur, mais ce serait au détriment de la correction des noms propres qui ne suivent pas forcément un modèle de langue. Comme la tâche qui nous intéresse est l'extrac-

tion de ces noms, les taux d'erreur mentionnés sont donc un peu surestimés. Par conséquent, nous concentrons notre analyse sur le taux d'erreur de caractères. Nous fournissons également, entre parenthèses, le taux d'erreur calculé à partir des zones de texte détectées seulement. Ce taux donne la vraie performance de la transcription des textes indépendamment de la détection de ceux ci.

L'utilisation de la combinaison permet une amélioration importante des performances. Le taux d'erreur sur les caractères (CER) chute de 8,6% à 4,6% sur tous les types de texte. Comme on peut le voir, les noms des personnes ont tendance à être plus lisibles par notre système (tous types de textes *vs* noms).

Pour les noms, on voit une réduction du taux d'erreur de 2,6% à 0,9% si on ne prend en compte que les boîtes détectées. Ce qui veut dire que la plupart des erreurs sur les noms sont dues soit à des boîtes qui ne sont pas correctement détectées, soit aux boîtes de la référence qui n'ont pas pu être alignées avec celles de l'hypothèse (trop d'erreurs dans l'hypothèse pour être sûr qu'elle corresponde à une des boîtes de texte de la référence).

Il est intéressant de noter que, dans notre corpus, certains noms sont écrits avec seulement 5 pixels de haut (par exemple les noms des joueurs d'une équipe de sport), tandis que la fonction d'une personne est écrite en 7 pixels de haut avec un fond uniforme. Dans ce dernier cas, la combinaison diminue le taux d'erreur en caractères de 2,8 % à 1,7 %. Par conséquent, même sur des textes de taille moyenne écrits sur un fond uniforme, la combinaison améliore la transcription.

Le temps de réponse (sur un Intel Xeon Core 2 Duo cadencé à 3 GHz, 4 Go de RAM) du système est de 728 secondes pour une vidéo de 2184 secondes (MPEG1, 352x288, 25 images/sec) : 441 secondes pour l'étape de détection et 287 secondes pour l'étape de transcription. L'efficacité dépend toutefois du nombre/durée des zones de texte trouvées.

3.4 Détection des noms de personnes dans les transcriptions du texte (corpus *REPERE*)

3.4.1 Détection des noms de personnes

A partir des transcriptions obtenues avec LOOV, nous utilisons une simple technique de détection des positions spatiales des cartouches. Cette technique compare chaque transcription avec une liste de 175000 noms de personnes célèbres, groupes de musique, personnages de fiction, etc. Nous avons constitué cette liste à partir d'une sélection des pages du site Wikipedia. La sélection a été effectuée en fonction des tags liés aux pages.

A chaque fois qu'une transcription correspond à un nom, nous ajoutons sa position spatiale à une liste. Les positions récurrentes dans cette liste nous permettent de déduire les positions spatiales des cartouches utilisés par l'émission pour introduire une personne.

Les boîtes de texte détectées à ces positions spatiales récurrentes ne contiennent pas toujours un nom. Un simple filtrage basé sur quelques règles linguistiques

(est-ce que le premier mot est un prénom, est-ce que c'est un nom célèbre, de combien de mots la transcription est-elle composée, etc) nous permet de supprimer les transcriptions ne contenant pas qu'un nom. Sur les 58h de l'ensemble d'apprentissage de la phase 1 du corpus *REPERE*, on obtient 4779 boîtes de texte candidates, 1315 après filtrage, 11 qui n'auraient pas dû être filtrées, 13 qui auraient dû être filtrées.

Nous avons évalué la qualité de la détection des noms écrits à l'écran pour introduire la personne correspondante (tableau 3.2). Ces résultats ont été calculés sur la plus grosse partie du corpus *REPERE* disponible, la partie apprentissage de la phase 1, mais aussi sur l'ensemble de test de cette phase.

Ensemble	#Noms dans la référence	#Noms dans l'hypothèse	#Noms en commun	Précision	Rappel	F1-mesure
Apprentissage	1378	1373	1352	98.5%	98.1%	98.3%
Test	186	179	178	99.4%	95.7%	97.5%

TAB. 3.2 – Qualité de détection des noms écrits à l'écran sur le corpus *REPERE*, phase 1, partie apprentissage + test. Évaluation sur les images annotées à l'aide du protocole du défi *REPERE*

Une correction a été appliquée pour corriger les erreurs de transcription. Elle est basée sur la liste de 175000 noms de personnes célèbres. Lorsque le ratio de la distance d'édition (entre 0 et 1) entre une transcription et un nom est supérieur à 0.9, nous corrigeons le nom. Nous avons corrigé 207 noms avec seulement 4 corrections erronées sur l'ensemble d'apprentissage.

Pour l'ensemble de test, nous avons aussi utilisé la liste des personnes présentes dans la partie apprentissage de la phase 1. L'utilisation de ces noms issus du même corpus nous permet d'avoir quelques noms de présentateurs et journalistes correctement orthographiés. On peut observer la très bonne qualité d'extraction des noms avec une très bonne précision de 98.5% et 99.5% avec un rappel de 98.3% et 95.7%. Les erreurs restantes sont principalement dues au filtrage et à une transcription erronée.

Les transcriptions (avant et après combinaison) et la liste des noms de personnes détectées dans ces transcriptions sont disponible à l'adresse `http://mrim.imag.fr/johann.poignant/` section téléchargement.

3.4.2 Identification des personnes basée sur les noms écrits

On utilise ensuite ces noms écrits pour identifier directement qui apparaît ou qui parle : c'est-à-dire qu'à chaque fois qu'un nom écrit a été détecté par cette méthode, nous avons considéré que la personne correspondante parlait et était visible à l'écran. On peut ainsi évaluer la capacité des noms écrits à identifier les personnes présentes dans les émissions de télévision.

La figure 3.14 montre la précision de l'identification avec les différentes erreurs possibles, dans la partie apprentissage du corpus *REPERE*. Cette précision

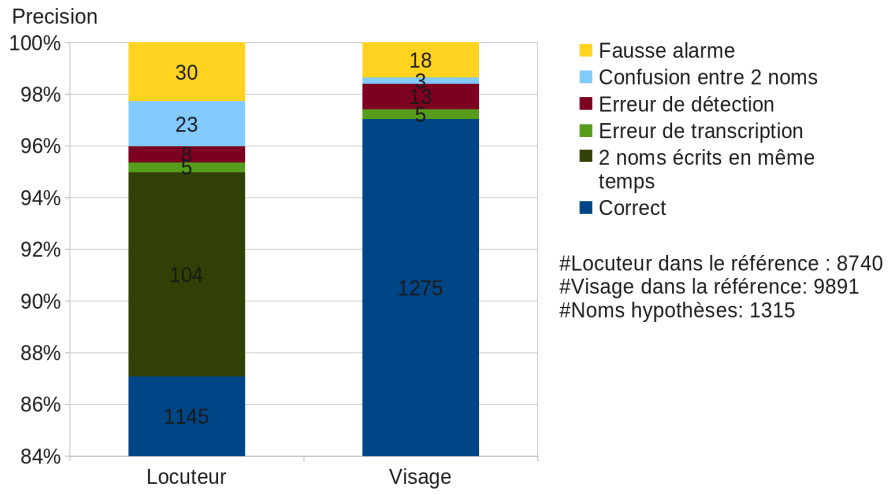


FIG. 3.14 – Noms écrits pour l'identification des locuteurs et des visages, partie apprentissage, images annotées du Corpus *REPERE*.

est calculée sur les images annotées. Ce graphique nous permet de voir la très bonne précision obtenue si on utilise les noms écrits pour identifier les personnes présentes dans les vidéos d'émissions de télévision.

Dans notre corpus, il arrive régulièrement que deux noms soient écrits en même temps. Comme un seul des deux peut identifier le locuteur sur l'image utilisée pour l'évaluation, l'autre génère automatiquement une erreur (104 des 1315 noms hypothèses). Ceci nous donne une précision de 95% de bonne identification lorsqu'il n'y a qu'un seul nom écrit. Ces erreurs n'existent pas pour l'autre tâche puisque les deux noms sont bien utilisés pour identifier un visage visible.

30 erreurs proviennent de fausses alarmes. Une partie d'entre elles sont des erreurs du système d'extraction des noms. Pour l'autre partie, sur les images évaluées, le locuteur n'était pas en train de parler (il y a de fortes chances qu'il ai parlé avant ou après les images annotées).

23 erreurs sont dues à des confusions. Ces erreurs proviennent principalement des émissions de débat où les interlocuteurs se coupent souvent la parole.

L'identification du visage par les noms écrits induit moins d'erreurs de confusions. Cela est dû à la plus grande proximité entre ces deux modalités. En effet, l'ajout des noms écrits en post-production est plus souvent corrélé aux personnes apparaissant que parlant.

A l'aide de LOOV et de cette simple technique d'extraction des noms, nous avons une très bonne brique de base pour nommer les clusters de personnes dans les émissions de télévision. Cette modalité ayant très peu été utilisée avec succès dans l'état de l'art par rapport aux noms cités, une comparaison de ces deux modalités d'extraction des noms de personnes s'impose. Dans le chapitre suivant, nous allons donc comparer la capacité des noms écrits et des noms prononcés à proposer les noms des personnes présentes dans les vidéos.


HOP!

Chapitre 4

Analyse des capacités de nommage des personnes par les noms écrits et par les noms cités

Comme nous l'avons vu dans le chapitre 1, les travaux de l'état de l'art ont utilisé principalement les noms prononcés comme source de noms dans les émissions télévisées. Les noms écrits ont assez peu été utilisés du fait d'un trop grand nombre d'erreurs dans leurs transcriptions.

Toutefois, ces dernières années ont vu une augmentation grandissante de la qualité de diffusion des émissions de télévision (résolution, compression, qualité d'incrustation des textes). Conformément à cette évolution, le corpus *REPERE* est lui aussi enregistré en très bonne qualité (720×576, MPEG2). Et, nous avons montré dans le chapitre précédent qu'il était possible d'extraire les noms écrits à l'écran avec assez peu d'erreurs de transcription. Donc, dans ces conditions, on peut penser qu'il est possible de les utiliser pour nommer les personnes avec une bonne précision.

Pour confirmer cette affirmation, nous allons comparer, dans ce chapitre, la capacité des noms écrits à l'écran et des noms cités à l'oral (issus des systèmes du LIMSI) à proposer les noms des personnes présentes dans les vidéos. Les résultats montrés dans la suite de ce chapitre sont comptabilisés sur l'ensemble d'apprentissage de la phase 1 du défi *REPERE*. Nous avons utilisé cet ensemble parce qu'il est le plus volumineux du corpus et donc les résultats donnés seront plus significatifs. Ce travail a fait l'objet de deux articles aux conférences CORIA 2013 [PBQ13] (article long) et INTERSPEECH 2013 [PBL+13].

La première partie de ce chapitre compare la différence de qualité d'extraction des noms écrits et prononcés. Ensuite, nous présentons la méthode de comptabilisation. Enfin, ces deux modalités seront comparées sous plusieurs angles :

- La tâche : qui parle vs qui apparaît
- Les modalités : noms écrits vs noms cités
- L'annotation : manuelle vs automatique
- La granularité : vidéo vs collection
- Le rôle des personnes : journaliste (R123) vs invité-autre (R45)

4.1 Comparaison de la qualité des systèmes

Avant de confronter les capacités de nommage de ces deux modalités, nous allons comparer la qualité d'extraction des noms que l'on obtient selon la nature de l'extraction (noms écrits ou noms prononcés). Cette première comparaison va utiliser le protocole d'évaluation du défi *REPERE*. Il y a donc des éléments à souligner :

- Comme spécifié dans le chapitre 2, l'annotation manuelle pour les noms écrits n'est pas complète (seulement une image toutes les 10 secondes en moyenne). L'évaluation ci-dessus porte donc seulement sur ces images. Un nom écrit peut apparaître sur deux images annotées successives, ce qui explique que le nombre de noms dans la référence soit supérieur à celui indiqué dans les tableaux du reste du chapitre (où nous ne compterons que les apparitions uniques : 1378 dans le tableau 4.1 à 1049 pour les tableaux des sections suivantes).
- Pour le nombre de noms écrits dans l'hypothèse, ce chiffre est supérieur (1407 sur les segments UEM, 2905 sur les vidéos complètes, voir tableau 4.2) parce que nous avons utilisé tous les noms extraits du signal vidéo, que ce soit sur les segments UEM ou les vidéos complètes.
- Pour les noms cités, l'annotation manuelle complète des segments UEM permet d'éviter ce type de différence. L'évaluation du tableau 4.1 porte donc sur l'ensemble du signal des segments UEM.

L'utilisation de LOOV et de la technique de détection des noms écrits à l'écran nous permet d'obtenir 98.1% (cf. tableau 4.1) des noms écrits (pour introduire une personne à l'écran) avec une précision de 98.5%. Les quelques erreurs restantes sont dues à des erreurs de transcription ou de filtrage (sélection des noms de personnes parmi les autres types de texte).

Modalités	#Noms dans la référence	#Noms dans l'hypothèse	#Noms en commun	Précision	Rappel	F1-mesure
Noms écrits	1378	1373	1352	98.5%	98.1%	98.3%
Noms cités	4264	2905	2133	73,5%	50%	59,5%

TAB. 4.1 – Qualité de détection des noms écrits à l'écran et des noms cités à l'oral, phase 1, partie apprentissage, segments UEM. Protocole d'évaluation du défi *REPERE*

Les systèmes qui traitent la bande son engendrent plus d'erreurs car nous travaillons sur une liste ouverte de personnes. Donc, nous n'avons pas connaissance des noms qui pourraient être prononcés. C'est pourquoi, on rencontre des erreurs :

- dues aux erreurs de la transcription.
- de détection des noms proches de mots de la langue courante (« Dupont », ou « du pont »).
- liées à la difficulté d'extraire le nom complet d'une personne alors qu'une partie seulement a été prononcée (par exemple, seulement le prénom).

Malgré une précision et un rappel inférieurs aux noms écrits, les noms cités apportent plus de noms hypothèses (c.f. tableau 4.2). En effet, nous pouvons remarquer qu'il y a environ 50% de plus de noms cités à l'oral que de noms écrits à l'écran ; que ce soit sur les vidéos complètes (avec le début et la fin de chaque émission) ou seulement sur les segments UEM (segments annotés). Cette proportion est respectée entre le nombre d'occurrences de citation des noms et le nombre de personnes différentes citées.

Modalités	Segments	#Occurrences de noms	#Personnes sans doublon
Noms écrits	Segments UEM	1407	458
	Vidéos complètes	2090	629
Noms cités	Segments UEM	2905	736
	Vidéos complètes	4922	1156

TAB. 4.2 – Nombre de noms hypothèses dans le corpus *REPERE*, phase 1, partie apprentissage

Dans la section suivante, nous allons voir si ce plus grand nombre d'hypothèses peut nous permettre de nommer plus de personnes présentes dans les vidéos.

4.2 Méthode d'analyse

Pour comparer ces deux sources de noms, nous avons utilisé deux métriques. La première, la proportion de personnes nommables, nous permet d'estimer la qualité intrinsèque d'une source à proposer le nom des personnes présentes ; indépendamment de la difficulté, liée à cette modalité, d'associer un nom à un cluster de personnes. La deuxième, le nombre d'occurrences de citation d'un nom, est une indication supplémentaire sur la manière dont est utilisée une modalité pour nommer les personnes présentes.

4.2.1 Proportion de personnes nommables

Le ratio de personnes nommables a été évalué pour chaque vidéo (intra-vidéo) :

$$Np_{intra} = \frac{\#\text{videos où } p \in Phr}{\#\text{videos où } p \in Pr} \tag{4.1}$$

Avec :
- p : une personne
- Pr : ensemble des p présents
- Phr : ensemble des p présents dans une vidéo et ayant leurs noms écrits/prononcés dans cette vidéo

Nous avons aussi évalué ce nombre avec une propagation des noms inter-vidéos :

$$Np_{inter} = \begin{array}{l} 1 \text{ Si } p \in Phr \\ 0 \text{ Sinon} \end{array} \tag{4.2}$$

En d'autres termes, le Np_{inter} d'une personne est égal à 1 si, dans au moins une vidéo, le nom de p a été écrit/prononcé quand elle parle ou est visible. Sinon il est de 0.

Donc, pour toutes les personnes, le score intra et inter vidéos est égal à :

$$N_{intra} = \frac{\sum_{p \in Pr} Np_{intra}}{\#p \in Pr} \tag{4.3}$$

$$N_{inter} = \frac{\sum_{p \in Pr} Np_{inter}}{\#p \in Pr} \tag{4.4}$$

Observons l'exemple ci-dessous. Il comporte trois vidéos (V_A, V_B, V_C) et cinq personnes (P_1 à P_5). Les noms de ces personnes peuvent être écrits ou prononcés dans chaque vidéo (N_1 à N_5).

	V_A	V_B	V_C
$Personnes$:	P_1, P_2, P_3	P_1, P_3, P_4	P_1, P_5
$Noms$:	N_1, N_2	N_3, N_5	N_5, N_4

Les méthodes de comptabilisation nous permettent de calculer les scores de chaque personne et les scores pour l'ensemble des personnes :

	P_1	P_2	P_3	P_4	P_5	$Globale$
Np_{intra}	1/3	1/1	1/2	0/1	1/1	$\rightarrow N_{intra} = 0.57$
Np_{inter}	1	1	1	0	1	$\rightarrow N_{inter} = 0.8$

Ainsi, on peut observer que : P_1 est présent dans les trois vidéos, mais il n'est nommable que dans une (V_1). Donc, son score Np_{intra} est égal à 1/3 et son score Np_{inter} est égal à 1. Le nom de P_4 n'est jamais prononcé ou écrit dans une vidéo où il est présent, donc cette personne n'est pas considérée comme nommable (Np_{intra}=Np_{inter}=0).

4.2.2 Nombre d'occurrences de citation d'un nom

En plus du rappel, nous allons aussi comptabiliser le nombre d'occurrences des noms écrits/cités (Occ) et le nombre d'occurrences de noms lorsque la personne correspondante parle ou est visible (Occ_{pv}). Un plus grand nombre d'occurrences peut aider les systèmes d'association nom-personne.

Nous utiliserons comme notations :

Occ	:	nombre d'occurrences des noms cités et/ou écrits
Occ_{pv}	:	nombre d'occurrences des noms cités et/ou écrits où la personne correspondant au nom parle ou est visible dans les segments UEM

L'annotation de l'image n'étant effectuée que toutes les 10 secondes sur les segments UEM, Occ_{pv} sera utilisé à titre indicatif pour comparer deux systèmes et ils seront donc sous-évalués pour les vidéos complètes.

Dans les tableaux suivants, nous utiliserons comme notation :

M_{UEM}	:	annotations manuelles sur les segments UEM
A_{UEM}	:	systèmes automatiques sur les segments UEM
A_{RAW}	:	systèmes automatiques sur les vidéos complètes (brutes)
$N_{cités}$	:	noms cités à l'oral
$N_{écrits}$	:	noms écrits dans un cartouche à l'écran

4.3 Noms cités ou écrits pour nommer les personnes présentes dans les vidéos

Comme nous avons pu le voir, les noms cités à l'oral proposent un plus grand nombre d'occurrences ainsi qu'un plus grand nombre de personnes différentes citées. En contrepartie, la probabilité que les personnes correspondant à ces noms soient présentes dans les vidéos est plus faible.

4.3.1 Personnes apparaissant ou parlant

Dans les tableaux 4.3 et 4.4, nous comparons les noms issus de la transcription de la parole ($N_{cités}$) et/ou écrits à l'écran ($N_{écrits}$) par rapport aux personnes apparaissant et/ou parlant dans les segments UEM. Ces noms sont produits à partir d'annotations manuelles (M_{UEM}) ou à partir de systèmes automatiques (A_{UEM}, A_{RAW}).

La proportion de personnes nommables par les noms écrits dans les annotations manuelles est légèrement sous-évaluée. En effet, seulement une image toutes les 10 secondes ou au moins une par plan a été annotée. L'annotation ne porte donc pas sur tous les noms écrits, ce qui explique le score supérieur du système automatique par rapport à celui des annotations manuelles.

4.3.1.1 Personnes apparaissant

Le tableau 4.3 présente la proportion de personnes apparaissant dont le nom a été cité/écrit ainsi que le nombre d'occurrences de ces noms. Dans les annotations manuelles, il y a plus d'occurrences de noms cités à l'oral (4273) qu'écrits à l'écran (1049). Par contre, lorsqu'un nom est écrit dans une vidéo, dans 99,1% des cas, la personne correspondant au nom apparaît à l'écran à un moment ou à un autre de la vidéo. Cette proportion est plus faible pour les noms cités à l'oral (60,3%).

L'utilisation de systèmes automatiques sur les segments UEM réduit le nombre (Occ) de noms cités de 4273 à 2905. Or, seulement 1435 occurrences (49.4%) de ces noms correspondent à des personnes visibles. L'utilisation conjointe des noms cités et des noms écrits, extraits de manière automatique, permet d'augmenter le nombre d'occurrences des noms de personnes apparaissant dans les segments UEM à 2767.

$N_{cités}$	$N_{écrits}$	Occ	Occ_{pv}	N_{intra}	N_{inter}
M_{UEM}	X	4273	2577 (60,3%)	59,1	66,2
X	M_{UEM}	1049	1040 (99,1%)	44,0	51,9
M_{UEM}	M_{UEM}	5322	3617 (68,0%)	71,9	78,5
A_{UEM}	X	2905	1435 (49,4%)	26,1	31,9
X	A_{UEM}	1407	1332 (94,7%)	49,5	57,0
A_{UEM}	A_{UEM}	4312	2767 (64,2%)	59,7	66,3

TAB. 4.3 – Nombre d'occurrences des noms et pourcentages des 724 personnes apparaissant nommables par les noms cités à l'oral et/ou écrits à l'écran

La proportion (N_{intra}) des personnes apparaissant à l'écran dont le nom a été cité dans les annotations manuelles (M_{UEM}=59,1%) est plus importante que celle dont le nom a été écrit (M_{UEM}=44%, A_{UEM}=49,5%). Cependant, les erreurs dans les noms cités extraits automatiquement abaissent N_{intra} à 26,1%. La combinaison des noms écrits et cités augmente le score pour les personnes apparaissant. Ce qui montre leur complémentarité, que ce soit avec les annotations manuelles (de 44% à 71,9%) ou avec les systèmes automatiques (de 49,5% à 59,7%). L'utilisation d'une propagation inter-vidéos augmente en moyenne le score N_{inter} de 7%.

4.3.1.2 Personnes parlant

Nous pouvons constater, dans le tableau 4.4, que les noms écrits extraits automatiquement peuvent nommer 73,5% des 555 locuteurs alors qu'ils ne peuvent nommer que 49,5% des 724 personnes apparaissant. Les noms écrits sont quasiment toujours utilisés pour introduire une personne qui parle et apparaît en même temps.

A contrario, les noms cités couvrent proportionnellement autant de locuteurs que de personnes apparaissant. Ils montrent donc leur utilité pour nommer les personnes apparaissant à l'écran alors que ces personnes ne parlent pas (personnes visibles dans un reportage de journal télévisé par exemple).

$N_{cités}$	$N_{écrits}$	Occ	Occ_{pv}	N_{intra}	N_{inter}
M_{UEM}	X	4273	1863 (43,6%)	62,2	66,5
X	M_{UEM}	1049	1022 (97,4%)	60,5	65,9
M_{UEM}	M_{UEM}	5322	2885 (54,2%)	80,4	83,6
A_{UEM}	X	2905	914 (31,5%)	26,7	30,8
X	A_{UEM}	1407	1348 (95,8%)	73,5	76,8
A_{UEM}	A_{UEM}	4312	2262 (52,5%)	75,8	78,7

TAB. 4.4 – Nombre d'occurrences des noms et pourcentages des 555 personnes parlant nommables par les noms cités à l'oral et/ou écrits à l'écran

Là aussi, l'utilisation conjointe des deux modalités augmente le score mais de façon moins importante que pour les personnes apparaissant (de 60.5% à 80.4% pour M_{UEM}) ; surtout lors de l'utilisation des systèmes automatiques (73,5% à 75,8% pour A_{UEM}). Une propagation inter-vidéos augmente moins les possibilités de nommage (+4% en moyenne) que pour les personnes apparaissant.

4.3.2 Détail par rôle de personnes

Nous avons vu dans le chapitre 2, pendant la présentation du corpus *REPERE*, qu'il y avait un déséquilibre du temps de présence entre les présentateurs/chroniqueurs/journalistes (rôles R123) et les invités/autres (rôles R45). Ainsi, sur la partie apprentissage de la phase 1 de ce corpus (dont sont issues les statistiques présentées dans ce chapitre) les personnes de R123 représentent 15% des locuteurs et 7% des personnes apparaissant alors qu'il couvrent 45% du temps de parole et 30% des apparitions de visage.

De plus, nous avons montré dans ce même chapitre que les modèles biométriques pour les personnes des rôles R123 pouvaient être utilisés à bon escient ; parce qu'il est facile d'avoir une connaissance a priori de leur présence dans les vidéos tout en limitant le nombre de modèles utilisés par les systèmes automatiques.

Par contre, les personnes des rôles 4 et 5 ne font pas partie de la distribution habituelle d'une émission. Il est donc difficile d'avoir un modèle biométrique leur correspondant sans multiplier le nombre de modèles utilisés. Par ailleurs, cela dégraderait la qualité de l'identification. Ce sont donc ces personnes qu'il est intéressant de nommer de manière non-supervisée.

Ainsi, nous détaillons donc dans le tableau 4.5, les possibilités de nommer les 808 personnes présentes (union des personnes parlant et apparaissant) en fonction du rôle qu'elles occupent dans les vidéos.

$N_{cités}$	$N_{écrits}$	Occ_{pv}		N_{intra}		N_{inter}	
		R123	R45	R123	R45	R123	R45
M_{UEM}	X	414	2353	78,9	55,2	86,9	61,7
X	M_{UEM}	91	952	23,0	40,6	35,7	47,9
M_{UEM}	M_{UEM}	505	3305	81,0	67,7	89,3	73,6
A_{UEM}	X	58	1396	13,9	24,7	16,7	30,6
X	A_{UEM}	174	1177	37,8	46,3	47,6	53,4
A_{UEM}	A_{UEM}	232	2573	42,9	56,3	52,4	62,5

Tab. 4.5 – Nombre d'occurrences des noms et pourcentages des personnes présentes nommables par les noms cités ou écrits en fonction de leurs rôles (R123 : 84 présentateur/chroniqueur/reporter, R45 : 728 Invité/autre)

Nous pouvons observer que les noms des 84 présentateurs / chroniqueurs / journalistes sont assez peu cités (Occ_{pv} pour M_{UEM}=414, A_{UEM}=58) ou écrits (Occ_{pv} pour M_{UEM}=91, A_{UEM}=174) par rapport à leurs temps de présence. Cependant, ils ont, pour la majorité, leurs noms prononcés dans les segments UEM (N_{intra} = 78.9% en M_{UEM}). Par contre, il est difficile pour des systèmes automatiques d'extraire ces noms parce qu'ils sont soit inconnus des systèmes, soit parce que les personnes sont juste citées par leur prénom. Comme il n'est pas toujours évident de pouvoir compléter ces prénoms pour obtenir une identité complète, le score N_{intra} diminue à 13.9% avec les systèmes automatiques.

Cette différence entre annotations manuelles et extraction automatique n'apparaît pas pour les noms écrits car les journalistes intervenants à l'oral et visibles à l'image, sont souvent introduits par leurs noms écrits alors que les journalistes en voix-off ne sont jamais présentés ainsi. Les personnes du rôle R123 sont donc plus difficilement nommables automatiquement alors qu'elles représentent une proportion importante du temps de présence.

En comparaison, les personnes de R45 sont plus nommables, quelle que soit la source automatique, que les personnes de R123 ($N_{écrits}$ 46,3% et 37,8%, $N_{cités}$ 24,7% et 13,9%). L'utilisation conjointe des deux modalités permet donc d'augmenter le nombre de personnes nommables et le nombre d'occurrences quel que

soit le type de rôle pris en compte et avec ou sans la propagation à d'autres vidéos.

4.3.3 Apport de l'utilisation des vidéos complètes

L'utilisation des vidéos complètes (A_{RAW}) par rapport à la seule utilisation des segments annotés (A_{UEM}) augmente le nombre d'occurrences de citation des noms de personnes apparaissant ou parlant (Occ_{pv} de A_{UEM}=2805 à A_{RAW}=3476), sans pour autant augmenter significativement le nombre de personnes présentes nommables dans les segments UEM (A_{UEM}=55,1% à A_{RAW}=56,7%). Par contre, ce nombre d'occurrences supplémentaires peut faciliter l'association noms-personnes.

$N_{cités}$	$N_{écrits}$	Occ	Occ_{pv}	N_{intra}	N_{inter}
M_{UEM}	X	4273	2767 (64,8%)	57,7	64,4
X	M_{UEM}	1049	1043 (99,4%)	39,0	46,9
M_{UEM}	M_{UEM}	5322	3810 (71,6%)	69,2	75,4
A_{UEM}	X	2905	1454 (50,1%)	23,7	29,3
X	A_{UEM}	1407	1351 (96,0%)	45,5	53,0
A_{UEM}	A_{UEM}	4312	2805 (65,1%)	55,1	61,6
A_{RAW}	X	4922	1755 (35,7%)	24,8	30,4
X	A_{RAW}	2090	1721 (82,3%)	47,3	54,6
A_{RAW}	A_{RAW}	7012	3476 (49,6%)	56,7	62,7

TAB. 4.6 – Apport des vidéos complètes, pour le nommage des 808 personnes apparaissant et/ou parlant dans les segments UEM

Les pourcentages des Occ_{pv} pour les A_{RAW} sont sous-évalués. L'annotation ne portant que sur les segments UEM, nous ne pouvons pas affirmer qu'un nom cité ou écrit ne correspond pas à une personne présente en dehors des segments UEM.

Si on détaille, encore une fois par type de rôle, le score N_{intra} (tableau 4.7), on voit que l'utilisation des vidéos complètes apporte plus d'informations pour les personnes des rôles R123 que pour celles des rôles R45. En effet, le nom des présentateurs/chroniqueurs est souvent cité/écrit en début d'émission alors que pour les invités/autres, il est cité/écrit au moment de l'intervention.

$N_{cités}$	$N_{écrits}$	Occ_{pv}		R_{intra}	
		R123	R45	R123	R45
A_{RAW}	X	78(+20)	1677(+281)	17.8(+3.9)	25.5(+0.8)
X	A_{RAW}	226(+52)	1495(+318)	41.8(+4)	47.8(+1.5)
A_{RAW}	A_{RAW}	304(+72)	3172(+599)	47.1(+4.2)	57.6(+1.3)

TAB. 4.7 – Apport des vidéos complètes en fonction du rôle des personnes pour le nommage des 808 personnes apparaissant ou parlant dans les segments UEM. Entre parenthèses apparaît l'augmentation en absolu par rapport aux données du tableau 4.6, ligne A_{UEM}.

4.3.4 Détail par type d'émission

Ce corpus étant composé de sept émissions différentes, nous pouvons observer les variations de la proportion de personnes nommables (voir tableau 4.8). Les deux journaux télévisés ont un comportement assez similaire alors que les personnes présentes dans l'émission d'actualité people sont plus difficilement nommables (beaucoup de personnes au second plan, les personnes ne parlant pas français ont un doubleur, les personnes de groupes de musique sont introduites par le nom du groupe et non le nom de chacun des membres, etc).

Émission	Type	Noms cités		Noms écrits	
		N_{intra}	N_{inter}	N_{intra}	N_{inter}
BFM story	Journal télévisé	31.3	32.6	53.9	57.5
LCP info	Journal télévisé	32.9	39.3	51.0	56.8
Planète showbiz	Actu people	17.6	21.7	32.7	37.1
Pile et face	Débat	58.3	61.1	100.0	100.0
Entre les lignes	Débat	51.9	61.9	42.9	42.9
Ça vous regarde	Débat	35.7	35.7	53.3	53.6
Top questions	Questions à l'Assemblée	31.2	40.2	59.4	69.0
Totalité		24.8	30.4	47.3	54.6

TAB. 4.8 – Détail par type d'émission pour le nommage des 808 personnes apparaissant et/ou parlant dans les segments UEM. Les noms hypothèses sont extraits des vidéos complètes

Les trois émissions de débat ont des résultats variables. Dans l'émission « Pile et face », seulement trois personnes sont présentes (un présentateur, deux invités - aucun reportage donc aucune personne supplémentaire). Leur nom est toujours écrit dans la vidéo (de multiples fois pour les invités) ce qui explique le résultat de 100% pour les noms écrits. Pour les noms prononcés, on retrouve un peu moins de deux tiers des personnes nommables, a priori, ce sont les invités interpelés par le présentateur. L'émission « Entre les lignes » invite des chroniqueurs à parler d'un sujet d'actualité, encadrés par un présentateur. Ces chroniqueurs sont redondants d'une émission à l'autre, ce qui explique l'augmentation importante (de 51,9% à 61,9%) obtenue avec les noms prononcés lorsque l'on utilise une propagation inter-vidéos. Il est à noter que sur cette émission, les noms prononcés obtiennent un résultat supérieur aux noms écrits. A contrario, dans « Ça vous regarde », les invités changent à chaque émission. Il n'y a donc quasiment aucune augmentation lors de la propagation des noms aux autres vidéos.

Dans les « Questions à l'Assemblée », de nombreuses personnes sont visibles au second plan (député assis juste derrière celui qui pose la question ou ministre proche de celui qui répond) mais comme elles ne jouent pas de rôle dans l'émission, elles ne sont donc pas introduites à l'oral ou par leur nom écrit. Malgré tout, dans le protocole d'évaluation du défi *REPERE* ces personnes sont à identifier. La propagation des noms aux autres vidéos permet là aussi de rendre ces personnes nommables si elles sont intervenues dans une autre vidéo.

4.3.5 Affiliation des noms hypothèses aux personnes à l'aide d'un « oracle au voisinage »

Jusqu'à présent, nous avons considéré qu'à partir du moment où un nom était cité ou écrit, la personne correspondant à ce nom pouvait être nommée quel que soit le moment où elle apparaissait/parlait dans la vidéo (ce qui correspond à l'utilisation d'un oracle au niveau de la vidéo).

Cependant, les systèmes de l'état de l'art se restreignent aux tours de parole contigus pour effectuer l'association d'un nom à une personne. Nous allons donc remplacer « l'oracle au niveau de la vidéo » par un « oracle au voisinage ». C'est-à-dire qu'une personne sera nommable si son nom est cité ou écrit dans le voisinage direct du moment où elle apparaît/parle.

Selon les travaux de l'état de l'art, l'oracle au voisinage a un comportement différent pour chacune des deux modalités :

- Un nom écrit pourra nommer seulement les personnes apparaissant/parlant pendant qu'il apparaît à l'écran.
- Un nom prononcé pourra nommer les personnes apparaissant/parlant dans les tours de parole précédents, courants ou suivants.

Dans cette section, nous allons donc comparer la capacité d'association des noms écrits ou cités aux bonnes personnes à l'aide de cet oracle au voisinage.

$N_{cités}$	$N_{écrits}$	Oracle au niveau de la vidéo			Oracle au voisinage		
		Occ_{pv}	N_{intra}	N_{inter}	Occ_{pv}	N_{intra}	N_{inter}
M_{UEM}	X	2767	57,7	64,4	1580	51,8	58,9
X	M_{UEM}	1043	39,0	46,9	977	38,4	45,9
A_{UEM}	X	1454	23,7	29,3	632	20,9	26,4
X	A_{UEM}	1351	45,5	53,0	1269	45,4	52,5

TAB. 4.9 – Proportion des 808 personnes parlant ou apparaissant nommables à l'aide d'un oracle au voisinage, segments UEM.

Le tableau 4.9 nous montre qu'il est plus facile d'utiliser un nom écrit pour identifier une personne présente. En effet, on peut constater que lorsqu'on restreint l'association des noms cités aux personnes présentes dans les tours de parole adjacents, le score de nommage réduit de 2,8% à 5,9% selon le système et la propagation utilisés. Alors qu'il n'y a que très peu ou pas de différence pour les noms écrits (réduction de 0,1% à 1,0%).

Le nombre d'occurrences de noms utilisables réduit lui aussi fortement pour les noms cités (de 2767 à 1580 pour M_{UEM} et de 1454 à 632 pour A_{UEM}) alors qu'il ne réduit que très peu pour les noms écrits (de 1043 à 977 pour M_{UEM} et de 1351 à 1269 pour A_{UEM}).

Pour les noms cités (comme nous l'a déjà montré l'état de l'art), il faut sélectionner les noms à utiliser ; alors que pour les noms écrits, la quasi totalité des occurrences sont utilisables pour identifier les personnes directement présentes.

4.4 Conclusion

Les noms prononcés et écrits à l'écran sont des sources d'informations importantes pour obtenir les noms des personnes présentes dans les émissions de télévision. Les noms prononcés bénéficient d'un plus grand nombre d'occurrences de citation par rapport aux noms écrits.

En revanche, les erreurs de détection et de transcription des systèmes automatiques réduisent le nombre de personnes nommables obtenu pour cette modalité. A contrario, l'augmentation de la qualité des vidéos permet aux systèmes automatiques d'extraction des noms écrits de générer très peu d'erreurs de transcription. Il y a donc une marge d'évolution plus importante pour les noms cités que pour les noms écrits.

Il est important de souligner que les noms cités sont dépendants d'un modèle de langue pour leur extraction (transcription de la parole et détection des entités nommées). Même si les noms écrits ont besoin d'un modèle de caractères pour effectuer la transcription, il est beaucoup plus facile de créer ce modèle qu'un modèle de langue (pour LOOV nous avons utilisé le modèle de caractères fournit par défaut avec le logiciel `Tesseract`).

Sur le corpus *REPERE*, les noms cités extraits automatiquement peuvent permettre de nommer environ deux fois moins de personnes que les noms écrits. Les noms écrits sont principalement utilisés pour introduire à la fois des personnes apparaissant et parlant en même temps. En revanche, les noms prononcés peuvent aussi introduire des journalistes parlant en voix-off ou encore des personnes apparaissant mais ne parlant pas. De plus, nous avons pu voir que les présentateurs /chroniqueurs/journalistes sont difficilement nommables, ce qui nous oriente vers l'utilisation de modèles biométriques pour ces catégories de personnes.

Un dernier point à noter est que l'association des noms écrits aux bonnes personnes est intrinsèquement plus simple qu'avec les noms cités. Malgré ces différences de résultats, ces deux modalités restent très complémentaires. En outre, la propagation d'un nom à d'autres vidéos augmente systématiquement et de manière importante la proportion de personnes nommables.

Chapitre 5

Méthodes d'identification des personnes dans les flux télévisés basées sur les noms écrits

Maintenant que nous avons toutes nos briques de base et une bonne connaissance des capacités de chacune, nous proposons plusieurs méthodes pour identifier les personnes présentes dans les flux télévisés à l'aide des noms écrits.

La première section de ce chapitre est consacrée à la tâche d'identification des locuteurs à l'aide des noms écrits. Pour ce faire, nous avons d'abord commencé par développer plusieurs méthodes de **nommage tardif** avec une propagation des noms écrits sur des clusters de locuteurs issus d'un système de diarization (optimisé pour avoir le plus faible taux d'erreur de diarization). Ces méthodes proposent différents niveaux de « remise en cause » des clusters de locuteurs issus de la diarization.

Ceci nous a amené à proposer une stratégie de **nommage intégré** où le critère d'arrêt du regroupement est choisi en fonction de la tâche cible (minimisation de l'erreur d'identification et non de l'erreur de diarization).

Ensuite, dans le **nommage précoce**, l'information issue des noms écrits est utilisée pour contraindre le processus de regroupement en locuteurs.

La seconde section de ce chapitre présente une adaptation de notre méthode de nommage précoce pour la tâche d'identification des visages. Cette adaptation prend en compte la possibilité que plusieurs visages peuvent apparaître en même temps à l'écran.

Les résultats présentés dans ce chapitre sont obtenus sur le corpus de test de la phase 1 du défi *REPERE*. Comme nous avons utilisé le protocole de ce défi, l'évaluation porte donc sur les 1252 images annotées pour la qualité de l'identification et sur tout le signal annoté pour la qualité de la diarization.

5.1 Identification non supervisée des locuteurs

Pour l'identification des locuteurs sans l'aide de modèles biométriques, la solution la plus utilisée dans l'état de l'art est un nommage tardif utilisant un regroupement en locuteurs et une source de noms hypothèses. Aucun article de l'état de l'art n'avait jusqu'à présent tenté de faire ce nommage tardif à l'aide des noms écrits à l'écran exclusivement. Comme nous pouvons extraire les noms écrits avec une très bonne qualité de transcription, nous pouvons proposer des méthodes qui utilisent exclusivement ces noms.

Nous allons d'abord présenter 3 méthodes de nommage tardif effectuant l'identification du locuteur avec seulement les noms écrits à l'écran comme source d'identité. Ces méthodes (`NT1` à `NT3`) ont fait l'objet d'une publication à la conférence INTERSPEECH en 2012 [PBL+12]. Plusieurs variantes de l'approche `NT3` sont aussi proposées.

La dernière méthode présentée dans cette section est le nommage précoce (`NP`). Elle correspond à un regroupement où les noms sont au préalable associés aux tours de parole ce qui permet de nommer directement les clusters et de contraindre le regroupement. Elle a fait l'objet d'une publication au workshop SLAM en 2013 [PBB+13].

Il est à noter que nous avons fait l'hypothèse qu'un seul locuteur parlait dans chacun des segments audios (aucune parole superposée).

5.1.1 Nommage tardif `NT`

Une première solution pour nommer les locuteurs de manière non-supervisée dans les vidéos est d'utiliser une méthode de fusion tardive entre deux briques indépendantes l'une de l'autre :

- La sortie d'un système de diarization réglé pour avoir la plus faible erreur (DER).
- Les noms écrits à l'écran pour introduire la personne correspondante.

On retrouve le schéma classique (figure 5.1), présenté dans le chapitre 1, où les deux briques de base sont réunies par l'étape d'association.

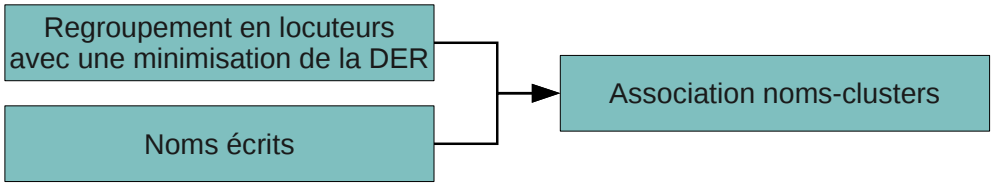


Fig. 5.1 – Schéma classique d'un nommage tardif

Nous avons essayé plusieurs stratégies pour l'étape d'association. Pour en illustrer les différences, prenons comme cas d'étude l'exemple de la figure 5.2.

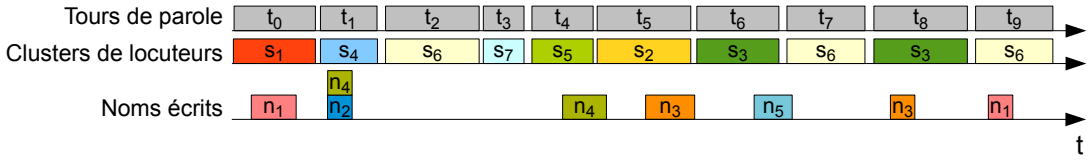


FIG. 5.2 – Exemple d'une chronologie avec les différentes segmentations de la parole et des noms écrits

Cette chronologie présente (sur la ligne) :

1. La segmentation en tours de parole $\mathcal{T} = \{t_1, \ldots, t_K\}$.
2. Le regroupement en clusters correspondants $\mathcal{S} = \{s_1, \ldots, s_L\}$.
3. L'affichage des noms écrits à l'écran $\mathcal{N} = \{n_1, \ldots, n_M\}$.

Un graphe peut être une autre représentation des liens de co-occurrences existant entre les tours de parole, les clusters de locuteurs et les noms écrits de notre exemple (voir figure 5.3).

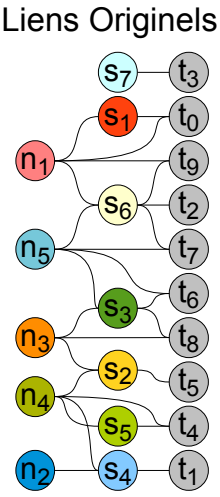


FIG. 5.3 – Graphe représentant les liens de co-occurrences entre modalités

L'objectif d'un nommage tardif est de trouver la fonction d'association optimale m, entre les tours de parole et les noms, définie comme suit :

$$m : \mathcal{T} \to \mathcal{N}$$
$$t \mapsto \begin{cases} n & \text{si le nom du tour de parole } t \text{ est } n \in \mathcal{N} \\ \varnothing & \text{si il n'est pas dans } \mathcal{N} \end{cases} \tag{5.1}$$

Nous proposons trois méthodes d'association tardive entre les noms écrits et les tours de parole. La première (`NT1`) ne remet pas en cause le regroupement en clusters. La suivante (`NT2`) propose d'extraire certains tours de parole d'un cluster pour les nommer différemment de leur cluster initial. La troisième méthode (`NT3`) propose de re-fusionner des clusters qui ne l'ont pas été lors du regroupement.

5.1.1.1 `NT1` : association 1-à-1

Cette première méthode repose sur l'hypothèse forte que la diarization fournit des clusters parfaits. Par conséquent, pour identifier les locuteurs, il suffit de trouver l'association 1-à-1 $f : \mathcal{S} \to \mathcal{N} \cup \varnothing$ qui maximise la durée de co-occurrence entre les clusters de locuteurs et les noms écrits à l'écran :

$$f = \underset{f}{\operatorname{argmax}} \sum_{s \in \mathcal{S}} \mathbb{K}\left(s, f(s)\right) \tag{5.2}$$

Où $\mathbb{K}(s, n)$ est la durée totale des segments où les locuteurs s parlent et un nom n est affiché à l'écran simultanément. $f(s) = \varnothing$ représente les noms des locuteurs s qui demeurent inconnus et dans ce cas $\mathbb{K}(s, \varnothing) = 0$.

L'algorithme dit Hongrois (également connu sous le nom d'algorithme d'attribution Munkres) est utilisé pour résoudre ce problème en temps polynomial [Kuh55].

Le résultat de l'exemple peut-être observé sur les figures 5.4 et 5.5.

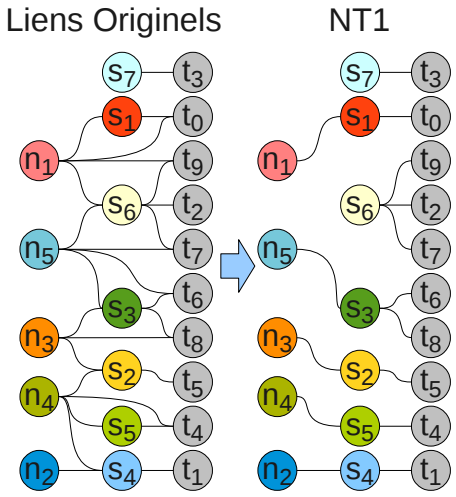


FIG. 5.4 – Graphe obtenu par la méthode `NT1`

On peut voir que les clusters s_6 et s_7 n'ont pas été nommés par cette première méthode.

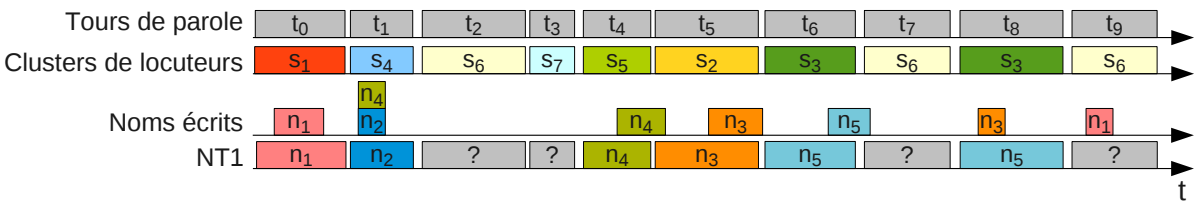


FIG. 5.5 – Chronologie obtenue par la méthode `NT1`

5.1.1.2 `NT2` : identification directe puis association 1-à-1

La seconde approche (notée `NT2`), remet en question une première fois la diarization. Cette remise en question est basée sur l'observation suivante : quand un et un seul nom n est écrit à l'écran dans un cartouche, le locuteur du tour de parole co-occurrant à une forte probabilité (96.6% sur l'ensemble de test de la phase 1 du corpus *REPERE*) de correspondre à ce nom.

Au vu de cette information, nous avons découpé le travail en deux étapes :

1. Les tours de parole co-occurrents avec un seul nom n sont nommés par ce nom.
2. La méthode `NT1` est appliquée sur les tours de parole restants.

Le résultat de l'exemple peut être vu sur les figures 5.6 et 5.7.

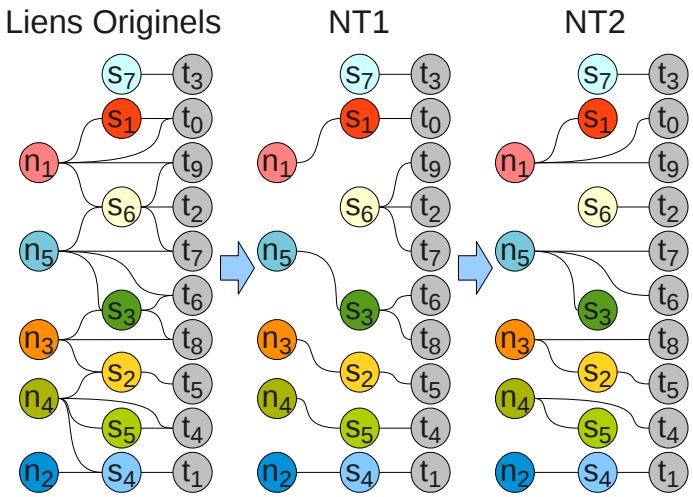


FIG. 5.6 – Graphe obtenu par la méthode `NT2`

Les clusters s_6 et s_3 sont découpés par la première étape de notre méthode. En effet, les tours de parole t_7, t_8 et t_9 sont nommés différemment du cluster auquel ils appartiennent.

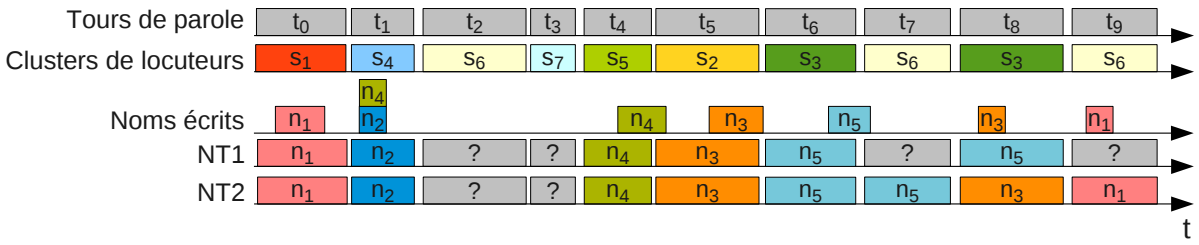


FIG. 5.7 – Chronologie obtenue par la méthode `NT2`

Cette méthode va nommer plus de tours de parole que la méthode `NT1`, ce qui devrait augmenter le rappel. La précision peut aussi augmenter car un nom écrit identifie avec une forte probabilité le locuteur co-occurrent.

5.1.1.3 `NT3` : identification directe puis association 1-à-n

Notre troisième approche (notée `NT3`) remet en cause une deuxième fois la diarization. Elle part de l'hypothèse qu'elle peut produire des clusters sur-segmentés. C'est-à-dire que plusieurs clusters peuvent correspondre à un seul locuteur. Ce qui est, a priori, le cas dans notre exemple pour les clusters s_2 et s_3. Nous avons donc remplacé l'alignement 1-à-1 par un alignement 1-à-n, où un nom peut correspondre à plusieurs clusters de locuteurs.

Avant d'effectuer cet alignement, nous avons, comme pour la méthode `NT2`, étiqueté directement certains tours de parole. Ensuite, les autres tours sont étiquetés selon le critère suivant :

$$t(s) = \underset{n \in \mathcal{N}}{\operatorname{argmax}} \, \mathrm{TF}\,(s, n) \cdot \mathrm{IDF}\,(n) \tag{5.3}$$

Où le coefficient *Term-Frequency Inverse Document Frequency* (TF-IDF) [RJ76, FTZ04], rendu célèbre par la communauté de recherche d'information, est adapté à notre problème :

$$\mathrm{TF}\,(s, n) = \frac{\text{durée du nom } n \text{ dans le cluster } s}{\text{durée totale de tous les noms dans le cluster } s} \tag{5.4}$$

$$\mathrm{IDF}\,(n) = \frac{\#\text{ clusters de locuteurs}}{\#\text{ clusters de locuteurs co-occurrents avec } n} \tag{5.5}$$

La partie IDF n'a que très peu d'influence. Elle ne jouera un rôle que si deux noms nomment le même cluster avec le même score de TF et qu'un des deux noms nomme aussi un autre cluster. La partie IDF influencera le score pour nommer le cluster avec le nom le moins utilisé. Son utilité n'est donc que très marginale.

Le résultat de l'exemple peut-être vu sur les figures 5.8 et 5.9.

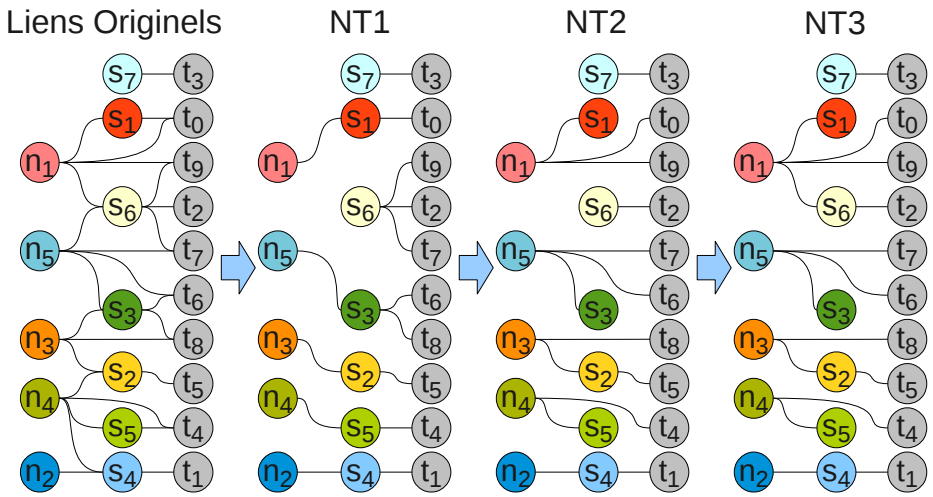


Fig. 5.8 – Graphe obtenu par la méthode `NT3`

NT3 va nommer le cluster s_6 par le nom n_1 en plus du cluster s_1. Tous les autres clusters et tours de parole restent nommés de la même manière.

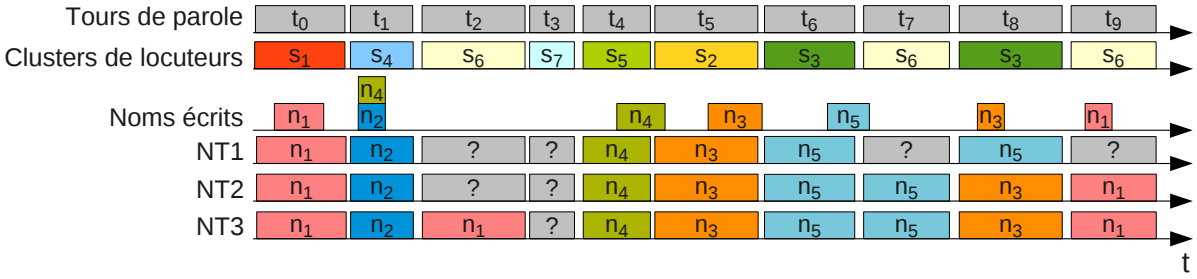


FIG. 5.9 – Chronologie obtenue par la méthode NT3

Comme cette méthode va fusionner des clusters, elle devrait augmenter le rappel mais elle peut aussi réduire la précision.

5.1.1.4 NT3$^{\ominus}$: Ré-alignement temporel entre noms écrits et tours de parole

Deux raisons nous ont orienté vers un ré-alignement temporel entre la modalité voix et les noms écrits. La première vient de l'utilisation de décodeurs vidéos utilisant des stratégies différentes de décodage, ce qui peut engendrer un décalage temporel entre les informations provenant de la bande son et les informations provenant de l'image.

La seconde vient de la segmentation des noms écrits qui ne correspond pas toujours à la segmentation de la voix, par exemple, lorsqu'une personne coupe la parole à une autre et que le nom de la première personne n'a pas disparu.

Pour éviter la propagation d'un nom sur un mauvais cluster, nous avons réduit la portée temporelle des noms écrits au tour de parole le plus co-occurrent.

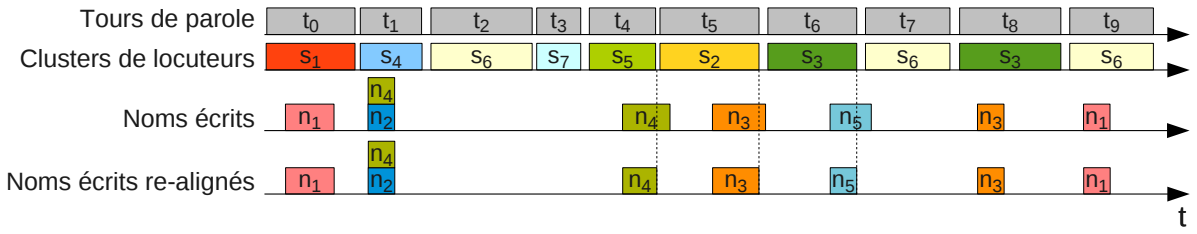


FIG. 5.10 – Réduction de la portée temporelle des noms écrits au tour de parole le plus co-occurrent

Dans notre exemple (figure 5.10), la segmentation des trois noms n_3, n_4 et n_5 co-occurrant les tours de parole t_4, t_5, t_6 a été réduite. Les noms n_4 et n_5 ne co-occurrent plus avec respectivement les tours de parole t_5 et t_7.

Sur la figure 5.11, on voit que le tour de parole t_7 n'est plus nommé directement par le nom n_5 .

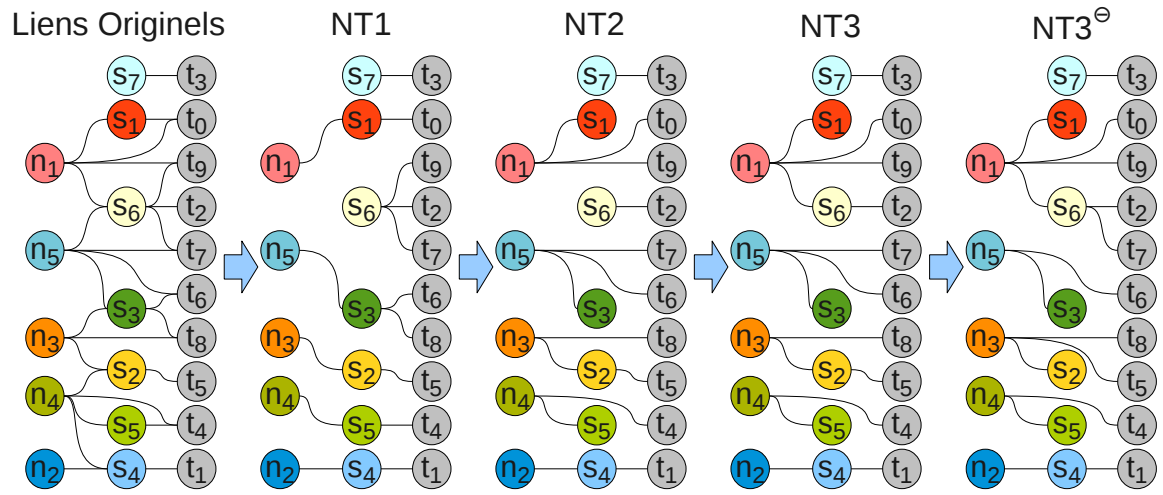


FIG. 5.11 – Graphe obtenu par la méthode NT3⊖

Quand on regarde la chronologie, il y a effectivement de fortes chances pour que le nom n_5 désigne le tour de parole t_6 seulement et non le tour de parole t_7. Il ne faut donc pas nommer ce tour de parole à l'aide de ce nom et surtout ne pas propager le nom sur le cluster dont t_7 fait partie.

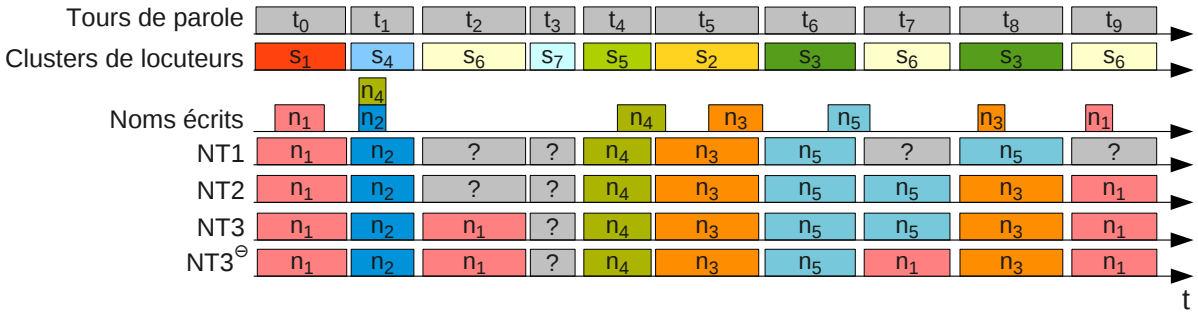


FIG. 5.12 – Chronologie obtenue par la méthode NT3⊖

Avec ce ré-alignement des noms par rapport à la segmentation audio, nous devrions éviter de propager les noms sur de mauvais clusters. Ce qui devrait donc augmenter la précision.

5.1.1.5 NT3⊖+NA : Ajout de l'information des noms prononcés des allocutaires

Dans notre exemple, le tour de parole t_3 n'a pas pu être nommé par les méthodes précédentes. Plusieurs solutions pourraient être utilisées pour identifier les tours de parole encore inconnus. Par exemple, à partir de modèles biométriques ([EKLMP12]) ou encore à l'aide de quelques annotations ([PMT10, PTM11]).

Pour pallier ce manque, nous avons utilisé les noms prononcés extraits par le système du LIMSI. En effet, leur système nous indique en plus si un nom prononcé peut correspondre à un allocutaire (voir chapitre 2) sans pour autant nous préciser s'il correspond au locuteur courant, suivant ou précédent.

Dans un premier temps, nous avons donc étiqueté chaque tour de parole avec les noms des allocutaires par le locuteur précédent ou suivant (voir figure 5.13, ligne « `NA` propagés »). Ensuite, nous avons appliqué l'algorithme Hongrois (méthode `NT1`) pour trouver le meilleur alignement entre les noms et les clusters de locuteurs (ligne « `NT1` pour les `NA` »).

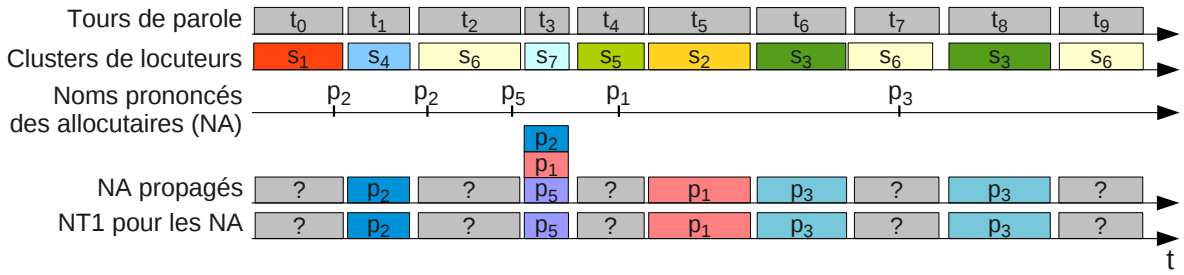


FIG. 5.13 – Segmentation obtenue par la méthode `NT1` pour les noms des allocutaires

Puis, dans un deuxième temps, nous avons utilisé cet alignement pour nommer les tours de parole encore inconnus après l'application de la méthode $\texttt{NT3}^{\ominus}$.

Dans la figure 5.14, on retrouve toutes les segmentations proposées par ces 5 méthodes. On peut voir que seul le tour de parole t_3 a été nommé à l'aide d'un nom prononcé désigné comme allocutaire. L'ajout de cette information devrait augmenter le rappel.

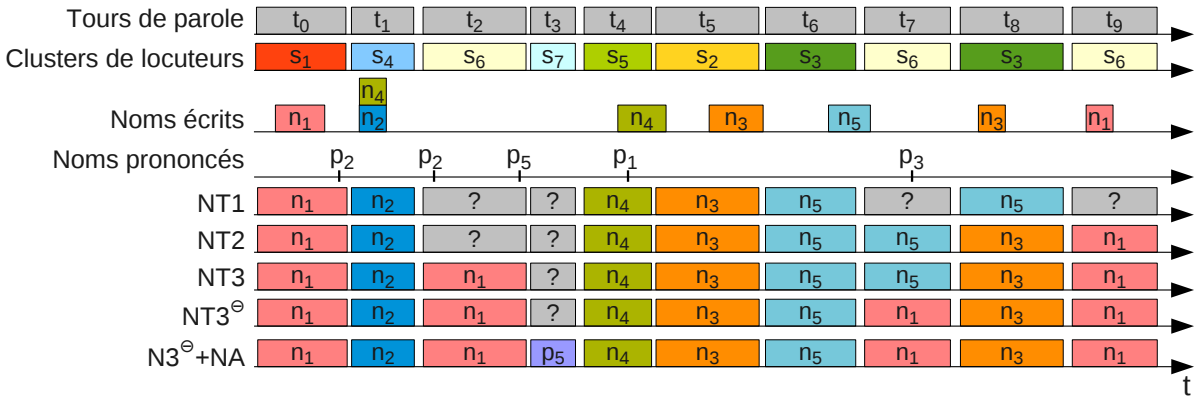


FIG. 5.14 – Résultats de nos méthodes de nommage tardif pour notre exemple

Le choix de la meilleure méthode est d'abord dépendant de la qualité du regroupement en locuteurs. Par ailleurs, il dépend aussi de la tâche visée : est-ce que l'on veut essayer de nommer le maximum de tours de parole en acceptant que certains soient mal identifiés ou l'importance de la précision est-elle primordiale (par exemple pour créer des modèles biométriques de manière non-supervisée) ?

5.1.1.6 Comparaison des résultats des nommages tardifs

Dans le tableau 5.1, on retrouve les résultats de toutes nos méthodes de nommage tardif (à partir d'une diarization BIC+CLR) par les noms écrits. Nous avons ajouté, sur les deux premières lignes, les résultats obtenus par les systèmes supervisés basés sur des modèles biométriques.

Méthode	$\%P$	$\%R$	$\%F$	%EGER
GMM-UBM	52.9	49.8	51.3	49.6
GSV-SVM	60.5	54.2	57.2	44.2
NT1	**90.7**	67.6	77.5	29.0
NT2	89.7	68.9	**78.0**	28.2
NT3	84.3	**70.0**	76.5	28.2
NT3$^{\ominus}$	88.2	69.2	77.6	**28.1**
NT1 pour les NA	23.9	10.2	14.3	82.2
NT3$^{\ominus}$+NA	83.2	71.1	76.7	27.0

TAB. 5.1 – Comparaison des résultats des méthodes de nommage tardif des locuteurs (diarization BIC+CLR), ensemble de test de la phase 1 du corpus *REPERE*.

La première information importante à noter est que toutes les solutions de nommage non-supervisé surpassent très largement les systèmes supervisés utilisant des modèles biométriques. Cela est dû principalement au manque de couverture des modèles par rapport aux personnes parlant dans l'ensemble de test (sur 111 locuteurs, seulement 65 ont un modèle malgré les 626 modèles construits).

La différence entre les méthodes NT1, NT2, NT3 et NT3$^{\ominus}$ en terme de F1-mesure et de EGER est assez minime. Cependant, on peut voir une différence de comportement sur le rappel et la précision.

Le nommage NT1 obtient la meilleure précision puisque l'on nomme seulement les clusters les plus co-occurrents avec les noms. On voit que l'identification directe (NT2) des tours de parole par les noms co-occurrents (lorsqu'un seul nom est écrit à l'écran) augmente le rappel de 67.6% à 68.9% mais engendre une réduction de 90.7% à 89.7% de la précision.

L'association NT3 augmente encore le rappel puisque l'on peut nommer plusieurs clusters de locuteurs à partir du même nom, mais on va avoir une réduction importante de la précision (de 89.7% à 84.3%). Le ré-alignement temporel entre les noms et les tours de parole (NT3$^{\ominus}$) permet d'éviter cette réduction mais ce ré-alignement n'est pas parfait et le rappel est aussi légèrement réduit.

« NT1 pour les NA » correspond aux résultats obtenus à partir d'un alignement 1-à-1 entre les noms prononcés identifiés comme allocutaires par le LIMSI et les clusters de locuteurs. Les résultats sont assez faibles, mais ils permettent de nommer quelques clusters supplémentaires encore inconnus après NT3$^{\ominus}$, ce qui augmente le rappel à 71.1% (NT3$^{\ominus}$+NA) mais décroit la précision de 88.1% à 83.2%.

5.1.2 Nommage intégré : (NI)

Dans la section précédente, nous avons proposé différentes méthodes pour nommer les clusters de locuteurs après diarization. Toutefois, une des limites de ces méthodes est que le seuil d'arrêt du regroupement est optimisé pour minimiser l'erreur de diarization (DER), alors que l'objectif final est la minimisation de l'erreur d'identification. En outre, l'optimisation de la DER ne conduit pas forcément à l'optimisation de l'identification.

En effet, minimiser la DER revient à trouver un juste équilibre entre pureté et couverture des clusters, ce qui conduit donc forcément à la fusion et à la non fusion erronée de clusters. Nous avons essayé de pallier ces deux types d'erreurs avec nos méthodes de nommage tardif, mais il est très difficile de redécouper un cluster a posteriori.

De plus, il arrive régulièrement que les personnes soient introduites plusieurs fois par leur nom écrit à l'écran. Il n'est donc pas utile de fusionner deux clusters correspondants à une seule personne s'ils sont tous deux nommés par deux occurrences du même nom. Par conséquent, le « nommage intégré » est une simple extension de la méthode de « nommage tardif » où le critère d'arrêt de la diarization est choisi pour minimiser l'erreur d'identification.

Nous montrerons dans la section dédiée aux résultats que le seuil optimal pour l'identification est généralement plus haut que celui qui minimise l'erreur de diarization (DER), c'est-à-dire que le regroupement est arrêté plus tôt. La sortie de diarization obtenue avec ce seuil contient donc un plus grand nombre de clusters mais ceux-ci sont plus purs. Pour l'association noms-clusters, nous avons sélectionné la méthode $\texttt{NT3}^{\ominus}$ décrite précédemment. En effet, il est plus intéressant de nommer par le même nom deux clusters plus purs qu'un seul moins pur.

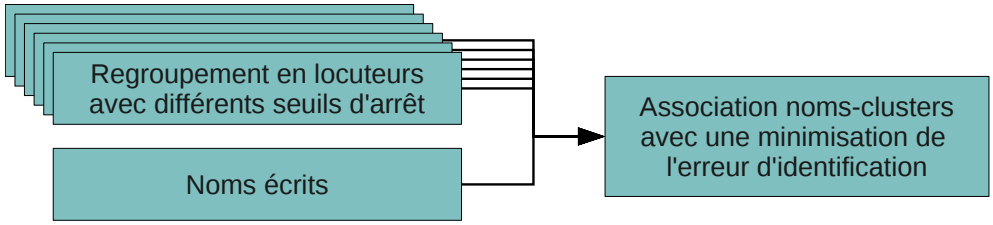


FIG. 5.15 – Nommage intégré

En pratique, comme montré dans la figure 5.15, nous gardons les sorties d'un regroupement pour de multiples seuils, auxquels nous appliquons le nommage tardif $\texttt{NT3}^{\ominus}$. Le seuil sélectionné est celui qui optimise l'identification (plus de détails seront donnés sur le corpus utilisé pour le réglage du seuil dans la section dédiée aux résultats).

Comme le clustering est arrêté plus tôt, les clusters sont plus petits et donc les noms seront propagés sur moins de tours de parole, ce qui va augmenter la précision. En revanche, c'est au détriment de certains tours de parole qui ne pourront être nommés et cela risque donc de faire baisser le rappel.

5.1.3 Nommage précoce (NP)

Pour augmenter le rappel tout en gardant une bonne précision, il faut donc pouvoir poursuivre le regroupement en évitant de fusionner des clusters nommés différemment. C'est pourquoi, l'intégration de l'information issue des noms pendant ce processus devient une évidence. Nous avons donc modifié le schéma d'association noms-clusters (voir figure 5.16) avec l'intégration des contraintes apportées par les noms pendant le processus de diarization.

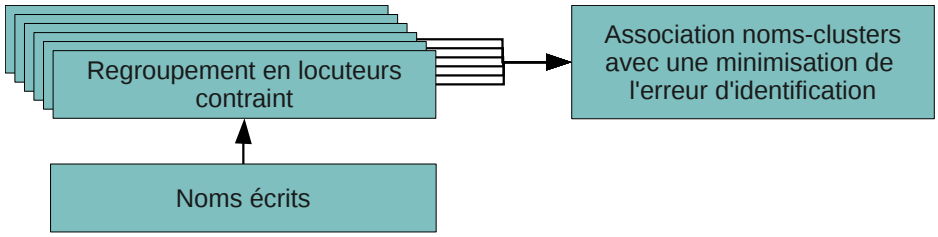


FIG. 5.16 – Nommage précoce

Nous avons profité du fait que lorsqu'un ou plusieurs noms sont écrits à l'écran, il y a une forte probabilité que le nom du locuteur courant corresponde à un des noms écrits à l'écran. Nous avons utilisé cette information pour à la fois nommer les clusters mais aussi contraindre le regroupement (empêcher la fusion de clusters nommés différemment).

Notre méthode de nommage précoce pour l'identification des locuteurs est découpée en quatre étapes :

- **Initialisation du regroupement** : avant d'effectuer le regroupement des tours de parole en clusters de locuteur, nous avons créé des liens entre les deux modalités.
- **Contraintes sur le regroupement** : au cours du regroupement hiérarchique basé sur une matrice de similarité des tours de parole, nous avons empêché certaines fusions pour éviter d'avoir des tours de parole d'un même cluster avec des noms différents.
- **Mise à jour après chaque fusion** : la fusion de deux clusters de parole peut changer les liens d'association entre les noms et les clusters. Il faut aussi recalculer les scores de similarité entre le nouveau cluster (créés par la fusion) et les autres clusters.
- **Association finale entre noms et clusters** : l'association finale va choisir la meilleure association noms-clusters.

Initialisation du regroupement

Nous définissons d'abord l'ensemble des noms $\mathcal{N}$ et des occurrences de noms $\mathcal{O}$:

$$\begin{array}{rcl} \mathcal{N} & = & \{a, b, \ldots, n\} \\ \mathcal{O} & = & \{o_i\} \end{array} \tag{5.6}$$

Ces deux ensembles sont reliés à l'aide de l'application $h\colon \mathcal{O} \rightarrow \mathcal{N}$, définie par :

$$h(o_i) \in \mathcal{N} \tag{5.7}$$

Nous définissons aussi l'ensemble des tours de parole $\mathcal{T}$:

$$\mathcal{T} = \{t_1, t_2, \ldots, t_M\} \tag{5.8}$$

Le regroupement va fusionner des tours de parole en cluster, donc nous définissons l'ensemble $\mathcal{G}$ des clusters de tours de parole. Un cluster correspondant à un sous-ensemble de $\mathcal{T}$. Comme avant le regroupement, il n'y a qu'un seul tour de parole par cluster, alors $\mathcal{G}$ correspond à l'ensemble des singletons de $\mathcal{T}$:

$$\mathcal{G} = \{\{t\}, t \in \mathcal{T}\} \tag{5.9}$$

Ensuite, nous allons créer des liens entre ces deux modalités avec la fonction $f : \mathcal{G} \rightarrow P(\mathcal{O})$ avec $P(\mathcal{O})$ l'ensemble des parties de $\mathcal{O}$, définie par

$$f(g) = \{o \in \mathcal{O} \mid o \text{ co-occurre avec } g\} \tag{5.10}$$

Ce qui nous permet de diviser l'ensemble $\mathcal{G}$ des clusters en deux sous-ensembles :

$$\begin{array}{rcl} \mathcal{K} & = & \{g \in \mathcal{G} \mid f(g_i) \neq \{\varnothing\}\} \\ \mathcal{U} & = & \mathcal{G} \backslash \mathcal{K} \end{array} \tag{5.11}$$

Il est important de préciser que, pour chaque élément de $\mathcal{O}$, l'étiquette porte sur le segment de parole le plus co-occurrent avec le nom détecté. Donc chaque élément de $\mathcal{O}$ correspond à un seul cluster alors qu'un cluster peut correspondre à plusieurs éléments de $\mathcal{O}$.

Maintenant que des liens ont été créés entre les deux modalités, nous pouvons effectuer le regroupement hiérarchique des éléments de l'ensemble $\mathcal{G}$ à partir d'une matrice de similarité entre les tours de parole.

Le but de ce regroupement est de trouver les classes d'équivalence qui minimisent l'erreur d'identification, mais aussi de réduire l'ensemble d'arrivée d'un cluster de parole g dans la fonction f à des occurrences du même nom :

$$card(\{h(o) \mid o \in f(g)\}) = 1 \tag{5.12}$$

Contraintes sur le regroupement

Nous avons utilisé les liens entre les clusters de parole et les occurrences de noms pour contraindre ce regroupement. Ainsi, deux clusters g_1 et g_2 de $\mathcal{K}$ (donc des clusters déjà nommés) ne pourront pas fusionner si :

$$\nexists (o_1 \in f(g_1), o_2 \in f(g_2)) \mid h(o_1) = h(o_2) \tag{5.13}$$

C'est-à-dire s'ils n'ont pas un nom en commun dans l'ensemble des noms avec lesquels ils co-occurrent.

Mise à jour après chaque fusion

A chaque itération du regroupement, la fusion de deux clusters g_1 et g_2 en un cluster g_{12} va modifier la fonction qui relie $\mathcal{G}$ à $\mathcal{O}$. Trois cas de figure se présentent quant à la fonction f :

- Les deux clusters appartiennent à $\mathcal{K}$, alors :

$$f(g_{12}) = \{o_1 \in f(g_1), o_2 \in f(g_2) \mid h(o_1) = h(o_2)\} \tag{5.14}$$

- Seulement le cluster g_1 (respectivement g_2) appartient à $\mathcal{K}$ alors

$$f(g_{12}) = f(g_1) \text{ (respectivement } f(g_{12}) = f(g_2)) \tag{5.15}$$

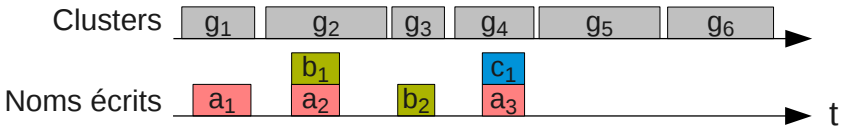
- Aucun cluster n'appartient à $\mathcal{K}$, alors la fonction f reste inchangée.

Après chaque fusion, il faut recalculer le score de similarité entre le nouveau cluster g_{12} et tous les autres cluster g de $\mathcal{G}$. Ce nouveau score correspond à la moyenne des scores de similarité entre les éléments de chaque cluster :

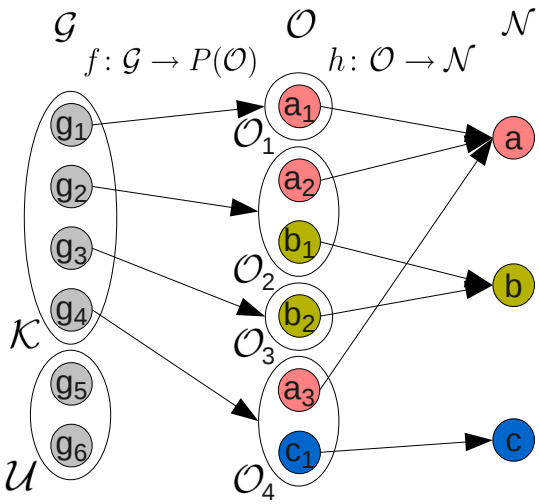
$$score(g_{12}, g) = \frac{\sum_{t_1 \in g_{12}, t_2 \in g} score(t_1, t_2)}{card(g_{12}) * card(g)} \tag{5.16}$$

Exemple pour les contraintes et la mise à jour des ensembles

Prenons un exemple avec $\mathcal{K} = \{g_1, g_2, g_3, g_4\}$ et $\mathcal{U} = \{g_5, g_6\}$. 3 noms sont affichés $\mathcal{N} = \{a, b, c\}$ avec $a = \{a_1, a_2, a_3\}$, $b = \{b_1, b_2\}$ et $c = \{c_1\}$.



Une autre représentation est donnée ci-dessous avec les deux fonctions f et h :



Les co-occurrences nous permettent de définir que :

$$f(g_1) = \{a_1\} \qquad f(g_2) = \{a_2, b_1\} \qquad f(g_3) = \{b_2\} \qquad f(g_4) = \{a_3, c_1\}$$

Des exemples de fusion des classes suivantes donnent comme résultat :

$\cup$ classes	$f : \mathcal{G} \rightarrow P(\mathcal{O})$	Ensembles $\mathcal{K}$ et $\mathcal{U}$
$g_5 \cup g_6 \rightarrow g_{56}$		$\mathcal{K} = \{g_1, g_2, g_3, g_4\}$ et $\mathcal{U} = \{g_{56}\}$
$g_1 \cup g_6 \rightarrow g_{16}$	$f(g_{16}) = \{a_1\}$	$\mathcal{K} = \{g_{16}, g_2, g_3, g_4\}$ et $\mathcal{U} = \{g_5\}$
$g_2 \cup g_6 \rightarrow g_{26}$	$f(g_{26}) = \{a_2, b_1\}$	$\mathcal{K} = \{g_1, g_{26}, g_3, g_4\}$ et $\mathcal{U} = \{g_5\}$
$g_1 \cup g_2 \rightarrow g_{12}$	$f(g_{12}) = \{a_1, a_2\}$	$\mathcal{K} = \{g_{12}, g_3, g_4\}$ et $\mathcal{U} = \{g_5, g_6\}$
$g_1 \cup g_3$ $g_3 \cup g_4$	Fusion interdite	

Association finale entre noms et clusters

Lorsque que le critère d'arrêt est atteint, pour chacun des g de $\mathcal{K}$ qui n'ont qu'un seul nom associé ($card(\{h(o) \mid o \in f(g)\}) = 1$), on nomme directement g par le nom. Pour les autres clusters appartenant à $\mathcal{K}$, on sélectionne le nom qui a le meilleur score TF.IDF du cluster (voir section 5.1.1.3).

Dans les faits, sur le corpus *REPERE*, seule l'émission « Pile et face » utilise régulièrement l'affichage de deux noms simultanément, mais ces noms peuvent être affichés seuls à un autre moment de la vidéo. Donc, dans la majorité des cas, le regroupement va produire des clusters associés à un seul nom.

5.1.4 Comparaison des nommages tardifs (NT), intégrés (NI) et précoces (NP)

Pour comparer les trois méthodes d'intégration des noms écrits avec la même qualité de diarization, nous utilisons un regroupement en locuteurs sur une matrice de distance BIC, non mise à jour après chaque fusion. Les performances du nommage tardif sont donc inférieures à celles présentées précédemment.

Apprentissage du critère d'arrêt

Ces trois méthodes ont besoin d'un critère d'arrêt du regroupement. Nous avons donc utilisé l'ensemble d'entraînement de la phase 1 du défi *REPERE* pour apprendre ce seuil. Toutefois, dans l'idée d'être moins dépendants d'une grande quantité d'annotations manuelles, nous n'avons pas utilisé la totalité des 24 heures de cet ensemble, mais seulement un sous ensemble de trois heures.

Dans le tableau 5.2 on peut voir la variation du seuil appris sur la distance BIC pour les trois méthodes, à partir de 100 sous-ensembles de trois heures sélectionnés aléatoirement. Ces sous-ensembles ont été choisis pour correspondre à l'ensemble de test (même durée, même équilibre dans la proportion des émissions, même nombre de vidéos pour chaque émission).

Méthode de nommage	Médiane	Min	Max	Écart type
NT3$^{\ominus}$: plus basse DER	1540	1440	1680	54
NI : plus basse EGER	1620	1520	1740	44
NP : plus basse EGER	1260	300	1640	277

TAB. 5.2 – Seuil sur le regroupement appris sur 100 sous-ensembles, pour minimiser la DER ou l'EGER.

Ce tableau nous montre deux stratégies différentes par rapport à NT3$^{\ominus}$:

- NI arrête plus tôt son regroupement hiérarchique. Donc pour chaque locuteur, NI produit plusieurs clusters mais ils sont nommés par les différentes occurrences de nom correspondant au locuteur.
- La contrainte de NP pendant le regroupement permet d'éviter de mauvaises fusions et donc d'arrêter ce regroupement plus tard.

L'écart type pour NP est plus élevé, en comparaison avec les deux autres méthodes. Il est possible de l'interpréter par une sensibilité inférieure de NP au choix du seuil.

Pour le reste de la section, nous avons utilisé, pour chaque méthode, le seuil médian appris sur l'ensemble d'entraînement. Les résultats sont calculés sur l'ensemble de test.

Comparaison pour l'identification

La figure 5.17 montre l'évolution du taux d'EGER par rapport au seuil sélectionné. Un seuil plus petit signifie que le regroupement hiérarchique est arrêté plus tard, cette figure peut donc être lue de droite à gauche.

Les courbes pour NT3$^{\ominus}$ et NI se superposent car les méthodes ne diffèrent que dans le choix du seuil : ⓐ correspond à la minimisation de la DER (nommage tardif) alors que ⓑ correspond à la minimisation de l'EGER (nommage intégré).

A contrario, le nommage précoce suit une courbe très différente. ① montre l'impact de la contrainte ajoutée par les noms écrits (interdiction de fusionner des clusters avec des noms différents). ⓒ correspond au seuil appris qui minimise l'EGER pour cette méthode.

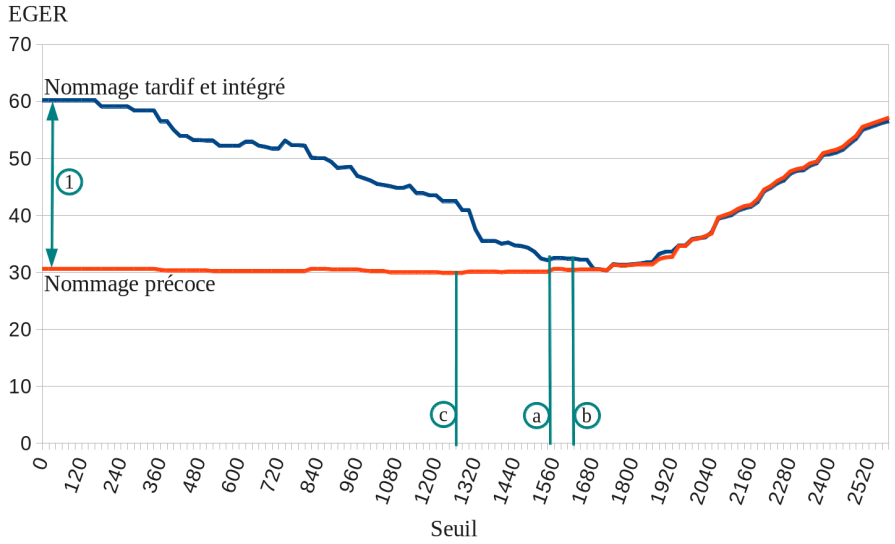


FIG. 5.17 – Influence du critère d'arrêt (ⓐ, ⓑ, ⓒ appris sur l'ensemble d'apprentissage) sur l'erreur d'identification de l'ensemble de test, pour les trois stratégies de nommage des locuteurs, ensemble de test de la phase 1 du corpus *REPERE*.

Le tableau 5.3 résume les performances des trois méthodes (basé sur une matrice BIC ce qui explique les résultats inférieurs à ceux présentés précédemment). Le nommage intégré a une EGER un peu plus haute, mais la différence est vraiment très faible. Cette méthode a une meilleure précision en raison de son seuil plus élevé.

En ce qui concerne NP, la contrainte sur le regroupement permet de garder la même précision (80,4%) avec un seuil beaucoup plus bas. La méthode permet de fusionner correctement des clusters supplémentaires et augmente donc le rappel à 68,3%. Pour NI et NP, il est possible d'optimiser d'autres métriques, par exemple, on peut imaginer définir une précision et une durée de parole suffisante pour construire des modèles de locuteurs.

Méthode	Seuil	%P	%R	%F	%EGER
NT3$^{\ominus}$	ⓐ 1540	80.4	66.0	72.5	32.1
NI	ⓑ 1620	81.5	65.3	72.5	32.4
NP	ⓒ 1260	80.4	68.3	73.9	29.9

TAB. 5.3 – EGER, précision et rappel des 3 méthodes, ensemble de test de la phase 1 du corpus *REPERE*. Seuil appris sur l'ensemble d'entraînement (valeur médiane à partir de 100 sous-ensembles de trois heures).

Comparaison pour la qualité du regroupement

La figure 5.18 montre l'évolution de la DER en fonction du seuil. La ligne bleue « avant nommage » correspond à un regroupement en locuteurs classique (sans les noms écrits). La DER est différente après le nommage tardif et intégré (ligne rouge) à cause des modifications apportées par ces nommages. La ligne jaune correspond au tracé de la DER après le nommage précoce.

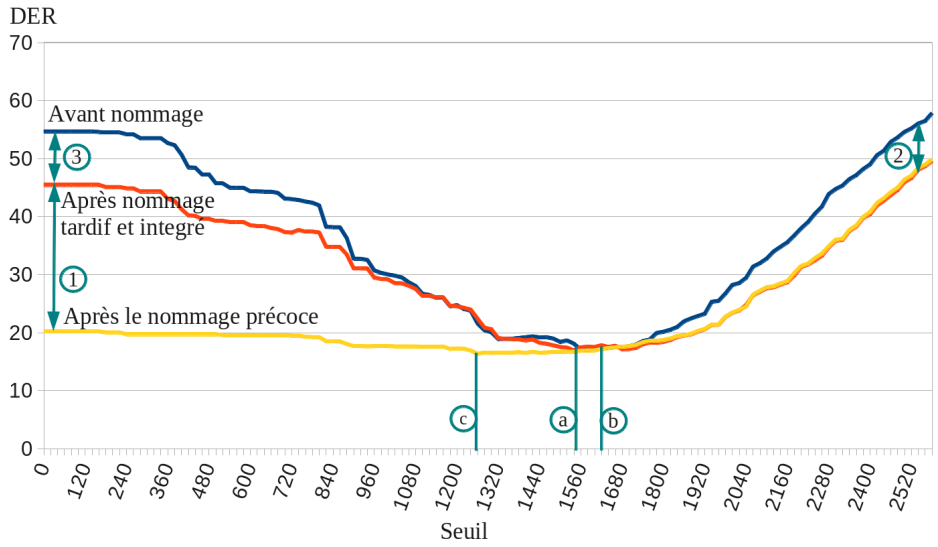


FIG. 5.18 – Influence du critère d'arrêt sur le taux d'erreur de diarization sur le jeu de test, avant et après nommage, ensemble de test, phase 1 du corpus *REPERE*.

② et ③ montrent l'influence de l'étape de nommage direct des tours de parole par les noms écrits. Au début du regroupement ②, les tours de parole qui ont le même nom sont fusionnés. A la fin du regroupement ③, des tours de parole sont nommés différemment même s'ils appartenaient à un même cluster. ① montre l'effet des contraintes sur le nommage précoce qui empêchent les clusters avec des noms différents de fusionner.

	Seuil	DER
Diarization sans les noms écrits	ⓐ 1540	18.11
Après nommage tardif et intégré	ⓑ 1620	17.51
Après nommage précoce	ⓒ 1260	16.37

TAB. 5.4 – DER en fonction du seuil d'arrêt du regroupement agglomératif, ensemble de test de la phase 1 du corpus *REPERE*.

ⓐ correspond au seuil appris pour minimiser la DER. Nous obtenons une DER de 18,11% pour la diarization sans les noms écrits. (voir le tableau 5.4). Le nommage intégré a une DER très légèrement inférieure (17,5%), malgré son seuil plus élevé. Ce seuil produit plus de clusters mais ceux ci sont plus purs. Certains d'entre eux peuvent être fusionnés grâce à leurs noms associés identiques. Le nommage précoce montre une très légère variation de la DER (de 18,7% à 20,2%, avec un minimum de 16,37%) sur la plage de seuil [0-1800] : il apparait donc beaucoup moins sensible au choix du seuil (voir figure 5.18).

Sensibilité à l'ensemble de réglage utilisé

Pour comparer la sensibilité des trois méthodes au changement d'ensemble de réglage pour le choix du seuil, nous avons aléatoirement choisi 100 sous-ensembles différents et sélectionné pour chacun d'eux le meilleur seuil à appliquer sur l'ensemble de test. Nous obtenons donc une plage de seuil pour chacune des méthodes.

L'axe abscisse de la figure 5.19 résume la plage de variation du seuil optimal des 100 sous-ensembles d'apprentissage (par exemple 1440-1680 pour la stratégie de nommage tardif). L'axe des ordonnées indique le taux d'EGER moyen correspondant et son écart-type sur l'ensemble de test.

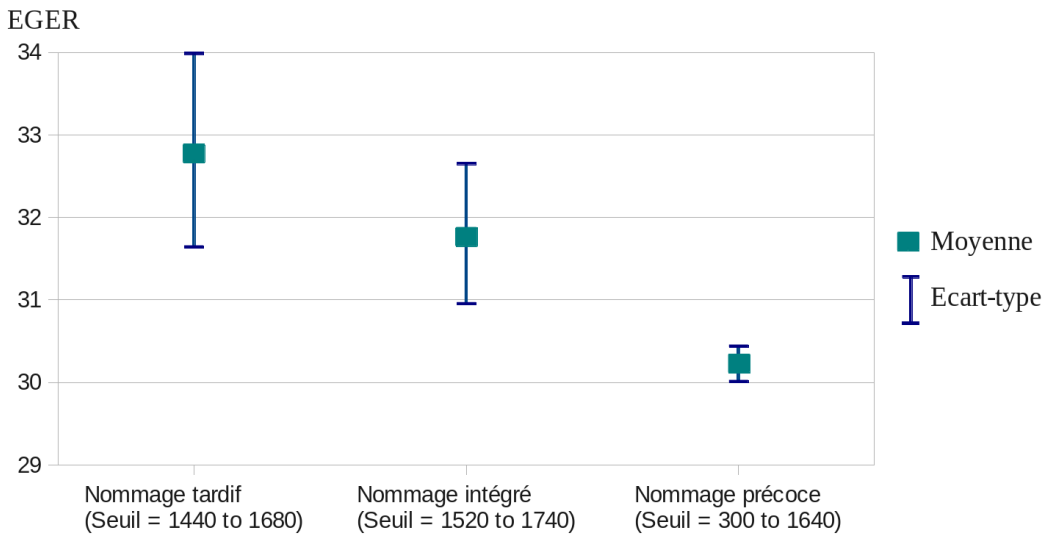


FIG. 5.19 – Moyenne et écart type de l'EGER sur le jeu de test en fonction des sous-ensembles utilisés pour apprendre le seuil d'arrêt du regroupement, phase 1 du corpus *REPERE*

Cette figure souligne que les stratégies de nommage tardif et intégré sont plus dépendantes de l'ensemble d'apprentissage. Leurs taux d'erreurs d'identification (EGER) respectifs présentent un écart type de 1,2% et 0,8%, tandis que l'écart type de l'EGER du nommage précoce n'est que de 0,2%. D'autant que la plage de variabilité des seuils optimaux sur les 100 sous-ensembles est beaucoup plus grande.

Dépendance du seuil à l'émission

Le corpus de test est composé de sept émissions différentes. Même si un seuil global peut être appris, nous étudions également l'utilisation d'un seuil dépendant de l'émission et présentons les résultats de cette expérience sur la figure 5.20. Le seuil d'un oracle correspond à la meilleure performance possible dans le cas où un oracle est capable de prédire le meilleur seuil sur l'ensemble de test pour chaque vidéo.

La robustesse d'une stratégie de nommage peut être déduite de la différence entre les seuils appris sur l'ensemble d'apprentissage (*seuil global* et *seuil par émission*) et le seuil optimal (*seuil d'un oracle*).

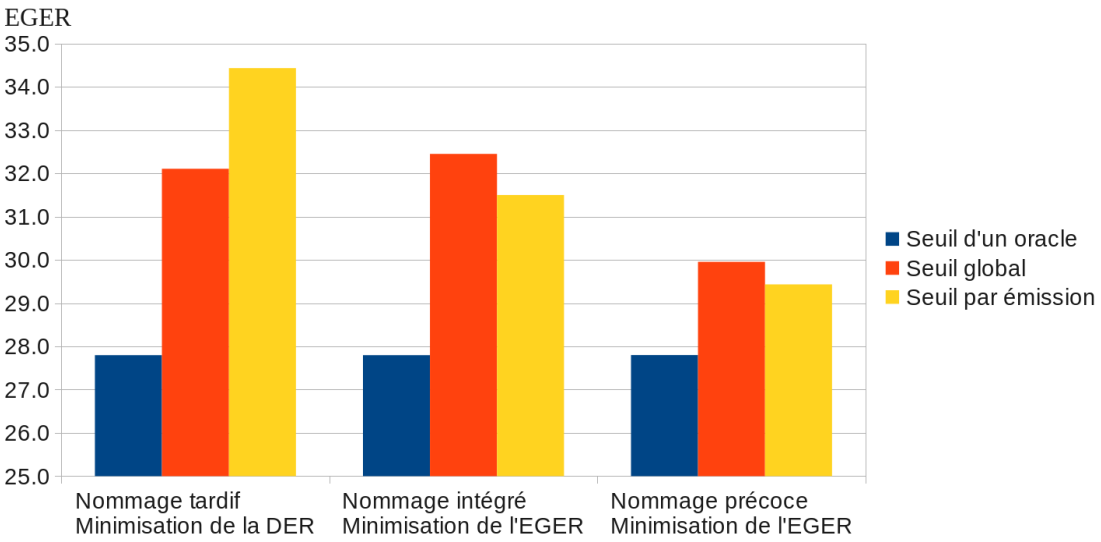


FIG. 5.20 – Taux d'erreur d'identification (EGER) pour un seuil oracle global ou dépendant de l'émission, ensemble de test de la phase 1 du corpus *REPERE*.

La figure 5.20 montre qu'il y a une différence de comportement entre la minimisation DER (nommage tardif) et la minimisation de l'EGER (nommage intégré ou nommage précoce). La minimisation de la DER vise à associer un cluster spécifique à chaque orateur, qu'il soit nommable ou non. Alors que, la minimisation de l'EGER essaie d'associer un nom à chaque locuteur. Les locuteurs anonymes après le processus de nommage peuvent être dans le même cluster ou répartis en plusieurs groupes. Ils n'ont aucune influence sur la valeur finale du taux d'erreur d'identification (EGER).

Parmi les émissions du corpus *REPERE*, certaines contiennent de nombreux intervenants (jusqu'à 18 pour les journaux télévisés de *BFMStory*) dont les noms ne sont généralement affichés qu'une seule fois. D'autres, comme le débat *Pile et Face*, n'ont que trois locuteurs (deux invités et le présentateur), les invités ont leurs noms affichés 24 fois en moyenne sur toute la durée de chaque émission. Pour ce type de d'émission, le seuil optimal de DER est de 1300 alors que celui de l'EGER est 1560. Puisque les noms des locuteurs sont écrits plusieurs fois, il est inutile d'essayer d'obtenir exactement un cluster par locuteur. En outre, un locuteur peut être divisé en plusieurs clusters plus petits du moment qu'ils sont nommés correctement.

Enfin, nous voulons souligner que les performances de l'oracle sont presque identiques pour les trois stratégies (27,796% pour `NT` et `NI`, 27,799% pour `NP`). Cependant, puisque le nommage précoce est moins sensible au choix du seuil, il conduit à de bien meilleures performances d'identification, très proches de l'oracle qui choisit le seuil optimal pour chaque vidéo sur l'ensemble de test.

5.2 Adaptation du nommage précoce pour identifier les visages

Nous avons adapté la méthode de nommage précoce à l'identification des visages. Pour les locuteurs, nous avons considéré que, lorsqu'un seul nom était écrit à l'écran, il identifiait avec certitude le locuteur courant et que lorsque deux noms apparaissaient en même temps le locuteur était nommé par l'un des deux noms.

Pour les visages, c'est un peu différent puisque plusieurs visages peuvent apparaître en même temps. L'association entre les noms et les visages n'est plus aussi déterministe. Prenons comme exemple les images de la figure 5.21.

(a) (b)

(c) (d)

FIG. 5.21 – Exemple d'images du corpus *REPERE*

Lorsqu'un seul nom et un seul visage apparaissent, comme dans l'image 5.21.a, l'association est évidente. Si deux noms et deux visages apparaissent comme dans l'image 5.21.b, un nom correspond a priori à l'un des deux visages et l'autre nom à l'autre visage, sans pour autant savoir quelles sont les bonnes associations.

Par contre, si plus de visages que de noms apparaissent (comme dans l'image 5.21.c), chacun des noms correspond a priori à l'un des visages mais on ne sait pas lequel, et les autres visages restant correspondent forcement à des noms non affichés.

La figure 5.21.d montre que la proximité spatiale, dans l'image, d'un nom et d'un visage n'est pas un critère toujours discriminant puisque le nom écrit dans le cartouche correspond au visage de droite qui est plus loin que le visage de gauche.

Ces quatre exemples nous montrent que, contrairement au nommage précoce pour les locuteurs, quel que soit le cas de figure, toutes les occurrences de noms doivent être utilisées pour nommer une séquence de visages (une séquence de visages correspond aux images de visage d'une même personne qui apparaissent sur des trames successives. Pour plus de détails voir le chapitre 2).

Nous allons décrire, dans la suite de cette section, la manière dont nous avons adapté la méthode de nommage précoce aux particularités de l'identification des visages.

Comme pour la section précédente, notre méthode se décompose en quatre étapes :

- Initialisation du regroupement
- Contraintes sur le regroupement
- Mise à jour après chaque fusion
- Association finale entre noms et clusters

Initialisation du regroupement

Comme pour les locuteurs, nous définissons d'abord l'ensemble des noms $\mathcal{N}$ et des occurrences de noms $\mathcal{O}$:

$$\begin{array}{rcl} \mathcal{N} & = & \{a, b, \ldots, n\} \\ \mathcal{O} & = & \{o_i\} \end{array} \tag{5.17}$$

Ces deux ensembles sont reliés à l'aide de l'application $h\colon \mathcal{O} \rightarrow \mathcal{N}$, définie par :

$$h(o_i) \in \mathcal{N} \tag{5.18}$$

Nous définissons aussi l'ensemble des séquences de visages :

$$\mathcal{V} = \{v_1, v_2, \ldots, v_N\} \tag{5.19}$$

Le regroupement va fusionner des séquences de visages en clusters, donc nous définissons l'ensemble $\mathcal{G}$ des clusters de séquences de visages. Un cluster correspond à un sous-ensemble de $\mathcal{V}$. Comme avant le regroupement, il n'y a qu'une seule séquence de visages par cluster. Donc $\mathcal{G}$ correspond à l'ensemble des singletons de $\mathcal{V}$:

$$\mathcal{G} = \{\{v\}, v \in \mathcal{V}\} \tag{5.20}$$

Ensuite, nous allons créer des liens entre ces deux modalités avec la fonction $f\colon \mathcal{G} \rightarrow P(\mathcal{O})$ avec $P(\mathcal{O})$ l'ensemble des parties de $\mathcal{O}$, définie par

$$f(g) = \{o \in \mathcal{O} \mid o \text{ co-occurre avec } g\} \tag{5.21}$$

Ce qui nous permet de diviser l'ensemble $\mathcal{G}$ des clusters en deux sous-ensembles :

$$\begin{array}{rcl} \mathcal{K} & = & \{g \in \mathcal{G} \mid f(g_i) \neq \{\varnothing\}\} \\ \mathcal{U} & = & \mathcal{G} \backslash \mathcal{K} \end{array} \tag{5.22}$$

Pour l'instant, il n'y a pas de différence entre le nommage précoce pour les visages et le nommage précoce pour les locuteurs. Par contre, à partir de maintenant, nous allons utiliser deux particularités : la première est que plusieurs séquences de visages peuvent apparaître en même temps et la seconde est que nous voulons utiliser toutes les occurrences de noms apparaissant dans une vidéo.

Nous allons diviser l'ensemble $\mathcal{K}$ en deux sous-ensembles :

- $\mathcal{K}_1$: ensemble des clusters dont on est sûr d'avoir le nom. Pour être sûr d'avoir le nom d'un cluster g, il faut qu'il y ait autant de noms que de clusters co-occurrants (comme par exemple dans la figure 5.21.a (1 nom - 1 visage) ou dans la figure 5.21.b (2 noms - 2 visages)).

$$\mathcal{K}_1 = \{g \in \mathcal{K} \mid card(\mathcal{X}_g) \geq card(\mathcal{Y}_g)\} \tag{5.23}$$

  Avec :

$$\begin{array}{rcl} \mathcal{X}_g & = & \{o \in \mathcal{O} \mid o \text{ co-occurre avec } g\} \\ \mathcal{Y}_g & = & \{g \in \mathcal{K} \mid g \text{ co-occurre avec un élément } \mathcal{X}_g\} \end{array} \tag{5.24}$$

- $\mathcal{K}_0$: ensemble des clusters qui co-occurrent avec un nom et qui n'appartiennent pas à $\mathcal{K}_1$, c'est à dire ceux pour lesquels $card(\mathcal{X}_g) < card(\mathcal{Y}_g)$ (comme dans les figures 5.21.c et 5.21.d). Pour ceux-ci, nous ne pouvons pas dire quelles séquences de visages correspondent aux noms affichés et quelles autres correspondent à un autre nom.

$$\mathcal{K}_0 = \mathcal{K} \backslash \mathcal{K}_1 \tag{5.25}$$

Maintenant que des liens ont été créés entre les deux modalités, nous pouvons effectuer le regroupement hiérarchique des éléments de l'ensemble $\mathcal{G}$ à partir d'une matrice de distance entre les séquences de visages.

Comme pour `NP` pour les locuteurs, le but de ce regroupement est de trouver les classes d'équivalence qui minimisent l'erreur d'identification, mais aussi de réduire l'ensemble d'arrivée d'un cluster g dans la fonction f à des occurrences du même nom :

$$card(\{h(o) \mid o \in f(g)\}) = 1 \tag{5.26}$$

Contraintes sur le regroupement

La première contrainte que l'on peut imposer à ce regroupement est la même que pour les locuteurs. Nous avons utilisé les liens entre les clusters de visages et les occurrences de noms pour contraindre ce regroupement : deux clusters g_1 et g_2 de $\mathcal{K}_1$ ne pourront pas fusionner si :

$$\nexists (o_1 \in f(g_1), o_2 \in f(g_2)) \mid h(o_1) = h(o_2) \tag{5.27}$$

C'est-à-dire qu'ils ne pourront pas fusionner s'ils n'ont pas un nom en commun dans l'ensemble des noms avec lesquels ils co-occurrent. Cette contrainte est la même que celle que nous avons utilisé pour le nommage précoce des locuteurs.

Les deux contraintes suivantes utilisent les particularités liées à la modalité visage citées ci-dessus.

La seconde contrainte n'utilise pas les noms mais les co-occurrences entre les visages. Elle interdit la fusion de deux clusters dont des séquences de visages co-occurrent.

Nous posons d'abord une fonction surjective p: $\mathcal{G} \rightarrow P(\mathcal{G})$ ($P(\mathcal{G})$ correspond à l'ensemble des parties de $\mathcal{G}$). Cette fonction va d'un cluster g_1 vers les autres clusters qui ont une séquence de visages qui co-occurre avec une des séquences de visages de g_1. Elle est définie par :

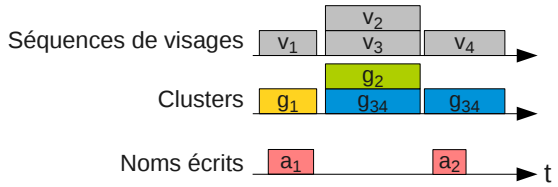
$$p(g_1) = \{g_2 \in \mathcal{G} \mid \exists (v_1 \in g_1, v_2 \in g_2) \text{ avec } v_1 \text{ et } v_2 \text{ co-occurrent}\} \tag{5.28}$$

La contrainte revient donc à empêcher le regroupement de deux clusters g_1 et g_2 si $g_2 \in p(g_1)$. En d'autres termes, on ne regroupe pas deux clusters qui contiennent des séquences de visages apparaissant au même moment car les visages sont forcement différents.

La troisième contrainte découle de la seconde, elle empêche deux clusters g_1 et g_2 de fusionner si g_1 est nommé avec certitude et que g_2 co-occurre avec un cluster g_3 nommé lui aussi avec certitude par le même nom que g_1.
Deux clusters g_1 et g_2 ne pourront pas fusionner si $\exists g_3 \in p(g_2)$ tel que :

$$\begin{array}{rrcl} & card(\{h(o) \mid o \in f(g_1)\}) & = & 1 \\ \text{et} & card(\{h(o) \mid o \in f(g_3)\}) & = & 1 \\ \text{et} & \{h(o) \mid o \in f(g_1)\} & = & \{h(o) \mid o \in f(g_3)\} \end{array} \tag{5.29}$$

Pour comprendre cette dernière contrainte, prenons l'exemple ci-dessous :



La séquence de visages v_1 est identifiée directement par a_1. De même, v_4 est associée à a_2. Une première fusion a déjà eu lieu entre les séquences v_3 et v_4, ce qui a produit le cluster g_{34}. Nous avons donc les clusters g_1 et g_{34} identifiés par le nom a. Cette contrainte va empêcher la fusion de g_1 et g_2 puisque g_2 a un cluster co-occurrant (g_{34}) portant le même nom que g_1.

Mise à jour après chaque fusion

A chaque itération du regroupement, la fusion de deux clusters g_1 et g_2 en un cluster g_{12} va modifier la fonction qui relie $\mathcal{G}$ à $\mathcal{O}$. Quatre cas de figure se présentent quant à la fonction f :

- Les deux clusters appartiennent à $\mathcal{K}_1$, alors :

$$f(g_{12}) = \{o_1 \in f(g_1), o_2 \in f(g_2) \mid h(o_1) = h(o_2)\} \tag{5.30}$$

- Les deux clusters appartiennent à $\mathcal{K}_0$, alors :

$$f(g_{12}) = f(g_1) \cup f(g_2) \tag{5.31}$$

- Seulement le cluster $g_1 \in \mathcal{K}_1$ ou bien $g_1 \in \mathcal{K}_0$ et $g_2 \in \mathcal{U}$, alors :

$$f(g_{12}) = f(g_1) \text{ (respectivement } f(g_{12}) = f(g_2)) \tag{5.32}$$

- Aucun n'appartient à $\mathcal{K}$, alors la fonction f reste inchangée.

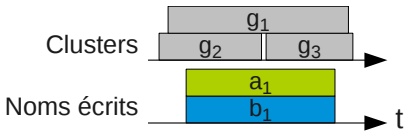
A l'issue de la fusion, si le nouveau cluster g_{12} est associé à un seul nom n ($\{h(o) \mid o \in f(g_{12})\} = \{n\}$) avec certitude ($g_{12} \in \mathcal{K}_1$), les clusters qui co-occurrent avec g_{12} ne peuvent pas être nommés par le nom n. Donc, pour chacun des clusters $g_3 \in \mathcal{K}$ dans $p(g_{12})$:

$$f(g_3) = f(g'_3) \backslash \{o \in f(g'_3) \mid h(o) = n\} \tag{5.33}$$

Avec $f(g'_3)$ égale à l'ensemble des occurrences de noms associées à g_3 avant la fusion.

Une deuxième vérification doit être effectuée à l'issue de chaque fusion : chaque cluster $g \in \mathcal{K}$ dont on est sûr d'avoir le nom appartient à $\mathcal{K}_1$. Pour être sûr d'avoir le nom d'un cluster g, il faut qu'il co-occurre avec x occurrences de noms et que ces occurrences de noms co-occurrent avec autant ou moins de clusters.

Si on prend l'exemple de la chronologie dans la figure ci-dessous, la fusion de g_2 et g_3 est autorisée puisqu'ils ne co-occurrent pas. A l'issue de cette fusion, il n'y a plus que deux clusters pour deux occurrences de noms, donc les deux clusters doivent faire partie de $\mathcal{K}_1$.



Après chaque fusion, comme pour les locuteurs, il faut recalculer le score de similarité entre le nouveau cluster g_{12} et tous les autres clusters g de $\mathcal{G}$. Ce nouveau score correspond à la moyenne des scores de similarité entre les éléments de chaque cluster :

$$score(g_{12}, g) = \frac{\sum\limits_{v_1 \in g_{12}, v_2 \in g} score(v_1, v_2)}{card(g_{12}) * card(g)} \tag{5.34}$$

Association finale entre noms et clusters

Une fois le critère d'arrêt atteint, les clusters de $\mathcal{K}_1$ associés à un seul nom ($card(\bigcup_{o \in f(g)} h(o)) = 1$) sont nommés par ce nom. Pour les autres, on va nommer en priorité les associations noms-clusters avec la plus grande durée de co-occurrence en respectant les contraintes décrites ci-dessus, et ce jusqu'à ne plus avoir, si possible, d'occurrences de noms non utilisées.

Résultats

Une séquence de visages est identifiable seulement si on peut calculer un descripteur pour elle. Le système de détection des visages de KIT a détecté 27898 séquences de visages sur l'ensemble de test de la phase 1 du défi *REPERE*. Toutefois, nous n'avons pu utiliser que 9050 d'entre elles dans le nommage précoce. Pour les autres, il n'y avait pas une confiance suffisante dans le maillage appliqué sur les visages pour pouvoir extraire un descripteur.

Pour avoir une idée précise du résultat maximal que peut atteindre notre système, nous avons reporté sur la première ligne du tableau 5.5 les scores obtenus par un oracle qui identifie correctement toutes les séquences de visages qui ont un descripteur. On voit que seulement 60% des séquences de visages pourront être nommées et que quelques faux positifs (correspondant à des fausses détections de visages) réduiront la précision de l'identification si nous essayons de toutes les nommer.

La seconde ligne rappelle les scores d'un système supervisé utilisant des modèles de visages construits sur l'ensemble d'apprentissage de la phase 1 du défi *REPERE*.

Méthode	$\%P$	$\%R$	$\%F$	%EGER
Oracle	97.2%	60.0%	74.2%	31.5%
Supervisé : KNN	59.9%	32.2%	41.9%	68.4%
Non-supervisée : `NP` seuil global	85.1%	42%	56.3%	49.8%
Non-supervisée : `NP` seuil par émission	89.4%	44.7%	59.5%	47.6%

TAB. 5.5 – Comparaison de l'identification des 1449 visages de l'ensemble de test de la phase 1 du corpus *REPERE* pour la méthode de nommage précoce par rapport à un oracle et un système utilisant des modèles biométriques.

Pour le nommage précoce (`NP`), nous avons appris sur l'ensemble d'entraînement deux types de seuils pour arrêter le regroupement. Un seuil global où toutes les vidéos ont été utilisées et un seuil par émission.

Les 557 modèles biométriques appris sur les 24 heures annotées de l'ensemble d'entraînement ne permettent d'obtenir qu'un rappel de 32.2% sur les 60% possible. Alors qu'une solution non supervisée obtient 42% et 44.7% de rappel selon le type de seuil fixé, tout en gardant une bonne précision (85.1% et 89.4%).

La figure 5.22 montre la distribution des distances entre les séquences de visages. En bleu, si les séquences appartiennent à la même personne, en rouge à des personnes différentes, sur l'ensemble d'apprentissage de la phase 1 du corpus *REPERE*.

Entre les valeurs de distance égale à 11-12, les deux distributions se croisent et au dessus de 30, il n'y plus que 2944 comparaisons positives sur un total de 116252 positives. Pour ce même seuil, près de la moitié des comparaisons négatives ont une distance inférieure à 30 (300520 sur 632492).

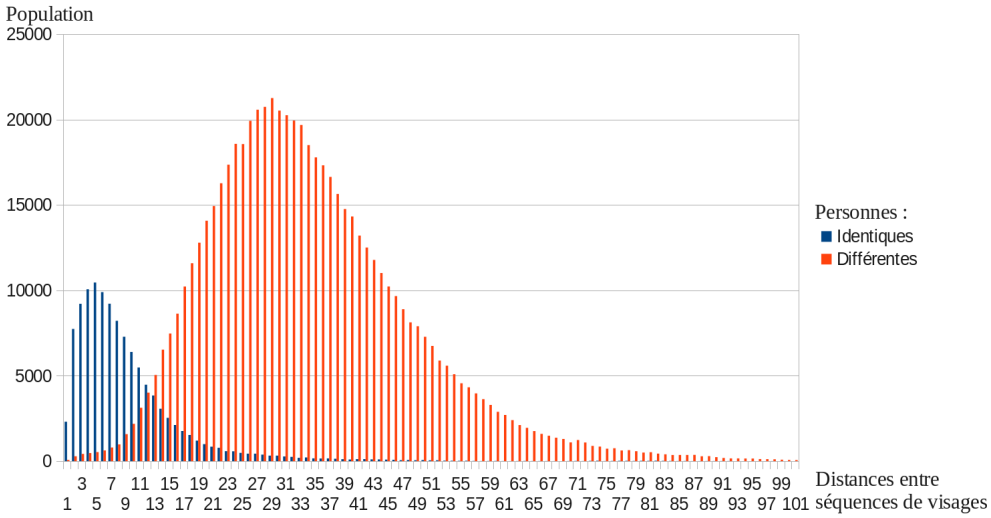


FIG. 5.22 – Distribution des distances entre séquences de visages, en bleu les séquences de la même personne, en rouge pour des personnes différentes, sur l'ensemble d'apprentissage du corpus *REPERE*.

Nous avons poursuivi le regroupement jusqu'à un seuil de 30 pour voir comment se comportait le taux d'EGER (voir figure 5.23)

Sur les deux ensembles, les deux courbes ont un tracé proche montrant une rapide diminution du taux d'EGER avec la fusion des séquences de personnes identiques pour les faibles distances. Puis, aux alentours d'un seuil de 14, les courbes s'aplanissent et ensuite changent de pente mais sans augmenter fortement le taux d'EGER. Comme pour le nommage précoce des locuteurs, le choix du seuil est facilité par cette zone plane dans les courbes

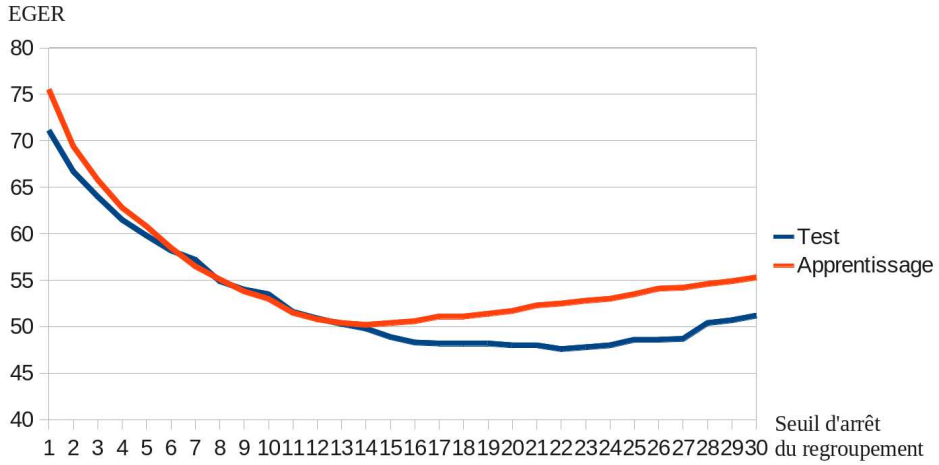


FIG. 5.23 – Évolution du taux d'EGER en fonction du seuil sur l'ensemble d'apprentissage et de test de la phase 1 du corpus *REPERE*

5.3 Conclusion

Dans ce chapitre, nous avons proposé plusieurs méthodes de nommage des locuteurs sans l'aide de modèles biométriques. Les premières proposent différentes solutions de nommages tardifs entre des noms écrits et des clusters de locuteurs, pour ensuite intégrer de plus en plus l'information issue de ces noms pendant le regroupement (nommage intégré puis nommage précoce).

Cette dernière méthode, avec quelques adaptations, s'applique bien au nommage des séquences de visages mais nous ne pouvons identifier que les séquences de visages pour lesquels nous avons un descripteur.

Dans le chapitre suivant nous proposons une solution pour augmenter le nombre de séquences de visages nommables basées sur un nommage précoce de clusters multi-modaux (voix et visages) par les noms écrits.

Chapitre 6

Nommage précoce de clusters multi-modaux pour identifier les locuteurs et les visages

Pour augmenter le taux de rappel des visages reconnus, il faut augmenter le nombre de séquences de visages nommables. Jusqu'à présent, nous nous sommes limités aux séquences ayant un descripteur. Toutefois, il y a des visages de profil ou ayant la tête penchée qui correspondent à des visages à nommer qui n'ont pas de descripteurs, nous n'avons donc pas pu les intégrer dans le nommage précoce pour les visages décrit précédemment.

Il faut donc utiliser une autre source d'informations pour nommer ces visages. Comme l'a montré *Bendris et al.* [BFC+13], on peut utiliser l'identité du locuteur pour reconnaitre un des visages à l'écran. Pour cela, les auteurs utilisent une stratégie en plusieurs étapes (voir le chapitre 2 pour plus de détails) avec d'abord une identification directe des visages et des locuteurs par les noms écrits et les noms prononcés. Puis une propagation intra-modale des noms des visages vers les visages encore anonymes. Et enfin, une propagation inter-modale des noms des locuteurs vers les visages encore anonymes, après les deux premières étapes.

Nous avons préféré adapter notre méthode de nommage précoce à des clusters multi-modaux parce qu'elle nous semble mieux intégrer l'information issue des noms écrits. Donc l'idée est d'effectuer un regroupement des tours de parole et des séquences de visages en clusters de personnes, comme celui proposé par *Khoury et al.* [KSJ12]. Dans cet article, il est montré qu'une diarization multi-modale donne de meilleurs résultats qu'une diarization mono-modale. Comme pour nos méthodes de nommage précoce de clusters mono-modaux, nous voulons contraindre ce regroupement avec les noms écrits et aussi nommer les clusters au cours du processus de regroupement.

Dans la figure 6.1, on retrouve l'architecture globale de notre proposition de nommage précoce de cluster multi-modaux.

La première étape sélectionne les séquences de visages à nommer (section 6.1). Ensuite, nous normalisons les scores des matrices D_{visage} et S_{voix} pour qu'ils

soient comparables (section 6.2). Pour calculer des scores d'association entre les séquences visages et les tours de parole co-occurrant, nous utilisons les sorties d'un classifieur perceptron multi-couches entraîné à partir de l'activité des lèvres mais aussi d'autres caractéristiques visuelles et temporelles (section 6.3). Tous ces scores sont intégrés dans une matrice multi-modale, sur laquelle nous appliquons notre méthode de nommage précoce (section 6.4)

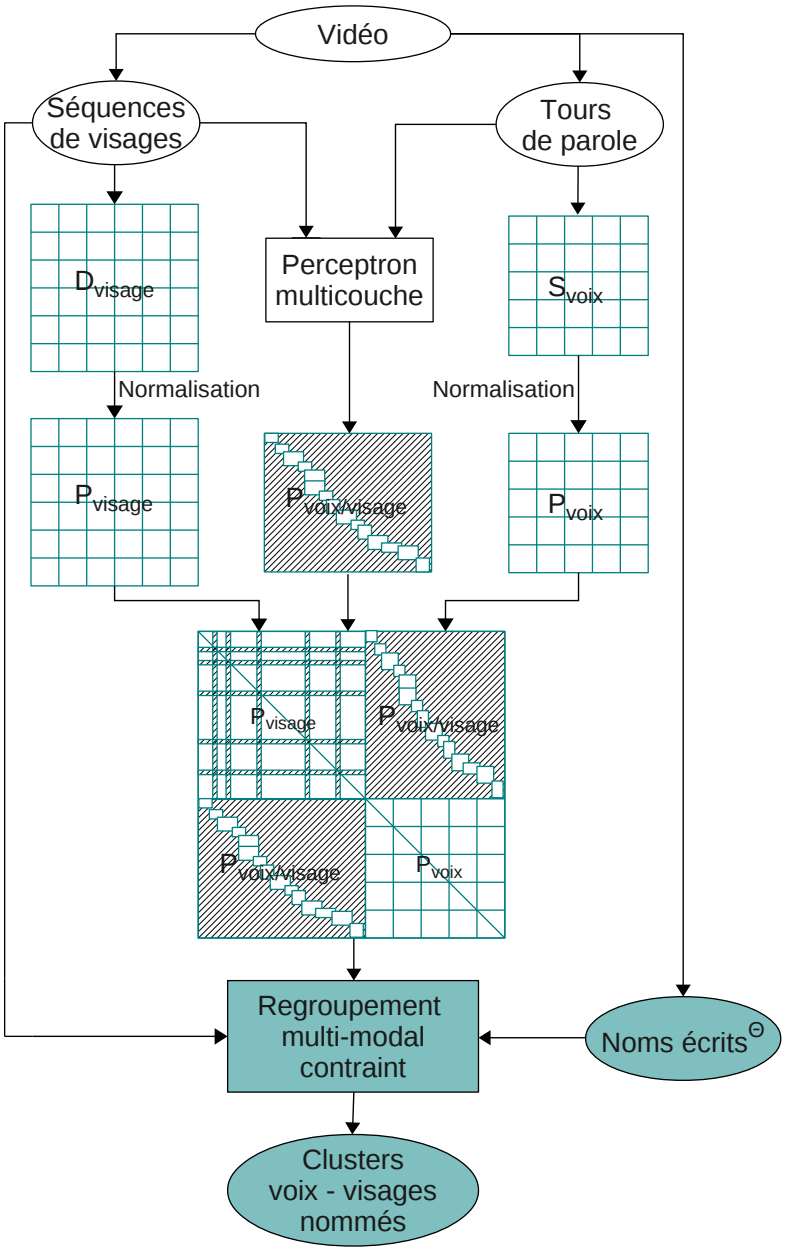


FIG. 6.1 – Schéma global de notre nommage précoce des clusters multi-modaux

6.1 Sélection des visages à nommer

Parmi tous les visages extraits par le détecteur de KIT, il y en a de nombreux qui sont à l'arrière plan de la vidéo. Comme ces « petites têtes » ne jouent que très rarement un rôle intéressant, il ne faut pas les nommer. Nous avons appris un seuil (pour chaque type d'émission) sur la taille minimum du carré englobant le visage (fourni par KIT). Cette idée de filtrage concorde avec l'évaluation du défi *REPERE* qui limite l'identification aux visages supérieurs à une aire de 2000 pixels carrés (pour le détourage du visage).

Nous avons utilisé un deuxième filtre sur la durée d'apparition des séquences de visages qui permet d'éviter de nommer des fausses détections. Les seuils sur ces filtres ont été fixés sur l'ensemble d'apprentissage. Dans les faits, nous conservons les visages dont le carré à un côté supérieur de 40 à 80 pixels et une durée d'affichage minimal comprise entre 0.4 et 1.9 secondes selon le type d'émission.

Nous retrouvons dans le tableau 6.1 les résultats d'identification des visages obtenus par deux oracles qui choisissent le nom correct pour les visages qui ont un descripteur ($Oracle_d$, comme déjà présenté dans le chapitre précédent) ou les visages sélectionnés ($Oracle_s$).

Méthode	$\%P$	$\%R$	$\%F$	%EGER
$Oracle_d$	97.2%	60.0%	74.2%	31.5%
$Oracle_s$	97.6%	68.5%	80.5%	25.5%

Tab. 6.1 – Identification des visages par un oracle qui n'utilise que les visages qui ont un descripteur ($Oracle_d$) ou par un oracle qui n'utilise que les visages filtrés selon le protocole donné ci-dessus ($Oracle_s$), ensemble de test de la phase 1 du défi *REPERE*.

On peut voir que, si on utilise tous les visages filtrés, on peut augmenter le rappel de 60.0% à 68.5%. L'EGER minimum que l'on peut obtenir diminue en conséquence.

6.2 Normalisation des matrices mono-modales

Pour fusionner la matrice de distance entre séquences de visages et la matrice de proximité entre tours de parole, il faut normaliser les scores. On peut voir, sur la figure 6.2, la distribution des scores contenus dans ces deux matrices sur l'ensemble d'entraînement du corpus *REPERE*, avec en bleu les distances/proximités des séquences/tours de parole de la même personne et en rouge de deux personnes différentes.

Nous avons décidé d'approximer ces distributions par une densité de probabilité de la loi normale. Ensuite, pour chaque modalité, nous calculons séparément la probabilité a postériori que deux séquences de visages/tours de parole correspondent à la même personne.

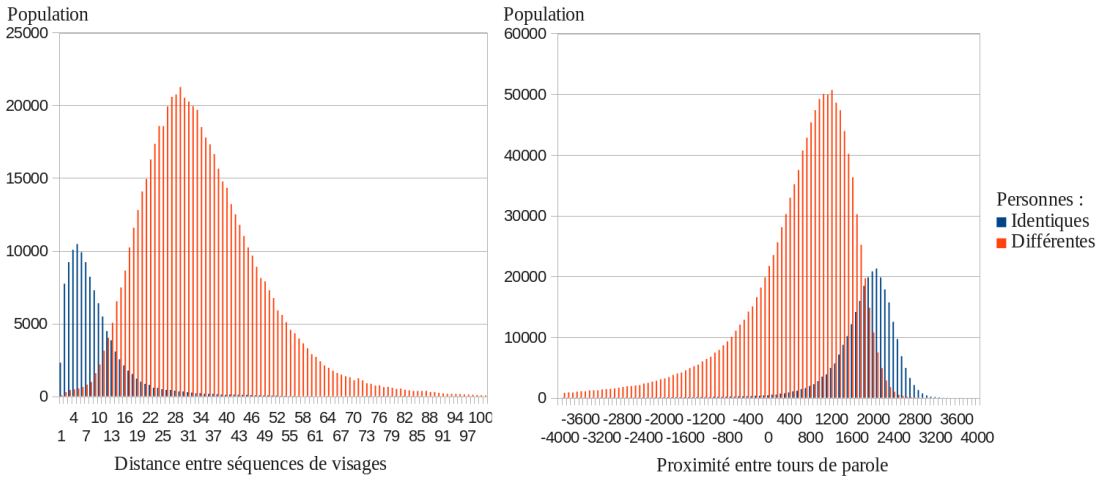


FIG. 6.2 – Distribution des distances entre séquences de visage et tours de parole identiques (bleu) ou différents (rouge) sur l'ensemble d'entraînement de la phase 1 du corpus *REPERE*.

Nous utilisons directement ces probabilités a postériori pour remplir la matrice de similarité. C'est sur cette matrice que l'on effectue le regroupement.

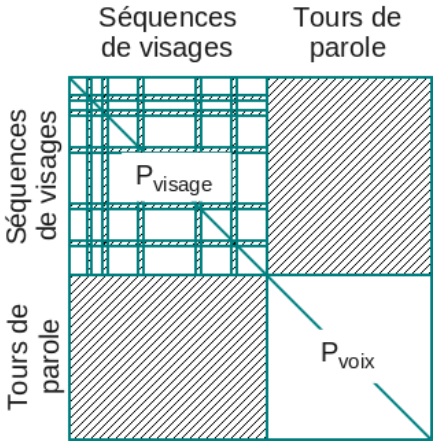


FIG. 6.3 – Matrice multi-modale sans liens entre les deux modalités.

Il est important de noter que la sous-matrice P_{visage} n'est pas pleine puisque certaines séquences de visages que nous avons sélectionnées n'ont pas de descripteurs. Elles ne peuvent donc pas être comparées aux autres séquences de visages.

6.3 Score d'association entre tours de parole et séquences de visages

Pour effectuer un regroupement en clusters multi-modaux dans cette matrice, il faut rajouter des liens entre les séquences de visages et les tours de parole. La solution la plus couramment utilisée dans l'état de l'art est d'observer l'activité des lèvres. Si les lèvres d'un visage bougent, il y a de fortes chances pour que le locuteur corresponde à ce visage.

Nous utilisons donc un détecteur très simple basé sur une variation de l'histogramme des couleurs dans la région des lèvres avec un écart temporel de 80 millisecondes (2 trames à 25 images/seconde). Or, sur certain visages il est difficile de positionner avec précision les lèvres, donc nous utilisons d'autres caractéristiques spatio-temporelles comme la taille et la centralité du visage, la durée de recouvrement entre le tour de parole et la séquence de visages, etc.

Pour prendre une décision à partir de ces caractéristiques, nous utilisons un classifieur de type perceptron multi-couches (à l'aide de l'outil Weka[1]). Ce classifieur a été entraîné à partir d'exemples positifs (le locuteur et le visage correspondent à la même personne dans l'annotation manuelle) et d'exemples négatifs issus de la base d'apprentissage. Le perceptron nous retourne un score d'association entre 0 et 1 pour chaque paire (visage, tour de parole) qui co-occurre sur l'ensemble de test. Plus de détails sur l'extraction des descripteurs sont disponibles dans l'annexe B.

Nous complétons notre matrice multi-modale avec ces scores d'association entre les tours de parole et les séquences de visages. Cette matrice n'est toujours pas pleine, mais les séquences de visages qui n'ont pas de descripteur mais qui correspondent à des visages parlant d'après notre classifieur peuvent maintenant être nommées via l'identité des locuteurs.

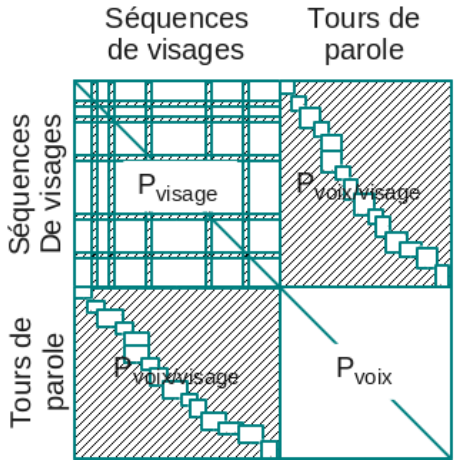


FIG. 6.4 – Matrice multi-modale avec des liens entre les deux modalités.

[1] www.cs.waikato.ac.nz/ml/weka

6.4 Regroupement hiérarchique contraint

Maintenant que nous avons notre matrice multi-modale avec des liens entre les voix et les visages, nous pouvons effectuer notre regroupement hiérarchique.

Nous avons utilisé la même méthode que pour le nommage précoce de clusters de visages avec tout de même quelques adaptations. Nous ne re-décrivons pas la méthode complète mais seulement les différences pour chacune des 4 étapes :

Initialisation du regroupement

Avant le regroupement, l'ensemble des clusters multi-modaux $\mathcal{G}$ correspond à l'ensemble des singletons de $\mathcal{T} \cup \mathcal{V}$ avec $\mathcal{T}$ l'ensemble des tours de parole et $\mathcal{V}$ l'ensemble des séquences de visages.

La construction des ensembles $\mathcal{K}_1$, $\mathcal{K}_0$ et $\mathcal{U}$ est la même que pour le nommage précoce des visages. les clusters contenant un tour de parole co-occurrant avec un nom sont forcément dans $\mathcal{K}_1$, alors que les autres sont dans $\mathcal{U}$.

Contraintes sur le regroupement

Les trois contraintes sont les mêmes que pour le nommage précoce des visages.

Mise à jour après chaque fusion

La fonction f est modifiée de la même manière que le nommage précoce des visages.

Après chaque fusion, les deux mêmes vérifications sont effectuées (deux clusters avec des séquences de visages co-occurrant ne peuvent pas être nommés par le même nom, chaque cluster $g \in K$ dont on est sûr d'avoir le nom appartient à $\mathcal{K}_1$).

La principale différence par rapport aux deux précédents nommages précoces vient de la mise à jour des scores de similarité entre clusters. En effet, il faut tenir compte de tous les scores disponibles pour connaitre la proximité de deux clusters g_1 et g_2. Dans la suite, $s(e_1, e_2)$ correspond au score de similarité entre les éléments (tour de parole ou séquence de visages) e_1 et e_2 dans l'une des trois matrices P_{visage}, P_{voix} ou $P_{voix/visage}$.

Il faut considérer quatre ensembles de scores distincts entre deux clusters g_1 et g_2 :

- L'ensemble des scores entre tours de parole des deux clusters :

$$\mathcal{S}_t(g_1, g_2) = \{s(t_1, t_2) \mid t_1 \in (g_1 \cap \mathcal{T}), t_2 \in (g_2 \cap \mathcal{T})\} \tag{6.1}$$

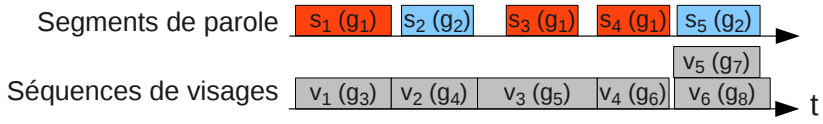
- L'ensemble des scores entre séquences de visages des deux clusters :

$$\mathcal{S}_v(g_1, g_2) = \{s(v_1, v_2) \mid v_1 \in (g_1 \cap \mathcal{V}), v_2 \in (g_2 \cap \mathcal{V})\} \tag{6.2}$$

- L'ensemble des scores d'association entre tours de parole et séquences de visages des deux clusters :

$$\mathcal{S}_{tv}(g_1, g_2) = \{s(e_1, e_2)\} \text{ tel que : } \begin{cases} e_1 \in (g_1 \cap \mathcal{T}) \text{ et } e_2 \in (g_2 \cap \mathcal{V}) \\ \text{ou} \quad e_1 \in (g_1 \cap \mathcal{V}) \text{ et } e_2 \in (g_2 \cap \mathcal{T}) \end{cases} \tag{6.3}$$

- L'ensemble des scores (obtenus par transitivité) entre tours de parole d'un cluster et séquences de visages de l'autre cluster via une séquence de visages tiers. C'est-à-dire que le score entre deux clusters de modalités différentes est issu des relations directes (les scores de co-occurrence issus de la section 6.3) mais aussi des relations indirectes par transitivité via une autre séquence de visages. Pour mieux comprendre, prenons l'exemple dans la figure ci-dessous :



Le score entre g_1 et g_3 ne dépend pas que de la relation directe (s_1, v_1) mais aussi de la relation indirecte (v_1, s_3) via v_3 et de (v_1, s_4) via v_4.
Pour le score entre g_2 et g_4, il faut procéder de la même manière, mais en choisissant la meilleure transitivité entre v_5 et v_6 pour relier s_5 à v_2.
L'utilisation de cette transitivité est intéressante dans les cas où par exemple les scores $s(s_5, v_5)$ et $s(s_5, v_6)$ sont proches, dans ce cas la proximité entre v_2 et (v_5 ou v_6) sera déterminante pour fusionner le tour s_5 de parole avec le bon visage.

Pour remplir l'ensemble des scores transitifs $\mathcal{S}_{tr}(g_1, g_2)$ nous allons d'abord sélectionner les tours de parole t_{tr} de $(g_1 \cup g_2) \cap \mathcal{T}$ qui ne co-occurrent avec aucune séquence de visages de $(g_1 \cup g_2) \cap \mathcal{V}$.
Ensuite, pour chacun de ces tours de parole t_{tr} nous trouvons la meilleure séquence de visages $v_{tr} \notin (g_1 \cup g_2)$ qui peut-être utilisée comme transition. C'est-à-dire celle qui a le plus gros score $s(t_{tr}, v_{tr})$ (celui-ci devant être supérieur à 0.5) mais aussi le plus gros score avec les séquences de visages du même cluster que t_{tr}.
Ainsi, l'ensemble des scores transitifs entre les deux clusters correspond à :

$$\mathcal{S}_{tr} = \{ \max_{v_{tr} \in (\mathcal{V} \setminus (g_1 \cup g_2))} (\frac{s(t_{tr}, v_{tr}) + \frac{\sum_{d \in \mathcal{S}_v(g', \{v_{tr}\})} d}{card(\mathcal{S}_v(g', \{v_{tr}\})}}{2})) \} \text{ tel que :} \tag{6.4}$$

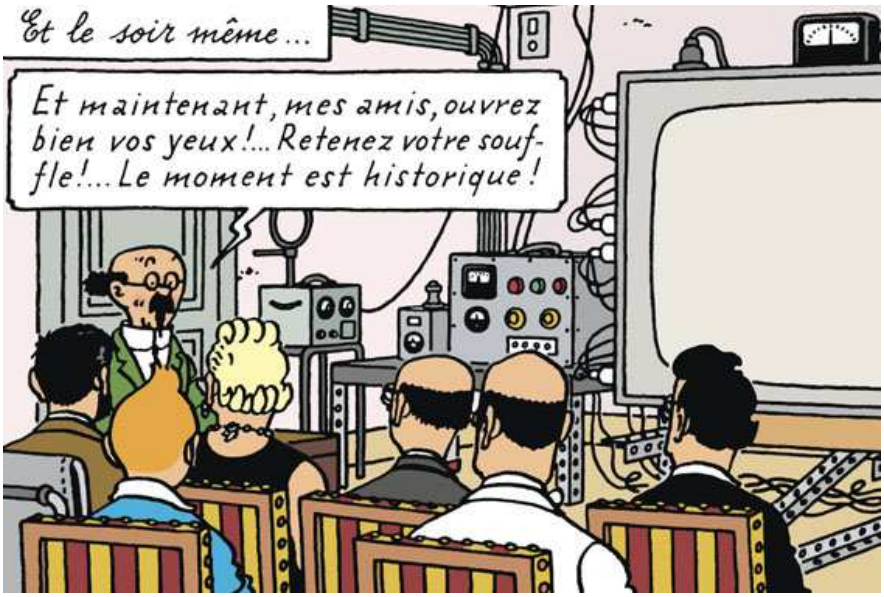
$$\begin{aligned} & t_{tr} \in (g_1 \cup g_2) \cap \mathcal{T} \\ \text{et} \quad & t_{tr} \text{ ne co-occurre aucun } v \in (g_1 \cup g_2) \cap \mathcal{V} \\ \text{et} \quad & \text{si } t_{tr} \in g_1 \text{ alors } g' = g_2 \text{ sinon } g' = g_1 \\ \text{et} \quad & s(t_{tr}, v_{tr}) > 0.5 \end{aligned} \tag{6.5}$$

Le score final entre deux clusters correspond à la somme des scores des différents ensembles divisée par le nombre total de scores dans ces ensembles :

$$s(g_1, g_2) = \frac{\sum_{s \in \mathcal{S}_t} s + \sum_{s \in \mathcal{S}_v} s + \sum_{s \in \mathcal{S}_{tv}} s + \sum_{s \in \mathcal{S}_{tr}} s}{card(\mathcal{S}_t \cup \mathcal{S}_v \cup \mathcal{S}_{tv} \cup \mathcal{S}_{tr})} \tag{6.6}$$

Association finale entre noms et clusters

Une fois le critère d'arrêt atteint, les clusters de $\mathcal{K}_1$ associés à un seul nom ($card(\{h(o) \mid o \in f(g)\}) = 1$) sont nommés par ce nom. Pour les autres clusters nommables, on va nommer en priorité les clusters les plus proches de clusters déjà nommés, en respectant les contraintes décrites ci-dessus, et ce jusqu'à ne plus avoir, si possible, d'occurrences de noms non utilisées.



6.5 Résultats du nommage précoce de clusters multi-modaux

Dans la figure 6.5, on retrouve le tracé du taux d'EGER pour l'identification des locuteurs (courbes bleue et rouge) et le tracé du taux d'EGER pour l'identification des visages (courbes jaune et verte) en fonction du critère d'arrêt utilisé pour arrêter le regroupement. Un seuil plus petit signifie que le regroupement est arrêté plus tard (cette figure se lit de préférence de droite à gauche).

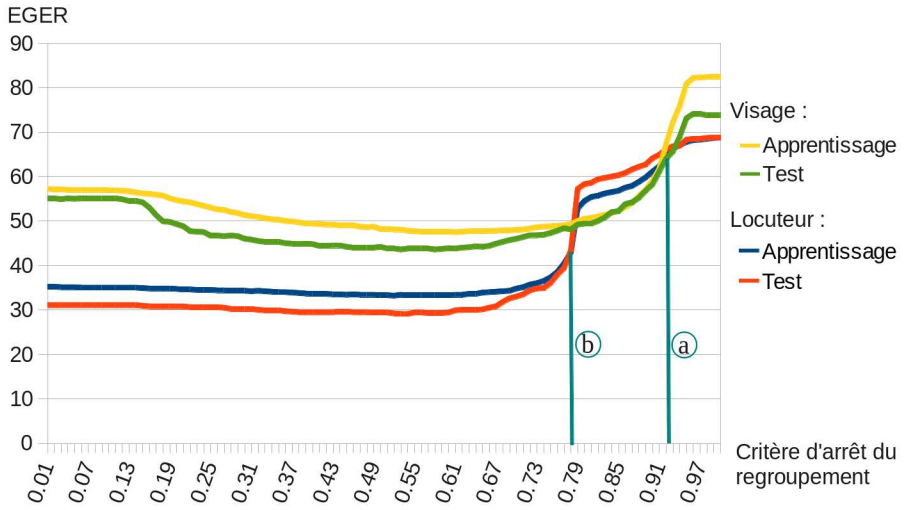


FIG. 6.5 – Évolution du taux d'EGER d'identification des locuteurs (courbes bleue et rouge) et des visages (courbes jaune et verte) en fonction du critère d'arrêt du regroupement sur l'ensemble d'apprentissage et de test de la phase 1 du corpus *REPERE*

Sur cette figure, on voit que les courbes d'EGER calculées sur l'ensemble de test suivent la même forme que celles calculées sur l'ensemble d'apprentissage. La large zone plane au centre des courbes permet d'être assez peu sensible au choix du critère d'arrêt.

Sur ces courbes, en lisant de droite à gauche, on voit une rapide diminution du taux d'EGER (ⓐ et ⓑ) qui correspond à la propagation intra-modale des noms écrits (d'un tour de parole vers un autre tour de parole, idem pour les séquences de visages). Cette réduction n'apparait pas au même moment pour les deux tâches (plus tardive pour l'identification des locuteurs ⓑ). Cela peut s'expliquer par le recouvrement plus important des deux distributions des scores de similarité (entre tours de parole de la même personne ou de personnes différentes) pour le locuteur (voir figure 6.2) par rapport à celles pour les visages. Comme ce recouvrement est plus important, il donne des probabilités a postériori plus faibles pour les locuteurs que pour les visages.

On remarque tout de même, avant ⓑ, une lente baisse du taux sur la courbe bleue et sur la rouge. Avant, des noms se sont propagés de séquences de visages à d'autres séquences mais aussi vers des tours de parole grâce aux scores d'association entre les deux modalités (ce qui explique cette lente diminution).

Le tableau 2 montre tous les résultats de nos méthodes de nommage précoce de clusters mono-modaux et multi-modaux. Pour séparer l'apport des clusters multi-modaux et l'apport lié à l'ajout des visages nommables supplémentaires, nous avons appliqué notre méthode multi-modale avec les mêmes visages que ceux utilisés dans le nommage précoce des clusters mono-modaux pour les visages (ligne indiquant « Avec un descripteur ») ou les visages que nous avons sélectionné dans la section 6.1 (ligne indiquant « Sélectionnés »).

Modalités à identifier	Méthode `NP` avec des clusters	Séquences de visages	$\%P$	$\%R$	$\%F$	%EGER
Visages	mono-modaux	Avec un descripteur	85.1	42	56.3	49.8
	multi-modaux	Avec un descripteur	85.9	44.9	59.0	47.7
	multi-modaux	Sélectionnées	85.9	49.0	62.4	43.8
Locuteurs	mono-modaux	Avec un descripteur	80.4	68.3	73.9	29.9
	multi-modaux	Avec un descripteur	83.5	67.6	74.7	30.4
	multi-modaux	Sélectionnées	83.7	68.8	75.5	29.2

TAB. 6.2 – Résultats des méthodes de nommage précoce de clusters mono-modaux (avec un seuil global) et multi-modaux, ensemble de test de la phase 1 du corpus *REPERE*.

Le choix du seuil d'arrêt du regroupement a été fait sur l'ensemble d'apprentissage pour minimiser l'EGER, avec un seuil global quelque soit le type d'émission.

Le but du nommage précoce de clusters multi-modaux est surtout d'augmenter le nombre de visages nommés. On voit effectivement le rappel augmenter de 42% à 49.0%, ce qui diminue le taux d'EGER correspondant. Par ailleurs, un autre effet intéressant à noter est que la multi-modalité permet aussi de faire baisser ce taux d'EGER pour une même précision (49.8% à 47.7%).

Pour les locuteurs, il y a peu de changement avec seulement une légère amélioration de la précision pour à peu près le même rappel.

6.5.1 Intégration des modèles biométriques des personnes des rôles R123

Un des défauts de notre méthode de nommage précoce est que les noms des personnes des rôles R123 sont assez peu souvent écrits. Donc les contraintes pendant le regroupement ne peuvent pas beaucoup éviter les fusions erronées des clusters de personnes des rôles R123 avec d'autres clusters nommés par un nom écrit (autre que le nom d'une personne des rôles R123).

L'utilisation des noms des modèles biométriques des rôles R123 va permettre de corriger ce défaut. Si on arrive à nommer les clusters correspondant à ces personnes par les modèles biométriques, les contraintes sur le regroupement empêcheront les mauvaises fusions citées ci-dessus.

Les systèmes d'identification supervisés du LIMSI (locuteurs) et de LEAR (visages) nous fournissent un score entre chaque tour de parole/séquence de visages et les modèles biométriques construits sur la base d'apprentissage. Nous

nous sommes limités aux modèles des personnes appartenant aux rôles R123 de la même émission que celles ciblées puisque ce sont ces personnes qui ont le plus de chance d'être présentes.

Il est important de noter que des expériences complémentaires nous ont montré que l'utilisation de tous les modèles construits sur la base d'apprentissage n'apporte pas de meilleurs résultats.

Pour utiliser cette source d'informations, nous initialisons d'abord notre méthode de nommage précoce de clusters multi-modaux. Ensuite, avant de procéder au regroupement, nous nommons avec certitude les tours de parole/séquences de visages avec le nom du modèle biométrique le plus proche si le score fournit par le système supervisé est supérieur à un seuil (seuil appris sur l'ensemble d'apprentissage). Enfin, le regroupement va propager les noms écrits et les noms des modèles biométriques en respectant les trois contraintes présentées précédemment. L'utilisation de cette propagation pour les noms des modèles biométriques nous permet d'être assez stricts sur le choix des tours de parole/séquences de visages que l'on nomme directement par les modèles biométriques.

Dans le tableau 6.3, on retrouve les résultats d'identification des locuteurs et des visages avec l'intégration de l'information issue des modèles biométriques.

Modalités à identifier	Modèle R123 pour les Locuteurs	Modèle R123 pour les Visages	$\%P$	$\%R$	$\%F$	%EGER
Visages			85.9	49.0	62.4	43.8
		X	90.6	56.4	69.5	36.9
	X		86.6	53.2	65.9	40.0
	X	X	89.4	55.8	68.7	37.4
Locuteurs			83.7	68.8	75.5	29.2
		X	87.5	73.2	79.8	24.8
	X		89.4	79.3	84.0	19.5
	X	X	88.9	79.7	84.1	19.2

TAB. 6.3 – Résultats du nommage précoce de clusters multi-modaux avec l'ajout des modèles biométriques de visage et/ou de locuteurs des personnes des rôles R123, ensemble de test de la phase 1 du corpus *REPERE*.

La première et la cinquième ligne du tableau reprennent les résultats du nommage précoce des clusters multi-modaux sans l'utilisation des modèles biométriques. On peut voir que, pour les deux tâches, les résultats sont toujours meilleurs quel que soit le type de modèles biométriques utilisés.

Ce tableau permet aussi d'apprécier la multi-modalité de notre méthode. En effet, si on ajoute l'information des modèles d'une modalité et que l'on évalue la tâche d'identification de l'autre modalité, on voit que les résultats sont bien meilleurs que si l'on n'avait pas utilisé de modèles biométriques. Ce qui montre que les noms des modèles biométriques de visages (respectivement des locuteurs) sont propagés vers les tours de parole (respectivement les séquences de visages).

Enfin, on observe que l'utilisation conjointe des modèles biométriques des locuteurs et des visages n'apporte pas ou peu d'amélioration par rapport à l'utilisation seule des modèles de la tâche ciblée.

6.5.2 Comparaison avec les résultats de la campagne d'évaluation *REPERE*

Nous avons reporté dans le tableau 6.4 les résultats obtenus par les trois consortiums (A, B, C) lors de la campagne de la phase 1 du défi *REPERE*. Ces résultats ont été obtenus en évaluant les sorties des systèmes des consortiums A, B et C ainsi que les systèmes proposés dans ce chapitre[2].

Modèles biométriques	Modalités à identifier	Consortium	$\%P$	$\%R$	$\%F$	%EGER
Sans	Visages	A	68.7	50.4	58.2	46.1
		B (NP mono-modal)	79.3	48.5	60.2	46.4
		C	93.4	45.7	61.4	47.3
		NP multi-modal	85.9	49.0	62.4	43.8
Sans	Locuteurs	A	70.0	68.5	69.2	31.8
		B (NT3$^{\ominus}$+NA)	76.9	73.0	74.9	26.3
		C	81.5	59.3	68.6	37.3
		NP multi-modal	83.7	68.8	75.5	29.2
Avec	Visages	A	74.8	55.0	63.4	41.5
		B ([BPF^{+}13])	78.0	63.2	69.8	36.7
		C	89.2	55.1	68.1	39.8
		NP multi-modal	89.4	55.8	68.7	37.4
Avec	Locuteurs	A	79.2	77.4	78.3	22.8
		B ([BPF^{+}13])	85.9	82.0	83.9	17.6
		C	87.8	81.1	84.3	17.7
		NP multi-modal	88.9	79.7	84.1	19.2

TAB. 6.4 – Résultats des 3 consortiums pour l'identification des locuteurs et des visages **sans** et **avec** modèles biométriques, corpus *REPERE* phase 1.

Les 8 premières lignes reprennent les résultats pour une tâche **sans** l'utilisation de modèles biométriques. Nous avons utilisé lors de cette campagne le système de nommage précoce de clusters mono-modaux pour l'identification des visages. On voit ici que notre méthode de nommage précoce de clusters multi-modaux obtient un meilleur taux d'EGER, par rapport aux trois consortiums.

Pour l'identification du locuteur, nous avons utilisé lors de la campagne la méthode NT3$^{\ominus}$+NP qui utilise une diarization BIC+CLR. Notre méthode nommage précoce de clusters multi-modaux obtient un moins bon taux d'EGER, mais nous pensons que si le regroupement de notre méthode est plus proche d'une méthode de l'état de l'art (remplacement de la matrice BIC par une diarization BIC+CLR ou *i-vector*) nous devrions obtenir de meilleurs résultats.

Les 8 dernières lignes reprennent les résultats pour une tâche **avec** l'utilisation de modèles biométriques. Malgré l'utilisation des modèles biométriques des personnes de R123 uniquement et d'une diarization sur un matrice BIC, nous obtenons des résultats assez similaires à ceux des trois consortiums avec systématiquement une meilleure précision.

[2]Avec les références de l'archive version 9 de l'évaluation de l'ensemble de test de la phase 1, pour une raison inexpliquée, les résultats obtenus sont légèrement différents des résultats officiels fournis par les organisateurs de la campagne


C'est une boule de feu!...
Une ÉNORRRME boule de feu!
?

Chapitre 7

Conclusion et perspectives

7.1 Conclusion

Les travaux présentés dans cette thèse s'inscrivent dans le domaine de l'identification des personnes dans les flux télévisés et plus particulièrement dans le nommage des clusters de personnes par une source de noms interne à la vidéo.

Dans la littérature, les approches qui n'utilisent pas de modèles biométriques pour identifier les personnes dans les documents audio-visuels se sont orientées vers différentes sources pour obtenir le nom des personnes présentes. Ces sources de noms sont dépendantes de la nature du média ciblé. Pour les émissions de télévision, les précédents travaux ont principalement utilisé les noms prononcés et les noms écrits.

Pour les noms prononcés, outre le fait que leur extraction soit sujette à des erreurs de transcription et de détection des noms dans ces transcriptions, il n'est pas évident de savoir qui ils désignent : le locuteur courant, suivant ou précédent, ou un visage visible au moment de la citation ou dans le plan suivant ou précédent, ou encore une personne qui n'est pas présente. Les noms écrits étaient, par le passé, difficilement utilisables à cause des trop nombreuses erreurs de transcription, dues à la mauvaise qualité des images. Cela a donc orienté les travaux de l'état de l'art vers des méthodes d'association noms-personnes qui prennent en compte cette incertitude liée aux sources de noms.

Ces dernières années, la qualité des vidéos et l'incrustation des textes surimposés à l'image se sont beaucoup améliorées. Ce qui nous a permis d'extraire les noms écrits à l'écran avec très peu d'erreurs de transcription.

La comparaison des capacités de ces deux sources de noms à proposer le nom des personnes présentes dans les émissions de télévisions montre que les noms prononcés proposent un plus grand nombre d'occurrences de citation par rapport aux noms écrits ; mais que les erreurs d'extraction réduisent le nombre de personnes nommables dans les vidéos. A contrario, le peu d'erreurs que produit l'extraction automatique des noms écrits permet de nommer deux fois plus de personnes qu'avec les noms prononcés. Un point important à noter est que l'association des noms écrits aux bonnes personnes est intrinsèquement plus simple que pour les noms prononcés.

Ceci nous a permis de proposer plusieurs méthodes de nommage des locuteurs sans l'aide de modèles biométriques. Les différentes méthodes de nommage tardif que nous proposons, nomment des clusters de locuteurs par les noms écrits à l'écran. Nous avons ensuite intégré de plus en plus l'information issue de ces noms pendant le regroupement (nommage intégré puis nommage précoce).

Chacune de ces méthodes a ses avantages selon le type de système de diarization disponible :

- Si on dispose seulement d'une sortie de diarization et qu'elle est de très bonne qualité : l'association 1-à-1 entre noms écrits et clusters de locuteurs (`NT1`) apportera les meilleurs résultats. Les autres méthodes de nommage tardif que nous proposons `NT2` (ajout du nommage direct des tours de parole), `NT3` (association 1-à-n), $\texttt{NT3}^{\ominus}$ (Ré-alignement temporel des noms écrits) et $\texttt{NT3}^{\ominus}\texttt{+NA}$ (ajout des noms prononcés) corrigeront une diarization comportant des erreurs.
- Si on dispose d'une boîte noire de diarization sur laquelle on peut régler le critère d'arrêt : le nommage intégré (`NI`) va maximiser la métrique liée à la tâche finale plutôt qu'une minimisation de l'erreur de diarization.
- Si on dispose d'un système complet de diarization modifiable : le nommage précoce (`NP`) est le plus prometteur. Il permet à la fois de maximiser la métrique liée à la tâche finale, et de contraindre la diarization pour éviter les erreurs de fusion pendant le processus de regroupement hiérarchique. Ces contraintes permettant d'être beaucoup moins sensible au choix du critère d'arrêt.

Cette dernière méthode, avec quelques adaptations, s'applique bien au nommage des séquences de visages. Toutefois, nous ne pouvons identifier que celles pour lesquelles nous avons un descripteur. En effet, de nombreux visages ne peuvent pas être nommés par cette méthode parce que le système d'extraction du descripteur propre à chaque visage n'a pas une confiance suffisante dans l'alignement des points caractéristiques.

Pour palier ce défaut, nous avons adapté notre méthode de nommage précoce à la propagation des noms écrits sur des clusters multi-modaux voix-visage, ce qui permet de nommer les visages parlant, pour lesquels nous n'avons pas de descripteur, via l'identité du locuteur.

L'avantage de cette méthode est qu'elle nomme au cours d'un processus unique les locuteurs et les visages. De plus, il est très facile d'intégrer les sorties de systèmes d'identification mono-modaux.

Que ce soit avec ou sans l'information issue des modèles biométriques des présentateurs-reporters-journalistes, notre nommage précoce de clusters multi-modaux obtient des résultats comparables aux meilleurs systèmes qui ont participé à la phase 1 du défi *REPERE*, et ce malgré l'utilisation d'une simple matrice de distance BIC entre tours de parole.

7.2 Perspectives

Projets à court terme :

Comme nous l'avons indiqué, notre méthode de nommage précoce utilise une matrice de distance BIC sans mise à jour de cette distance après chaque fusion entre le nouveau cluster et les autres clusters. Nous voudrions la remplacer par une métrique plus à l'état de l'art (BIC+CLR, ou *i-vector*). Toutefois, se pose le problème de la normalisation des distances si on veut qu'elle reste comparable à des distances venant d'autres modalités.

Un autre évolution possible est le remplacement du détecteur d'activité des lèvres par un système basé sur le flou optique [BCC10]. Cela devrait nous permettre de mieux discriminer les visages parlant des autres.

En ce qui concerne les visages qui n'ont pas de descripteur, plusieurs travaux de l'état de l'art proposent d'extraire un descripteur sur les vêtements. Néanmoins, là aussi, se pose le problème d'intégrer des distances entre visages basées sur un très bon descripteur de visages et des distances basées sur un descripteur sur les vêtements moins discriminant.

Une dernière perspective à court terme est l'utilisation de matrices de distance multi-modales et inter-vidéos pour propager les noms d'une vidéo aux autres. Nous avons en effet montré dans le chapitre 4 qu'on augmentait le nombre de personnes nommables avec une propagation inter-vidéos des noms. Nous pensons donc pouvoir mieux nommer les personnes récurrentes sur plusieurs vidéos d'une même émission et qui ont peu souvent leur nom écrit (les présentateurs ou certains chroniqueurs par exemple).

Projets à moyen et long terme :

Détection des présentateurs par leurs rôles

Dans cette thèse, nous avons évalué nos systèmes sur une tâche d'identification globale quel que soit le rôle des personnes. Nous avons donc essayé de nommer toutes les personnes y compris les personnes des rôles R123. Toutefois, il est évident qu'identifier les invités ou les personnes interviewées est plus porteur d'informations que d'identifier le présentateur. Donc, dans une tâche où seules les personnes des rôles R45 doivent être reconnues, il n'est plus nécessaire d'utiliser les modèles biométriques pour les présentateurs-journalistes.

Plusieurs articles de l'état de l'art proposent de détecter les présentateurs non pas en les reconnaissant mais en identifiant les personnes qui jouent ce rôle. En effet, les présentateurs sont rarement nommables et ce sont les personnes les plus présentes dans les vidéos donc il est très facile de propager de mauvais noms sur les clusters leurs correspondants. Le but de la détection des clusters jouant le rôle de présentateur est d'écarter du regroupement quelques tours de parole/séquences de visages.

Utilisation conjointe des noms écrits et prononcés

Dans le chapitre 4, nous avons montré que l'utilisation conjointe des noms écrits et des noms prononcés augmentait le nombre de personnes nommables mais aussi le nombre d'occurrences de noms utilisables. Nous avons aussi montré que les noms prononcés désignent d'autres personnes comme celles ne parlant pas dans la vidéo ou encore des personnes n'apparaissant pas (voix-off de journaliste par exemple).

Il est donc intéressant d'utiliser ces occurrences de noms pour, à la fois, nommer plus de personnes mais aussi pour éviter de propager des noms sur des clusters qui auraient dû rester anonymes sans cette source. A contrario de notre méthode `NT3`$^{\ominus}$`+NA` qui intègre les noms prononcés dans un nommage tardif seulement après un nommage par les noms écrits, nous pensons qu'il faudrait intégrer les noms prononcés, malgré leurs faibles précisions, dans un processus unique tel que notre méthode de nommage précoce.

Construction de modèles biométriques automatiquement

Notre système de nommage précoce offre une précision aux alentours de 85%. Elle pourrait être augmentée par un élagage de certains tours de parole/séquences de visages des clusters nommés. Ces annotations automatiques issues d'une vidéo peuvent permettre de construire des modèles biométriques pour identifier à la fois les personnes de la même vidéo mais aussi d'autres vidéos, ce qui correspond à une manière de propager les noms d'une vidéo aux autres. Il serait aussi intéressant d'évaluer la qualité de ces modèles par rapport à des modèles issus d'annotations manuelles.

Apprentissage actif

Comme l'ont proposé *Pham et al.* dans [PMT10, PTM11], quelques annotations manuelles peuvent remplacer l'utilisation des modèles biométriques des personnes des rôles de R123. En effet, il n'est pas nécessaire d'annoter toute la vidéo pour construire leurs modèles biométriques. L'identification manuelle des quelques clusters qui leur correspondent pourrait donc être suffisante pour initier une boucle d'apprentissage actif. De plus, comme ces personnes sont récurrentes d'une vidéo à l'autre de la même émission, ces annotations peuvent être effectuées seulement sur quelques vidéos.

Utilisation de ressources externes

L'utilisation d'images ou de signaux de parole issus d'Internet pour construire des modèles biométriques ([LJH08]) est une autre proposition intéressante de l'état de l'art. En effet, ces ressources sont annotées indirectement et peuvent s'avérer être une source d'information importante. Cependant, elles peuvent être très bruitées pour les personnes peu connues. Par ailleurs, la multiplication des modèles biométriques n'est pas forcément une solution pertinente.

La presse web constitue une autre ressource externe qui peut aussi s'avérer intéressante. Grâce aux noms de personnes contenus dans ses articles, on pourrait former une liste de noms hypothèses assez complète pour les émissions de

journaux télévisés. La redondance des noms au sein de ces articles est aussi un indicateur de leur possible présence dans ce type de flux télévisés.

Utilisation du contexte

Pour finir, on peut aussi envisager de répondre à la question « qui parle » ou « qui apparaît » en lui ajoutant une dimension sémantique. En effet, les personnes présentes dans les flux télévisés sont dépendantes du sujet abordé. Une meilleure compréhension des concepts présents (lieux, objets, thèmes ...) peut nous aider à avoir une meilleure vision des personnes potentiellement présentes. Pour ce faire, on peut envisager un découpage en histoires de ces vidéos. Ensuite, il faudrait y détecter les concepts tant dans les éléments langagiers que dans les éléments visuels. Pour finir par les fusionner afin de comprendre les sujets abordés. Une fois que ce dernier est déterminé, on pourrait utiliser les ressources externes citées précédemment afin d'aider l'identification des personnes.

Annexe A : Corpus *REPERE*

Répartition des vidéos dans le corpus *REPERE*

Dans la figure 1, on peut observer la répartition des émissions sur le corpus *REPERE*. Il y a une grande disparité de la durée des segments UEM (2 minutes en moyenne pour Planète showbiz à 34 minutes pour BFM Story).

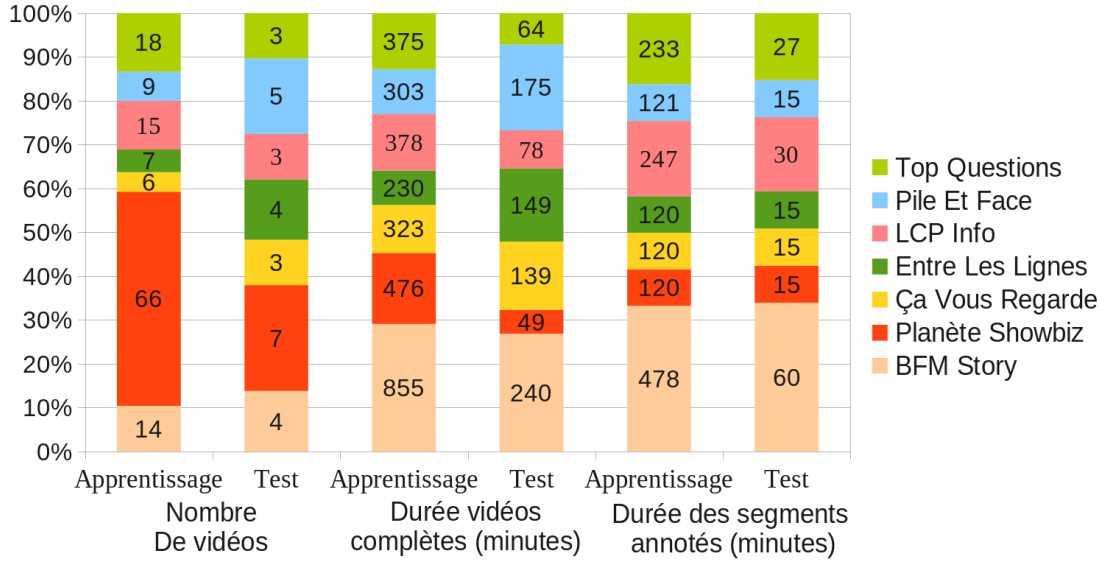


FIG. 1 – Répartition en nombre et en durée des vidéos sur l'ensemble d'entraînement et de test de la phase 1 du corpus *REPERE*

Répartition des modèles des personnes des rôles de R123

Le tableau 1 détaille le nombre de personnes visibles et de locuteurs des personnes des rôles R123 sur l'ensemble d'entraînement et de test de la phase 1 du corpus *REPERE* par type d'émission. La troisième et la huitième lignes correspondent au nombre de personnes/locuteurs en commun sur les deux ensembles. Plus ce nombre en commun est proche de celui de la ligne précédente, plus les personnes de l'ensemble du test pourront être couvertes par un modèle construit sur l'ensemble d'entraînement.

L'avant-dernière ligne de chaque partie correspond au nombre d'apparitions-durées de parole des personnes de R123 sur l'ensemble de test. La ligne en dessous donne le pourcentage couvert par une personne de l'ensemble d'apprentissage du même type d'émission.

	BFM story	Planète showbiz Culture et vous	Ça vous regarde	Entre les lignes	LCP info	Pile et face	Top questions
# pers visibles sur l'ensemble d'entraînement	22	6	1	9	10	1	2
# pers visibles sur l'ensemble de test	7	5	1	6	2	1	1
# pers visibles en commun	4	4	1	5	2	1	1
# d'apparition sur le test	133	38	21	181	47	17	11
% des apparitions couvertes	87	82	100	90	100	100	100
# locuteurs sur l'ensemble d'entraînement	39	11	4	9	21	2	2
# locuteurs sur l'ensemble de test	12	4	2	6	9	1	1
# locuteurs en commun	8	4	1	5	8	1	1
Durée en secondes de la parole sur le test	1548	456	355	939	848	184	53
% de la durée couverte	85	100	87	98	90	100	100

TAB. 1 – Nombre de personnes uniques et nombre d'apparitions-durées de parole des personnes visibles/parlants des rôles R123 sur l'ensemble d'apprentissage, de test et en commun, par émission, phase 1 du corpus *REPERE*

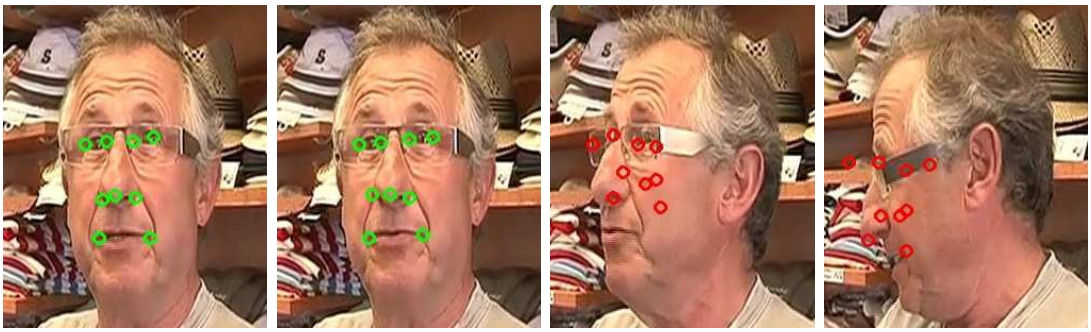
Ce tableau montre que la majorité des apparitions et du temps de parole des personnes de R123 sur l'ensemble de test de la phase 1 du *corpus* est couverte par les modèles construits sur l'ensemble d'apprentissage. Il est à noter que l'on va utiliser un peu plus de modèles de locuteurs que de modèles de visages à cause des voix-off des journalistes.

Annexe B : Extraction des scores d'association entre tours de parole et séquences de visages

Pour trouver le visage qui a le plus de chances de correspondre au locuteur courant nous utilisons un détecteur très simple basé sur une variation de l'histogramme des couleurs dans la région des lèvres, avec un écart temporel de 80 millisecondes (2 trames à 25 images/seconde).

Pour effectuer cette comparaison, nous avons utilisé la détection des lèvres effectuée par LEAR. En effet, lorsqu'ils alignent le maillage de neuf points sur le visage, deux d'entre eux sont utilisés pour la bouche.

Si on regarde l'exemple (voir figure 2) que nous avions utilisé dans le chapitre 2 pour présenter l'alignement du maillage, on voit que les deux premières images sont espacées de 80 millisecondes. Le maillage a été aligné avec un score de confiance suffisamment haut, pour être sûr que les coins des lèvres ont été localisés avec précision.



(a) (b) (c) (d)

FIG. 2 – Exemples d'alignement du maillage de 9 points : (a) et (b) bien alignés (vert), (c) et (d) mal alignés (rouge)

On peut voir une différence de couleur dans la zone entre les deux coins des lèvres, mais cette information n'est disponible que si le maillage a été bien aligné.

Sur les deux images (c) et (d), le score de confiance pour le maillage n'était pas suffisamment élevé. Et effectivement, les points correspondants aux coins des lèvres n'ont pas été positionnés aux bons endroits. Sur ces deux dernières images,

il est très difficile de détecter la position de la bouche, nous ne pouvons donc pas connaître l'activité des lèvres sur ces images.

Pour calculer un score global d'activité des lèvres entre une séquence de visages et un tour de parole, nous calculons d'abord un score pour chacune des images de la séquence. Si on prend comme exemple la chronologie présentée dans la figure 3, on voit un tour de parole t_1 (rectangle bleu) et une séquence de visages v_1 (cadre vert). Les barres verticales bleues ou rouges représentent les images de visage i de cette séquence. Celles en bleu correspondent aux images où le maillage a été correctement aligné. Pour celles en rouge, la confiance dans l'alignement était trop faible.

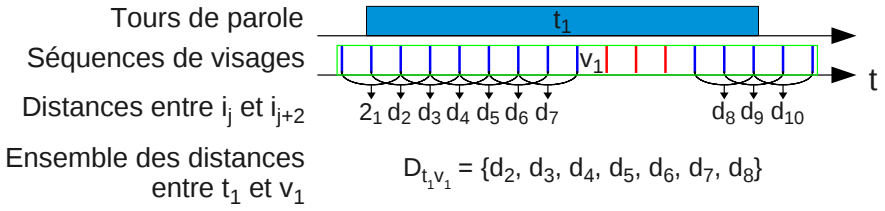


FIG. 3 – Chronologie pour le calcul de l'activité des lèvres

A partir de la zone centrale entre les deux coins des lèvres de l'image courante et celle de l'image 80 millisecondes plus tard (si le maillage a été correctement aligné aussi sur cette deuxième image), nous pouvons calculer une distance d_i entre les deux histogrammes de ces zones. La distance entre histogrammes est celle proposée par la librairie *opencv*[1].

Le score global entre la séquence de visages v_1 et le tour de parole t_1 correspond à la moyenne des distances des images qui co-occurrent avec le tour de parole :

$$d_{t_1 v_1} = \frac{\sum_{d \in \mathcal{D}} d}{card(\mathcal{D})} \tag{1}$$

Avec $\mathcal{D}$ qui est l'ensemble des d_i calculés pour les i_j qui co-occurrent avec t_1.

Ne vouloir utiliser que l'activité des lèvres va à l'encontre de notre hypothèse de base qui est d'essayer de nommer des séquences de visages qui n'ont pas de descripteurs. Si elles n'ont pas de descripteurs, c'est parce que le maillage n'a pas été positionné avec une confiance suffisante, donc nous ne pouvons pas connaître l'activité des lèvres. De ce fait, nous devons utiliser d'autres caractéristiques pour savoir quel visage on peut associer au locuteur courant.

Ainsi, pour savoir si une séquence de visages v correspond au locuteur du tour de parole courant t, nous avons utilisé comme caractéristiques :

- Le score d'activité des lèvres de v par rapport à t

[1] http://opencv.org/

- La taille moyenne des visages de v
- Est-ce que v a la plus grande taille moyenne des visages par rapport à toutes celles qui co-occurrent avec t.
- Est-ce que les visages de v sont centrés (au tiers ou deux tiers s'il y deux visages détectés, au centre de l'image sinon).
- L'orientation de v (de face ou de profil avec un angle en degrés).
- La durée de co-occurrence entre v et t.
- Est-ce que v est celle qui co-occurre le plus avec t.

Pour prendre une décision à partir de ces informations, nous avons utilisé un classifieur perceptron multi-couches entrainé sur la base d'apprentissage (à l'aide de l'outil Weka[2]). Ce classifieur a été entraîné à partir d'exemples positifs (le locuteur et le visage correspondent à la même personne dans l'annotation manuelle) et d'exemples négatifs. Il nous retourne un score de confiance entre 0 et 1 pour chaque paire (visage/tour de parole) qui co-occurre sur l'ensemble de test.

Nous avons propagé l'identité du locuteur (déterminée par un oracle) d'un tour de parole vers la séquence de visage qui a le meilleur score si celui-ci est supérieur à 0.5. Une séquence de visages ne pouvant être nommée qu'une seule fois. Nous présentons les résultats dans le tableau 2. Nous ne pouvons évidement nommer qu'une seule séquence de visage par tour de parole, ce qui explique le rappel de 39.5% par rapport au 68.5% possible (voir tableau 6.1). Par contre nous obtenons une bonne précision (87.3%).

Méthode	$\%P$	$\%R$	$\%F$	%EGER
Nom du locuteur pour le visage parlant	87.3%	39.5%	54.4%	51.6%

Tab. 2 – Résultats obtenus par la propagation du nom du locuteur (déterminée par un oracle) vers la séquence de visages avec le plus gros score en sortie du classifieur.

[2] www.cs.waikato.ac.nz/ml/weka

Bibliographie

[ABM08] Alexandre Allauzen and Hélène Bonneau-Maynard. Training and evaluation of pos taggers on the french multitag corpus. In *the 6th International Conference on Language Resources and Evaluation, LREC*, page , 2008.

[AGP10] Marios Anthimopoulos, Basilis Gatos, and Ioannis Pratikakis. A two-stage scheme for text detection in video images. *Image and Vision Computing*, 28 :1413–1426, 2010.

[BBF+10] Martin Bäuml, Keni Bernardin, Mika Fischer, Hazim Kemal Ekenel, and Rainer Stiefelhagen. Multi-pose face recognition for person retrieval in camera networks. In *7th International Conference on Advanced Video and Signal-Based Surveillance, AVSS*, pages 441–447, 2010.

[BCC10] Meriem Bendris, Delphine Charlet, and Gérard Chollet. Lip activity detection for talking faces classification in TV-Content. In *International Conference on Machine Vision*, 2010.

[BEN11] Meriem BENDRIS. *Indexation audio visuelle des personnes dans un contexte de télévision.* PhD thesis, Thèse en informatique de l'École TÉLÉCOM ParisTech, 2011.

[BFC+13] Meriem Bendris, Benoit Favre, Delphine Charlet, Géraldine Damnati, Rémi Auguste, Jean Martinet, and Gregory Senay. Unsupervised Face Identification in TV Content using Audio-Visual Sources. In *Proceedings of the 11th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 243–249, 2013.

[BFD12] Frédéric. Bechet, Benoît. Favre, and Géraldine Damnati. Detecting person presence in tv shows with linguistic and structural features. In *the 37th IEEE International Conference in Acoustics, Speech and Signal Processing, ICASSP*, pages 5077–5080, 2012.

[BP13] Hervé Bredin and Johann Poignant. Integer Linear Programming for Speaker Diarization and Cross-Modal Identification in TV Broadcast. In *the 14th Annual Conference of the International Speech Communication Association, INTERSPEECH*, page , 2013.

[BPF+13] Hervé Bredin, Johann Poignant, Guillaume Fortier, Makarand Tapaswi, Viet Bac Le, Achintya Sarkar, Claude Barras, Sophie Rosset, Anindya Roy, Qian Yang, Hua Gao, Alexis Mignon, Jakob Verbeek,

Laurent Besacier, Georges Quénot, Hazim Kemal Ekenel, and Rainer Stiefelhagen. QCompere at REPERE 2013. In *First Workshop on Speech, Language and Audio in Multimedia - the 14th Annual Conference of the International Speech Communication Association, INTERSPEECH*, page , 2013.

[BPT+12] Hervé Bredin, Johann Poignant, Makarand Tapaswi, Guillaume Fortier, Viet Bac Le, Thibault Napoleon, Hua Gao, Claude Barras, Sophie Rosset, Laurent Besacier, Jakob Verbeek, Georges Quénot, Frédéric Jurie, and Hazim Kemal Ekenel. Fusion of speech, faces and text for person identification in TV broadcast. In *Workshop on Information Fusion in Computer Vision for Concept Recognition, ECCV-IFCVCR*, pages 385–394, 2012.

[BRG+09] Guillaume Bernard, Sophie Rosset, Olivier Galibert, Eric Bilinski, and Gilles Adda. LIMSI participation in the QAst 2009 track : Experimenting on Answer Scoring. In *the 10th Workshop of the Cross-Language Evaluation Forum, CLEF*, pages 289–296, 2009.

[BTS13] Martin Bäuml, Makarand Tapaswi, and Rainer Stiefelhagen. Semi-supervised Learning with Constraints for Person Identification in Multimedia Data. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, page , 2013.

[Bun97] Horst Bunke. On a relation between graph edit distance and maximum common subgraph. *Pattern Recogn. Lett.*, 18(9) :689–694, 1997.

[BZMG06] Claude Barras, Xuan Zhu, Sylvain Meignier, and Jean-Luc Gauvain. Multi-Stage Speaker Diarization of Broadcast News. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5) :1505–1512, 2006.

[CG98] Scott Shaobing Chen and Ponani S. Gopalakrishnan. Speaker, Environment And Channel Change Detection And Clustering Via The Bayesian Information Criterion. In *DARPA Broadcast News Transcription and Understanding Workshop*, pages 127–132, 1998.

[CH04] Ming-yu Chen and Alexander G. Hauptmann. Searching for a specific person in broadcast news video. In *the IEEE 29th International Conference on Acoustics, Speech and Signal Processing, ICASSP*, page , 2004.

[CLG05] Leonardo Canseco, Lori Lamel, and Jean-Luc Gauvain. A Comparative Study Using Manual and Automatic Transcriptions for Diarization. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 415–419, November 2005.

[CMAQ05] Mbarek Charhad, Daniel Moraru, Stéphane Ayache, and Georges Quénot. Speaker identity indexing in audio-visual documents. In *the 3rd Workshop on Content-Based Multimedia Indexing, CBMI*, page , 2005.

[CNT10] Timothée Cour, Akash Nagle, and Benjamin Taskar. Talking pictures : temporal grouping and dialog-supervised person recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1014–1021, 2010.

[CRLG04] Leonardo Canseco-Rodriguez, Lori Lamel, and Jean-Luc Gauvain. Speaker diarization from speech transcripts. In *the 5th Annual Conference of the International Speech Communication Association, INTERSPEECH*, page , 2004.

[CSJT09] Timothée Cour, Benjamin Sapp, Chris Jordan, and Benjamin Taskar. Learning from ambiguously labeled images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 919–926, 2009.

[CSL02] Min Cai, Jiqiang Song, and Michael R. Lyu. A new approach for video text detection. In *the IEEE International Conference on Image Processing, ICIP*, pages 117–120, 2002.

[CST11] Timothée Cour, Benjamin Sapp, and Benjamin Taskar. Learning from Partial Labels. *Journal of Machine Learning Research*, pages 1501–1536, 2011.

[DEMM05] Paul Deléglise, Yannick Estève, Sylvain Meignier, and Téva Merlin. The LIUM speech transcription system : a CMU Sphinx III-based system for french broadcast news. In *the 6th Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 1653–1656, 2005.

[DKD+11] Najim Dehak, Patrick Kenny, Reda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-End Factor Analysis for Speaker Verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(4) :788–798, 2011.

[DLLP97] Thomas G. Dietterich, Richard H. Lathrop, and Tomas Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2) :31–71, 1997.

[DRMY12] Grégor. Dupuy, Mickael Rouvier, Sylvain Meignier, and Estève Yannick. I-vectors and ILP clustering adapted to cross-show speaker diarization. In *the 13rd Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 2174–2177, 2012.

[DT05] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 886–893, 2005.

[EIH07] Farshideh Einsele, Rolf Ingold, and Jean Hennebert. A HMM-based approach to recognize ultra low resolution anti-aliased words. In *the 2nd international conference on Pattern recognition and machine intelligence, PReMI*, pages 511–518, 2007.

[EKLMP12] Elie El-Khoury, Antoine Laurent, Sylvain Meignier, and Simon Petitrenaud. Combining transcription-based and acoustic-based speaker

identifications for broadcast news. In *the 37th IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pages 4377–4380, 2012.

[EMDM07] Yannick. Estève, Sylvain Meignier, Paul Deléglise, and Julie Mauclair. Extracting true speaker identities from transcriptions. In *the 8th Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 2601–2604, 2007.

[ESZ06] Mark Everingham, Josef Sivic, and Andrew Zisserman. Hello ! My name is... Buffy – Automatic naming of characters in TV video. In *the 17th British Machine Vision Conference, BMVC*, page , 2006.

[ESZ09] Mark Everingham, Josef Sivic, and Andrew Zisserman. Taking the bite out of automatic naming of characters in TV video. *Image and Vision Computing*, 27(5), 2009.

[FE04] Bernhard Fröba and Andreas Ernst. Face detection with the modified census transform. In *the Sixth IEEE international conference on Automatic face and gesture recognition*, Sixth IEEE International Conference on Automatic Face and Gesture Recognition, AFGR, pages 91–96, 2004.

[FM08] Jenny Rose Finkel and Christopher D. Manning. Enforcing transitivity in coreference resolution. In *the 46th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 45–48, 2008.

[Fou11] N. Fourour. *Identification et catégorisation automatiques des entités nommées dans les textes français*. PhD thesis, Thèse en informatique à l'université de Nantes, 2011.

[FTZ04] Hui Fang, Tao Tao, and ChengXiang Zhai. A formal study of information retrieval heuristics. In *the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–56, 2004.

[GCM+12] Aude Giraudel, Matthieu Carré, Valérie Mapelli, Juliette Kahn, Olivier Galibert, and Ludovic Quintard. The REPERE corpus : a multimodal corpus for person recognition. In *the 8th International Conference on Language Resources and Evaluation, LREC*, page , 2012.

[GLA98] Jean-Luc Gauvain, Lori Lamel, and Gilles Adda. Partitioning and Transcription of Broadcast News Data. In *the 5th International Conference on Spoken Language Processing, Incorporating The 7th Australian International Speech Science and Technology Conference, ICSLP*, pages 1335–1338, 1998.

[GLA02] Jean-Luc Gauvain, Lori Lamel, and Gilles Adda. The LIMSI Broadcast News Transcription System. In *Speech Communication*, pages 89–108, 2002.

[GMVS12] Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, and Cordelia Schmid. Face recognition from caption-based supervision. *International Journal of Computer Vision*, 96(1) :64–82, 2012.

[Gra] Graphviz. . http://www.graphviz.org/.

[HCWZ01] Xian-Sheng Hua, Xiang-Rong Chen, Liu Wenyin, and Hong-Jiang Zhang. Automatic Location of Text in Video Frames. In *ACM workshops on Multimedia : multimedia information retrieval, MIR*, pages 24–27, 2001.

[Hou99] Ricky Houghton. Named faces : putting names to faces. *IEEE Intelligent Systems*, 14 :45–50, 1999.

[Hou06] Michael E. Houle. *A generic query-based model for scalable clustering*. National Institute of Informatics technical report. 2006.

[HTF01] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The elements of statistical learning : data mining, inference, and prediction : with 200 full-color illustrations*. New York : Springer-Verlag, 2001.

[HvL11] Marijn Huijbregts and David A. van Leeuwen. Diarization-based Speaker Retrieval for Broadcast Television Archives. In *the 12nd Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 1037–1040, 2011.

[IOTM07] Ichiro Ide, Tomokazu Ogasawara, Takashi Takahashi, and Hiroshi Murase. Name Identification of People in News Video by Face Matching. In *the 3rd International Workshop on Computer Vision meets Databases, CVDB*, page , 2007.

[JJM+08] Vincent Jousse, Christine Jacquin, Sylvain Meignier, Yannick Estève, and Béatrice Daille. Etude pour l'amélioration d'un système d'identification nommée du locuteur. In *Les Journées d'Étude sur la Parole - Traitement Automatique des Langues Naturelles, JEP-TALN*, 2008.

[JLK09] Cheolkon Jung, Qifeng Liu, and Joongkyu Kim. A stroke filter and its application to text localization. *Pattern Recognition Letters*, 30 :114–122, 2009.

[JMJ+09] Vincent Jousse, Sylvain Meignier, Christine Jacquin, Simon Petitrenaud, Yannick Estève, and Béatrice Daille. Analyse conjointe du signal sonore et de sa transcription pour l'identification nommée de locuteur. In *Traitement Automatique des langues, TAL*, pages 201–225, 2009.

[Jou11] Vincent Jousse. *Identification nommée du locuteur : exploitation conjointe du signal sonore et de sa transcription*. PhD thesis, Thèse en informatique à l'université de Maine, 2011.

[JPRM+09] Vincent Jousse, Simon. Petit-Renaud, Sylvain. Meignier, Yannick. Estève, and Christine Jacquin. Automatic named identification of speakers using diarization and ASR systems. In *the 34th IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pages 4557–4560, 2009.

[KSJ12] Elie Khoury, Christine Sénac, and Philippe Joly. Audiovisual Diarization Of People In Video Content. *Multimedia Tools and Applications*, 2012.

[Kuh55] Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2) :83–97, 1955.

[KWRB11] Martin Köstinger, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Learning to recognize faces from videos and weakly related information cues. In *the 8th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS*, pages 23–28, 2011.

[LBF0] Viet Bac Le, Claude Barras, and Marc Ferràs. On the use of GSV-SVM for Speaker Diarization and Tracking. In *Odyssey - The Speaker and Language Recognition Workshop*, pages 146–150, 2010 .

[LCD+11] Lori Lamel, Sandrine Courcinous, Julien Despres, Jean-Luc Gauvain, Yvan Josse, Kevin Kilgour, Florian Kraft, Viet Bac Le, Hermann Ney, Ilya Nußbaum-Thom, Markus Oparin, Tim Schlippe, Ralf Schlüter, Tanja Schultz, Thiago Fraga da Silva, Sebastian Stüker, Martin Sundermeyer, Bianca Vieru, Ngoc Thang Vu, Alexander Waibel, and Cécile Woehrling. Speech Recognition for Machine Translation in Quaero. In *The International Workshop on Spoken Language Translation, IWSLT*, page , 2011.

[LCY10] Thomas Lavergne, Olivier Cappé, and François Yvon. Practical very large scale CRFs. In *the 48th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 504–513, 2010.

[LE98] Rainer Lienhart and Wolfgang Effelsberg. Automatic Text Segmentation and Text Recognition for Video Indexing. *ACM/Springer Multimedia Systems*, pages 69–81, 1998.

[LFF+09] Anan Liu, Jinghao Fei, Jianping Fan, Lin Pang, Yongdong Zhang, and Jintao Li. Confusion network based Video OCR post-processing approach. In *IEEE international conference on Multimedia and Expo, ICME*, pages 137–140, 2009.

[lin] lingpipe. . http://www.alias-i.com/lingpipe/.

[LJH08] Chunxi Liu, Shuqiang Jiang, and Qingming Huang. Naming faces in broadcast news video by image google. In *the 16th ACM International Conference on Multimedia, ACMMM*, pages 717–720, 2008.

[LSHN07] Duy-Dinh Le, Shin'ichi Satoh, Michael E. Houle, and Dat Phuoc Tat Nguyen. Finding important people in large news video databases using multimodal and clustering analysis. In *the 23rd IEEE International Conference on Data Engineering Workshop, ICDEW*, pages 127–136, 2007.

[MLP98] Oded Maron and Tomás Lozano-Pérez. A Framework for Multiple-Instance Learning. In *Advances in Neural Information Processing Systems*, pages 570–576, 1998.

[MME06a] Julie Mauclair, Sylvain Meignier, and Yannick Estève. Indexation en locuteur : utilisation d'informations lexicales. In *Les Journées d'Étude sur la Parole, JEP*, 2006.

[MME06b] Julie Mauclair, Sylvain Meignier, and Yannick Estève. Speaker diarization : about whom the speaker is talking ? In *IEEE Odyssey 2006 - The Speaker and Language Recognition Workshop*, page , 2006.

[MNM07] Chengyuan Ma, Patrick Nguyen, and Mahajan Milind. Finding speaker identities with a conditional maximum entropy model. In *the 32nd IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pages 261–264, 2007.

[MR11] Dinarelli Marco and Sophie Rosset. Models Cascade for Tree-Structured Named Entity Detection. In *the 5th International Joint Conference on Natural Language Processing, IJCNLP*, pages 1269–1278, 2011.

[OD06] Derya Ozkan and Pınar Duygulu. Finding people frequently appearing in news. In *the 5th international conference on Image and Video Retrieval*, pages 173–182, 2006.

[ON03] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29 :19–51, 2003.

[PBB+13] Johann Poignant, Hervé Bredin, Laurent Besacier, Georges Quénot, and Claude Barras. Towards a better integration of written names for unsupervised speakers identification in videos. In *First Workshop on Speech, Language and Audio in Multimedia - the 14th Annual Conference of the International Speech Communication Association, INTERSPEECH*, page , 2013.

[PBL+12] Johann Poignant, Hervé Bredin, Viet Bac Le, Laurent Besacier, Claude Barras, and Georges Quénot. Unsupervised speaker identification using overlaid texts in TV broadcast. In *the 13rd Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 2650–2653, 2012.

[PBL+13] Johann Poignant, Laurent Besacier, Viet Bac Le, Sophie Rosset, and Georges Quénot. Unsupervised naming of speakers in broadcast TV : using written names, pronounced names or both ? In *the 14th Annual Conference of the International Speech Communication Association, INTERSPEECH*, page , 2013.

[PBQ13] Johann Poignant, Laurent Besacier, and Georges Quénot. Nommage non-supervisé des personnes dans les émissions de télévision : une revue du potentiel de chaque modalité. In *la 10ème COnférence en Recherche d'Information et Applications, CORIA*, page , 2013.

[PBQT12] Johann Poignant, Laurent Besacier, Georges Quénot, and Franck Thollard. From text detection in videos to person identification.

In *IEEE International Conference on Multimedia and Expo, ICME*, pages 854–859, 2012.

[PJME10] Simon Petitrenaud, Vincent Jousse, Sylvain Meignier, and Yannick Estève. Reconnaissance Automatique de Locuteurs à l'aide de Fonctions de Croyance. In *le 17e congrès francophone Reconnaissance des Formes et Intelligence Artificielle (RFIA'10)*, pages 4557–4560, 2010.

[PMT10] Phi The Pham, Marie-Francine Moens, and Tinne Tuytelaars. Naming persons in news video with label propagation. In *IEEE international conference on Multimedia and Expo, ICME*, pages 1528–1533, 2010.

[Poi11] Johann Poignant. Détection et reconnaissance de texte dans les documents vidéos, Et leurs apports à la reconnaissance de personnes. In *la 6ème Rencontres Jeunes Chercheurs en Recherche d'Information - 8è COnférence en Recherche d'Information et Applications, RJCRI-CORIA*, pages 409–414, 2011.

[PRJME10] Simon Petit-Renaud, Vincent Jousse, Sylvain Meignier, and Yannick Estève. Identification of speakers by name using belief functions. In *the 13th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Methods, IPMU*, pages 179–188, 2010.

[PSM+08] Rohit Prasad, Shirin Saleem, Ehry MacRostie, Premkumar Natarajan, and Michael Decerbo. Multi-frame combination for robust videotext recognition. In *the 33rd IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pages 1357–1360, 2008.

[PTM11] Phi The Pham, Tinne Tuytelaars, and Marie-Francine Moens. Naming people in news videos with label propagation. *IEEE MultiMedia*, 18(3) :44–55, 2011.

[PTQB11] Johann Poignant, Franck Thollard, Georges Quénot, and Laurent Besacier. Text detection and recognition for person identification in video. In *the 9th Workshop on Content-Based Multimedia Indexing, CBMI*, pages 245–248, 2011.

[QMB03] Georges Quénot, Daniel Moraru, and Laurent Besacier. CLIPS at TRECvid : Shot Boundary Detection and Feature Detection. In *Workshop TRECVID*, pages 35–40, 2003.

[RJ76] Stephen E. Robertson and Karen Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3) :129–146, 1976.

[RM12] Mickael Rouvier and Sylvain Meignier. A Global Optimization Framework For Speaker Diarization. In *Odyssey - The Speaker and Language Recognition Workshop*, page , 2012.

[RSC+98] Douglas A. Reynolds, Elliot Singer, Beth A. Carlson, Gerald C. O'Leary, Jack McLaughlin, and Marc A. Zissman. Blind clustering

of speech utterances based on speaker and language characteristics. In *The 5th International Conference on Spoken Language Processing, Incorporating The 7th Australian International Speech Science and Technology Conference, ICSLP*, page , 1998.

[SEZ09] Josef Sivic, Mark Everingham, and Andrew Zisserman. "Who are you ?" - Learning person specific classifiers from video. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1145–1152, 2009.

[SK97] Shin'ichi Satoh and Takeo Kanade. Name-It : association of face and name in video. In *Conference on Computer Vision and Pattern Recognition, CVPR*, pages 368–, 1997.

[SLS04] Xiaodan Song, Ching-Yung Lin, and Ming-Ting Sun. Cross-modality automatic face model training from large video databases. In *Conference on Computer Vision and Pattern Recognition Workshop, CVPRW*, pages 91–, 2004.

[SLXC11] Jitao Sang, Chao Liang, Changsheng Xu, and Jian Cheng. Robust movie character identification and the sensitivity analysis. In *IEEE international conference on Multimedia and Expo, ICME*, pages 1–6, 2011.

[SNK97] Shin'ichi Satoh, Yuichi Nakamura, and Takeo Kanade. Name-It : naming and detecting faces in video by the integration of image and natural language processing. In *The Fifteenth International Joint Conference on Artifical Intelligence, IJCAI*, pages 1488–1493, 1997.

[SNK99] Shin'ichi Satoh, Yuichi Nakamura, and Takeo Kanade. Name-It : naming and detecting faces in news videos. *IEEE Multimedia*, 6 :22–35, 1999.

[SP00] Jaakko J. Sauvola and Matti Pietikäinen. Adaptive document image binarization. *Pattern Recognition*, pages 225–236, 2000.

[SS99] Robert E. Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3) :297—-336, 1999.

[SX12] Jitao Sang and Changsheng Xu. Robust Face-Name Graph Matching for Movie Character Identification. *IEEE Transactions on Multimedia*, 14(3-1) :586–596, 2012.

[TBS12] Makarand Tapaswi, Martin Bäuml, and Rainer Stiefelhagen. "Knock ! Knock ! Who is it ?" probabilistic person identification in TV-series. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 2658–2665, 2012.

[Tra06] Sue E. Tranter. Who really spoke when ? finding speaker turns and identities in broadcast news audio. In *the 31st IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pages 1013–1016, 2006.

[WJC02] Christian Wolf, Jean-Michel Jolion, and Françoise Chassaing. Text Localization, Enhancement and Binarization in Multimedia Documents. In *the 16th International Conference on Pattern Recognition, ICPR*, pages 1037–1040, 2002.

[WKRB11a] Paul Wohlhart, Martin Köstinger, Peter M. Roth, and Horst Bischof. Learning Face Recognition in Videos from Associated Information Sources. In *Workshop of the Austrian Association for Pattern Recognition*, page , 2011.

[WKRB11b] Paul Wohlhart, Martin Köstinger, Peter M. Roth, and Horst Bischof. Multiple instance boosting for face recognition in videos. In *the 33rd annual Symposium of the German Association for Pattern Recognition, DAGM*, pages 132–141, 2011.

[XHC+01] Jie Xi, X.-S. Hua, Xiang-Rong Chen, Liu Wenyin, and Hong-Jiang Zhang. A video text detection and recognition system. *IEEE international conference on Multimedia and Expo, ICME*, page 222, 2001.

[YCH04] Jun Yang, Ming-yu Chen, and Alexander G. Hauptmann. Finding person x : Correlating names with visual appearances. In *the 3rd International Conference on Image and Video Retrieval, CIVR*, pages 270–278, 2004.

[YH04] Jun Yang and Alexander G. Hauptmann. Naming every individual in news video monologues. In *the 12nd ACM International Conference on Multimedia, ACMMM*, pages 10–16, 2004.

[YH05] Qixiang Ye and Qingming Huang. A New Text Detection Algorithm in Images/Video Frames. In *the 5th Pacific Rim Conference on Advances in Multimedia Information Processing, PCM*, pages 858–865, 2005.

[YYH05] Jun Yang, Rong Yan, and Alexander G. Hauptmann. Multiple instance learning for labeling faces in broadcasting news video. In *the 13th ACM international conference on Multimedia, ACMMM*, pages 31–40, 2005.

[ZXCL09] Yi-Fan Zhang, Changsheng Xu, Jian Cheng, and Hanqing Lu. Naming faces in films using hypergraph matching. In *IEEE international conference on Multimedia and Expo, ICME*, pages 278–281, 2009.

[ZXL08] Yi-Fan Zhang, Changsheng Xu, and Hanqing Lu. Automatic character identification in feature-length films. In *IEEE international conference on Multimedia and Expo, ICME*, pages 1469–1472, 2008.

[ZXLH09] Yi-Fan Zhang, Changsheng Xu, Hanqing Lu, and Yeh-Min Huang. Character identification in feature-length films using global face-name matching. *Journal IEEE Transactions on Multimedia*, 11(7) :1276–1288, November 2009.

Table des figures

Liste des tableaux

Bibliographie personnelle

1. J. Poignant, L. Besacier, V. B. Le, S. Rosset, and G. Quénot. Unsupervised naming of speakers in broadcast TV : using written names, pronounced names or both ? In the 14rd Annual Conference of the International Speech Communication Association, INTERSPEECH, 2013.
2. J. Poignant, L. Besacier, and G. Quénot. Nommage non-supervisé des personnes dans les émissions de télévision : une revue du potentiel de chaque modalité. In CORIA 2013, papier long (oral), Neuchatel, Suisse, Apr. 2013.
3. J. Poignant, H. Bredin, L. Besacier, G. Quénot, and C. Barras. Towards a better integration of written names for unsupervised speakers identification in videos. In First Workshop on Speech, Language and Audio in Multimedia, SLAM, 2013.
4. H. Bredin and J. Poignant. Integer Linear Programming for Speaker Diarization and Cross-Modal Identification in TV Broadcast. In the 14rd Annual Conference of the International Speech Communication Association, INTERSPEECH, 2013.
5. H. Bredin, J. Poignant, G. Fortier, M. Tapaswi, V.-B. Le, A. Roy, C. Barras, S. Rosset, A. Sarkar, Q. Yang, H. Gao, A. Mignon, J. Verbeek, L. Besacier, G. Quénot, H. K. Ekenel, and R. Stiefelhagen. QCOMPERE at REPERE 2013. In First Workshop on Speech, Language and Audio in Multimedia, SLAM, 2013.
6. J. Poignant, H. Bredin, V.-B. Le, L. Besacier, C. Barras, and G. Quénot. Unsupervised Speaker Identification using Overlaid Texts in TV Broadcast. In Interspeech 2012, Portland, Oregon (USA), Sept. 2012.
7. J. Poignant, F. Thollard, G. Quénot, and L. Besacier. From text detection in videos to person identification. In Proceedings of the 2012 IEEE International Conference on Multimedia and Expo (ICME 2012), Melbourne, Australia, July 2012.
8. H. Bredin, J. Poignant, M. Tapaswi, G. Fortier, V. Bac Le, T. Napoleon, H. Gao, C. Barras, S. Rosset, L. Besacier, J. Verbeek, G. Quénot, F. Jurie, and H. Kemal Ekenel. Fusion of Speech, Faces and Text for Person Identification in TV Broadcast. In ECCV 2012, Workshop on Information Fusion in Computer Vision for Concept Recognition, Firenze, Italy, Oct. 2012.
9. J. Poignant, F. Thollard, G. Quénot, and L. Besacier. Text Detection and Recognition for Person Identification in Video. In 9th Internationnal Workshop on Content-Based Multimedia Indexing, pages 245–248, Madrid, Spain, June 2011.
10. J. Poignant. Détection et reconnaissance de texte dans les documents vidéos, Et leurs apports à la reconnaissance de personnes. In RJCRI - CORIA 2011 : 6è Rencontres Jeunes Chercheurs en Recherche d'Information - 8è COnférence en Recherche d'Information et Applications, pages 409–414, Avignon, France, Mar. 2011.

Résumé

Ce travail de thèse a pour objectif de proposer plusieurs méthodes d'identification non-supervisées des personnes présentes dans les flux télévisés à l'aide des noms écrits à l'écran. Comme l'utilisation de modèles biométriques pour reconnaître les personnes présentes dans de larges collections de vidéos est une solution peu viable sans connaissance a priori des personnes à identifier, plusieurs méthodes de l'état de l'art proposent d'employer d'autres sources d'informations pour obtenir le nom des personnes présentes.

Ces méthodes utilisent principalement les noms prononcés comme source de noms. Cependant, on ne peut avoir qu'une faible confiance dans cette source en raison des erreurs de transcription ou de détection des noms et aussi à cause de la difficulté de savoir à qui fait référence un nom prononcé.

Les noms écrits à l'écran dans les émissions de télévision ont été peu utilisés en raison de la difficulté à extraire ces noms dans des vidéos de mauvaise qualité. Toutefois, ces dernières années ont vu l'amélioration de la qualité des vidéos et de l'incrustation des textes à l'écran. Nous avons donc ré-évalué, dans cette thèse, l'utilisation de cette source de noms.

Nous avons d'abord développé LOOV (pour Lig Overlaid OCR in Vidéo), un outil d'extraction des textes sur-imprimés à l'image dans les vidéos. Nous obtenons avec cet outil un taux d'erreur en caractères très faible. Ce qui nous permet d'avoir une confiance importante dans cette source de noms.

Nous avons ensuite comparé les noms écrits et les noms prononcés dans leurs capacités à fournir le nom des personnes présentes dans les émissions de télévisions. Il en est ressorti que deux fois plus de personnes sont nommables par les noms écrits que par les noms prononcés extraits automatiquement. Un autre point important à noter est que l'association entre un nom et une personne est intrinsèquement plus simple pour les noms écrits que pour les noms prononcés.

Cette très bonne source de noms nous a donc permis de développer plusieurs méthodes de nommage non-supervisé des personnes présentes dans les émissions de télévision. Nous avons commencé par des méthodes de nommage tardives où les noms sont propagés sur des clusters de locuteurs. Ces méthodes remettent plus ou moins en cause les choix fait lors du processus de regroupement des tours de parole en clusters de locuteurs. Nous avons ensuite proposé deux méthodes (le nommage intégré et le nommage précoce) qui intègrent de plus en plus l'information issue des noms écrits pendant le processus de regroupement.

Pour identifier les personnes visibles, nous avons adapté la méthode de nommage précoce pour des clusters de visages. Enfin, nous avons aussi montré que cette méthode fonctionne aussi pour nommer des clusters multi-modaux voix-visage.

Avec cette dernière méthode, qui nomme au cours d'un unique processus les tours de paroles et les visages, nous obtenons des résultats comparables aux meilleurs systèmes ayant concouru durant la première campagne d'évaluation *REPERE*.

Abstract

In this thesis we propose several methods for unsupervised person identification in TV broadcasts. As the use of biometric models to recognize people in large video collections is not a viable option without a priori knowledge of people present in this videos, several methods of the state-of-the-art propose to use other (multimodal) sources of information.

These methods mainly use the names pronounced as source of names. However, we can not have a good confidence in this source due to transcription or name detection errors and also due to the difficulty of knowing to who refers a pronounced name.

The names written on the screen in TV broadcast have not be used in the past due to the difficulty of extracting these names from low quality videos. However, improvement of the videos quality observed recently, lead us to re-evaluate, in this thesis, the use of overlaid names for person identification.

We first developed LOOV (for LIG Overlaid OCR in Video) : this tool extracts overlaid texts written in video with a very low character error rate. This allows us to have an important confidence in this source of information.

We then compared the written names and pronounced names in their ability to provide the names of a person present in TV broadcast. We found that, when an automatic name extraction is used, twice more persons are nameable by written names than by pronounced names. Another important point to note is that the association between a name and a person is inherently easier for written names than for pronounced names.

With this excellent source of information, we were able to develop several unsupervised naming methods of speakers in TV broadcasts. We started with late naming methods where names are propagated onto speaker clusters. These methods question differently the choices made during the diarization process. We then proposed two methods (integrated naming and early naming) that incorporate more information from written names during the speaker diarization process.

To identify people that appear on screen, we adapted the previously proposed early naming, for faces clusters. It is also shown that this method also works for multi-modal speakers-faces clusters.

Finally, we propose a unified approach which names speakers and faces during a single pass. This unified approach obtains performance comparable (or even better) to the best systems presented during the *REPERE* evaluation campaign (dedicated to person identification in videos).


FIN