Nommage non supervisé des personnes dans les émissions de télévision. Utilisation des noms écrits, des noms prononcés ou des deux ?

Johann Poignant, Laurent Besacier, Georges Quénot

Nommage non supervisé des personnes dans les émissions de télévision

Utilisation des noms écrits, des noms prononcés ou des deux ?

Johann Poignant [1,2], **Laurent Besacier** [1,2], **Georges Quénot** [2,1]

[1] *Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France*
[2] *CNRS, LIG, F-38000 Grenoble, France*

Prénom.Nom@imag.fr

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 prononcés ou écrits peuvent nous fournir une liste de noms hypothèses. Nous proposons une comparaison du potentiel de ces deux modalités (noms prononcés ou écrits) afin d'extraire le nom des personnes parlant et/ou apparaissant. Les noms prononcés proposent un plus grand nombre d'occurrences de citation mais les erreurs de transcription et de détection de ces noms réduisent de moitié le potentiel de cette modalité. Les noms écrits bénéficient d'une amélioration croissante de la qualité des vidéos et sont plus facilement détectés. Par ailleurs, l'affiliation aux locuteurs/visages des noms écrits reste plus simple que pour les noms prononcés.*

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 they are easy to find. Moreover, the affiliation to speakers/faces of names written is simpler than for names pronounced.*

MOTS-CLÉS : *identification des personnes, multi-modalité, ROC*[1]*, RAP.*[2]

KEYWORDS: *person identification, multi-modality, OCR*[1]*, ASR*[2]*.*

---

1. ROC : Reconnaissance optique des caractères, OCR : optical caracter recognition
2. RAP : Reconnaissance automatique de la parole, ASR : automatic speech recognition

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. Elle peut par exemple s'appuyer sur la reconnaissance des visages ou des locuteurs. Cependant, la construction de modèles biométriques des personnes parlant ou apparaissant dans des vidéos nécessite des annotations manuelles très coûteuses pour l'entraînement de ces modèles. De plus, ils 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. Cette solution peut répondre à plusieurs besoins :
- **L'extraction automatique d'échantillons de parole ou d'images de visages** pour entraîner des modèles biométriques. C'est une tâche « orientée précision », dont le but est d'obtenir des signaux étiquetés précisément et en quantité suffisante pour construire des modèles biométriques.
- **L'identification automatique des visages/voix** pour l'annotation complète de vidéos peut remplacer l'annotation manuelle ou minimiser l'effort de post-annotation.
- **L'indexation automatique** pour proposer plusieurs extraits vidéos en réponse à une requête. La liste retournée doit être un équilibre entre précision et exhaustivité.

Les approches d'identification sont très dépendantes des paramètres suivants :
- **Tâche** : Qui parle et/ou qui apparaît ?
- **Média à cibler** : est-ce un flux télévisé ? Une vidéo de fiction ? Autre ?
- **Granularité** : doit-on regarder au niveau du plan, de la vidéo ou de la collection ?
- **Sources de noms utilisées** : sous-titres, noms prononcés ou écrits, externes, etc.
- **Annotations préexistantes** : oui (manuelle), non (système automatique).
- **Rôle des personnes** : présentateur, journaliste, invité, interviewé, etc.

Nous nous sommes intéressés plus particulièrement aux flux télévisés, parce qu'ils contiennent une variété de personnes à identifier importante. Ces personnes peuvent apparaître dans une ou plusieurs vidéos et dans des contextes variés (journaux, débats, etc.).

Pour répondre aux deux questions, la majorité des travaux de l'état de l'art sont effectués dans le même cadre, décrit dans la figure 1, avec quelques adaptations inhérentes aux sources d'informations utilisées.

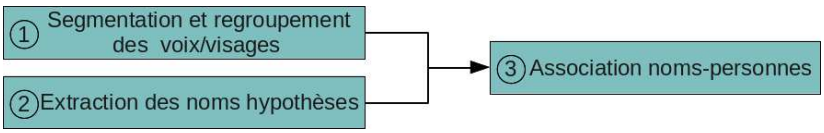

Figure 1. Cadre général majoritairement utilisé dans l'état de l'art pour l'annotation automatique des personnes

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 tours de parole et/ou de visages. Idéalement un cluster correspond à une personne et une personne par cluster.

Plusieurs sources de noms sont issues d'annotations manuelles coûteuses, comme les sous-titres. Elles sont à écarter parce que c'est justement le coût d'annotation que nous voulons éviter. Par contre, deux autres sources intrinsèques aux flux télévisés sont disponibles : les noms prononcés dans la piste audio et les noms écrits en surimpression dans la piste image, en général dans un cartouche[3]. La figure 2 montre un extrait de journal télévisé avec le présentateur, un journaliste et une personne interviewée. Une corrélation est visible entre la prononciation ou l'écriture d'un nom et la présence auditive ou visuelle de cette personne.

Figure 2. Noms prononcés et noms écrits pour le nommage des personnes

La dernière étape du cadre général, 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 dans les plans adjacents ou à aucune personne présente. Un nom issu du guide des programmes télévisés fait référence à une personne présente dans la vidéo sans spécifier le moment où elle est présente. C'est donc cette étape qui va choisir le lien entre un nom et une personne (les flèches dans notre exemple).

L'illustration de la figure 2, 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. Une personne n'est en effet 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 pallier ce manque.

3. Cartouche : positions spatiales de l'image utilisées par l'émission pour écrire un nom en vue d'introduire la personne correspondante dans la piste image ou dans la piste audio.

– Le nom prononcé « Nathalie Koscuisko-Morizet » n'introduit pas la personne correspondante. L'étape d'association noms-personnes doit, en plus de choisir les bons liens, choisir si un nom hypothèse doit être utilisé ou non.

Les noms prononcés et les noms écrits apportent donc tous les deux des informations pertinentes pour répondre aux deux questions : « Qui parle et/ou qui apparaît ? ». Les travaux menés jusqu'à présent utilisaient principalement les noms prononcés. Les noms écrits étaient assez peu utilisables du fait de la mauvaise qualité des images et des systèmes de détection et de transcription. Mais l'évolution récente 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 prononcés et des noms écrits pour l'identification des personnes (voix et/ou visages) dans les émissions de télévision. Cet article est une extension de (Poignant, Besacier, Quénot, 2013b; Poignant, Besacier, Le *et al.*, 2013a).

Nous commençons par un tour d'horizon de la littérature portant sur le nommage des personnes dans les documents radios et vidéos. Ensuite nous poursuivons par une présentation du corpus REPERE sur lequel sera basé notre comparaison. Puis nous comparons la qualité d'extraction des noms prononcés/écrits à l'aide de systèmes automatiques. Enfin, nous observons le nombre de noms hypothèses correspondant aux personnes présentes dans les vidéos, avec une association noms-personnes réalisée par un oracle quel que soit le moment de citation du nom ou seulement dans les tours de parole/visages dans les plans contigus au moment de la citation. Cette comparaison porte sur l'utilisation des noms extraits automatiquement avec des méthodes à l'état de l'art, mais aussi sur des noms extraits manuellement pour connaître les capacités de ces deux modalités.

2. État de l'art

La littérature concernant le nommage des personnes dans les émissions de télévision ou de radio a d'abord été divisée entre deux communautés. La première s'est intéressée à la reconnaissance des visages. La seconde s'est concentrée sur l'identification du locuteur. Ces deux dernières années plusieurs travaux ont tenté de répondre aux deux questions en même temps. Les noms prononcés ont été utilisés majoritairement dans la littérature du fait de la mauvaise qualité de transcription des noms écrits.

2.1. Identification de « qui parle » à l'aide des noms prononcés

Pour le nommage des locuteurs, les premiers travaux ont été proposés par Canseco *et al.* dans (Canseco-Rodriguez *et al.*, 2004), (Canseco *et al.*, 2005) et (Charhad *et al.*, 2005). Les auteurs utilisent des patrons linguistiques définis manuellement pour déterminer à qui fait référence un nom prononcé : au locuteur courant (« Bonjour, ici John Chan »), suivant (« Nous allons écouter Candy Crowley ») ou précédent (« Nous venons d'entendre Candy Crowley »). Cet article montre qu'un nom prononcé peut désigner plusieurs locuteurs, l'étape d'association doit donc choisir la bonne association

pour cette source de noms. Tranter (2006) va remplacer les règles définies manuellement par une phase d'apprentissage de séquences de n-grammes avec des probabilités associées.

Mauclair *et al.* (2006) ont utilisé un arbre de classification sémantique (SCT) 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.* (2007) ont fait une comparaison entre les SCT et les n-grammes. Ils en 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.* (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 locuteurs). Ils ont aussi montré une augmentation du taux d'erreur d'identification des locuteurs de 19,5 % à 70 % (en nombre de locuteurs) lors de l'utilisation de transcriptions automatiques de la parole à la place de transcriptions manuelles.

En 2010, Petitrenaud *et al.*, dans (PetitRenaud *et al.*, 2010a; Petitrenaud *et al.*, 2010b), 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.* (2012) ont utilisé ces fonctions de croyance avec des briques de base automatiques et/ou manuelles (diarization, transcription et détection des entités nommées). De plus, ils ont ajouté des informations issues des scores d'un système de reconnaissance du locuteur à base de modèles biométriques GMM-UBM. Ces scores ont été transformés en fonctions de croyance pour s'intégrer aux précédents travaux. Ils ont remarqué une augmentation du taux d'erreur 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 32,7 %.

2.2. Identification de « qui apparaît » à l'aide des noms prononcés et des noms écrits

Le système Name-It (Satoh, Kanade, 1997a) a été le premier à introduire dans la littérature le principe d'associer un nom (extrait à l'aide d'un dictionnaire dans les transcriptions manuelles de la parole) et un visage visible sur la base de leurs co-occurrences. L'affiliation est effectuée entre un nom et les visages visibles dans une fenêtre autour du moment de citation du nom. La redondance des cooccurrences per-

met 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 prononcés, les noms écrits mais avec un taux d'erreur au mot de 52 %. Cette modalité ne peut donc être que très peu utilisée.

Houghton (1999) propose de construire une base de données de visages nommés. L'association nom-visage est effectuée à partir des noms écrits extraits par un système de reconnaissance des caractères. Mais le taux d'erreur de mots de 65 % dans les transcriptions a obligé l'auteur à avoir recours à un dictionnaire de noms pour les corriger, ce qui réduit le taux d'erreur à 45 %. Ce taux encore élevé restreint toujours l'utilisation de cette modalité. En outre, cela ne lui permet pas de nommer des personnes « hors dictionnaire ».

Pour parer cette difficulté, Yang *et al.* (2004) proposent d'utiliser à la fois les noms écrits et les noms prononcés comme dans (Satoh *et al.*, 1999) mais, avec l'utilisation d'un modèle d'association noms-visages 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.

D'autres travaux (Song *et al.*, 2004; Yang *et al.*, 2005) ont préféré adapter un classifieur MIL[4](Dietterich *et al.*, 1997) pour trouver à quel visage associer un nom prononcé, mais cette méthode n'est applicable que pour les noms dont on sait que les personnes correspondantes apparaissent dans les vidéos.

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. Ensuite, une requête sur Internet leur permet d'extraire des images de visages correspondant aux noms hypothèses et de construire des modèles biométriques à partir de ces images. Enfin, ils comparent les visages extraits des 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) et plus difficilement pour une tâche de reconnaissance du locuteur.

2.3. Utilisation des noms écrits extraits avec très peu d'erreurs

L'augmentation de la qualité des vidéos ces dernières années 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 (extraits à l'aide de l'outil LOOV (Poignant, Besacier *et al.*, 2012a)) dans les émissions de télévision. Nous avons proposé plusieurs méthodes de propagation des noms écrits sur des clusters de locuteurs (Poignant, Bredin *et al.*, 2012b). Ce qui nous

4. MIL : multiple instance learning.

a permis d'obtenir une précision de 98,9 % lorsque la segmentation en locuteurs est parfaite et de 89,1 % à 91 % (selon la méthode) pour une diarization automatique, avec un rappel supérieur à 70 %. Nous avons ensuite étendu ces travaux avec l'utilisation des noms prononcés, remplacé l'étape de diarization par un clustering PLNE, effectué un regroupement des tours de parole contraint par l'information des noms écrits et adapté certaines méthodes pour l'identification des visages (Bredin *et al.*, 2012 ; 2013b; Poignant, Bredin *et al.*, 2013c; Bredin, Poignant, 2013a).

En 2013, Bendris *et al.* (2013) ont proposé d'utiliser les noms écrits et les noms prononcés pour identifier qui apparaît dans les émissions de télévisions. Ils ont d'abord nommé les visages et les tours de parole. Puis, à partir de règle de priorité, ils ont propagé d'abord les noms écrits puis les noms prononcés mais aussi le nom des tours de parole identifiés vers les visages détectés comme parlant.

2.4. Étude récente sur la capacité de nommage des noms prononcés ou écrits

Une étude intéressante, proposée par Bechet *et al.* (2012), 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, voir section 3), 72 % des personnes ont leur nom prononcé. Parmi les 717 noms prononcés, 447 correspondent à une personne présente (précision de 62 %). Pour intégrer cette source d'informations dans un processus de décision multi-modale plus complet, les auteurs ont proposé de diminuer le nombre de noms utilisés pour augmenter la précision. Un classifieur adaboost (Schapire, Singer, 1999) 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, etc.

– 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 ? etc.).

Ces deux types d'informations ont permis au classifieur de sélectionner les noms ayant la plus grande probabilité de correspondre à une personne présente. Ce qui permet de réduire le nombre de noms inutiles (augmentation de la précision de 62 % à 68 %) mais qui a aussi entraîné 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 concernant la présence de la personne correspondant au nom : positive, négative ou incertaine. 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 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.

2.5. Bilan

Le premier enseignement que l'on peut tirer est que la qualité de l'association noms-clusters dépend de la source utilisée pour les noms. Moins cette source est sûre, plus la méthode d'association devra s'affranchir de cette incertitude.

Pour réaliser correctement l'association entre noms prononcés et personnes, beaucoup de travaux ont cherché à prendre en compte l'incertitude liée à cette modalité (erreurs de transcription, erreurs de détection des noms dans ces transcriptions, est ce qu'un nom correspond bien à une personne présente et si oui 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.* (2012) 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.

Pour les noms écrits, les trop nombreuses erreurs de transcription ont rendu difficile leur utilisation par le passé. Cependant, l'amélioration de la qualité des vidéos ces dernières années permet de les extraire désormais avec un très faible taux d'erreur sur les mots. Un dernier point à noter est que l'utilisation de cette modalité réduit l'incertitude liée à l'association d'un nom (un nom écrit correspond au locuteur courant et à un des visages visibles au moment de l'écriture du nom).

Bien que de nombreux travaux aient utilisé l'une ou l'autre de ces sources, voire les deux, aucune comparaison objective n'a été effectuée sur leur capacité à proposer les noms des personnes parlant ou apparaissant dans les flux télévisés.

3. Corpus REPERE

Le corpus REPERE (Giraudel *et al.*, 2012), sur lequel sera basée notre comparaison, a été constitué pour le défi du même nom qui s'intéresse à l'identification des personnes dans les flux télévisés. Il est composé de sept types d'émissions différentes incluant des journaux télévisés, des débats, etc. Les enregistrements ont été effectués au cours des années 2011 et 2012, au format 720×576 en MPEG-2, sur deux chaînes de télévision française (BFM et LCP). La très bonne qualité de ces enregistrements nous permet d'utiliser les textes écrits apparaissant dans la piste image. Ce corpus a été constitué en trois 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.

Tableau 1. Répartition des vidéos du corpus REPERE (phase 0 et phase 1)

Phase	Segment	Apprentissage	Développement	Évaluation
Phase 0	Vidéo complète	X	14 h.	13 h.
	Segment UEM	X	3 h.	3 h.
Phase 1	Vidéo complète	58 h.	13 h.	15 h.
	Segment UEM	24 h.	3 h.	3 h.

Pour le défi REPERE, ces vidéos ont été partiellement annotées, un ou plusieurs segments UEM[5] 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, 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 était supérieure à 2 000 pixels^2, 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.

Les vidéos brutes incluent des publicités ainsi que des émissions en dehors de celles ciblées. Ce temps de signal supplémentaire peut permettre d'extraire plus d'occurrences d'un nom ou encore d'avoir des noms peu souvent prononcés (les présentateurs sont souvent cités seulement en début de journal). On trouvera plus de détails sur le corpus REPERE et les annotations manuelles dans (Giraudel *et al.*, 2012).

Les résultats montrés par la suite sont comptabilisés sur l'ensemble d'apprentissage de la phase 1 du défi *REPERE* parce qu'il est le plus volumineux du corpus et est donc le plus significatif. Le tableau 2 résume la répartition des émissions sur cet ensemble. On peut voir une grande disparité de la durée des segments UEM (2 minutes en moyenne pour Planète showbiz à 34 minutes pour BFM Story).

La figure 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 des quatre personnes visibles ont été identifiées alors qu'une seule est le sujet principal de l'image.

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

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

Émissions	Type d'émission	#Vidéos	Durée (en minutes)	
			Vidéo complète	Segment UEM
BFM Story	Journaux télévisés	14	854	478
Planète Showbiz	Actualités people	66	1 019	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

(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 3. Exemples d'images du corpus REPERE

Sur le corpus phase 1, partie apprentissage, 724 personnes différentes ont été identifiées par leurs visages et 555 par leurs voix, pendant l'annotation manuelle. 1 907 des 11 703 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 1 440 minutes annotées (466 tours de parole sur les 14 782). 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

Pour détecter les noms écrits introduisant une personne, nous avons tout d'abord besoin d'un système de détection et de transcription des textes surimposés dans la piste image. Nous avons utilisé le système LOOV (LIG Overlaid OCR in Video (Poignant, Besacier *et al.*, 2012a)), 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, MPEG-1) avec un taux d'erreur en caractère de 4,6 % pour tout type de texte et de 2,6 % pour les noms écrits.

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, 175 000 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 (4 779 boîtes de texte candidates, 1 315 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 175 000 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 supérieur à 0,9, nous corrigeons le nom. Nous avons corrigé 207 noms avec seulement 4 corrections erronées.

4.2. Noms prononcés

L'extraction des noms prononcés a été effectuée par le LIMSI avec d'abord leur système de transcription de la parole française (Lamel *et al.*, 2011). Il utilise la même technique statistique de modélisation et de décodage que le système « LIMSI English BN » (Gauvain *et al.*, 2002). 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 (Gauvain *et al.*, 1998). Ensuite,

un regroupement des segments en clusters est effectué, où idéalement un cluster représente une personne. 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 36 000 mots) pendant la première campagne d'évaluation du défi *REPERE*. Il est à noter qu'aucun modèle acoustique ou modèle de langage n'a été re-entraîné sur les données du corpus *REPERE*.

Sur ces transcriptions, le LIMSI a appliqué deux systèmes basés sur leurs expériences au sein du projet *Quaero* (Marco, Rosset, 2011). Le premier utilise des modèles de CRF spécifiques à l'aide de Wapiti (Lavergne *et al.*, 2010) entraînés 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).

Ils utilisent les mêmes caractéristiques que celles de (Marco, Rosset, 2011) :

– Un ensemble de fonctions standards comme les mots préfixes et suffixes de longueur de 1 à 5, ainsi que certains paramètres (*Est ce que le mot commence par une majuscule? Est ce que le mot ne contient pas de caractère alphanumérique?*, etc.).

– Des caractéristiques morphosyntaxiques (outil *tagger* (Allauzen, Bonneau-Maynard, 2008)).

– Des caractéristiques extraites à partir de la sortie d'un analyseur multi-niveaux, utilisées dans un système de questions-réponses (Bernard *et al.*, 2009), qui contiennent des informations morphosyntaxiques détaillées ainsi que des informations sémantiques au même niveau que les entités nommées.

4.3. 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 la section dédié au corpus, 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 d'occurrences de noms dans la référence soit supérieur à celui indiqué dans les tableaux du reste de l'article (où nous ne compterons que les apparitions uniques : 1 378 dans le tableau 3 et 1 049 pour la deuxième ligne des tableaux 5, 6 et 9).

– Le nombre de noms écrits extraits automatiquement est supérieur à ceux de la référence (1 407 sur les segments UEM, 2 090 sur les vidéos complètes, voir tableau 4)

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 prononcés, l'annotation manuelle complète des segments UEM permet d'éviter ce type de différence. L'évaluation du tableau 3 porte donc sur l'ensemble du signal des segments UEM.

Nous obtenons 98,1 % (cf. tableau 3) des noms écrits (pour introduire une personne visible) avec une précision de 98,5 % (sur les images annotées). 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).

Tableau 3. Qualité d'extraction des noms écrits et des noms prononcés selon l'annotation sur la partie apprentissage du corpus REPERE

Modalités	Signal annoté	#Occ de noms dans la réf	Précision	Rappel	F1-mesure
Noms écrits	1 image / 10 sec	1 378	98,5 %	98,1 %	98,3 %
Noms prononcés	Segments UEM	4 264	73,5 %	50 %	59,5 %

Les scores inférieurs pour les noms prononcés sont dus aux erreurs :

– De transcription de la parole.

– De détection des noms proches de mots de la langue courante.

– 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 ces performances plus faibles, l'extraction des noms prononcés produit plus d'occurrences (c.f. tableau 4). En effet, nous pouvons remarquer qu'il y a environ 50 % de plus de noms prononcés que de noms écrits ; 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.

Tableau 4. Nombre d'occurrences de noms extraits automatiquement sur la partie apprentissage du corpus REPERE

Modalités	Segments	#Occ de noms	#Personnes sans doublon
Noms écrits	Segments UEM	1 407	458
	Vidéos complètes	2 090	629
Noms prononcés	Segments UEM	2 905	736
	Vidéos complètes	4 922	1 156

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

5. 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 capacité d'une source à proposer le nom des personnes présentes ; indépendamment de la difficulté d'association noms-personnes liée à cette modalité. La seconde, 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.

5.1. Proportion de personnes nommables

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

$$Np_{intra} = \frac{\text{\#videos où } p \in Phr}{\text{\#videos où } p \in Pr} \tag{1}$$

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 :

$$Np_{inter} = \begin{array}{l} 1 \text{ Si } p \in Phr \\ 0 \text{ Sinon} \end{array} \tag{2}$$

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-vidéo et inter-vidéos est égal à :

$$N_{intra} = \frac{\sum_{p \in Pr} Np_{intra}}{\#p \in Pr} \tag{3}$$

$$N_{inter} = \frac{\sum_{p \in Pr} Np_{inter}}{\#p \in Pr} \tag{4}$$

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).

5.2. Nombre d'occurrences de citation d'un nom

En plus du rappel, nous allons aussi comptabiliser le nombre d'occurrences des noms écrits/prononcés (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 prononcés et/ou écrits
Occ_{pv}	:	nombre d'occurrences des noms prononcés 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 prononcés
$N_{écrits}$	:	noms écrits dans un cartouche dans la piste image

6. Noms prononcés ou écrits pour nommer les personnes présentes dans les vidéos

Comme nous avons pu le voir, les noms prononcés 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.

6.1. 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 ($N_{écrits}$) 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.

6.1.1. Personnes apparaissant

Le tableau 5 présente la proportion de personnes apparaissant dont le nom a été prononcé/écrit ainsi que le nombre d'occurrences de ces noms. Dans les annotations manuelles, il y a plus d'occurrences de noms prononcés (4 273) que de noms écrits (1 049). Par contre, lorsqu'un nom est écrit dans une vidéo, dans 99,1 % des cas, la personne correspondant au nom est visible à un moment ou à un autre de la vidéo. Cette proportion est plus faible pour les noms prononcés (60,3 %).

L'utilisation de systèmes automatiques sur les segments UEM réduit le nombre (Occ) de noms prononcés de 4 273 à 2 905. Or, seulement 1 435 occurrences (49,4 %) de ces noms correspondent à des personnes visibles. L'utilisation conjointe des noms prononcés 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 2 767.

Tableau 5. Nombre d'occurrences des noms et pourcentages des 724 personnes apparaissant nommables par les noms prononcés et/ou écrits

$N_{cités}$	$N_{écrits}$	Occ	Occ_{pv}	N_{intra}	N_{inter}
M_{UEM}		4 273	2 577 (60,3 %)	59,1	66,2
	M_{UEM}	1 049	1 040 (99,1 %)	44,0	51,9
M_{UEM}	M_{UEM}	5 322	3 617 (68,0 %)	71,9	78,5
A_{UEM}		2 905	1 435 (49,4 %)	26,1	31,9
	A_{UEM}	1 407	1 332 (94,7 %)	49,5	57,0
A_{UEM}	A_{UEM}	4 312	2 767 (64,2 %)	59,7	66,3

La proportion (N_{intra}) des personnes visibles dont le nom a été prononcé 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 prononcés extraits automatiquement abaissent N_{intra} à 26,1 %. La combinaison des noms écrits et prononcés 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 %.

6.1.2. Personnes parlant

Nous pouvons constater, 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 apparaît en même temps.

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

Tableau 6. Nombre d'occurrences des noms et pourcentages des 555 personnes parlant nommables par les noms prononcés et/ou écrits

$N_{cités}$	$N_{écrits}$	Occ	Occ_{pv}	N_{intra}	N_{inter}
M_{UEM}		4 273	1 863 (43,6 %)	62,2	66,5
	M_{UEM}	1 049	1 022 (97,4 %)	60,5	65,9
M_{UEM}	M_{UEM}	5 322	2 885 (54,2 %)	80,4	83,6
A_{UEM}		2 905	914 (31,5 %)	26,7	30,8
	A_{UEM}	1 407	1 348 (95,8 %)	73,5	76,8
A_{UEM}	A_{UEM}	4 312	2 262 (52,5 %)	75,8	78,7

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.

6.2. Détail par rôle de 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 lisibilité. Les trois premières ont été regroupées dans le rôle R123 : présentateur/journaliste, les deux dernières dans le rôle R45 : 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.

Tableau 7. Répartition de la présence des personnes en fonction de leurs rôles : ***R123*** *: Présentateur/journaliste,* ***R45*** *: Invité/autre.*

Rôle	#Personnes parlant	Temps de parole	#Tours de parole	#Personnes apparaissant	#Apparitions à l'image
R123	84 (15 %)	632 (45 %)	6 149 (42 %)	48 (7 %)	2 935 (30 %)
R45	475 (85 %)	783 (55 %)	8 378 (58 %)	680 (93 %)	6 861 (70 %)

Seulement 48 des 84 personnes de R123 sont visibles et inversement 475 personnes de R45 parlent alors que 680 sont visibles. 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 y donc un déséquilibre du temps de présence entre les présentateurs/chroniqueurs/journalistes (rôles R123) et les invités/autres (rôles R45).

Le tableau 8 détaille les possibilités de nommer les 808 personnes présentes (union des personnes parlant et apparaissant) en fonction du rôle qu'elles occupent dans les vidéos.

Tableau 8. Nombre d'occurrences des noms et pourcentages des personnes présentes nommables par les noms prononcés ou écrits en fonction de leurs rôles (R123 : 84 présentateur/chroniqueur/reporter, R45 : 728 invité/autre)

$N_{cités}$	$N_{écrits}$	Occ_{pv}		N_{intra}		N_{inter}	
		R123	R45	R123	R45	R123	R45
M_{UEM}		414	2 353	78,9	55,2	86,9	61,7
	M_{UEM}	91	952	23,0	40,6	35,7	47,9
M_{UEM}	M_{UEM}	505	3 305	81,0	67,7	89,3	73,6
A_{UEM}		58	1 396	13,9	24,7	16,7	30,6
	A_{UEM}	174	1 177	37,8	46,3	47,6	53,4
A_{UEM}	A_{UEM}	232	2 573	42,9	56,3	52,4	62,5

Nous pouvons observer que les noms des 84 personnes de R1 sont assez peu prononcés (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 oralement et visibles à l'image, sont souvent introduits par leurs noms écrits alors que les journalistes en voix off ne sont jamais présentés ainsi.

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 %). Les personnes du rôle R123 sont donc plus difficilement nommables automatiquement alors qu'elles représentent une proportion importante du temps de présence. L'utilisation de quelques modèles biométriques correspondant aux personnes de R123 reste donc une solution intéressante puisqu'il est facile d'avoir une connaissance a priori de leur présence dans les vidéos.

6.3. 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}=2 805 à A_{RAW}=3 476), 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.

Tableau 9. Apport des vidéos complètes, pour le nommage des 808 personnes apparaissant et/ou parlant dans les segments UEM

$N_{cités}$	$N_{écrits}$	Occ	Occ_{pv}	N_{intra}	N_{inter}
M_{UEM}		4 273	2 767 (64,8 %)	57,7	64,4
	M_{UEM}	1 049	1 043 (99,4 %)	39,0	46,9
M_{UEM}	M_{UEM}	5 322	3 810 (71,6 %)	69,2	75,4
A_{UEM}		2 905	1 454 (50,1 %)	23,7	29,3
	A_{UEM}	1 407	1 351 (96,0 %)	45,5	53,0
A_{UEM}	A_{UEM}	4 312	2 805 (65,1 %)	55,1	61,6
A_{RAW}		4 922	1 755 (35,7 %)	24,8	30,4
	A_{RAW}	2 090	1 721 (82,3 %)	47,3	54,6
A_{RAW}	A_{RAW}	7 012	3 476 (49,6 %)	56,7	62,7

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 prononcé 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 10), 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 prononcé/écrit en début d'émission alors que pour les invités/autres, il est prononcé/écrit au moment de l'intervention.

Tableau 10. 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 9, ligne A_{UEM}

$N_{cités}$	$N_{écrits}$	Occ_{pv}		R_{intra}	
		R123	R45	R123	R45
A_{RAW}		78(+20)	1 677(+281)	17,8(+3,9)	25,5(+0,8)
	A_{RAW}	226(+52)	1 495(+318)	41,8(+4)	47,8(+1,5)
A_{RAW}	A_{RAW}	304(+72)	3 172(+599)	47,1(+4,2)	57,6(+1,3)

6.4. 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 11). 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 de groupes de musique sont introduites par le nom du groupe et non le nom de chacun des membres, etc.).

Tableau 11. 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

Émission	Type	Noms prononcés		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

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 oralement 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.

6.5. 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 prononcé 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 prononcé 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 dans l'image.

– Un nom prononcé pourra nommer les personnes apparaissant/parlant dans les tours de parole ou plans précédents, courants ou suivants.

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

Tableau 12. Proportion des 808 personnes parlant ou apparaissant nommables à l'aide d'un oracle au voisinage, segments UEM

$N_{cités}$	$N_{écrits}$	Oracle au niveau de la vidéo			Oracle au voisinage		
		Occ_{pv}	N_{intra}	N_{inter}	Occ_{pv}	N_{intra}	N_{inter}
M_{UEM}		2 767	57,7	64,4	1 580	51,8	58,9
	M_{UEM}	1 043	39,0	46,9	977	38,4	45,9
M_{UEM}	M_{UEM}	3 810	69,2	75,4	2 557	64,6	71,4
A_{UEM}		1 454	23,7	29,3	632	20,9	26,4
	A_{UEM}	1 351	45,5	53,0	1 269	45,4	52,5
A_{UEM}	A_{UEM}	2 805	55,1	61,6	1 901	53,6	60,5

Le tableau 12 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 au voisinage, 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 prononcés (de 2 767 à 1 580 pour M_{UEM} et de 1 454 à 632 pour A_{UEM}) alors qu'il ne réduit que très peu pour les noms écrits (de 1 043 à 977 pour M_{UEM} et de 1 351 à 1 269 pour A_{UEM}).

Pour les noms prononcés, il faut aussi sélectionner les noms à utiliser alors que, pour les noms écrits, la quasi-totalité des occurrences sont utilisables pour identifier les personnes directement présentes.

7. Conclusion

Deux sources intrinsèques aux vidéos peuvent fournir les noms des personnes présentes dans les flux télévisés : les noms prononcés et les noms écrits. Les premiers bénéficient d'un plus grand nombre d'occurrences de citation par rapport aux seconds. 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 de progression plus importante pour les noms prononcés que pour ceux écrits.

Il est important de souligner que les noms prononcés 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).

Sur le corpus *REPERE*, les noms prononcés extraits automatiquement peuvent permettre de nommer environ deux fois moins de personnes que les noms écrits. Ces derniers sont principalement utilisés pour introduire des personnes apparaissant et parlant en même temps. En revanche, ceux prononcés peuvent aussi introduire des journalistes parlant en voix off ou encore des personnes apparaissant mais ne parlant pas.

Cette étude montre qu'il y a une forte différence entre les présentateurs/chroniqueurs/journalistes et les invités/autres. Les premiers représentent un temps de présence important alors qu'ils ne sont que peu nombreux. Les seconds, malgré qu'ils soient huit fois plus nombreux dans ce corpus, représentent le temps de présence restant. Sur des corpus de taille plus importante, cet écart devrait être plus marqué.
La seconde différence provient de la difficulté à nommer les présentateurs/chroniqueurs/journalistes, notamment en utilisant les noms extraits automatiquement. Ce qui doit orienter les travaux futurs vers l'utilisation de modèles biométriques pour ces personnes, puisque leur présence dans les flux télévisés est a priori connue.

Le corpus *REPERE* a permis de montrer des différences notables sur la proportion de personnes nommables, selon le type d'émission visé. Certaines d'entre elles n'utilisent que peu les noms écrits ou prononcés pour introduire les personnes présentes. Par exemple, 17,6 % de personnes sont nommables dans « Planète showbiz » à l'aide des noms prononcés et 32,7 % à l'aide des noms écrits. Alors que l'émission « Pile et face » présente systématiquement tous les intervenants à l'aide des noms écrits.

Les différences entre ces deux sources montrent qu'il est important de développer des méthodes d'identification non supervisées intégrants à la fois les noms écrits (qui permettent de nommer facilement le locuteur courant et un des visages visibles avec certitude) mais aussi le grand nombre d'occurrences des noms prononcés, avec l'incertitude inhérente à cette source. Ces méthodes ont tout intérêt à utiliser un corpus entier pour profiter de la redondance de présence des personnes sur plusieurs vidéos, mais aussi à adopter des stratégies spécifiques à chaque type d'émission.

Remerciements

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

Bibliographie

Allauzen A., Bonneau-Maynard H. (2008). Training and evaluation of pos taggers on the french multitag corpus. In *LREC*, p. 3373–3377.

Bechet F., Favre B., Damnati G. (2012). Detecting person presence in tv shows with linguistic and structural features. In *ICASSP*, p. 5077–5080.

Bendris M., Favre B., Charlet D., Damnati G., Auguste R., Martinet J. (2013). Unsupervised face identification in tv content using audio-visual sources. In *CBMI*, p. 243–249.

Bernard G., Rosset S., Galibert O., Bilinski E., Adda G. (2009). Limsi participation in the qast 2009 track: Experimenting on answer scoring. In *CLEF*, p. 289–296.

Bredin H., Poignant J., Tapaswi M., Fortier G., Le V. B., Napoleon T. (2012). Fusion of speech, faces and text for person identification in tv broadcast. In *ECCV-IFCVCR*, p. 385–394.

Bredin H., Poignant J. (2013a). Integer linear programming for speaker diarization and cross-modal identification in tv broadcast. In *Interspeech*, p. 1467–1471.

Bredin H., Poignant J., Fortier G., Tapaswi M., Le V. B., Sarkar A. (2013b). Qcompere at repere 2013. In *SLAM*, p. 49–54.

Canseco-Rodriguez L., Lamel L., Gauvain J.-L. (2004). Speaker diarization from speech transcripts. In *INTERSPEECH*, p. 1272–1275.

Canseco L., Lamel L., Gauvain J.-L. (2005). A comparative study using manual and automatic transcriptions for diarization. In *ASRU*, p. 415–419.

Charhad M., Moraru D., Ayache S., Quénot G. (2005). Speaker identity indexing in audio-visual documents. In *Cbmi.*

Dietterich T. G., Lathrop R. H., Lozano-Pérez T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, vol. 89, n° 1-2, p. 31–71.

El-Khoury E., Laurent A., Meignier S., Petitrenaud S. (2012). Combining transcription-based and acoustic-based speaker identifications for broadcast news. In *ICASSP*, p. 4377–4380.

Estève Y., Meignier S., Deléglise P., Mauclair J. (2007). Extracting true speaker identities from transcriptions. In *INTERSPEECH*, p. 2601–2604.

Gauvain J.-L., Lamel L., Adda G. (1998). Partitioning and transcription of broadcast news data. In *ICSLP-AISSTC*, p. 1335–1338.

Gauvain J.-L., Lamel L., Adda G. (2002). The limsi broadcast news transcription system. In *Speech communication*, p. 89–108.

Giraudel A., Carré M., Mapelli V., Kahn J., Galibert O., Quintard L. (2012). The repere corpus : a multimodal corpus for person recognition. In *LREC*, p. 1102–1107.

Houghton R. (1999). Named faces: putting names to faces. *IS*, vol. 14, p. 45–50.

Jousse V., PetitRenaud S., Meignier S., Estève Y., Jacquin C. (2009). Automatic named identification of speakers using diarization and asr systems. In *ICASSP*, p. 4557–4560.

Lamel L., Courcinous S., Despres J., Gauvain J.-L., Josse Y., Kilgour K. (2011). Speech recognition for machine translation in quaero. In *IWSLT*, p. 121–128.

Lavergne T., Cappé O., Yvon F. (2010). Practical very large scale crfs. In *ACL*, p. 504–513.

Liu C., Jiang S., Huang Q. (2008). Naming faces in broadcast news video by image google. In *ACMMM*, p. 717–720.

Marco D., Rosset S. (2011). Models cascade for tree-structured named entity detection. In *IJCNLP*, p. 1269–1278.

Mauclair J., Meignier S., Estève Y. (2006). Speaker diarization: about whom the speaker is talking? In *Odyssey*.

PetitRenaud S., Jousse V., Meignier S., Estève Y. (2010a). Identification of speakers by name using belief functions. In *IPMU*, p. 179–188.

Petitrenaud S., Jousse V., Meignier S., Estève Y. (2010b). Reconnaissance automatique de locuteurs á l'aide de fonctions de croyance. In *RFIA*, p. 4557–4560.

Poignant J., Besacier L., Quénot G., Thollard F. (2012a). From text detection in videos to person identification. In *IEEE ICME*, p. 854–859.

Poignant J., Bredin H., Le V. B., Besacier L., Barras C., Quénot G. (2012b). Unsupervised speaker identification using overlaid texts in tv broadcast. In *INTERSPEECH*.

Poignant J., Besacier L., Le V. B., Rosset S., Quénot G. (2013a). Unsupervised naming of speakers in broadcast tv: using written names, pronounced names or both ? In *INTERSPEECH*.

Poignant J., Besacier L., Quénot G. (2013b). Nommage non-supervisé des personnes dans les émissions de télévision: une revue du potentiel de chaque modalité. In *CORIA*, p. 5-20.

Poignant J., Bredin H., Besacier L., Quénot G., Barras C. (2013c). Towards a better integration of written names for unsupervised speakers identification in videos. In *SLAM*, p. 84-89.

Satoh S., Kanade T. (1997a). Name-it: association of face and name in video. In *CVPR*.

Satoh S., Nakamura Y., Kanade T. (1997b). Name-it: naming and detecting faces in video by the integration of image and natural language processing. In *IJCAI*, p. 1488–1493.

Satoh S., Nakamura Y., Kanade T. (1999). Name-it: naming and detecting faces in news videos. *IEEE Multimedia*, vol. 6, p. 22–35.

Schapire R. E., Singer Y. (1999). Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, vol. 37, n° 3, p. 297–336.

Song X., Lin C.-Y., Sun M.-T. (2004). Cross-modality automatic face model training from large video databases. In *CVPRW*, p. 91–.

Tranter S. E. (2006). Who really spoke when? finding speaker turns and identities in broadcast news audio. In *ICASSP*, p. 1013–1016.

Yang J., Hauptmann A. G. (2004). Naming every individual in news video monologues. In *ACMMM*, p. 10–16.

Yang J., Yan R., Hauptmann A. G. (2005). Multiple instance learning for labeling faces in broadcasting news video. In *ACMMM*, p. 31–40.