Identification de personnes dans des flux multimédia

F. Béchet[1], M. Bendris[1], D. Charlet[2], G. Damnati[2], B. Favre[1], M. Rouvier[1], R. Auguste[3], B. Bigot[4], R. Dufour[4], C. Fredouille[4], G. Linarès[4], G. Senay[4], P. Tirilly[3] and J. Martinet[3]

[1] LIF, Aix-Marseille Université, [2] Orange Labs, [3] CRIStaAL, Université de Lille, [4] LIA, Université d'Avignon

RÉSUMÉ. *Cet article présente un système d'identification de personnes dans des flux multimédia. Ce système a été engagé dans le défi REPERE, co-organisé par l'ANR et la DGA et qui s'est terminé en 2014. La tâche principale du défi consistait à identifier des individus apparaissant dans au moins une des modalités portées par la vidéo, qu'il s'agisse de locuteurs audibles ou de visages visibles à l'écran. Un des verrous scientifiques majeurs de cette tâche est lié à la combinaison des modalités audio et vidéo. Cet article présente une stratégie pour la reconnaissance des personnes basée sur une identification du locuteur reposant sur des descripteurs de haut niveau, modélisant différents aspects de la scène filmée : la transcription et l'analyse des textes incrustés, l'identification du type de la scène filmée (reportage, plateau, ...), le nombre de personnes présentes, la disposition des caméras... Nos expériences sur le corpus REPERE montrent l'intérêt de l'approche proposée.*

ABSTRACT. *This paper describes a multi-modal person recognition system for video broadcast developed for participating to the REPERE challenge, that was organized jointly by the DGA and the ANR (French Research National Agency). The main track of this challenge targets the identification of all persons occurring in a video either. The main scientific issue addressed by this challenge is the combination of audio and video information extraction processes for improving the extraction performance in both modalities. In this paper, we present a strategy for speaker identification based on enriching the speaker diarization by features related to the "understanding" of the video scenes: text overlay transcription and analysis, automatic situation identification (TV set, report), the amount of people visible, TV set disposition and even the camera when available. Experiments on the REPERE corpus show interest of the proposed approach.*

MOTS-CLÉS : *Reconnaissance de personnes$_1$, Indexation multimédia$_2$.*

KEYWORDS: *Person recognition$_1$, Multimedia indexing$_2$.*

1. Introduction

L'identification de personnes dans des émissions de télévision est une tâche naturellement multimodale. De fait, les personnes peuvent être identifiées de manière biométrique (visage, voix), ou leur nom peut être cité dans les paroles prononcées ou dans les incrustations de texte à l'écran. Le défi *REPERE*[1] (REconnaissance de PERsonnes dans des Emissions audiovisuelles) a pour objectif l'identification des personnes dans les programmes télévisés en utilisant des informations multimodales (Giraudel *et al.*, 2012). Ce challenge offre un corpus de vidéos annotées manuellement en locuteurs, transcription de la parole, textes incrustés et annotation des visages. Les annotations sur les images sont effectuées environ toutes les 10 secondes sur des *keyframes*. Les systèmes participant au challenge devaient générer une liste de noms de personnes présentes à chaque instant de la vidéo en utilisant des informations biométriques et contextuelles.

Les systèmes ayant participé à l'édition précédente de la campagne étaient basés sur la fusion de systèmes mono-modaux, en particulier la segmentation en locuteurs, le suivi de visages, la transcription automatique de la parole, la reconnaissance de caractères et l'identification de locuteur/visage (Bredin *et al.*, 2013 ; Favre *et al.*, 2013). Les principales composantes de ces systèmes sont les modèles biométriques de locuteurs et visages tirant parti de dictionnaires d'identités.

Les résultats obtenus durant l'édition 2013 du défi *REPERE* montrent que cette approche porte ses fruits dans les situations non ambigues où une seule personne parle et est visible à l'écran, permettant la propagation de son identité d'une modalité à l'autre. Néanmoins, elle ne permet de traiter qu'un petit nombre de cas, ignorant les personnes dont l'identité n'est pas présentée explicitement et les nombreux cas où plusieurs personnes sont à l'écran. De plus, maintenir des dictionnaires de modèles biométriques est très coûteux et peu réaliste dans un contexte industriel. Le système présenté dans cet article repose sur la prédiction de l'identité des locuteurs, qui est relativement fiable, et la propagation de cette hypothèse issue de l'audio dans la modalité visuelle grâce à l'analyse de scènes et la représentation globale de différents aspects de la scène, qu'on appellera *compréhension multimodale*. Cet enrichissement des hypothèses locuteurs utilise des caractéristiques liées à la compréhension de scènes complexes : analyse des textes incrustés, identification de type de scène (plateau, reportage), prédiction du nombre de personnes visibles, reconnaissance de la disposition du plateau et identification de caméras lorsque c'est possible.

La section 2 présente l'état de l'art en identification multimodale de personnes ; La section 3 décrit notre système de reconnaissance multi-modale de locuteurs ; la section 4 décrit les caractéristiques visuelles utilisées pour construire une représentation sémantique de scène et comment elle est intégrée à l'identification de locuteur. La section 5 présente quelques résultats contrastifs sur les corpus de la campagne et l'article

1. `http://www.defi-repere.fr`

se termine par une conclusion et quelques perspectives d'extension ou d'amélioration du système proposé.

2. Reconnaissance multi-modale de personnes : le défi *REPERE*

L'analyse des contenus de vidéos est un problème qui a été très souvent abordé de manière séparée par les communautés du traitement d'images et de la parole. Par exemple, on trouve du côté *parole* des travaux sur la segmentation en thèmes (Guinaudeau *et al.*, 2012) et la reconnaissance du rôle des locuteurs (Damnati et Charlet, 2011) n'utilisant que des informations provenant de l'audio et des transcriptions de parole. Parallèlement, la communauté *image* a exploré l'analyse de scènes et la reconnaissance d'actions (Gaidon *et al.*, 2013), en se basant uniquement sur les images pour identifier des situations particulières ou modéliser globalement la scène avec des cadres et des objectifs applicatifs assez variés.

En plus de ces études monomodales, la segmentation vidéo a été explorée à l'aide de caractéristiques audio et vidéo (Dumont et Quénot, 2012 ; Wang *et al.*, 2010). Ces approches limitent cependant la coopération entre les modalités à une fusion précoce (au niveau des observations), ou tardive (au niveau des modèles) des décisions monomodales.

Néanmoins, pour les tâches plus complexes que la segmentation vidéo, il faut utiliser des modèles multimodaux plus avancés. Le défi *REPERE* illustre bien cette problématique (Kahn *et al.*, 2012). La tâche principale du challenge est l'identification de personnes dans des vidéos (débats, nouvelles, reportages, etc). Cette tâche implique un certain nombre de tâches secondaires qui ont, elles aussi, été évaluées dans le cadre du défi.

Les personnes à identifier peuvent être connues (politiciens, musiciens, acteurs et présentateurs TV). Dans ce cas, il est facile de créer des modèles biométriques à priori. Il peut aussi s'agir d'invités liés à l'actualité et non couverts par les dictionnaires d'identités. Pour ces derniers, une analyse poussée du contexte dans lequel elles apparaissent est nécessaire, notamment en tirant parti des incrustations de texte et de la transcription des contenus parlés.

Dans l'édition 2014 de la campagne, le corpus de test était constitué d'environ 10 heures de télévision provenant de neuf émissions : *BFMStory, CaVousRegarde, LCPInfo, EntreLesLignes, PileEtFace, TopQuestion, RuthElkief, CultureEtVous, LCPActu*. Alors que les personnes visibles sont annotées environ toutes les dix secondes, l'identité des locuteurs est, quand à elle, disponible sur la toute la durée des émissions.

Pour la modalité image, la tâche dépasse le problème classique d'identification de visage, toutes les personnes apparaissant dans une trame clé devant être identifiées, que le visage soit visible ou non (il faut identifier les personnes de dos, dont le visage est partiellement couvert, ou dont le visage est trop petit pour appliquer les méthodes d'identification biométrique (Wallace *et al.*, 2011)). Dans notre corpus, environ 10%

des personnes ne peuvent pas être identifiées avec des modèles biométriques pour ces raisons. La table 1 détaille la distribution des personnes non-identifiables par des modèles biométriques par émission.

Tableau 1. *Distribution des occurrences de personnes dans les keyframes qui ne peuvent pas être identifiées avec des modèles biométriques (visages tournés, petites têtes).*

Emission	Visage tournés	Petites têtes
BFMStory	4.56	9.86
CaVousRegarde	2.52	16.80
LCPInfo	17.11	11.14
EntreLesLignes	12.45	22.28
PileEtFace	9.26	17.13
Total	9.30	14.24

Dans la modalité audio, il n'y a pas une telle ambiguité : il n'y a pas de phénomène d'occlusion, mais plusieurs locuteurs peuvent parler en même temps. Néanmoins, même avec une identification parfaite des locuteurs, propager cette identité aux visages n'est pas trivial. Par exemple, (Bendris *et al.*, 2009) a montré que le locuteur courant était visible 60% du temps et que les personnes visibles parlent 30% du temps, sur un corpus d'émissions de divertissement.

Cette absence de synchronie entre les modalités audio et vidéo est une des difficultés majeures du traitement multimédia et, ici, elle complique considérablement le processus d'identification des personnes. Pour traiter cette difficulté particulière, une possibilité est de se reposer sur l'analyse contextuelle et la compréhension de scènes pour attribuer les identités. C'est sur cette approche que se fondent les travaux présentés dans cet article.

Nous présentons dans la suite un système qui met en oeuvre cette approche, d'abord en décrivant un module d'identification de locuteur fondé sur l'audio et la reconnaissance de noms dans les incrustations textuelles, puis nous décrivons comment l'analyse de scènes vidéo peut étendre ce système pour prendre des décisions sur la modalité image.

3. De la segmentation en locuteurs à la reconnaissance de personnes

Dans notre système, la stratégie d'identification du locuteur combine deux processus disjoints. Le premier est la segmentation et regroupement en locuteurs, qui consiste à extraire les parties parlées du flux, à les partitionner en segments mono-locuteurs puis à regrouper les segments produits par un même locuteur. A l'issue de cette première phase, des groupes de segments attribués au même locuteur sont constitués mais l'identité des locuteurs reste inconnue. On trouvera une description plus complète du système de diarization dans (Charlet *et al.*, 2013). La seconde phase cherche à identifier le locuteur associé à chacun de ces groupes ; concrètement, il s'agit d'affecter

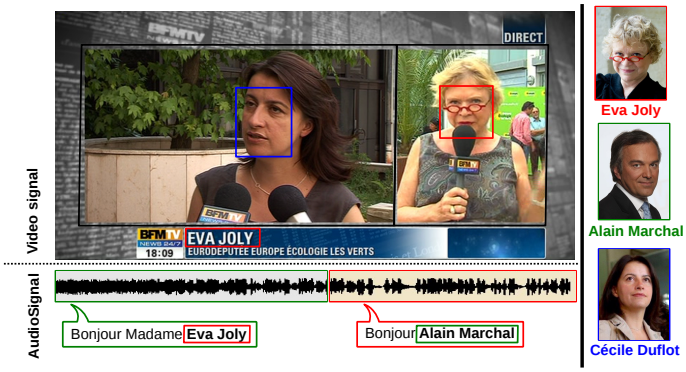

Figure 1. *Sources d'information pour l'identification de personnes dans une émission télévisée : locuteur, noms cités, incrustations de texte, visages à l'écran.*

un nom connu à chacun des groupes issu de la segmentation et regroupement en locuteurs.

Le processus de nomage des groupes utilise les trois sources potentielles d'information pouvant permettre la reconnaissance d'une identité : le module d'identification du locuteur, qui intègre un certain nombre de modèles à priori de locuteurs connus ; le système de reconnaissance des textes incrustés et le système de détection des noms cités dans la parole. Ces trois modules sont présentés dans les sections suivantes.

3.1. *Identification du locuteur*

Le système d'identification du locuteur est fondé sur la plateforme Alizé (Larcher *et al.*, 2013) et utilise une modélisation des locuteurs par *i-vector*, qui est devenue l'approche standard des systèmes à l'état de l'art. Pour *REPERE 2014*, nous avons entraîné 533 modèles pour les locuteurs les plus fréquents du corpus d'apprentissage *REPERE* ainsi que les personnalités politiques très présentes dans les médias dans la période couverte par les données de test. Les modèles sont entraînés sur au moins 30 secondes de parole. Pour les locuteurs plus fréquents, nous avons créé un modèle pour chaque 2 minutes 30 de parole disponibles. Un segment de test est nommé à l'aide de l'identité associée au modèle de score maximal.

Le score est calculé par la distance cosine entre l'*i-vector* du modèle et celui du test, suivi d'une normalisation de type *Within Class Covariance Normalization* (WCCN) (Hatch *et al.*, 2006). Le système d'identification est appliqué à deux horizons temporels à des fins de fusion : à chaque tour de parole, ou sur l'ensemble des segments regroupés par le module de regroupement en locuteurs.

3.2. *Noms incrustés*

Dans les émissions de type *débat* et *journal*, le nom des locuteurs est souvent incrusté à l'écran lors de leur première apparition, puis périodiquement. Notre module de détection de noms incrustés utilise la reconnaissance optique de caractères pour détecter le texte, le transcrire et retrouver des noms susceptibles d'identifier la personne visible à l'écran.

Le module est basé sur une reconnaissance de caractères dans des vidéos développée par Orange Labs. Ce module utilise un réseau de neurones convolutionnel pour identifier les caractères (Elagouni *et al.*, 2013). Un ensemble de règles sur la taille du cartouche de texte, sa position et les mots reconnus permet de filtrer les titres, lieux, dates, etc... qui n'identifient pas les personnes. Étant donné que le nom exact doit être retrouvé, les hypothèses sont normalisées par rapport à une liste statique d'environ 400k noms de personnes construite à partir de Wikipédia et de bases de connaissances, ainsi que des listes dynamiques collectées dans les *news* et sur le web à la date de l'émission.

La correspondance entre les hypothèses et la liste est faite grâce à des automates transducteurs pondérés, en composant les sorties de l'OCR avec un automate de distorsion (insertion, suppression, substitution) et un automate associant les noms de la liste à une forme normalisée. Le plus court chemin dans cette automate donne une hypothèse de nom, les noms comportant trop de distorsions étant rejetés.

3.3. *Détection de noms dans la parole*

Le module de détection de noms dans la parole prend comme entrée la même liste de noms de personnes que pour les textes incrustés et les retrouve dans les transcriptions automatiques des segments de parole issus de la segmentation en locuteurs. La recherche est réalisée en deux étapes. La première met en oeuvre un Système de Reconnaissance Automatique de la Parole continue à grand vocabulaire (SRAP). Une première difficulté est liée à la diversité des formes qu'un nom de personne peut prendre. Nous utilisons un certain nombre de règles permettant d'associer une forme orthographique à un nom référencé dans les listes de personnes constituées à priori. Malgré un vocabulaire de reconnaissance d'environ 90 000 mots, la couverture des noms propres par le SRAP est imparfaite et un deuxième processus recherche les noms hors du vocabulaire de reconnaissance. La détection se fait donc en trois étapes :

1) Recherche de noms complets (prénom et nom) dans la transcription : la transcription est produite par un système dont l'architecture est décrite dans (Linarès *et al.*, 2007). Les modèles acoustiques ont été appris sur environ 300 heures de parole annotée. Les modèles de langage sont des modèles 4-grammes, appris sur environ 10^9 mots issus du corpus Gigaword et de différentes sources de données textuelles antérieures à 2010 auxquelles ont été ajoutés les corpus ETAPE et REPERE, qui fournissent des données postérieures à cette date.

2) Recherche d'hypothèses partielles à partir des hypothèses complètes de la liste : nous avons utilisé un mécanisme assez simple basé sur des règles de ré-affection et la distance d'édition entre une hypothèse et les noms de références. Une hypothèse d'association est retenue lorsque qu'elle satisfait les règles de ré-affectation ou lorsque la distance d'édition est inférieure à un seuil fixé à priori.

3) Alignement phonétique des noms dans le réseau de confusion produit par la transcription automatique : ce module cherche à détecter la séquence phonétique d'un nom dans le flux de parole. Plutôt que de chercher dans des treillis de phonèmes issus des treillis de mots produits par le système de reconnaissance, la recherche est réalisée dans des réseaux de confusions phonétiques, approche qui s'est montrée plus efficace sur ce type de tâches (Bigot *et al.*, 2013).

Le résultat de ce traitement est une liste de noms de personnes détectées avec leur temps d'apparition et des scores de confiance.

4. Compréhension multimodale de scènes

La principale originalité du système présenté ici est l'intégration d'un module de *compréhension multimodale* dont le rôle est de traiter les indications ambigües. Ce module opère à quatre niveaux : métadonnées, audio, image, analyse de scènes vidéos. Plusieurs stratégies de prise de décision, basées sur l'ensemble ou une partie de ces niveaux, ont été développées et intégrées au système qui a été engagé dans le défi.

4.1. *Descripteurs de compréhension multimodale*

4.1.1. *Metadata*

Les méta-descripteurs représentent l'ensemble de la connaissance à priori qui peut être collectée : la chaîne de télévision, le programme TV, la date de diffusion... L'ensemble des listes de noms de personnes qui ont été décrites dans les sections précédentes sont considérées comme des méta-descripteurs. A chaque nom, peut être associé des descripteurs biographiques tels que la date de naissance et - éventuellement- de décès, la nationalité, un ensemble de thématiques associées, un historique de présence dans les émissions télévisées du corpus d'apprentissage. Enfin, lorsque l'information est disponible, nous utilisons la structure globale de l'émission vue comme un ensemble structuré de segments de styles particuliers (reportage, interview, débat, présentateur, etc...).

4.1.2. *Paramètres audio*

Outre les hypothèses relatives à l'identité du locuteur telles qu'elles sont présentées dans la section 3.1, les informations de genre et de rôle des locuteurs sont extraites dans chaque groupe de locuteurs.

4.1.3. *Extraction des textes incrustés*

Les textes incrustés sont extraits en suivant la technique décrite dans la section 3.2. De plus, les logos étant d'excellents marqueurs de la structure en chapitre des émissions, nous cherchons à les détecter de façon à localiser les transitions inter-chapitres.

4.1.4. *Paramètres visuels*

Deux types de descripteurs visuels sont utilisés : le visage et les vêtements, ces derniers étant caractérisés par des histogrammes de couleurs et une caractérisation de la scène. Une combinaison des ces deux descripteurs a également été utilisée. Les descripteurs de visage et de vêtements sont considérés comme des identifiants uniques de la personne dans une émission.

Le processus d'extraction commence par détecter les visages selon la méthode de Viola et Jones (Viola et Jones, 2004). Cette détection est réalisée sur chaque trame vidéo. Les visages détectés sont ensuite catégorisés par les histogrammes de couleurs des vêtements associés. Cet histogramme est estimé en isolant un rectangle centré sur le visage détecté, rectangle de taille proportionnelle à celle du visage. L'histogramme de couleur HSV est ensuite calculé et les similarités des vêtements sont évaluées par un simple cosinus entre les histogrammes. Le descripteur de bustes est calculé de manière similaire en utilisant les détections de visage. La détection permet d'initialiser un masque pour extraire le buste d'une personne à chaque trame. Les images successives de bustes sont utilisées pour calculer un histogramme spatio-temporel pour chaque occurrence de personne (Auguste *et al.*, 2012). Une mesure de similarité spécifique permet de générer une matrice de similarités entre les occurrences vidéo de personnes.

L'identification de scènes est un sujet de recherche actif dans le domaine du traitement d'images, par exemple pour la segmentation d'émissions sportives ou pour l'analyse de vidéos de surveillance. Nous suivons une méthode relativement standard qui repose sur trois descripteurs visuels : le type de plan, le rôle des personnes à l'écran et des identificateurs de caméra dans la situation "studio". Le système utilise des paramètres visuels HOG et RGB pour chaque plan. Il a été entraîné avec la librairie *Liblinear 2*. Le système ne repose pas sur la détection de visages ou de personnes pour déterminer les rôles, mais utilise plutôt des régularités au niveau trame pour identifier globalement l'ensemble des rôles impliqués dans le plan.

4.2. *Processus de décision*

Le rôle du processus de décision est de combiner l'ensemble des descripteurs extraits pour répondre à ces deux questions élémentaires : qui parle et qui voit-t-on à l'écran, pour chacune des trames de la vidéo. La principale source d'information vient de la segmentation en locuteurs qui a été présentée dans la section précédente. Le processus de décision opère à partir de ces segments extraits, en trois étapes successives :

– à chaque segment, affectation d'une identité issue du module d'identification du locuteur, des textes incrustés ou détectés dans le flux de parole,

– recherche de la présence à l'écran de l'identité hypothétique affectée au segment,

– détection des personnes non parlantes mais visibles à l'écran, à l'intérieur ou à l'extérieur d'un segment parlé. Affectation d'une identité avec ces personnes détectées.

Ce processus s'appuie sur les descripteurs suivants :

- métadonnées : noms de l'émission, durée du segment, nombre d'hypothèses de noms, genre, état (mort ou encore vivant), parole, thèmes, etc.

- reconnaissance des caractères : confiance dans l'hypothèse issue de l'OCR, OCR étendue au segment, au groupe de segments, au groupe de plans.

- identification du locuteur : présence de l'hypothèse dans les 10-meilleures hypothèses d'identification locuteur, meilleure hypothèse, score d'identification locuteur, genre du locuteur et concordance genre détecté-nom, rôle du locuteur.

- reconnaisseur de parole : temps depuis la dernière détection d'un nom dans les transcriptions.

- descripteurs de scène : type de plan, rôle visuels, numéro de caméra.

Quatre stratégies de prises de décisions ont été développées pour produire le système final, qui a été engagé comme système primaire dans *REPERE 2014* :

- S1 : système à base de règles similaire à celui présenté dans (Favre *et al.*, 2013), auquel ont été ajoutés des descripteurs de scènes.

- S2 : les règles sont remplacées ici par un classifieur entrainé sur le corpus *REPERE*. Trois méthodes de classification sont utilisées pour ré-ordonner les hypothèses de noms : icsiboost (Favre *et al.*, 2007), bonzaiboost (Raymond, 2007) avec trois niveaux d'arbre de décisions, et C4.5 avec 8 niveaux d'arbres de décisions. L'idée ici est que ces classifieurs sont susceptibles de mieux modéliser les interactions locales entre paramètres. En phase de test, nous n'utilisons qu'un de ces classifieurs, le choix étant réalisé sur le corpus de développement.

- S3 : ce système, similaire à celui décrit dans (Bendris *et al.*, 2014), tire avantage de la propagation des contraintes de la modalité locuteur vers les visages. Nous utilisons un algorithme de programmation linéaire en nombres entiers qui réalise le regroupement des visages à partir de descripteurs multimodaux.

- S4 : le système d'identification des caméras permet d'identifier les personnes filmées à partir de la position relatives des personnes et des caméras qui les filment. Cette information est utilisée dans la propagation des noms. Ce système est décrit dans (Rouvier *et al.*, 2014).

La conception du système principal repose sur la mise en compétition de ces différentes stratégies sur un corpus de développement, pour chaque condition (canal, type émission, modalité). Les résultats obtenus sont présentés dans la section suivante.

5. Expériences

5.1. *Résultats officiels*

Les systèmes participants au défi *REPERE* ont été évalués en termes de taux d'erreur global (EGER, *Estimated Global Error Rate*), calculé comme le nombre de noms insérés ou supprimés divisé par le nombre de noms de la référence. Les résultats sont donnés sur l'édition 2014 de la phase de test de la campagne, pour les modalités locuteur, visage et la métrique officielle basée sur les deux modalités. De plus, les systèmes sont évalués dans deux conditions : avec et sans modèles biométriques.

Condition	**Locuteur**	**Visage**	**Locuteur+visage**
Modèles biométriques applicables	18.7	37.4	28.9
Pas de modèles biométriques	30.9	39.4	35.5

Tableau 2. *EGER pour le système primaire, par modalité et en fonction de la possibilité d'utiliser des modèles biométriques.*

Les résultats du système présenté ici sont détaillés dans la table 2. Il est notable que, bien que les modèles biométriques impactent fortement le système d'identification du locuteur, leur couverture reste relativement faible. L'apport de ces modèles est cependant moins net sur la modalité visuelle, les descripteurs de scène n'étant pas biométriques et restant efficients dans toutes les conditions.

		Speaker		**Head**		
Show	**Type**	**S1**	**S2**	**S1**	**S3**	**S4**
BFMStory	actualités/deb.	18.6	**20.5***	**49.4***	50.0	-
CultureEtVous	tabloïd	35.4	**34.4***	**77.5***	83.9	-
RuthElkief	actualités/deb.	**25.3***	26.7	**44.2***	54.4	-
CaVousRegarde	débat	15.8	**17.8***	42.5	54.0	**27.4***
EntreLesLignes	débat	**16.1***	13.7	36.1	33.3	**20.8***
LCPActu	actualités	**12.0***	11.1	34.1	**31.8***	51.2
LCPInfo	actualités	**17.8***	16.8	50.6	**43.5***	43.4
PileEtFace	débat	6.2	**7.1***	16.2	15.7	**8.9***
TopQuestion	politique	**9.6***	9.6	**62.2***	69.2	-

Tableau 3. *Résultats par émission, pour chaque sous système (S_i) par modalité, avec les modèles biométriques. Les sous systèmes utilisés dans le primaire sont notés par une étoile ; ils ont été sélectionnés de façon empirique, en fonction des performances observées sur un corpus de développement. S_1 est le système à base de règles, S_2 est le système à base de classifieur automatique, S_3 est celui avec propagation de contraintes et S_4 est le système basé sur la position des caméras.*

Le système primaire résulte d'une sélection de sous-systèmes dépendants de l'émission. La liste des sous-systèmes sélectionnés est reportée dans la table 3. Cette sélection est réalisée de façon empirique, à partir d'un critère de performance estimé sur le corpus de développement (condition signalée par des étoiles dans la table 3).

Les résultats montrent que cette sélection semble être sensible aux différences entre corpus de développement et de test, qui limite la robustesse du processus de sélection.

5.2. *Résultats contrastifs*

Dans cette section, nous nous concentrons sur les résultats contrastifs susceptibles de mettre en évidence l'intérêt des descripteurs de scènes pour l'identification multimodale de personnes. Quatre configurations sont comparées : mono-modale avec seulement la segmentation et l'identification des locuteurs (SpkId), multi-modale sans identification du locuteur (en utilisant uniquement l'OCR et les descripteurs de scène), multi-modale avec les descripteurs de scène, et avec tous les descripteurs. Dans la configuration mono-modale, une première étape regroupe les locuteurs puis l'identification est appliquée sur chacun des groupes obtenus.

Le mode multi-modal sans identification des locuteurs correspond à la sortie du système à base de règles S1, sans aucun modèle biométrique. Les deux derniers systèmes correspondent au système S2 (à base de classification automatique).

Condition	**Tous les locuteurs**	**Têtes parlantes seulement**
SpkID	35.2	36.0
OCR+Scene	30.9	20.2
SpkID+OCR	21.2	14.7
SpkID+OCR+Scene	19.2	12.2

Tableau 4. *EGER sur la modalité* **Locuteurs** *en fonction du sous-ensemble de descripteurs, sur toutes les instances ou seulement sur les locuteurs visibles.*

Les résultats détaillés en table 4 montrent que, même si l'identification des locuteurs (*SpkID*) et l'OCR (*OCR*) constituent l'ossature du système, les descripteurs de scène (*Scene*) apportent une information complémentaire dont le système tire bénéfice. Ces gains sont encore plus nets si on ne considère que les situations dans lesquelles les locuteurs sont à la fois visibles et audibles.

La table 5 montre une même tendance sur la tâche de détection des visages, qui consiste à identifier la personne dont le visage est à l'écran sans modèle biométrique des visages. Le tableau montre les résultats de l'utilisation de la reconnaissance du locuteur pour la reconnaissance des visages, de l'utilisation de l'OCR seule pour l'identification des visages, et en ajoutant les descripteurs de scènes à ces paramètres. Les résultats montrent l'intérêt de ces descripteurs, ainsi que celle du module d'identification du locuteur pour la reconnaissance de visages. Le gain est particulièrement important lorsque l'individu est visible et audible, ce qui confirme l'intérêt de ces descripteurs dans une approche multi-modale de l'identification de personnes.

6. Conclusion

Nous avons présenté un système de reconnaissance de personnes dans des flux

Condition	Tous les locuteurs	Têtes parlantes seulement
SpkID	72.6	23.0
OCR	80.5	49.7
OCR+Scene	39.4	15.6
SpkID+OCR+Scene	37.4	13.1

Tableau 5. *EGER sur la modalité* **visage***, en fonction du sous-ensemble de descripteurs, sur toutes les instances ou seulement sur les locuteurs visibles. Ces résultats contrastifs sont obtenus avec le système S1.*

multimédia, développé dans le cadre du défi *REPERE*. L'approche proposée repose sur un système d'identification du locuteur enrichi de descripteurs issus de l'extraction et de l'analyse des textes incrustés et d'un module de "compréhension" de la scène. Ce module modélise globalement la scène filmée par l'extraction d'indicateurs de haut niveau tels que le nombre de personnages présents à l'image, la position des caméras, la situation filmée... Les résultats obtenus montrent l'intérêt de cette modélisation globale de la scène filmée, à la fois pour la reconnaissance des locuteurs, la reconnaissance des visages (avec une réduction d'environ 2% de l'*EGER*) et pour l'objectif principal du défi qui était l'identification multi-modale des personnes.

7. Bibliographie

Auguste R., Aissaoui A., Martinet J., Djeraba C. *et al.*, « Les histogrammes spatio-temporels pour la ré-identification de personnes dans les journaux télévisés », *Compression et Représentation des Signaux Audiovisuels (CORESA)*, 2012.

Bendris M., Charlet D., Chollet G., « Introduction of quality measures in audio-visual identity verification », *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, p. 1913-1916, 2009.

Bendris M., Favre B., Charlet D., Damnati G., Auguste R., « Multiple-view constrained clustering for unsupervised face identifiction in TV-broadcast », *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2014.

Bigot B., Senay G., Linarès G., Fredouille C., Dufour R., « Combining Acoustic Name Spotting and Continuous Context Models to improve Spoken Person Name Recognition in Speech », *InterSpeech*, ISCA, Lyon, France, 2013.

Bredin H., Poignant J., Fortier G., Tapaswi M., Le V.-B., Roy A., Barras C., Rosset S., Sarkar A., Yang Q., Gao H., Mignon A., Verbeek J., Besacier L., Quénot G., Ekenel H. K., Stiefelhagen R., « QCompere @ REPERE 2013 », *Speech, Language and Audio in Multimedia (SLAM)*, p. 49-54, 2013.

Charlet D., Barras C., Lienard J., « Impact of Overlapping Speech Detection on Speaker Diarization for Broadcast News and Debates », *2013 IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*, Vancouver, Canada, p. 7707-7711, 2013.

Damnati G., Charlet D., « Robust speaker turn role labeling of tv broadcast news shows », *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, p. 5684-5687, 2011.

Dumont E., Quénot G., « Automatic story segmentation for tv news video using multiple modalities », *International Journal of Digital Multimedia Broadcasting*, 2012.

Elagouni K., Garcia C., Mamalet F., Sébillot P., « Text recognition in multimedia documents : a study of two neural-based OCRs using and avoiding character segmentation », *International Journal on Document Analysis and Recognition (IJDAR)*, 2013.

Favre B., Damnati G., Bechet F., Bendris M., Charlet D., Auguste R., Ayache S., Bigot B., Delteil A., Dufour R., Fredouille C., Linarès G., Martinet J., Senay G., Tirilly P., « PERCOLI : a person identification system for the 2013 REPERE challenge », *Speech, Language and Audio for Multimedia (SLAM)*, p. 55-60, 2013.

Favre B., Hakkani-Tür D., Cuendet S., « Icsiboost », , http ://code.google.come/p/icsiboost, 2007.

Gaidon A., Harchaoui Z., Schmid C., « Temporal Localization of Actions with Actoms », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, p. 2782-2795, 2013.

Giraudel A., Carré M., Mapelli V., Kahn J., Galibert O., Quintard L., « The REPERE Corpus : a Multimodal Corpus for Person Recognition », *International Conference on Language Resources and Evaluation (LREC)*, 2012.

Guinaudeau C., Gravier G., Sébillot P., « Enhancing lexical cohesion measure with confidence measures, semantic relations and language model interpolation for multimedia spoken content topic segmentation », *Computer Speech & Language*, vol. 26, n^{o} 2, p. 90-104, 2012.

Hatch A. O., Kajarekar S., Stolcke A., « Within-class Covariance Normalization for SVM-based Speaker Recognition », *International Conference on Spoken Language Processing (ICSLP)*, p. 1471-1474, 2006.

Kahn J., Galibert O., Quintard L., Carré M., Giraudel A., Joly P., « A presentation of the REPERE challenge », *International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2012.

Larcher A., Bonastre J.-F., Fauve B. e. a., « ALIZE3.0 - Open Source Toolkit for State-of-the-Art Speaker Recognition », *Interspeech*, ISCA, 2013.

Linarès G., Nocera P., Massonié D., Matrouf D., « The LIA Speech Recognition System : From 10xRT to 1xRT », *Text, Speech and Dialogue (TSD)*, vol. 4629, p. 302-308, 2007.

Raymond C., « Bonzaiboost », , http ://bonzaiboost.gforge.inria.fr, 2007.

Rouvier M., Favre B., Bendris M., Charlet D., Damnati G., « Scene understanding for identifying persons in TV shows : beyond face authentication », *International Workshop on Content-Based Multimedia Indexing (CBMI)*, IEEE, p. 1-6, 2014.

Viola P., Jones M. J., « Robust Real-Time Face Detection », *Int. J. Comput. Vision*, vol. 57, p. 137-154, May, 2004.

Wallace R., McLaren M., McCool C., Marcel S., « Inter-session variability modelling and joint factor analysis for face authentication », *International Joint Conference on Biometrics (IJCB)*, IEEE, p. 1-8, 2011.

Wang X., Xie L., Ma B., Chng E. S., Li H., « Modeling broadcast news prosody using conditional random fields for story segmentation », *The Asia-Pacific Signal and Information Processing Association (APSIPA)*, vol. 1, p. 253-256, 2010.