Nommage non-supervisé des personnes dans les émissions de télévision : une revue du potentiel de chaque modalité

Johann Poignant, Laurent Besacier, Georges Quénot [1]

UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP / CNRS, LIG UMR 5217, Grenoble, F-38041, France

RÉSUMÉ. *L'identification de personnes dans les émissions de télévision est un outil précieux pour l'indexation de ce type de vidéos. Mais l'utilisation de modèles biométriques n'est pas une option viable sans connaissance a priori des personnes présentes dans les vidéos. Les noms cités à l'oral ou écrits à l'écran peuvent nous fournir une liste de noms hypothèses. Nous proposons une comparaison du potentiel de ces deux modalités (noms cités ou écrits) afin d'extraire le nom des personnes parlant et/ou apparaissant. Les noms cités à l'oral proposent un plus grand nombre d'occurrences de citation mais les erreurs de transcriptions et de détections de ces noms réduisent de moitié le potentiel de cette modalité. Les noms écrits à l'écran bénéficient d'une amélioration croissante de la qualité des vidéos et sont plus facilement détectés. L'affiliation aux locuteurs/visages des noms écrits reste plus simple que pour les noms cités à l'oral.*

ABSTRACT. *Persons identification in TV broadcast is a valuable tool for indexing these videos. But the use of biometric models is an unsustainable option without a priori knowledge of people present in the videos. The names pronounced or written on the screen can provide us a list of hypotheses names. We propose a comparison of the potential of these two modalities (names pronounced or written) to extract the true names of the speakers and/or faces. The names pronounced offer many instance of citation but transcription and detection errors of these names halved the potential of this modality. The names written benefits of the video quality improvement and there are easy to find. The affiliation to speakers/faces of names written is simpler than for names pronounced.*

MOTS-CLÉS : *Identification des personnes, multi-modalité, OCR, ASR*

KEYWORDS: *Person identification, multi-modality, OCR, ASR*

1. Ce travail a été en partie réalisé dans le cadre du programme Quaero et du projet QCompere, respectivement financés par OSEO et l'ANR.

1. Introduction

Avec l'augmentation grandissante du nombre de contenus audio-visuels disponibles de nos jours, l'identification automatique des personnes devient un outil très précieux pour la recherche et la navigation dans ce type de données. Cette identification peut par exemple s'appuyer sur la détection des visages ou la reconnaissance des locuteurs. Cependant, une identification biométrique des personnes parlant ou apparaissant dans des vidéos nécessite des annotations manuelles très coûteuses pour l'entraînement des modèles. Ces modèles doivent être adaptés aux conditions réelles acoustiques ou de prise d'images pour une meilleure efficacité.
Comme on ne peut pas considérer l'annotation manuelle de chaque nouvelle source vidéo comme une option viable, une alternative intéressante est l'utilisation des approches non supervisées pour nommer les personnes présentes dans les documents multimédias. À cette fin, nous pouvons classifier automatiquement chaque visage ou tour de parole en personnes anonymes (i.e. clustering des locuteurs et des visages) et utiliser une source d'information fournissant le vrai nom des personnes pour au moins une partie des clusters. Deux modalités intrinsèques à la vidéo, et plus particulièrement aux émissions de télévisions, sont capable de nous fournir les vrais noms des personnes apparaissant/parlant dans ces émissions : (i) les noms extraits de la transcription de la parole et (ii) les noms écrits à l'écran par la chaîne pour introduire une personne, c'est-à-dire les noms écrits dans un cartouche [1].
Dans la figure 1, nous pouvons voir un exemple extrait d'un journal télévisé avec le présentateur, une journaliste et une personne interviewée. Dans cet exemple, les flèches représentent les liens de citation des noms à l'oral ou des noms écrits avec les personnes correspondantes présentes dans la vidéo. Effectivement, dans cet exemple, il y a une corrélation entre la citation ou l'écriture d'un nom et la présence auditive ou visuelle de cette personne dans les tours de parole ou les scènes contiguës.

Figure 1 – Noms cités à l'oral et écrits à l'écran pour le nommage des personnes

Le nommage des personnes dans les émissions de télévisions en utilisant des systèmes automatiques peut répondre à plusieurs tâches :

Annotation : Automatique, elle peut aider/compléter/remplacer les annotations manuelles. Tâche orientée rappel et précision.
Création de modèles : l'extraction automatique de segments audio ou d'images de

1. Cartouche : position spatiale du texte utilisé par la chaîne pour écrire un nom en vue d'introduire la personne, correspondante au nom, à l'écran ou dans la bande son

visages peut permettre de construire des modèles biométriques de personnes. Tâche orientée précision pour obtenir les modèles les plus purs possibles, tout en ayant suffisamment de temps de signal pour générer ces modèles.
Recherche d'information : La réponse à une requête propose plusieurs segments où une personne est présente dans une vidéo. Tâche orientée sur la maximisation du rappel en nombre de personnes tout en ayant une précision importante.

Le sujet de cet article se concentre sur ce dernier point. Les noms cités à l'oral et écrits à l'écran apportant tous les deux des informations pertinentes pour répondre à cette question. Les travaux menés jusqu'à présent utilisaient principalement les noms cités à l'oral. Les noms écrits à l'écran étaient assez peu utilisables du fait de la mauvaise qualité des images et des systèmes de détection et de transcription. Mais l'évolution de la qualité des vidéos d'émissions de télévision disponibles doit nous faire réévaluer l'utilisation de cette modalité. Nous proposons donc une étude comparative des capacités d'utilisation des noms cités à l'oral et des noms écrits à l'écran pour l'identification des personnes (voix et/ou visages) dans les émissions de télévisions.

Cet article commencera par un tour d'horizon de la littérature portant sur le nommage des personnes dans les documents radios et vidéos et plus particulièrement sur les méthodes d'extraction des noms hypothèses (les méthodes de clustering de personnes et d'association noms-personnes étant en dehors du sujet de cet article). Ensuite nous poursuivrons par une présentation du corpus REPERE. Puis nous comparerons la qualité d'extraction des noms cités/écrits à l'aide de systèmes automatiques. Enfin, nous observerons le nombre de noms hypothèses correspondant aux personnes présentes dans les vidéos, avec une affiliation nom-personne quel que soit le moment de citation du nom ou avec une affiliation seulement dans les tours de parole/visages contigües au moment de citation.

2. État de l'art

La littérature concernant le nommage des personnes dans les émissions de télévision ou de radio est divisée entre deux communautés. La première s'est intéressée à la reconnaissance des visages. La seconde s'est concentrée sur l'identification du locuteur. La majorité des articles de l'état de l'art utilisent sensiblement le même cadre :

– Clustering des personnes (voix et/ou visages).

– Extraction des noms hypothèses pour chaque personne ou inversement.

– Association noms hypothèses/personnes.

Dans la suite, nous allons nous concentrer sur les méthodes d'extraction des noms hypothèses pour chaque personne. Les noms cités à l'oral sont utilisés majoritairement dans l'état de l'art du fait de la mauvaise qualité de transcription des noms écrits à l'écran. Pour utiliser les noms cités ou écrits, trois étapes sont nécessaires. Chacune de ces étapes peut être soit manuelle, soit automatique :

– Détection et transcription de la parole ou du texte écrit à l'écran.

– Détection des noms de personnes dans les transcriptions.

– Affiliations de chaque nom hypothèse aux personnes (voix, visages) candidates

Le système Name-It (Satoh *et al.,* 1997a) a le premier introduit dans la littérature le principe d'associer un nom et un visage apparaissant à l'écran basé sur leurs co-occurrences. Les noms sont détectés à l'aide d'un dictionnaire de noms dans les transcriptions manuelles de la parole. Dans cet article les auteurs attribuent une liste de trois noms classés en fonction d'un score à chacun des clusters et inversement (un nom pour trois visages). L'affiliation est effectuée entre un nom et les visages apparaissant à l'écran dans une fenêtre autour du moment de citation du nom. La redondance des co-occurrences permet d'avoir des scores d'association. Deux évolutions de ce travail ont été proposées par les auteurs. Dans (Satoh *et al.,* 1997b) ils extraient les noms des transcriptions de la parole manuelle à l'aide d'informations lexicales et grammaticales. Dans (Satoh *et al.,* 1999), ces mêmes auteurs utilisent en plus des noms cités à l'oral, les noms écrits à l'écran mais avec un taux d'erreur au mot de 52%. Cette modalité ne peut donc être que très peu utilisée. Houghton (Houghton, 1999) propose de construire une base de données de visages nommés. L'association nom-visage est effectuée à partir des noms écrits à l'écran extraits par un système de reconnaissance des caractères ; mais le taux d'erreur dans les transcriptions a obligé l'auteur à avoir recours à un dictionnaire de noms pour les corriger, ce qui ne lui permet pas de nommer les personnes hors dictionnaire. Pour parer cette difficulté, Yang et al. (Yang *et al.,* 2004) proposent d'utiliser à la fois les noms écrits à l'écran et les noms cités à l'oral comme dans (Satoh *et al.,* 1999) mais, avec l'utilisation d'un modèle d'affiliation des noms construit à partir d'une base annotée à la place des règles manuelles. Là aussi les auteurs ont utilisé une liste fermée de noms hypothèses ne permettant pas d'affilier un nom inconnu du dictionnaire à un visage. Dans (Liu *et al.,* 2008), les auteurs transforment le problème d'affiliation. Dans une première étape, ils extraient les noms de la transcription automatique de la parole et d'un système de vidéo OCR [2]. Une requête sur internet leur permet d'extraire des images de visages correspondants aux noms hypothèses et de construire des modèles biométriques à partir de ces images. Enfin, ils comparent les visages extraits de la vidéos aux modèles de personnes. Cette méthode peut fonctionner pour les visages de personnes connues mais moins pour les personnes peu connues sur internet (la requête peut retourner beaucoup de photos de visages erronés) ou pas pour une tâche de reconnaissance du locuteur. On peut aussi citer d'autres travaux (Everingham *et al.,* 2009), (Zhang *et al.,* 2009) portant sur les séries télévisées et les films où le script et les sous-titres sont disponibles mais pour les émissions de télévision, ces informations ne sont peu ou pas disponibles.

En ce qui concerne le nommage des locuteurs, les premiers travaux ont été proposés par Canseco et al. dans (Canseco-Rodriguez *et al.,* 2004) et (Canseco *et al.,* 2005). Les auteurs utilisent des patterns linguistiques définis manuellement pour déterminer à qui fait référence un nom cité : au locuteur courant ("Bonjour, ici John Chan"), suivant ("Nous allons écouter Candy Crowley") ou précédent ("Nous venons d'entendre Candy Crowley"). Tranter et al. (Tranter, 2006) vont remplacer les règles définies

2. OCR : reconnaissance optique des caractères (Optical Character Recognition)

manuellement par une phase d'apprentissage de séquences de n-grammes avec des probabilités associées. Mauclair et al.(Mauclair *et al.,* 2006) ont utilisé un arbre de classification sémantique pour associer un nom au locuteur précédent, courant, suivant ou à un autre locuteur. Des règles sont apprises automatiquement, à partir d'une base annotée, pour donner des probabilités d'affiliations entre un nom et les locuteurs contigus. Estève et al. (Estève *et al.,* 2007) ont fait une comparaison de ces deux techniques. Ils ont conclu que les arbres de classification sémantique sont moins sensibles que les séquences de n-grammes à l'utilisation de transcriptions automatiques de la parole. Jousse et al.(Jousse *et al.,* 2009) 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 locuteur). Ils ont aussi montré une réduction des performances de 19,5% à 70% (taux d'erreur d'identification des locuteurs, en nombre de locuteurs) lors de l'utilisation de transcriptions automatiques de la parole à la place de transcriptions manuelles. La grande majorité de ces travaux ont été appliqués sur des émissions de radio (corpus ESTER). Une des spécificités de la radio est que les personnes s'interpellent plus souvent par leurs noms que dans les émissions de télévision, ce qui facilite la tâche pour les émissions de radio. Plus récemment, dans (Poignant *et al.,* 2012b) nous avons proposé 3 méthodes de propagation des noms écrits sur les clusters de locuteurs. Ces méthodes non-supervisées, intrinsèquement multi-modales, obtiennent de bien meilleurs résultats qu'une solution supervisée mono-modale. Nous avons aussi montré que les méthodes d'affiliation automatiques des noms écrits aux clusters obtenaient une précision de 98,9% lorsque la segmentation en locuteurs est parfaite.

L'utilisation des noms cités à l'oral extraits par des méthodes automatiques fait face à plusieurs difficultés : (i) erreurs de transcription des noms inconnus : les systèmes de reconnaissance de la parole utilisent des dictionnaires de noms pour la transcription des entités nommées. Ils sont donc susceptibles de transformer un nom en un autre ou de faire des erreurs dans la transcription. (ii) Erreur de détection des noms dans les transcriptions : de la même manière les détecteurs d'entités nommées utilisent des dictionnaires de noms. Ils sont donc susceptibles de manquer des noms ou de détecter un nom lorsqu'il n'y en a pas ou encore d'ajouter/supprimer des mots aux noms (nom = "ici John Chan"). (iii) Erreurs d'affiliations : à quelle instance (locuteur ou/et visage) associer un nom ? A l'instance courante, suivante, précédente, à une autre ?
L'utilisation des noms écrits à l'écran extraits par des méthodes automatiques fait face aux mêmes difficultés : (a) erreurs de transcription : avec l'augmentation de la qualité des vidéos, elles sont moins nombreuses que les erreurs de transcription de la parole. (b) Erreurs de détection des noms dans les transcriptions : chaque émission utilisant un gabarit avec des emplacements spécifiques pour écrire les textes. La difficulté réside dans la détection des positions spatiales des cartouches. (c) Erreurs d'affiliations : généralement un nom est écrit à l'écran pendant que la personne est présente/parle, donc l'affiliation est beaucoup plus simple. Il reste une difficulté pour l'affiliation des visages. Si plusieurs visages sont présents, à quel visage affilier le nom ?

3. Corpus REPERE

Dans cette section, nous allons présenter le corpus REPERE, corpus sur lequel sera basée notre comparaison des deux modalités. Ce corpus est composé de 7 émissions différentes enregistrées sur deux chaînes de télévision françaises, BFMTV et LCP. Le détail des émissions est visible dans le tableau 2, ce sont des émissions de journaux télévisés, de débat, etc. Les enregistrements sont 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 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 cet article.

Pour le défi REPERE, ces vidéos ont été partiellement annotées, un ou plusieurs segments UEM[3] 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 toutes les dix secondes a été annotée manuellement. Pour chaque plan, l'image annotée a été choisie en essayant de maximiser le nombre d'informations contenues (nombre de visages et bonne orientation des visages, présence de texte…).
Pour la modalité audio, une transcription 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, ont été identifiés. Le texte en surimpression a été lui aussi détouré, transcrit et les noms de personnes ont été étiquetés. Si le nom est écrit dans un cartouche, un marquage supplémentaire a été ajouté.

Phase	Segment	Apprentissage	Développement	Évaluation
Dryrun	Vidéo complète	X	14h	13h
	Segment UEM	X	3h	3h
Phase1	Vidéo complète	58h	13h	15h
	Segment UEM	24h	3h	3h

Tableau 1 – Répartition du nombre d'heures du corpus REPERE sur les deux premières phases

Nous pouvons voir que les vidéos brutes sont plus longues que les segments annotés car des publicités ainsi que des émissions en dehors de celles ciblées peuvent être contenues dans les vidéos brutes. Ces vidéos brutes 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). On trouvera plus de détails sur le corpus REPERE et les annotations manuelles dans (Giraudel *et al.*, 2012).

3. Segments UEM : segments à traiter pour l'évaluation (Unpartitioned Evaluation Map)

Dans le tableau 2, on peut voir la répartition des émissions sur le corpus avec une grande disparité de la durée des segments UEM (2 minutes en moyenne pour Planète showbiz à 34 minutes pour BFM Story).

Émissions	Type d'émission	#Vidéos	Durée (en minutes)	
			Vidéo complète	Segment UEM
BFM Story	Journaux télévisés	14	854	478
Planète Showbiz	Actualités people	66	1019	120
Ça Vous Regarde	Débats	6	277	120
Entre Les Lignes	Débats	7	276	120
LCP Info	Journaux télévisés	15	378	247
Pile Et Face	Débats	9	303	120
Top Questions	Questions à l'Assemblée	18	396	238

Tableau 2 – Répartition du nombre de vidéos et de la durée des émissions sur le corpus de la phase 1, partie apprentissage

La figure 2 présente quelques exemples d'images extraites du corpus REPERE.

(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 2 – Exemples d'images 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.

Sur le corpus phase 1, partie apprentissage, 724 personnes différentes ont été identifiées par leur visage et 555 par leur voix, pendant l'annotation manuelle. 1907 des 11703 occurrences de visages n'ont pas pu être identifiées ainsi que 255 locuteurs. Ces locuteurs correspondent à 25 minutes de temps de parole sur les 1440 minutes annotées(466 tours de parole sur les 14782). Pour la suite de l'article, nous ne nous intéressons qu'aux personnes qui ont pu être nommées pendant l'annotation.

4. Systèmes automatiques d'extraction de noms

4.1. *Noms écrits à l'écran*

Pour détecter les noms écrits à l'écran introduisant une personne, nous avons besoin tout d'abord d'un système de détection et de transcription des textes surimposés à l'écran. Nous avons utilisé le système LOOV (Poignant *et al.*, 2012a) (LIG Overlaid OCR in Video), développé dans le cadre du défi lié au corpus REPERE. Ce système commence par une détection du texte en trois étapes. La première trouve les boîtes de textes candidates à l'aide d'une détection grossière sur toutes les images. Ensuite, les coordonnées sont affinées localement. Enfin, un suivi temporel supprime des fausses alarmes. Après une adaptation des images (augmentation de la résolution, binarisation des images…) pour un logiciel OCR standard (Tesseract de Google), une combinaison de plusieurs transcriptions pour une même boîte de texte permet d'augmenter la qualité de transcription. Ce système a été évalué sur un autre corpus de journaux télévisés avec des vidéos à basse résolution (352*288, MPEG1) avec un taux d'erreur en caractère de 4,6% pour tout type de texte et de 2,6% pour les noms écrits à l'écran.

A partir des transcriptions, nous utilisons une simple technique de détection des positions spatiales des cartouches. Cette technique compare chaque transcription avec une liste de noms de personnes célèbres (liste issue de Wikipedia, 175000 noms). A chaque fois qu'une transcription correspond à un nom célèbre, nous ajoutons sa position spatiale à une liste. Les positions récurrentes dans cette liste nous permettent de trouver 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 la transcription est un nom célèbre, de combien de mots la transcription est-elle composée…) nous permet de filtrer les transcriptions ne contenant pas qu'un nom (4779 boîtes de texte candidates, 1315 après filtrage, 11 n'auraient pas dû être filtrées, 13 auraient dû être filtrées).

Une correction est appliquée pour corriger les erreurs de transcription. Elle est basée sur une large liste de 175000 noms de personnes célèbres (issue de Wikipedia). Lorsque le ratio de la distance d'édition (entre 0 et 1) entre une transcription et un nom est supérieure à un 0,9, nous corrigeons le nom. Nous avons corrigé 207 noms avec seulement 4 corrections erronées.

4.2. *Noms cités à l'oral*

Nous avons utilisé les transcriptions fournies par le LIMSI à partir de leur système de transcription automatique de la parole (ASR[4]) utilisé pour le défi REPERE. Ce système (Gauvain *et al.,* 2002), à l'état de l'art, repose sur une technique de modélisation statistique. Après une segmentation et un clustering basé sur un mélange de gaussiennes, la modélisation acoustique est effectuée à l'aide de modèles de Markov cachés à densités continues et de statistiques n-grammes de mots pour le modèle de langage. La transcription est obtenue après plusieurs passes de décodage où chaque itération est utilisée pour l'adaptation des modèles acoustiques.

Ensuite un détecteur d'entités nommées (Dinarelli *et al.,* 2011) a été appliqué sur ces transcriptions obtenues automatiquement. Ce détecteur d'entités nommées utilise une structure arborescente à plusieurs niveaux, où les entités de base sont combinées pour en obtenir de plus complexes. Ces arbres sont construits à partir de CRF (conditional random fields) et d'approches basées sur l'analyse syntaxique.

4.3. *Comparaison de la qualité des systèmes*

Les résultats montrés dans la suite de cet article sont comptabilisés sur le corpus d'apprentissage de la phase 1 du défi REPERE. L'utilisation de LOOV et de cette technique de détection des noms écrits à l'écran nous permet d'obtenir 97,7% (cf. tableau 3) des noms écrits pour introduire une personne à l'écran, avec une précision de 95,7%. Les quelques erreurs restantes sont dues à des erreurs de transcription ou de filtrage. Le système qui traite la bande son engendre plus d'erreurs. La principale difficulté réside dans la transcription et la détection des noms qui ne font pas partie du dictionnaire initial. En effet nous travaillons sur une liste ouverte de personnes donc nous n'avons pas connaissance des noms qui pourraient être cités à l'oral.

Modalités	#Noms dans la référence	#Noms dans l'hypothèse	#Noms en commun	Précision	Rappel	F1-mesure
Noms écrits	1378	1407	1346	95,7%	97,7%	96,7%
Noms cités	4264	2905	2133	73,5%	50%	59,5%

Tableau 3 – Qualité de détection des noms écrits à l'écran et des noms cités à l'oral, phase 1, partie apprentissage, segments UEM

Malgré la précision et le rappel inférieurs des noms cités par rapport aux noms écrits à l'écran, cette modalité apporte plus de noms hypothèses (cf. tableau 4). Nous pouvons remarquer qu'il y a environ 50% de plus de noms cités à l'oral par rapport aux noms écrits à l'écran, que ce soit sur les vidéos complètes (avec le début et la fin de chaque émission) ou sur seulement les segments UEM. Cette proportion est respectée entre le nombre d'occurrences de citation des noms ou le nombre de personnes différentes citées.

4. ASR : Reconnaissance automatique de la parole (Automatic speech transcription)

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

Tableau 4 – Nombre de noms hypothèses, 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.

5. Noms cités ou écrits pour nommer les personnes présentes dans les vidéos

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

5.1. *Méthode de comptabilisation*

Le rappel a été comptabilisé avec une propagation intra-vidéos et inter-vidéos, nous définissons :

p : une personne
Pr : ensemble des p présentant ou parlant
Ph : ensemble des p ayant leurs noms écrits ou cités
Phr : ensemble des p ayant leurs noms écrits ou cités dans une vidéo et présents dans cette même vidéo

Le rappel intra et inter-vidéos pour une personne correspond à :

$$Rp_{intra} = \frac{\text{\#videos où } p \in Phr}{\text{\#videos où } p \in Pr} \qquad Rp_{inter} = \begin{cases} 1 \text{ Si } p \in Phr \\ 0 \text{ Sinon} \end{cases}$$

Le rappel intra-vidéo d'une personne p correspond au rapport du nombre de vidéos où le nom de p est écrit/cité lorsque p est présent par le nombre de vidéos où p est présent. Le rappel inter-vidéos de p est égal à 1 si, dans au moins une vidéo, le nom de p est écrit alors que p est présent dans cette vidéo, à 0 sinon.

Le rappel total intra et inter-vidéos, en pourcentage, correspond à la moyenne du rappel des personnes apparaissant/parlant dans le corpus :

$$R_{intra} = \frac{\sum_{p \in Pr} Rp_{intra}}{\#p \in Pr} \qquad R_{inter} = \frac{\sum_{p \in Pr} Rp_{inter}}{\#p \in Pr}$$

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

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 de la vidéo

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}	:	système automatique sur les segments UEM
A_{RAW}	:	système automatique sur les vidéos complètes
$N_{cités}$	:	noms cités à l'oral
$N_{écrits}$	:	noms écrits dans un cartouche à l'écran

5.2. *Personnes apparaissant ou parlant*

Dans les tableaux 5 et 6, 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}).

Le rappel des noms écrits dans les annotations manuelles est légèrement sous-évalué. Seulement une image toutes les 10 secondes ou au moins une par plan a été annotée (l'annotateur a pour consigne de choisir l'image dans le plan avec le plus d'informations). L'annotation ne porte donc pas sur tous les noms écrits, ce qui explique le rappel supérieur du système automatique par rapport aux annotations manuelles.

5.2.1. *Personnes apparaissant*

Le tableau 5 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, mais 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}	R_{intra}	R_{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

Tableau 5 – Nombre d'occurrences des noms et rappel en pourcentage des noms cités à l'oral et écrits à l'écran par rapport aux noms des 724 personnes apparaissant

La proportion (R_{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%). Mais, les erreurs dans les noms cités extraits automatiquement abaissent R_{intra} à 26,1%. La combinaison des noms écrits et cités augmente le rappel des personnes apparaissant, ce qui montre leur complémentarité, que ce soit avec les annotations manuelles (+27.9%) ou avec l'utilisation de systèmes automatiques (+10.2%). L'utilisation d'une propagation inter-vidéos augmente en moyenne le rappel R_{inter} de 7%.

5.2.2. *Personnes parlant*

Nous pouvons remarquer, dans le tableau 6, 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 apparait 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}	R_{intra}	R_{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

Tableau 6 – Nombre d'occurrences des noms et rappel en pourcentage des noms cités à l'oral et écrits à l'écran par rapport aux noms des 555 personnes parlant

Nous remarquons que, là aussi, l'utilisation conjointe des deux modalités augmente le rappel mais de façon moins importante que pour les personnes apparaissant (+19.6% pour M_{UEM}), surtout lors de l'utilisation des systèmes automatiques (+2.3% pour A_{UEM}). Une propagation inter-vidéos augmente aussi moins le rappel (+4% en moyenne) que dans le cas des personnes apparaissant.

5.3. *Détail par rôle des personnes*

Pour le corpus REPERE, cinq types de catégories différentes ont été définies pour classer les personnes (présentateur, chroniqueur, reporter, invité, autre). Au vu des résultats détaillés, nous avons fait un regroupement des catégories ayant un comportement similaire pour une meilleure lisibilitée. Les trois premières ont été regroupées dans le rôle R1 : Présentateur/journaliste, les deux dernières dans le rôle R2 : Invité/autre. Le tableau 7 détaille la répartition de présence des personnes dans les vidéos en fonction de leurs rôles. Un rôle a été affecté à chacune des personnes identifiées dans les vidéos, une personne pouvant avoir des rôles différents selon l'émission.

Rôle	#Personnes parlant	Temps de parole	#Tours de parole	#Personnes apparaissant	#Apparition à l'écran
R1	84 (15%)	632 (45%)	6149 (42%)	48 (7%)	2935 (30%)
R2	475 (85%)	783 (55%)	8378 (58%)	680 (93%)	6861 (70%)

Tableau 7 – Répartition de la présence des personnes en fonction de leurs rôles, corpus de la phase 1, partie apprentissage. **R1** : Présentateur/journaliste, **R2** : Invité/autre.

Seulement 48 des 84 personnes de R1 sont visibles et inversement 475 personnes de R2 parlent alors que 680 sont visibles. Les personnes de R1 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. Dans le tableau 8 nous détaillons les possibilités de nommer les personnes présentes en fonction du rôle qu'elles occupent dans les vidéos.

$N_{cités}$	$N_{écrits}$	Occ_{pv}		R_{intra}		R_{inter}	
		R1	R2	R1	R2	R1	R2
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

Tableau 8 – Nombre d'occurrences des noms et rappel en pourcentage des noms cités à l'oral ou écrits à l'écran des personnes apparaissant ou parlant en fonction de leurs rôles (R1 : 84 Présentateur/journaliste, R2 : 728 Invité/autre)

Nous pouvons voir que le nom des 84 présentateurs/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). Et tout particulièrement pour les noms cités de R1 extraits automatiquement. En effet, les présentateurs/journalistes sont souvent cités par leurs prénoms. De plus, leurs noms sont difficiles à transcrire parce que inconnus du système automatique. Les personnes du rôle R1 sont donc assez difficiles à nommer automatiquement alors qu'elles représentent 45% du temps de parole et 30% des visages apparaissant.

Concernant le rappel, 78.9% des personnes de R1 ont leurs noms cités mais seulement 13.9% d'entre elles ont pu être nommés par les systèmes automatiques. Les personnes de R2 sont plus nommées (noms écrits ou cités), dans les sorties automatiques, que les personnes de R1 ($N_{écrits}$ +8.5%, $N_{cités}$ +10.8%). L'utilisation conjointe des deux modalités permet d'augmenter le rappel et le nombre d'occurrences quel que soit le type de rôle pris en compte ou la propagation à d'autres vidéos.

5.4. *Apport de l'utilisation des vidéos complètes*

L'utilisation des vidéos complètes (A_{RAW}) 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 rappel sur le nombre de personnes présentes dans les segments UEM (A_{UEM}=55,1% à A_{RAW}=56,7%). Par contre, ce nombre d'occurrences supplémentaires peut faciliter l'association nom-personne.

$N_{cités}$	$N_{écrits}$	Occ	Occ_{pv}	R_{intra}	R_{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

Tableau 9 – Apport des vidéos complètes, pour le nommage des 808 personnes apparaissant 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.

5.5. *Affiliation des noms hypothèses aux personnes à l'aide d'un oracle*

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 quelque soit le moment où elle apparaissait/parlait dans la vidéo. Mais les systèmes de l'état de l'art se restreignent aux tours de parole contigus pour effectuer l'affiliation d'un nom à une personne. Dans cette section, nous allons comparer la capacité d'affiliation des noms écrits ou cités aux bonnes personnes à l'aide d'un oracle. Pour les noms écrits, l'oracle choisira d'affilier les bons noms aux bonnes personnes si ils sont présents

dans le tour de parole courant. Pour les noms cités à l'oral, le choix se fera sur les personnes apparaissant/parlant dans le tour de parole précédent, courant ou suivant.

$N_{cités}$	$N_{écrits}$	Occ	Occ_{pv}	R_{intra}	R_{inter}
M_{UEM}	X	4273	1580 (37,0%)	51,8	58,9
X	M_{UEM}	1049	977 (93,1%)	38,4	45,9
M_{UEM}	M_{UEM}	5322	2557 (48,0%)	64,6	71,4
A_{UEM}	X	2905	632 (21,8%)	20,9	26,4
X	A_{UEM}	1407	1269 (90,2%)	45,4	52,5
A_{UEM}	A_{UEM}	4312	1901 (44,1%)	53,6	60,5

Tableau 10 – Rappel des 808 personnes parlant ou apparaissant après affiliation des noms à l'aide d'un oracle. Segments UEM.

Les données de ce tableau 10 sont à comparer avec celles du tableau 9. On peut remarquer que lorsque l'on restreint l'affiliation des noms cités aux personnes présentes dans les tours de parole adjacents, le rappel réduit de 2,8% à 5,9% selon le système et la propagation utilisée, alors qu'il n'y a que très peu ou pas de différence pour les noms écrits (réduction de 0,1% à 1,0%). Malgré cette réduction, ce tableau permet de montrer la complémentarité de ces deux modalités. Quel que soit la propagation (intra-vidéo ou inter-vidéos) ou l'annotation utilisée (manuelle ou automatique), le rappel des personnes à l'aide des deux modalités est toujours plus important que celui des modalités seules.

6. Conclusion

Les noms cités à l'oral et écrits à l'écran sont des sources d'informations importantes pour obtenir les noms des personnes présentes dans les émissions de télévision. Malgré un plus grand nombre d'occurrences de citation des noms des personnes présentes dans les transcriptions manuelles de la parole par rapport aux noms écrits, les erreurs de détection et de transcription des systèmes automatiques réduisent le rappel obtenu pour cette modalité. L'amélioration de la qualité de transcription des noms écrits nous permet d'obtenir un rappel deux fois supérieur aux noms cités. Les noms écrits sont principalement utilisés pour introduire une personne apparaissant et parlant en même temps alors que les noms cités à l'oral peuvent aussi introduire des journalistes parlant en voix-off ou encore des personnes apparaissant mais ne parlant pas. Un dernier point à noter est que l'affiliation des noms écrits aux bonnes personnes est intrinsèquement plus simple que pour les noms cités. Malgré ces différences de résultats, ces deux modalités restent très complémentaires. Des méthodes de propagation non-supervisées doivent donc être développées pour réduire encore le besoin en annotations manuelles et faciliter l'indexation des émissions de télévision.

7. Bibliographie

Canseco L., Lamel L., Gauvain J.-L., « A Comparative Study Using Manual and Automatic Transcriptions for Diarization », *ASRU*, November, 2005.

Canseco-Rodriguez L., Lamel L., Gauvain J.-L., « Speaker diarization from speech transcripts », *INTERSPEECH*, 2004.

Dinarelli M., Rosset S., « Models Cascade for Tree-Structured Named Entity Detection », *IJCNLP*, 2011.

Estève Y., Meignier S., Deléglise P., Mauclair J., « Extracting true speaker identities from transcriptions », *INTERSPEECH*, p. 2601-2604, 2007.

Everingham M., Sivic J., Zisserman A., « Taking the Bite out of Automatic Naming of Characters in TV Video », *Image and Vision Computing*, 2009.

Gauvain J., Lamel L., Adda G., « The LIMSI Broadcast News Transcription System », *Speech Communication*, vol. 37, n° 1-2, p. 89-109, 2002.

Giraudel A., Carré M., Mapelli V., Kahn J., Galibert O., Quintard L., « The REPERE Corpus : a Multimodal Corpus for Person Recognition », *LREC*, 2012.

Houghton R., « Named Faces : Putting Names to Faces », *IEEE Intelligent Systems*, vol. 14, p. 45-50, 1999.

Jousse V., Petit-Renaud S., Meignier S., Estève Y., Jacquin C., « Automatic named identification of speakers using diarization and ASR systems », *ICASSP*, p. 4557-4560, 2009.

Liu C., Jiang S., Huang Q., « Naming faces in broadcast news video by image google », *ACM Multimedia*, p. 717-720, 2008.

Mauclair J., Meignier S., Estève Y., « Speaker diarization : about whom the speaker is talking ? », *IEEE Odyssey 2006*, 2006.

Poignant J., Besacier L., Quénot G., Thollard F., « From Text Detection in Videos to Person Identification », *IEEE ICME*, 2012a.

Poignant J., Bredin H., Le V., L.Besacier, C.Barras, G.Quénot, « Unsupervised Speaker Identification using Overlaid Texts in TV Broadcast », *Interspeech 2012*, 2012b.

Satoh S., Kanade T., Race T. P., « Name-It : Association of Face and Name in Video », *CVPR*, 1997a.

Satoh S., Nakamura Y., Kanade T., « Name-It : Naming and Detecting Faces in Video by the Integration of Image and Natural Language Processing », *IJCAI*, p. 1488-1495, 1997b.

Satoh S., Nakamura Y., Kanade T., « Name-It : Naming and Detecting Faces in News Videos », *IEEE Multimedia*, vol. 6, p. 22-35, 1999.

Tranter S. E., « WHO REALLY SPOKE WHEN ? FINDING SPEAKER TURNS AND IDENTITIES IN BROADCAST NEWS AUDIO », *ICASSP*, 2006.

Yang J., Hauptmann A. G., « Naming every individual in news video monologues », *ACM Multimedia*, p. 10-16, 2004.

Yang J., Yan R., Hauptmann A. G., « Multiple instance learning for labeling faces in broadcasting news video », *ACM Multimedia*, p. 31-40, 2005.

Zhang Y.-F., Xu C., Lu H., Huang Y.-M., « Character identification in feature-length films using global face-name matching », *Journal IEEE Transactions on Multimedia*, vol. 11, n° 7, p. 1276-1288, November, 2009.