Indexation audio-visuelle des personnes dans un contexte de télévision

Meriem Bendris

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2012-ENST-00xx

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « Signal et Image »

présentée et soutenue publiquement par

Meriem BENDRIS

le 7 juillet 2011

Indexation audio visuelle des personnes dans un contexte de télévision

Directeur de thèse : **Gérard CHOLLET**
Co-encadrement de la thèse : **Delphine CHARLET**

Jury

M. Bernard MERIALDO, Professeur, EURECOM, Sophia Antipolis	Président
Mme Régine ANDRE-OBRECHT, Professeur, IRIT, Toulouse	Rapporteur
M. Georges QUENOT, Professeur, LIG, Grenoble	Rapporteur
M. Josef SIVIC, Docteur, ENS, Paris	Examinateur
Mme Dijana PETROVSKA, Maître de conférence, TELECOM SudParis, Evry	Invitée
M. Kévin BAILLY, Docteur, TELECOM-ParisTech, Paris	Invité

TELECOM ParisTech
école de l'Institut Télécom - membre de ParisTech

T H È S E

Résumé

Le développement et l'amélioration du réseau Internet a permis de mettre un grand nombre de contenus télévisuels à disposition des utilisateurs. Afin de faciliter la navigation parmi ces vidéos, il est intéressant de développer des technologies pour indexer les personnes automatiquement. Les solutions actuelles proposent de construire l'index audio-visuel des personnes par combinaison des index audio et visuel obtenus de manière indépendante. Malheureusement, pour les émissions de télévision, il est difficile de détecter et de regrouper les personnes automatiquement à cause des nombreuses ambiguïtés dans l'audio, le visuel et leur association (interactivité des dialogues, variations de pose du visage, asynchronie entre la parole et l'apparence, etc). Les approches basées sur la fusion des index audio et visuel combinent les erreurs d'indexation issues de chaque modalité.

Les travaux présentés dans ce rapport exploitent la complémentarité entre les informations audio et visuelle afin de palier aux faiblesses de chaque modalité. Ainsi, une modalité peut appuyer l'indexation d'une personne lorsque l'autre est jugée peu fiable. Nous proposons une procédure de correction mutuelle des erreurs d'indexation de chaque modalité. D'abord, les erreurs sont détectées automatiquement à l'aide d'indicateurs de présence de visage parlant. Puis, la modalité qui a échoué est corrigée grâce à un schéma automatique.

Nous avons proposé en premier lieu un *système initial* d'indexation de visages parlants basé sur la détection et le regroupement du locuteur et du costume. Nous proposons une méthode de combinaison d'index basée sur la maximisation de la couverture globale des groupes de personnes. Ce système, évalué sur des émissions de plateaux, obtient une grande précision ($\sim 90\%$), mais un faible rappel (seulement 55% des visages parlants sont détectés).

Afin de détecter automatiquement la présence d'un visage parlant dans le processus de correction mutuelle, nous avons développé une nouvelle méthode de détection de mouvement des lèvres basée sur la mesure du degré de désordre de la direction des pixels autour de la région des lèvres. L'évaluation, réalisée sur le corpus de d'émission de plateaux, montre une amélioration significative de la détection des visages parlants comparé à l'état de l'art dans ce contexte. En particulier, notre méthode s'avère être plus robuste à un mouvement global du visage.

Enfin, nous avons proposé deux schémas de correction. Le premier est basé sur une modification systématique de la modalité considérée *a priori* la moins fiable. Le second compare des scores de vérification de l'identité non supervisée afin de déterminer quelle modalité a échoué et la corriger. Les modèles non supervisés des personnes sont appris à partir des ensembles homogènes de visages parlants obtenus automatiquement par le *système initial*. Les deux méthodes de correction conduisent à une amélioration significative des performances (+2 à 5% de la *F-mesure*).

Nous nous sommes également intéressé aux systèmes biométriques audio-visuels et particulièrement sur les techniques de fusion tardives pour la vérification d'identité. Nous avons proposé une méthode de fusion dépendante de la qualité du signal dans chaque modalité.

Abstract

With increasing internet use, the amount of multimedia content multiplies, making it necessary to develop technologies in order to enable users to browse through the multimedia data. One key element for browsing is the presence of people. However, structuring TV-Content in terms of people is a hard problem due to many difficulties in audio and visual modalities as well as in their association (short speaker turns, variations in facial expressions and pose, no synchronization between sequences of a person's appearance and sequences of his/her speech).

The goal underlying this dissertation is to structure TV-Content by person in order to allow users to navigate through sequences in which a particular indivisual appears. To this end, most methods propose indexing people separately by the audio and visual information and then associating the results of each in order to obtain a talking-face index. Unfortunately, this type of approach combines clustering errors present in each modality. Our work seeks to capitalise on interactions between the audio and visual modalities rather than treating them separately. We propose a mutual correction scheme for audio and visual clustering errors. First, the clustering errors are detected using indicators that suspect a talking-face presence (Step 1). Then, the incorrect label is corrected according to an automatic modification scheme (Step 2).

In more detail, first we proposed a *Baseline system* of talking faces indexing in which audio and visual indexes of people are generated independently by speaker and clothes clustering. Then, we proposed a fusion method based on maximizing global coverage of detected clusters. Results on a TV-show database show a high precision ($\sim 90\%$), but with a significant missed-detection rate (only 55% of talking faces sequences are detected).

In order to automatically detect a talking face presence (in the step 1), we exploited the fact that the lip-activity is strongly related to speech activity. We developed a new method for lip-activity detection in TV-Context based on the disorder of the directions of pixels. An evaluation is performed on manually annotated TV-Shows and significant improvement is observed compared to the state-of-the-art in TV-Contexts.

Next, the modification method is based on the paradigm that one modality (either audio or visual) is more reliable than the other. We proposed two modification schemes : one based on systematic correction of the supposedly less reliable modality *a priori* while the second proposes comparing unsupervised audio-visual model scores to determine which modality failed. The unsupervised models are trained from the homogeneous sets of talking faces obtained automatically by the *Baseline system*. Experiments conducted on a TV-show database show that the proposed correction schemes yield significant improvement in performance, mainly due to an important reduction of missed talking-faces.

We have investigated also on late fusion techniques for identity verification in biometric systems. We have proposed a fusion method based on the signal quality in each modality.

Remerciements

Je tiens tout d'abord à remercier *Régine André-Obrecht* et *Georges Quenot* pour avoir accepté d'évaluer mon rapport de thèse. Je remercie également *Bernard Merialdo* (président du jury), *Josef Sivic*, *Dijana Petrovska* et *Kévin Bailly* pour avoir accepté de participer à mon jury de thèse.

Je tiens à remercier particulièrement *Delphine Charlet* pour le temps qu'elle m'a consacré durant l'encadrement de ma thèse. Je la remercie de m'avoir laissé profiter des ses conseils et de son expérience.

Je voudrais également remercier *Gérard Chollet* pour son encadrement et ses précieuses remarques durant la thèse.

J'aimerais remercier spécialement *Géraldine Damnati* et *Elizabeth Godoy* pour leurs soutiens et leurs très précieux conseils.

Je remercie également *Denis Jouvet* pour tous ses bons conseils au début de ma thèse et *Kévin Bailly* pour toutes ses relectures à la fin.

Je remercie tous mes amis et collègues qui m'ont soutenu et avec qui j'ai passé de très agréables moments lors de nos activités extra-thèse. Et très particulièrement : Lizzy, Ronaldo et Annalisa, Sarah, Mounia, Fatou, Géraldine et tant d'autres. J'ai gardé d'agréables souvenirs de nos discussions et de nos sorties « découverte » de la côte de granit rose.

Je tiens à remercier ma mère ainsi que mes deux sœurs pour leurs soutiens et la confiance qu'elles ont eux pour mon travail. J'aurais tant voulu que papa soit présent dans cette aventure.

Je finirai par un petit mot à l'attention de mon compagnon qui a su être patient dans les moments les plus durs.

Meriem.

Table des matières

Introduction générale 2

I Structuration de contenus audio-visuels par personne dans un contexte de télévision 13

1 État de l'art 15
- 1.1 Principes de structuration . . . 15
  - 1.1.1 Étape de segmentation . . . 15
  - 1.1.2 Étape de regroupement . . . 15
- 1.2 Structuration basée sur l'information audio . . . 17
  - 1.2.1 Extraction des paramètres . . . 17
  - 1.2.2 Segmentation en tours de parole . . . 17
  - 1.2.3 Regroupement . . . 20
  - 1.2.4 Segmentation et regroupement conjoints . . . 21
  - 1.2.5 Limites des méthodes d'indexation en locuteurs . . . 22
- 1.3 Structuration basée sur l'information visuelle . . . 22
  - 1.3.1 Segmentation en plans . . . 23
  - 1.3.2 Détection des personnes dans un plan . . . 26
  - 1.3.3 Regroupement . . . 30
- 1.4 Fusion . . . 33
  - 1.4.1 Catégories de fusion . . . 33
  - 1.4.2 Fusion pour l'identification des personnes . . . 36
  - 1.4.3 Fusion pour la reconnaissance audiovisuelle de la parole . . . 37
  - 1.4.4 Fusion pour la détection de visages parlants . . . 37
  - 1.4.5 Fusion pour la structuration de documents audio-visuels . . . 39
- 1.5 Conclusion . . . 41

2 Contexte d'étude et corpus 43
- 2.1 Inventaire des corpus audio-visuels . . . 43
  - 2.1.1 Corpus audio . . . 43
  - 2.1.2 Corpus visuels . . . 44
  - 2.1.3 Corpus audio-visuel . . . 45
- 2.2 Présentation de la base de données *TSDB* . . . 46

2.2.1 Description du corpus . . . 46
2.2.2 Annotations . . . 47
2.3 Analyse du corpus *TSDB* . . . 50
2.4 Problématiques d'indexation des personnes dans les *Talk shows* . . . 51
2.4.1 Difficultés de l'information audio . . . 52
2.4.2 Difficultés de l'information visuelle . . . 52
2.4.3 Difficultés dans l'association des informations audio et visuelle 53
2.4.4 Avantages des *Talk shows* . . . 54
2.5 Conclusion . . . 54

3 Protocole d'évaluation **57**
3.1 Tour d'horizon . . . 57
3.1.1 Mesure de pureté . . . 57
3.1.2 Protocole d'évaluation *TRECVID* . . . 58
3.1.3 Protocole d'évaluation *ESTER* . . . 59
3.2 Proposition d'un protocole d'évaluation . . . 61
3.2.1 Notations . . . 61
3.2.2 Origines des erreurs . . . 61
3.2.3 Mesures de performances . . . 62

4 Système de structuration de documents audio-visuels **67**
4.1 Système basé sur l'information audio . . . 67
4.1.1 Traitements préliminaires . . . 68
4.1.2 Première phase : structuration et segmentation disjoints . . . 68
4.1.3 Seconde phase : structuration et segmentation conjoints . . . 68
4.1.4 Résultats et discussion . . . 69
4.2 Système basé sur l'information visuelle . . . 71
4.2.1 Détection des costumes . . . 72
4.2.2 Représentation du costume . . . 72
4.2.3 Sélection du meilleur costume . . . 73
4.2.4 Regroupement des costumes . . . 73
4.2.5 Résultats et discussion . . . 75
4.3 Appariement audio-visuel . . . 76
4.3.1 Recherche d'associations . . . 76
4.3.2 Fusion d'index . . . 77
4.3.3 Résultats et discussion . . . 78
4.4 Conclusion . . . 82

5 Activité visuelle de la parole **85**
5.1 Tour horizon . . . 85
5.2 Système de détection de l'activité des lèvres . . . 86
5.2.1 Description . . . 86
5.2.2 Extraction des caractéristiques du visage . . . 87
5.2.3 Sélection de la région des lèvres . . . 88
5.2.4 Mesure de l'activité visuelle de la parole . . . 89
5.3 Évaluation . . . 90
5.3.1 Expérience . . . 90
5.3.2 Protocole . . . 90
5.3.3 Résultats des alignements . . . 91
5.3.4 Résultats et discussion . . . 91
5.4 Mesure de l'activité visuelle de la parole dans le système de structuration 95
5.4.1 Système de structuration audio-visuel modifié . . . 95
5.4.2 Résultats et discussion . . . 97
5.4.3 Conclusion . . . 98

II Identification audio-visuelle des personnes **103**
Introduction . . . 105

6 Modèles de vérification de l'identité **109**
6.1 Présentation des systèmes de vérification de l'identité . . . 109
6.1.1 Système de vérification du locuteur . . . 109
6.1.2 Système de vérification de l'identité visuelle . . . 111
6.1.3 Fusion des scores . . . 114
6.2 Introduction de mesures de qualité dans la fusion . . . 115
6.2.1 Tour d'horizon . . . 115
6.2.2 Notre méthode de fusion . . . 116
6.3 Conclusion . . . 118

7 Application des modèles non-supervisés pour la structuration **121**
7.1 Principe . . . 121
7.2 Les modèles de vérification audio-visuelle de l'identité . . . 122
7.2.1 Vérification de l'identité audio . . . 123
7.2.2 Vérification de l'identité visuelle . . . 123
7.2.3 Vérification de l'identité audio-visuelle . . . 124
7.3 Les hypothèses de modification . . . 124
7.4 Améliorations basées sur la vérification d'identité de l'audio-visuelle . 126

7.4.1 Détection des erreurs de regroupement . . . 126
7.4.2 Schéma de modification basé sur la vérification d'identité de l'audio-visuelle . . . 126
7.5 Expériences . . . 127
7.6 Résultats et discussion . . . 128
7.7 Conclusion et perspectives . . . 129

8 Application des modèles dans les systèmes biométriques 131
8.1 Les systèmes Biométriques . . . 131
8.2 Base de données *Banca* . . . 133
8.3 Performances des modèles de personnes dans le système biométrique . 135
8.4 Introduction de mesures de qualité dans la fusion . . . 137
8.4.1 Les dépendances . . . 138
8.4.2 Les mesures de qualité dans *Banca* . . . 139
8.4.3 Détection des classes de dégradation du signal dans *Banca* . . 141
8.4.4 Estimation des paramètres de fusion . . . 142
8.4.5 Résultats . . . 143
8.5 Conclusion . . . 144

Conclusions générales et perspectives 146

***Annexe A* : Le système de structuration sur le *Grand Échiquier* 154**

***Annexe B* : Évaluation de l'outil *Stasm* sur *BioID* 159**

***Liste des publications* 166**

***Liste des acronymes* 170**

Introduction générale

Grâce aux nouvelles technologies sur internet, nous constatons une prolifération des contenus multimédia (vidéo à la demande, YouTube, *INA*, etc). La grande majorité de ces contenus est issue de la télévision. En effet, les chaînes de télévision mettent à disposition de leurs utilisateurs une partie des émissions en rediffusion grâce à des sites Internet de type « *replay* »ou des fournisseurs de vidéos d'informations comme *2424actu.fr* par Orange ou *Exalead*. Toutes ces nouvelles applications Internet représentent une très grande base de données de contenus issus de la télévision. Et bien qu'il existe plusieurs technologies disponibles pour collecter et stocker ces contenus, les technologies pour faciliter l'accès à ces données restent encore à développer. Il devient donc nécessaire de développer des technologies d'indexation automatique afin de faciliter la navigation dans ces contenus. Une des clés pertinentes de navigation dans ces contenus est l'annotation de personnes. L'indexation audio-visuelle des personnes a pour objectif de permettre à un utilisateur de localiser des séquences d'interventions télévisées d'une personne.

Intérêt

L'objectif de l'indexation des personnes dans des contenus de télévision est d'annoter automatiquement les interventions audio-visuelles des personnes. L'intérêt majeur d'indexer ces interventions est de faciliter la navigation dans le contenu par un moteur de recherche. Son application principale est de permettre à des utilisateurs de localiser automatiquement les interventions d'une certaine personnalité recherchée sans avoir à jouer toute la vidéo. L'indexation des personnes est également utilisée comme première phase de traitement pour la recherche d'autres informations sémantiques telles que la présence de personnes, la détection d'événements, le chapitrage par sujet de discussion à partir des identités des personnes intervenants (politiques, musiciens, sportifs, etc).

Contexte d'émissions de plateaux (*Talk Shows*)

L'indexation automatique de personnes dans un contexte de *Talk Shows* est un problème très difficile en raison de nombreuses ambiguïtés que présentent l'information audio, l'information visuelle ainsi que leur association. La figure 1 présente les caractéristiques d'interventions dans un contexte de télévision.

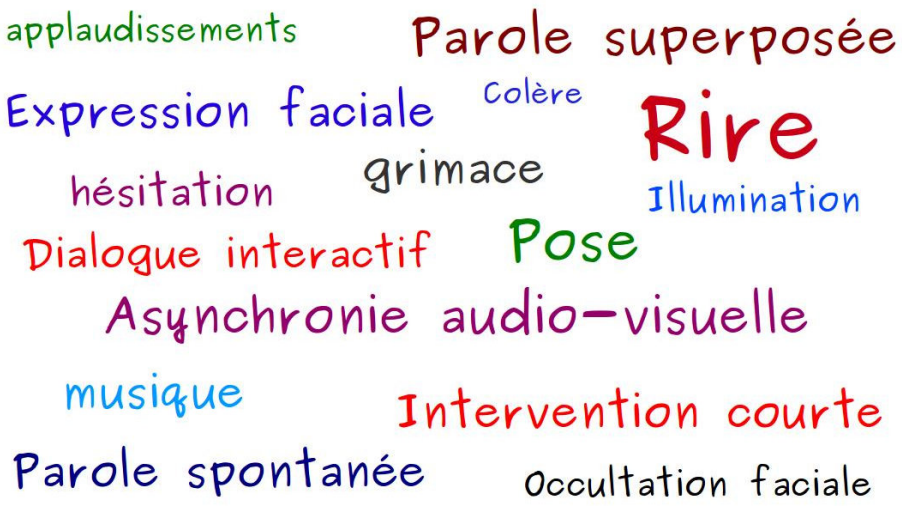


FIGURE 1 – Les difficultés de l'indexation des contenus TV

D'abord, dans ce contexte l'information audio se caractérise par de la parole spontanée et expressive avec un dialogue très interactif. Les tours de parole sont courts et souvent plusieurs personnes interviennent au même moment. Toutes ces difficultés rendent les techniques existantes d'indexation en locuteur peu fiables.

Concernant l'information visuelle, la plupart des méthodes sont basées sur la détection et l'identification du visage. Dans un contexte de télévision, les visages se présentent avec beaucoup de variations de pose, de conditions d'éclairage et d'expression faciale. Toutes ces ambiguïtés rendent difficile la détection et l'identification des personnes par un système automatique basé sur le visage. D'autres informations sujettes à moins de variabilités peuvent être utilisées afin de détecter et/ou d'identifier les personnes comme les habits, les cheveux ou l'arrière plan. La figure 2 présente des exemples de variations classiques d'apparence des personnes.

Enfin, l'association des deux modalités audio et visuelle consiste à déterminer la liaison entre chaque personne détectée par l'information audio avec le visage qui

FIGURE 2 – Exemples de variations classiques d'apparence des personnes

lui correspond. La plupart des méthodes de détection et d'identification des personnes dans les séquences vidéo considèrent que la personne visible au moment de la détection d'un locuteur est celle qui lui est associée. Malheureusement, dans un contexte de télévision, il n'est pas garanti que les plans visuels d'une personne soient synchronisés avec les séquences de ses interventions audio. En effet, les personnes peuvent parler et ne pas être filmées ou le contraire. Les plans contenant plusieurs visages constituent également une difficulté dans l'association des informations audio et visuelle. Lorsque plusieurs personnes sont détectées, il devient difficile de déterminer le « bon »visage à associer au locuteur.

Que chercher ?

L'intérêt de l'indexation des personnes dans un contexte de télévision est de faciliter la recherche d'une intervention d'une personnalité. Selon l'information recherchée, trois applications sont possibles :

1. La recherche d'interventions sonores des personnes : cas où l'utilisateur souhaite naviguer dans les interventions audio d'une personne. Ces séquences sonores peuvent être également un premier traitement dans la recherche d'informations sémantiques contenues dans le discours de la personne recherchée. Par exemple : rechercher la séquence d'intervention d'une personne qui parle d'un sujet précis.
2. La recherche des séquences d'apparition des personnes : cas où l'utilisateur souhaite naviguer dans les séquences d'apparitions d'une personne. À partir de ces séquences, d'autres informations visuelles peuvent être recherchées comme l'âge, les vêtements, la coupe de cheveux, etc.

3. La recherche des séquences de visage parlants : cas où l'utilisateur recherche les séquences d'apparition d'une personne lorsqu'elle parle.

Dans nos approches, nous avons considéré les trois applications possibles. Pour cela, nous avons fait le choix de construire de manière indépendante deux index : l'un basé sur l'information audio et l'autre sur l'information visuelle. L'index de visage-parlant est obtenu par combinaison tardive des deux index audio et visuel. Ainsi, selon l'application souhaitée par l'utilisateur, il est possible de naviguer dans le contenu par interventions sonores, visuelles ou les deux.

Comment chercher ?

Dans notre étude, nous souhaitons offrir à un utilisateur la possibilité de naviguer par personne dans un contenu et de retrouver des interventions spécifiques. Afin d'atteindre ces objectifs, deux grandes phases sont nécessaires : la structuration par personne et l'identification. Chacune de ces deux phases constitue un domaine de recherche à part entière. La figure 3 présente les résultats de chaque phase d'indexation des personnes.

Structuration audio-visuelle

La structuration d'un document audio-visuel par personne consiste à localiser les interventions des individus sans utiliser de dictionnaire pré-défini de leurs identités. Il s'agit de détecter les personnes automatiquement et de les regrouper dans des classes homogènes contenant chacun une seule personne. La structuration s'effectue dans un seul document et produit un index décrivant toutes les interventions audio-visuelles des personnes. Deux étapes sont nécessaires : la détection de personne et le regroupement des interventions détectées. La détection de personnes dans des contenus audio-visuels est une grande problématique de recherche. Les difficultés introduites par le contexte de télévision rendent difficile l'utilisation du visage seul pour détecter une personne. D'autres informations peuvent être utilisées telles que la silhouette ou le costume. L'étape de regroupement (appelé également *Clustering*) consiste à réunir les séquences de personnes détectées automatiquement dans des classes contenant chacune une seule personne. L'intérêt de la structuration est de permettre à l'utilisateur de naviguer dans le document audio-visuel même si l'on ne possède pas les informations nécessaires à l'identification de chaque personne (pas de modèles d'identification).

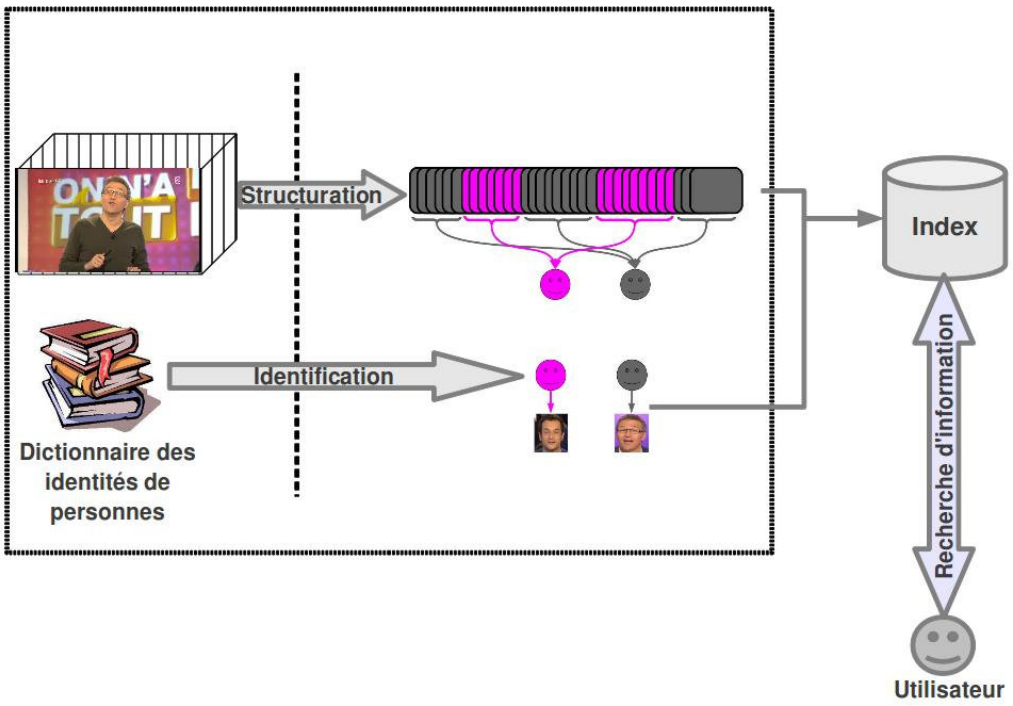


FIGURE 3 – Composants du système d'indexation de personnes. La phase de structuration permet d'obtenir un index permettant la navigation dans le contenu par personne. La phase d'identification permet d'associer des identités à l'index obtenu

Identification audio-visuelle

L'identification de personne consiste à déterminer l'identité d'un individu parmi une population. Dans un contexte de télévision, l'ensemble de la population est un ensemble ouvert. Il est donc impossible de disposer d'un modèle associé à chaque personne qui va potentiellement intervenir dans un contenu de télévision. Ainsi, en indexation, le problème d'identification de la personne reviendrait à vérifier si la personne détectée possède bien l'identité de la personne recherchée. La vérification de l'identité nécessite une collection d'exemples de la personne recherchée afin d'apprendre un modèle de vérification de l'identité.

Objectif

La thèse est effectuée au sein du centre de recherche de *France Télécom - OrangeLabs* dans le cadre du projet *Media search* qui vise à développer des nouvelles technologies pour faciliter l'accès et la navigation dans des contenus multimédias.

L'objectif principal de la thèse est d'étudier des solutions robustes aux difficultés de l'indexation des personnes dans des documents de type émissions de plateaux. Dans ce contexte, les modalités audio et visuelle sont souvent peu fiables. La philosophie adoptée durant notre étude est basée sur le principe de la complémentarité des deux modalités. Nous exploitons cette complémentarité afin d'améliorer les résultats de chaque système d'indexation mono-modal. Ainsi, l'information audio peut appuyer l'indexation d'une personne lorsque l'information visuelle est jugée peu fiable. De manière inverse, lorsque l'audio est jugé non fiable, l'information visuelle est utilisée afin de confirmer l'indexation. L'index de visages parlants résultant de la combinaison des deux modalités est amélioré à son tour grâce à la combinaison des améliorations « *à double sens* »obtenues dans chaque modalité.

Contributions

D'abord, nous avons développé un système de structuration des documents audio-visuels par personne en utilisant l'information audio et visuelle de manière indépendante, puis en combinant ces index par une fonction d'appariement afin d'obtenir les séquences de visages-parlants. La structuration basée sur l'information audio est effectuée par une méthode de détection et regroupement par la signature vocale des personnes. La structuration basée sur l'information visuelle est effectuée par la détection du visage et costume et le regroupement par la signature des costumes des personnes. Ensuite, nous avons proposé une méthode d'appariement basée sur la maximisation de la couverture globale des groupes de personnes détectées et regroupées. L'avantage de cette méthode est qu'elle prend en compte l'asynchronie, très récurrente dans le contexte d'études, entre les interventions sonores et visuelles des personnes. Un bon taux de précision sur les interventions audio-visuelles des personnes est obtenu comparé à la méthode proposée dans [Khoury et al., 2010]. Par contre, le taux de rappel obtenu est jugé subjectivement faible pour l'application (c-à-d pas assez d'interventions audio-visuelles ne sont détectées par émission). Ce faible rappel est principalement dû aux erreurs de regroupement du système audio et visuel dans certaines séquences ambiguës (difficultés introduites par le contexte de télévision).

Ensuite, ce système de structuration, considéré comme le système de référence,

est amélioré afin de retrouver les séquences de visages parlants non détectés à cause des ambiguïtés introduites par le contexte de télévision. Cette amélioration est basée sur la remise en cause de la fiabilité des annotations automatiques audio et visuelles de manière à ce que chaque modalité corrige les erreurs de structuration de l'autre modalité. Dans ce paradigme, nous avons d'abord travaillé sur la détection automatique des erreurs de structuration puis sur des procédures de corrections de ces erreurs.

D'abord, les erreurs de structuration sont localisées par la détection de présence de visages parlants. Pour cela, nous avons travaillé sur la détection automatique de l'activité visuelle de la parole. Nous avons proposé une nouvelle méthode basée sur la mesure du degré de désordre de la direction des pixels autour de la région des lèvres. Les résultats démontrent que dans un contexte de *Talk Shows*, la méthode proposée pour la détection de présence de visage parlant est plus robuste au mouvement du visage comparée à l'état de l'art.

Ensuite, nous proposons deux schémas de correction de la modalité qui a échoué. Dans le premiers schéma, nous considérons avoir connaissance *a priori* de la modalité la moins fiable. Dans le second schéma, considérant la grande précision du système de référence, des modèles non supervisés sont appris pour chaque personne à partir des séquences détectées automatiquement par le système de structuration. Ces modèles sont utilisés afin de vérifier l'identité de la personne dans chaque segment perdu dans lequel une présence de visage parlant est détectée. Ainsi, le score de vérification de l'identité audio-visuel détermine pour chaque segment la modalité la moins fiable (à corriger). Les résultats montrent une amélioration significative du rappel expliquée par la récupération d'une grande partie des visages parlants non détectés par le système de référence.

En indexation audio-visuelle des personnes, très peu de travaux ont été effectués sur des données de télévision. Durant la thèse, il n'existait pas de données publiques de contenus de télévision de type *Talk Shows* annotées en audio et en visuel. À notre connaissance, cela peut s'expliquer par deux raisons :

- Le domaine est relativement nouveau et nécessite des efforts de coordination de la part de la communauté scientifique. En novembre 2010, l'Agence Nationale de la Recherche (*ANR*) a lancé le projet *REPERE*[1] (REconnaissance de PERsonnes dans des Émissions audiovisuelles) afin d'encourager la recherche dans ce contexte et dont la première tâche est la constitution d'une base de données commune d'émissions de télévision (*JTs* et *Talk shows*).
- La seconde raison est qu'il est très difficile de négocier avec des chaînes de

1. www.agence-nationale-recherche.fr/programmes-de-recherche/appel-detail/

télévisions les droits de collecte et d'exploitation des émissions contenant des personnalités.

Au cours de la thèse, il a été nécessaire de constituer, à titre privé, une base de données adaptée à notre tâche d'indexation de *Talk shows*. Cette base de données comprend 5 épisodes de l'émission populaire « *On n'a pas tout dit* »diffusée sur la chaîne publique française *France 2*. Une annotation très fine des interventions sonores et visuelles des personnes est effectuée. Il a été également nécessaire de définir un protocole d'évaluation.

Au cours de nos expérimentations, nous nous sommes intéressés aux modèles de vérification de l'identité audio-visuelle. Nous avons soulevé le problème de la sensibilité des modèles de vérification à la qualité du signal. Nous avons proposé une méthode de fusion des scores de vérification de l'identité basée sur des mesures de qualité. L'intérêt est d'adapter la confiance accordée à chaque modalité (accorder plus d'importance au système basé sur l'image dans le cas où la séquence de parole est bruitée ou plus d'importance au système basé sur l'audio lorsque l'image n'est pas de bonne qualité). Nous avons mis en évidence le problème de la dépendance du seuil de décision à la qualité du signal que nous avons résolu par l'adaptation des paramètres de normalisation à la qualité du signal. Les expériences effectuées dans la base de donnée *Banca* [Bailly-Bailliére et al., 2003] montre une amélioration significative des résultats obtenus par cette méthode en comparaison avec les méthodes de fusion des scores normalisés utilisées dans les systèmes biométriques.

Organisation

Au cours de la thèse, nous avons distingué deux grandes phases de l'indexation de document audio-visuel : la structuration sans modèles de personne et la structuration avec modèles de personnes (vérification de l'identité). Suivant ces considérations, le rapport est organisé de la manière suivante :

La première partie du rapport est dédiée à la structuration des contenus de télévision en personne sans aucun dictionnaire d'identité de personne. D'abord, dans le chapitre 1, nous présentons un état de l'art de la structuration de document par personne. Ensuite, dans le chapitre 2, nous présentons la base de données de *Talk Shows TSDB* que nous avons constituée et sur laquelle nous avons évalué nos méthodes d'indexation audio-visuelles de personnes. Dans ce chapitre, nous effectuons une brève présentation des émissions collectées ainsi qu'une analyse détaillée du corpus. Dans le chapitre 3, nous présentons le protocole utilisé afin d'évaluer nos

méthodes. Le chapitre 4 est dédié à la présentation du système de structuration basé sur le regroupement de personnes en utilisant l'information audio, visuelle et leur combinaison. Le chapitre 5 présente le processus de détection d'activité visuelle de la parole et son intégration dans le système de structuration.

La seconde partie du rapport est dédiée à l'indexation des personnes basées sur des systèmes de vérification de l'identité. Dans le chapitre 6, nous présentons les modèles de vérification de l'identité utilisés dans nos expériences. Ces modèles peuvent être supervisés (appris sur des données annotées manuellement) ou non-supervisés (appris sur des données annotées automatiquement). Dans le chapitre 7, nous présentons l'application des modèles de vérification non-supervisée de l'identité pour l'amélioration de la structuration de document audio-visuel par personne. Dans le chapitre 8, nous présentons l'application des modèles de vérification de l'identité dans les systèmes biométriques pour l'identification des personnes. Cette application est basée sur un apprentissage supervisé des modèles des personnes (c.à.d que nous utilisons des collections de données annotées manuellement afin d'apprendre les modèles des personnes). Des expériences sont effectuées sur la base de données biométrique *Banca*

Première partie

Structuration de contenus audio-visuels par personne dans un contexte de télévision

Chapitre 1

État de l'art

La structuration des contenus audio-visuels par personne exploite la complémentarité des deux informations (audio et visuelle) afin de localiser automatiquement les séquences d'intervention des personnes. La plupart des travaux qui existent dans la littérature considèrent l'intervention audio-visuelle comme une combinaison de deux modalités indépendantes. Dans ce chapitre, nous détaillons l'état de l'art des méthodes de structuration de documents par personne basées sur l'information audio et visuelle de manière indépendante, puis leurs combinaisons. L'objet de ce chapitre est de faire un état de l'art général et de décrire les méthodes les plus utilisées.

1.1 Principes de structuration

Structurer un document audio-visuel par personnes consiste à détecter et regrouper automatiquement les interventions de chaque personne du document. Dans les méthodes de structuration, on distingue deux étapes : la segmentation et le regroupement.

1.1.1 Étape de segmentation

La segmentation d'un document audio-visuel consiste à détecter automatiquement les personnes intervenant de manière audio et/ou visuelle dans le contenu. L'objectif est de découper le document en segments homogènes contenant chacun une seule personne. En audio, les segments homogènes doivent contenir chacun un seul locuteur. Dans la modalité visuelle, les segments homogènes doivent doivent contenir les plans d'apparence des personnes.

1.1.2 Étape de regroupement

La phase de regroupement (appelée *Clustering*) consiste à regrouper les segments appartenant à chaque personne dans un seul *Cluster*. Les méthodes les plus utilisées sont basées sur l'approche classique de regroupement hiérarchique [Johnson, 1967].

Cette approche est basée sur la construction des *Clusters* de manière itérative. Il existe deux types de regroupement hiérarchique : *hiérarchique ascendant* et *hiérarchique descendant.*

Regroupement hiérarchique ascendant

Le regroupement hiérarchique ascendant est le plus utilisé pour la structuration. Le principe est de regrouper de manière itérative les plus proches segments selon une mesure de similarité. La figure 1.1 montre un exemple de regroupement d'une séquence audio segmentée préalablement en 9 segments contenant un seul locuteur. Il existe plusieurs approches pour mesurer la similarité entre les groupes :

- *Single linkage Clustering* : distance entre la paire d'éléments les plus proches, où chaque élément appartient à un groupe.
- *Complete linkage Clustering* : distance entre la paire d'éléments les plus éloignés, où chaque élément appartient à un groupe.
- *Average linkage Clustering* : distance entre l'élément moyen de chaque groupe.
- *Average group linkage Clustering* : moyenne des distances entre chaque paire d'éléments, où chaque paire est composée de deux éléments appartenant chacun à un groupe.
- Regroupement *Ward* : minimum de perte d'information [Ward, 1963].

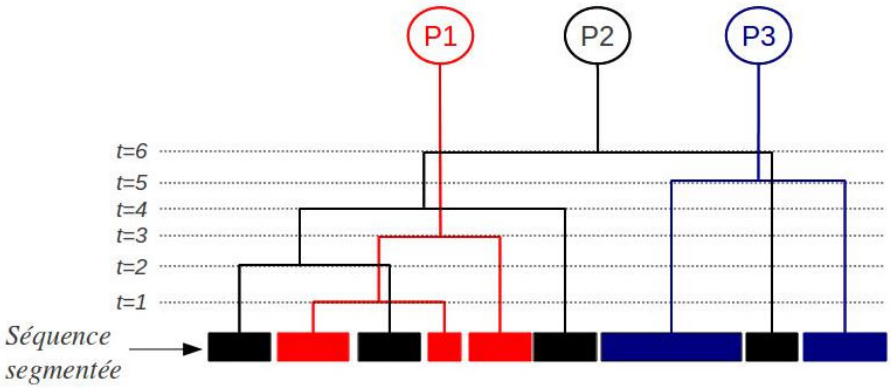


FIGURE 1.1 – Exemple de regroupement hiérarchique ascendant d'une séquence segmentée préalablement en 9 segments. À chaque itération, les deux segments les plus proches sont regroupés. L'algorithme s'arrête après 6 itérations, lorsque toutes les similarités mesurées entre les groupes construits est inférieur un seuil fixé *a priori*. Trois personnes distinctes résultent de l'algorithme de regroupement

Regroupement hiérarchique descendant

Les méthodes de regroupement hiérarchique descendants (appelées *top-down*) sont basées sur un processus de division des *Clusters*. Initialement, les segments sont associés à un seul groupe. Ensuite, à chaque itération, une procédure de division est appliquée sur chaque groupe. La procédure de division est comme suit :

- Initialement, chaque groupe de segments (*nœud parent*) est divisé arbitrairement en n sous-groupes (*nœuds enfants*) de manière à ce qu'il y ait approximativement le même nombre de segments dans chaque sous-groupe.
- Pour chaque segment, une distance à chaque sous-groupe est calculée. Le segment est assigné au sous-groupe le plus "*proche*".
- Ces itérations sont répétées tant qu'il y a du mouvement.

1.2 Structuration basée sur l'information audio

L'objectif de la structuration de documents par personne basée sur l'information audio (appelée également *indexation en locuteurs*) est de détecter les interventions sonores des personnes (tours de parole) et de les regrouper par personne. Souvent, l'indexation en locuteurs prend l'hypothèse que l'on ne possède pas de dictionnaire prédéfinie des locuteurs potentiellement présents dans le document audio. Cette hypothèse permet de traiter des documents contenant des locuteurs recherchés (à identifier ultérieurement) et inconnus (que l'on ne souhaite pas indexer). L'architecture générale d'un système d'indexation en locuteurs est divisée en 3 étapes distinctes : l'extraction des paramètres, la segmentation et le regroupement (voir figure 1.2).

1.2.1 Extraction des paramètres

La première étape consiste à extraire des paramètres acoustiques de la bande sonore contenant des interventions de plusieurs locuteurs. Les paramètres fréquemment extraits sont les coefficients *MFCC* (Mel Frequency Cepstral Coefficients) avec un nombre varié de coefficients et de complémentaires (dérivées premières et secondes).

1.2.2 Segmentation en tours de parole

Après extraction des paramètres acoustiques, l'étape de segmentation consiste à découper la séquence audio en petits segments homogènes supposés contenir chacun de la parole provenant d'un seul locuteur. Dans la littérature, deux niveaux de segmentations sont utilisés séparément ou de manière combinée :

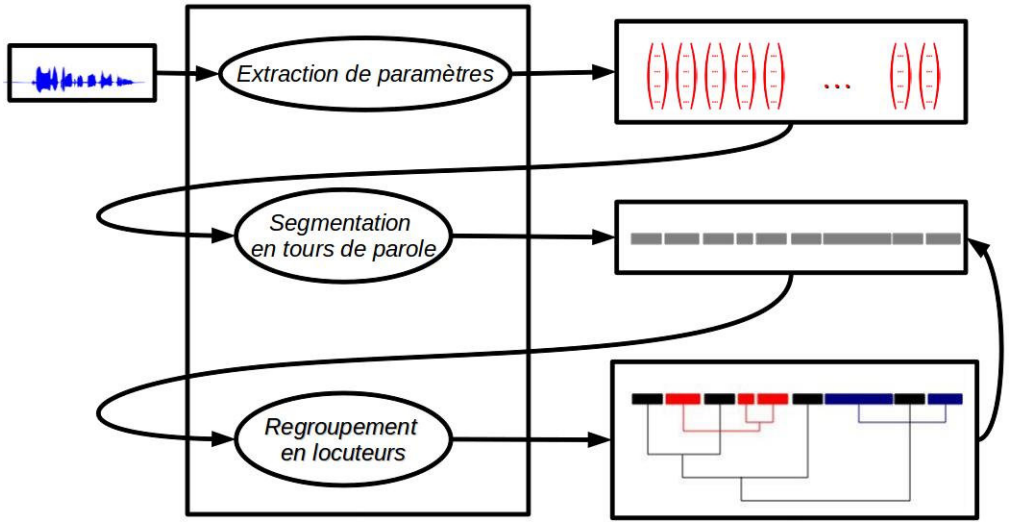


FIGURE 1.2 – Architecture générale d'un système d'indexation en locuteurs

Segmentation par détection de la parole et non-parole

La première étape de segmentation consiste à séparer les zones de parole des zones de non-parole. Les zones de non paroles peuvent être des segments de silence, de bruit, musique, rire, applaudissements, etc. Pour détecter les zones de silence, la méthode la plus simple est d'étudier les pics d'énergie. Pour détecter des zones de non parole plus difficiles (musique, rire, bruit, etc), il est nécessaire de modéliser plus finement le son. L'approche la plus fréquemment utilisée repose sur une modélisation *HMM* à deux états (parole et non-parole), où chaque état est un modèle de mélange de gaussiennes (*GMM*). Ensuite, un décodage *Viterbi* permet de faire la segmentation de la séquence en trouvant la séquence d'états (parole et non-parole) la plus probable. Afin d'affiner la segmentation, il est possible de modéliser plus finement les classes parole et non-parole. Dans [Reynolds and Torres-Carrasquillo, 2005], il est proposé de modéliser la parole par les classes : parole, parole bruité, parole sur musique et la non-parole par les classes : musique, silence et bruit.

Segmentation par détection de changements acoustiques

La seconde étape consiste à découper les segments classés comme parole en séquences homogènes contenant un seul locuteur. Les méthodes les plus utilisées sont basées sur la détection de changements acoustiques. Le principe est de mesurer à l'instant t une similarité entre deux fenêtres consécutives. Cette mesure est com-

parée à un seuil de décision de rupture (un changement de locuteur est annoté si la similarité entre deux séquences consécutives est faible). Plusieurs mesures de similarités sont proposées dans la littérature : la divergence de *Kullback-Leibler* [Siegler et al., 1997], le critère *Hotelling* [Bowen and Hansen, 2005], ou le rapport de vraisemblance généralisée ΔGLR qui mesure pour un vecteur de paramètres acoustiques $X(x_1, ..x_n)$ un rapport de vraisemblance entre la vraisemblance des deux hypothèses :

- H_1 : la séquence $X(x_1, ..x_n)$ est prononcée par une seule personne.
- H_2 : la séquence $X = X_1 \cup X_2$ est prononcée par deux personnes $X1(x_1, ..x_p)$ et $X2(x_{P+1}, ..x_n)$.

Soient M_1 et M_2 les modèles respectifs de X_1 et X_2. Le rapport de vraisemblance généralisé ΔGLR est calculé de la manière suivante :

$$\Delta GLR = log(\frac{Ł(X_1/M_1) \times Ł(X_2/M_2)}{Ł(X_1 \cup X_2/M_1 \cup M_2)}) \tag{1.1}$$

Dans le cas où X_1 et X_2 sont mono-gaussiens, le rapport ΔGLR est calculé de la manière suivante :

$$\Delta GLR = \frac{n}{2}\log(|\Sigma_X|) - \frac{p}{2}\log(|\Sigma_{X1}|) - \frac{n-p}{2}\log(|\Sigma_{X2}|) \tag{1.2}$$

où Σ est la matrice de covariance. Généralement, le rapport GLR est combiné avec une pénalité sur la complexité du modèle. Cette métrique, appelé critère BIC [Chen and Gopalakrishnan,] (Bayesien Information Criterion), est calculée de la manière suivante :

$$\Delta BIC = \Delta GLR - \frac{\lambda}{2}(d + \frac{d(d+1)}{2})\log(n) \tag{1.3}$$

où d est la dimension des paramètres acoustiques extraient et λ le poids de pénalité. Dans le cas ou le ΔBIC est positif (la différence entre les hypothèses est positive) l'hypothèse H_1 est adoptée (H_2 si négatif). Généralement, dans toutes les méthodes de regroupement en locuteurs basées sur un regroupent hiérarchique (*KL*, *Hotelling*, *GLR)*, les segments sont modélisés par des mono-gaussiennes à covariance pleine. Cela permet de calculer rapidement les mesures de similarités.

À l'issue de l'étape de segmentation, le document sonore que l'on souhaite structurer par locuteur est découpé en petits segments contenant chacun un seul locuteur (voir la figure 1.2). Plusieurs couches d'informations peuvent être rajoutées aux segments détectés afin de faciliter le regroupement des segments du même locuteur. Dans [Reynolds and Torres-Carrasquillo, 2005], il est proposé d'ajouter une classification des segments par genre (masculin ou féminin) et par bande passante (afin de détecter les conditions d'enregistrement). Cette démarche a pour but de fournir des informations afin d'optimiser les paramètres de regroupement pour chaque groupe.

1.2.3 Regroupement

Après avoir segmenté le document sonore en locuteurs, la phase de regroupement consiste à classer les segments de chaque locuteur dans un seul groupe. On considère que l'on ne possède aucune information à priori sur les locuteurs et que le traitement du document est hors ligne (c.à.d que l'on procède en connaissant l'ensemble des segments à regrouper).

Regroupement hiérarchique ascendant

La technique la plus utilisée est basée sur des algorithmes de classification *hiérarchique ascendante*. Les segments sont souvent modélisés par des représentations gaussiennes des paramètres acoustiques. Afin de mesurer la similarité entre deux segments, plusieurs métriques sont proposées dans la littérature : la divergence *Kullback-Leibler* [Goldberger and Roweis, 2004], le rapport de vraisemblance croisé (CLR) [Reynolds et al., 1998]. Dans [Siegler et al., 1997], le critère δBIC est utilisé comme mesure de similarité entre deux segments (regrouper si $\Delta BIC < 0$) et comme critère d'arrêt du regroupement (lorsque tous les groupes ont des ΔBIC positif).

Dans le processus de regroupement hiérarchique ascendant, au départ, on dispose de très peu de données pour chaque *Cluster*. C'est pourquoi la modélisation des segments est mono-gaussienne. Au cours du processus de regroupement, lorsque les *Clusters* sont jugés suffisamment importants, le processus de regroupement se poursuit avec des modélisation multi-gaussiennes. Dans [Zhu et al., 2005], les auteurs proposent de démarre par le processus de regroupement basé sur le critère *BIC*. Puis, une fois les *Clusters* jugés assez gros, les segments sont modélisés par des *GMM*. Le processus de regroupement se poursuit par un *CLR*. Cette méthode s'avère meilleure que le regroupement basé sur du le critère *BIC* classique.

Regroupement hiérarchique descendant

Très peu de méthodes de regroupement hiérarchique descendant sont utilisées pour le regroupement en locuteurs. Dans [Sue E. Johnson, 1998], la procédure de division est comme suit :

- Chaque groupe de segments est divisé arbitrairement en 4 sous-groupes de manière à ce qu'il y ai approximativement le même nombre de segments dans chaque sous-groupe.
- La moyenne et covariance de chaque sous-groupe est calculée.
- Pour chaque segment, une distance (basée sur la divergence gaussienne) à chaque sous-groupe est calculée. Le segment est assigné au sous-groupe le plus "*proche*".
- Si à l'issue de l'étape précédente un groupe contient très peu de segments (< 25 segments), il est supprimé et ses segments sont dispersés de manière arbitraire sur les autres sous-groupes.
- Ces itérations sont répétées tant qu'il y a du mouvement.

1.2.4 Segmentation et regroupement conjoints

Il existe des méthodes qui proposent d'effectuer la segmentation et le regroupement de manière simultanée. Ces méthodes sont basées sur des chaînes de Markov cachées *HMM* dans lequel chaque état représente un locuteur (modèle de voix), et les transitions représentent le passage d'un locuteur à un autre [Meignier et al., 2001, Deléglise et al., 2005, Wooters and Huijbregts, 2008]. La segmentation et regroupement sur la séquence audio est effectué par un algorithme *Viterbi* qui permet de découper la parole en segment contenant un seul locuteur (généré par un même état) et de détecter les changements de locuteurs (transitions). Malheureusement, cette méthode prend l'hypothèse que l'on possède une modélisation de la voix de chaque locuteur ce qui n'est pas souvent le cas.

Généralement, en indexation en locuteur, le *HMM* est initialisé par les résultats d'une première segmentation/regroupement disjoints [Wooters and Huijbregts, 2008, Deléglise et al., 2005]. Premièrement, la séquence audio est découpée en segments contenant chacun un seul locuteur. Ensuite, une méthode de regroupement est appliquée afin de classer les segments par locuteur. Puis un modèle de la voix est appris pour chaque *Cluster* détecté afin d'initialiser le *HMM*. Une nouvelle segmentation et une nouvelle attribution des segments aux états du *HMM* sont obtenues par décodage *Viterbi*. À partir de ces segments, un nouveau regroupement est effectué (basé sur le critère *BIC* dans [Wooters and Huijbregts, 2008] et *CLR* dans

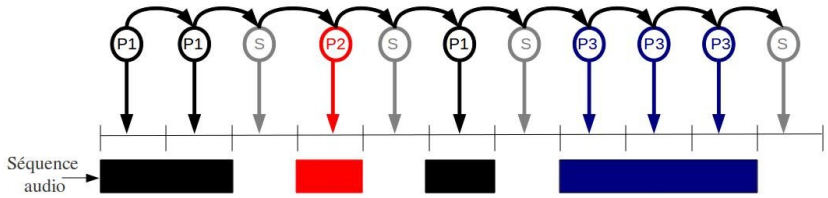


FIGURE 1.3 – Exemple de segmentation et regroupement conjoints d'une séquence audio par *HMM* à 4 états : 3 états locuteurs (P1), (P2), (P3) et un état représentant le silence (S)

[Deléglise et al., 2005]). le *HMM* évolue de manière itérative grâce à la nouvelle segmentation et regroupement. Le HMM réinitialisé peut ainsi re-segmenter la séquence. Ces étapes peuvent être appliquées jusqu'à ce que le découpage se stabilise. Les performances obtenues par cette méthode sont, à ce jour, les meilleurs dans l'état de l'art. Désormais, cette approche évolutive du *HMM* est considérée par la communauté parole comme l'approche de référence en indexation en locuteurs

1.2.5 Limites des méthodes d'indexation en locuteurs

Dans cette section, nous avons présenté les approches les plus utilisées pour la segmentation en locuteurs. Ces méthodes atteignent leurs limites lorsque l'on traite des contenus multimédias complexes (en particulier sur des contenus provenant de la télévision). La grande difficulté de la segmentation est qu'il devient difficile de mesurer une similarité sur des segments très courts ($< 2sc$). Les algorithmes de regroupement ne sont pas fiables lorsqu'il s'agit de regrouper les segments d'une personne à l'élocution spontanée et expressive (rire, colère, paroles superposés, etc). Des changements de tonalité dans la voix peuvent conduire à ce qu'un locuteur soit regroupé dans plusieurs *Clusters* différents. Par ailleurs, les séquences de double parole sont souvent regroupées dans un même *Cluster*.

1.3 Structuration basée sur l'information visuelle

L'objectif de la structuration de documents par personnes basée sur l'information visuelle est d'annoter automatiquement l'apparition de chaque personne dans la séquence vidéo et de regrouper toutes les apparitions d'une même personne. La figure 1.4 montre un exemple d'index visuel des personnes dans un extrait de l'émission *On n'a pas tout dit*.

FIGURE 1.4 – Exemple d'index visuel dans un extrait de l'émission "*On n'a pas tout dit*". L'objectif est de détecter automatiquement les personnes apparaissant dans un plan puis lier chaque personne détectée à ses apparitions dans les autres plans

Dans la littérature, il existe plusieurs méthodes de structuration de documents audio-visuels par personne en utilisant l'information visuelle. Les approches principales se décomposent généralement en 3 phases distinctes : la segmentation en plans, la détection de personne dans chaque plan et le regroupement des apparitions de chaque personne.

1.3.1 Segmentation en plans

La segmentation de séquences vidéo en plans consiste à déterminer automatiquement les transitions d'un plan à un autre dans une séquence vidéo. Un plan est une séquence d'enregistrement continue à partir d'une même caméra. Il existe deux types de transition : la transition brusque et progressive. Les transitions brusques correspondent aux points de montage (changement de caméra, de point de vue, etc). Plusieurs transitions progressives sont également utilisées dans les contenus audio-visuels telle que la transition "*en volet*" (glissement d'image laissant apparaître progressivement une image d'un autre plan) ou "*enchaînée*" (deux images superposées durant quelques trames où la luminosité de la première image diminue pendant que celle de la seconde augmente jusqu'à disparition de la première image). La figure 1.5 présente un exemple de transition brusque et de transition progressive de type "*enchaînée*" extraient de l'émission *On n'a pas tout dit*.

Le principe des approches développées pour la segmentation en plans est de considérer les images autour d'une transition comme des images ayant des signatures très différentes. Dans la littérature, les approches classiques sont divisées en trois

FIGURE 1.5 – Exemple de transitions enchaînée et de transition brusque dans l'émission "*On n'a pas tout dit*"

étapes [Boreczky and Rowe, 1996] : une représentation des l'images, un calcul de similarité entre deux images successives et une recherche de discontinuité (pics) dans les similarités calculées.

Représentation des images

Afin de mesurer la similarité entre deux images, il est nécessaire d'extraire la signature de chaque image. La signature d'une image est une représentation synthétique de l'information contenue dans celle-ci. Plusieurs représentations sont utilisées dans la littérature :

- *Au niveau des pixels :* la représentation la plus simple à utiliser pour représenter une image (bas niveau de représentation).
- *Histogrammes de niveaux de gris ou de couleurs :* la représentation la plus utilisée [Zhang et al., 1993, Cernekova et al., 2003] car elle est très facile à calculer et robuste aux mouvements de la caméra (translation, rotation, zoom, etc).
- *Vecteurs de mouvement :* le calcul du mouvement dans une séquence a pour but de détecter un mouvement causé par une transition. Dans [Shahraray, 1995], une corrélation spatiale est calculée entre deux vecteurs de mouvement calculés entre deux images consécutives afin de détecter une transition.

Malheureusement, certaines de ces représentations ne prennent pas en compte la distribution spatiale des caractéristiques extraites. Par exemple, si l'image est représentée par un histogramme de couleurs, la transition n'est pas détectée dans le cas où l'image du plan suivant a le même histogramme de couleurs mais celles-ci sont réparties différemment dans l'image. C'est pourquoi la plupart des algorithmes de segmentation en plans proposent d'extraire les caractéristiques d'une image par

bloc ou sur certaines régions d'intérêt dans l'image.

Calcul de similarités

Une fois que les caractéristiques de chaque image sont extraits, un calcul de similarité est effectué entre deux vecteurs de caractéristiques consécutifs. Plusieurs métriques sont utilisées pour mesurer la similarité (ou distance) : la moyenne de la différence entre pixels (*Mean Squared Distance - MSD*), la similarité *cosinus*, *Euclidien*, Test χ^2.

Recherche de discontinuités (Pics)

Après avoir calculé la similarité entre deux images consécutives, un algorithme de détection de changement doit décider si certaines discontinuités dans les valeurs de similarité sont des transitions significatives. La méthode la plus utilisée est de fixer un seuil sur la mesure de similarité au-dessous duquel on décide que les images sont trop différentes pour être des images consécutives d'un même plan [Cernekova et al., 2003]. La figure 1.6 présente les résultats du calcul de similarité (*1-Cosine*) entre deux images consécutives dans un extrait contenant 4 transitions brusques. Concernant les transition enchaînée, une analyse de la transition sur une fenêtre de temps permet de déterminer si l'image diffère significativement des précédentes [Boccignone et al., 2005].

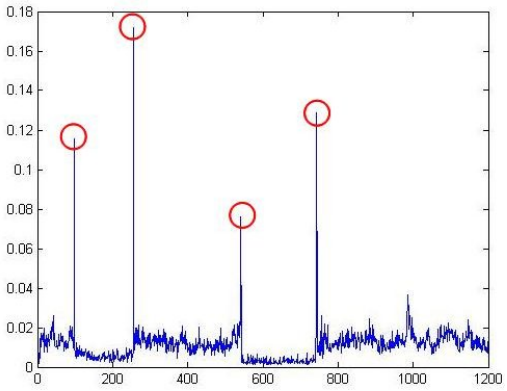


FIGURE 1.6 – Segmentation d'une séquence vidéo en plan. Exemple de détection de pics par calcul de similarité *1-Cosine* entre deux images consécutives

Le domaine de recherche en segmentation des contenus visuels en plans est relativement mûr. Pour cette raison, la tâche qui lui a été consacrée dans la campagne

d'évaluation *TRECVID* a été supprimée en 2008 [Smeaton et al., 2010].

1.3.2 Détection des personnes dans un plan

Après avoir segmenté la séquence vidéo en plans, la seconde étape consiste à détecter automatiquement les personnes dans chaque plan. Selon les applications, plusieurs informations visuelles peuvent être utilisées pour détecter les personnes dans une image (ou séquence d'images). La première information visuelle utilisée pour détecter une personne est le visage [Everingham et al., 2006]. Dans certaines applications spécifiques, d'autres informations visuelles peuvent être utilisées comme la silhouette ou une partie du corps (costumes, mains, bras, jambes, etc). Par exemple, en vidéo surveillance, il est souvent question de détecter des piétons dans un environnement extérieur en utilisant leurs silhouettes. Dans les émissions de plateaux, les informations dont nous disposons sont le visage et/ou une partie supérieure du corps. Dans cette partie, nous présentons les principales méthodes de détection des personnes dans les séquences vidéo basées sur le visage et le costume.

A- Détection du visage

Un travail de référence regroupant les méthodes existantes de détection du visage à partir d'une image est proposé dans Dans [Yang et al., 2002]. Ces méthodes sont regroupées en 3 grandes catégories :

Approches basées sur la connaissance humaine :

Ces méthodes se basent sur des règles simples déduites à partir de connaissances humaines sur les caractéristiques du visage et leurs relations géométriques. Dans [Chiang et al., 2003], un ensemble de règles à 3 niveaux sont proposées. Le premier est basé sur la détection de la peau du visage par l'analyse de la distribution des couleurs. Ensuite, les yeux et les lèvres sont localisés à partir de règles basées sur des observations de différence de couleurs avec la peau détectée. Puis des règles sur la disposition géométrique des yeux et lèvres détectés sont utilisées afin de supprimer des faux candidats. Les méthodes basées sur ces approches sont généralement faciles à mettre en œuvre, rapides dans le cas ou l'arrière plan est uniforme et robustes aux variations de pose. L'inconvénient est qu'il est difficile de déterminer des règles à partir de connaissances humaines. Dans le cas où les règles sont trop détaillées, on risque de ne pas détecter certains visages, alors qu'une description trop générale engendre beaucoup de faux candidats retenus.

Approches globales :

Dans la famille des méthodes globales, le visage est modélisé par ses propriétés globales. Le modèle de visage global est appris à partir d'une grande base de données. Puis, une méthode de classification est utilisée afin de détecter un visage dans une nouvelle image. Parmi ces méthodes de classification, on retrouve les réseaux de neurones [Féraud et al., 2001, Garcia and Delakis, 2002], *SVM* ou l'analyse par composantes principales [Turk and Pentland, 1991] (*Eigenfaces*). Dans toutes ces méthodes, différentes informations peuvent être utilisées pour représenter le visage : les intensités des pixels concaténées en un vecteur [Turk and Pentland, 1991], la couleur de la peau [McKenna et al., 1998], la texture [Dai and Nakano, 1996]. La méthode de détection du visage la plus utilisée pour sa robustesse et sa rapidité est l'algorithme *Adaboost* de *Viola&Jones* basé sur des descripteurs faibles de type *Haar* [Viola and Jones, 2001]. Le principe est de combiner itérativement plusieurs classificateurs faibles (simples fonctions à seuil calculées sur un seul descripteur), construits "*en cascade*" à différentes échelles.

Approches basées sur la mise en correspondance :

Dans ces approches, un modèle de visage est appris à partir d'exemple. Le visage est détecté par une mise en correspondance des formes (comparer chaque région candidate à modèle de visage). Deux techniques se distinguent : la technique "*locale*" qui consiste à modéliser chaque caractéristique du visage (yeux, nez, bouche, etc) de manière indépendante, et la technique "*globale*" qui consistent à modéliser le visage de manière globale.

Dans les méthodes "*locales*", chaque caractéristique du visage est associée à un modèle "prototype" (*Template*) [Luhong et al., 2000, Duffner and Garcia, 2005]. À partir d'une image candidate, pour chaque caractéristique du visage (le contour du visage, les yeux, le nez et la bouche), une corrélation avec son prototype est calculée. L'existence d'un visage est alors déterminée en combinant les valeurs des corrélations. L'inconvénient de ces méthodes est que les prototypes doivent être initialisés non loin du visage recherché car elles ne prennent pas en compte la géométrie des caractéristiques.

Dans les méthodes "*globales*", les approches les plus utilisées actuellement sont basées sur des modèles actifs de formes (*Active Shape Models*) [Cootes et al., 1995, Milborrow and Nicolls, 2008] qui construisent un modèle statistique géométrique du visage (modèle de forme) à partir des coordonnées des points caractéristiques (nez, bouche, les yeux ..). Le modèle de forme est appris à l'aide d'un ensemble de visages

annotés manuellement (données d'apprentissage de visages frontaux sans occultation sur lesquelles on a annoté manuellement les points caractéristiques) de la manière suivante : la première étape consiste à aligner les visages sur une référence arbitraire par une transformation géométrique (rotation, translation et mise à l'échelle). La deuxième étape consiste à calculer la "forme moyenne". Ces deux étapes sont répétées jusqu'à convergence en minimisant la distance euclidienne moyenne entre les points caractéristiques de forme. Ensuite, une analyse en composantes principales (ACP) est appliquée à la forme moyenne afin d'obtenir le modèle de visage. Pour extraire les caractéristiques d'un visage d'une nouvelle image, le modèle de visage appris est positionné sur le visage, puis itérativement déformé jusqu'à ce qu'il corresponde au visage de l'image. Une extension de la méthode ASM est proposée dans [Cootes et al., 2001, Matthews and Baker, 2004] appelée modèles actifs d'apparence (AAM) qui prend en compte l'information de texture en plus de la forme. Ces méthodes de détection de visage sont très efficaces et robustes aux changements de poses, de conditions d'illuminations et à certaines occultations. L'inconvénient est que pour détecter les caractéristiques du visage avec une grande précision, il est nécessaire de modélisation toutes les distorsions possibles du visage (ce qui demande une très grande base d'apprentissage avec des annotations manuelles très fines).

Suivi du visage dans une séquence d'images

L'intérêt principal du suivi du visage est d'exploiter la redondance de l'information apportée par la séquence d'images afin d'avoir une collection d'exemples de la personne dans un même plan. Cette collection est utilisée pour sélectionner le meilleur exemple [Bredin, 2007] ou pour en extraire une représentation moyenne [Everingham et al., 2006] de la personne.

Il existe deux approches principales pour le suivi du visage : le suivi du visage global et le suivi de points caractéristiques du visage. La première approche considère le visage comme un objet entier que l'on va détecter automatiquement dans une première image puis suivre dans les images suivantes. Souvent, l'algorithme utilisé est le *Mean SHIFT tracking* [Comaniciu et al., 2003] basé sur une maximisation de similarité entre deux régions (voir la figure 1.7). La seconde approche considère le visage comme un ensemble de points caractéristiques (yeux, nez, lèvres, sourcils, etc). Plusieurs méthodes de suivi de points sont proposées dans la littérature telles que l'algorithme KLT [Bourel et al., 2000] ou le recalage de points d'intérêts (*block matching*) [Spors and Rabenstein, 2001].

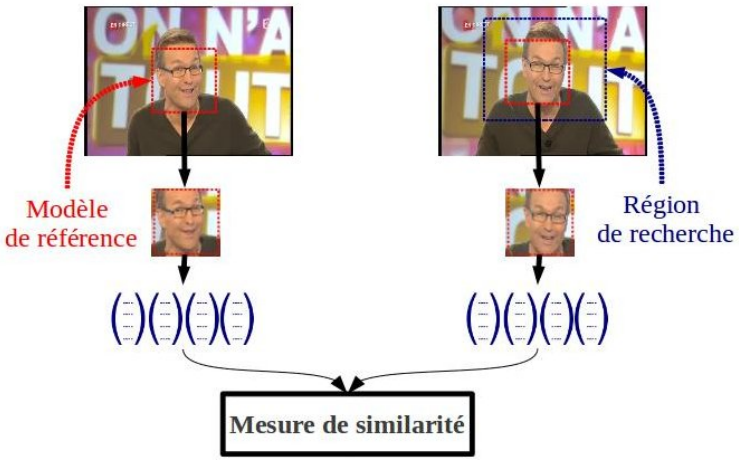


FIGURE 1.7 – Exemple de suivi du visage par l'algorithme *Mean Shift tracking*. La première étape consiste à choisir dans la première image le modèle de référence que l'on souhaite suivre (le visage détecté automatiquement). Ensuite, une recherche est effectuée dans l'image suivante sur une plus grande zone englobant le modèle de référence. L'objet est retrouvé en maximisant la similarité entre l'objet à suivre et toutes les régions candidates issues de la région de recherche

B- Détection du costume

Afin de détecter automatiquement les personnes dans une image, le costume est une information complémentaire ou alternative au visage. En effet, les méthodes basées sur la détection du visage sont très sensibles à la pose, aux conditions d'éclairage, expressions faciales, occultations (lunettes, coiffure, moustache, etc). La figure 1.8 présente l'apparence visuelle des personnes dans chaque plan extrait de l'émission "*On n'a pas tout dit*". Les costumes des personnes sont sujets à moins de variations que le visage.

Très peu de travaux existent pour la détection automatique de costume des personnes. La plupart des approches utilisent des méthodes de détection du visage ou de la silhouette, puis localisent le costume par déduction. Dans le but d'indexer des personnes dans des émissions de télévision, dans [Jaffre and Joly, 2004, Everingham et al., 2006], les histogrammes de couleurs des costumes portés par les personnes sont utilisés pour indexer les personnes. Le costume est localisé par un rectangle proportionnel au visage détecté automatiquement. L'inconvénient de cette méthode est qu'elle dépend du détecteur de visage. Afin de localiser directement

FIGURE 1.8 – Apparence visuelle des personnes dans l'émission "*On n'a pas tout dit*"

les costumes sans passer par un détecteur de visage ou de silhouette, des modèles de formes de costumes peuvent être appris selon le même principe que pour les détecteurs de visages [Jaffré, 2005]. Pour cela, il faut constituer une base de données contenant toutes les variabilités de forme et de texture des costumes, ce qui est très difficile à réaliser.

Il existe également des méthodes de détection des costumes basées sur la détection de la silhouette entière d'une personne, puis le costume est localisé par déduction spéciale. Par exemple, afin d'estimer la pose du corps des personnes dans des films, dans [Ferrari et al., 2008], les auteurs proposent de réduire la région de recherche en utilisant une méthode de détection de silhouette basée sur des histogrammes de gradient orienté (*HOG*) suivi d'une classification par *SVM* (humain ou pas) [Dalal and Triggs, 2004]. Malheureusement, les méthodes basées sur la détection de silhouette ne sont pas adaptées à notre étude, car nous traitons des *Talk Shows* où les personnes se présentent autour d'une table.

1.3.3 Regroupement

L'objectif est de construire un index d'apparence des personnes dans un document audio-visuel. Une fois que les personnes sont détectées dans tous les plans du document, il est nécessaire de regrouper les plans d'apparence de chaque personne. La plupart des méthodes considèrent le regroupement comme étant un problème de reconnaissance du visage. Ces méthodes nécessitent l'apprentissage de modèles super-

visés pour chaque personne. Il existe également des méthodes qui ne nécessitent pas une modélisation supervisée des personnes. Nous distinguons donc deux catégories de regroupement : regroupement supervisé et regroupement non supervisé.

Regroupement supervisé

Le regroupement supervisé considère que l'on possède un ensemble d'exemples pour chaque personne que l'on souhaite détecter dans le document audio-visuel. Le regroupement se réduit donc à un problème de vérification de l'identité des personnes où le but est de rechercher dans l'ensemble de toutes les personnes détectées automatiquement celles qui correspondent à la personne recherchée (dont on dispose d'exemples).

Dans [Arandjelovic and Zisserman, 2005], les auteurs proposent de détecter les séquences d'apparences de plusieurs acteurs dans les films "*Un jour sans fin*" et "*Pretty woman*" ainsi que la série "*Fawlty Towers*". La méthode proposée est basée sur une mesure de similarités entre un ensemble de visages *Query* et l'ensemble des visages détectés automatiquement durant toute la séquence vidéo. La mesure de similarité est une différence de pixels qui intègre une probabilité de présence d'occultation. Les évaluations montrent de très bon taux de précision et rappel. L'inconvénient de cette méthode est que la réponse est très dépendante de l'image *Query*.

Dans [Acosta et al., 2002], les auteurs proposent d'indexer des personnes dans des contenus de type *JTs*. Un modèle de visage basé sur la méthode des *Eigen Faces* [Turk and Pentland, 1991] est appris pour chaque personne à partir d'un ensemble d'images. Ensuite, chaque visage détecté automatiquement est projeté sur chaque *EigenFaces*, et le visage est identifié par le modèle qui minimise la perte d'information lors de la reconstruction du visage. Un très bon taux de bonne classification est obtenue dans les visages de *JTs*. L'inconvénient de cette méthode est qu'elle n'est pas très robuste à la pose du visage et aux variations d'expressions faciales.

Regroupement non supervisé

Le regroupement non supervisé considère que l'on ne possède pas de modèles des personnes à regrouper. La plupart des approches de regroupement de personne sont basées sur des algorithmes hiérarchiques ascendants où les groupes de personnes sont construits en regroupant itérativement les plus proches personnes déterminées par une mesure de similarité.

Dans [Eickeler et al., 2001], les auteurs proposent de regrouper les personnes par une méthode de *K-Moyenne*. Cette méthode prend l'hypothèse que le nombre de

personnes K à indexer dans le document est connu. Chaque classe est modélisée par un *HMM* à deux dimensions. D'abord, les visages détectés sont séparés arbitrairement en K groupes dans lesquels un *2D-HMM* est appris. Ensuite, les modèles appris sont utilisés pour prédire les identités des personnes détectées. Ces deux phases sont réitérées tant qu'il y a un changement dans les groupes. Cette méthode n'utilise pas de modèles appris à priori mais suppose que l'on connaît le nombre de personnes à détecter.

Dans [Everingham et al., 2006], une première indexation des apparences des personnes dans deux épisodes de la série "*Buffy contre les vampires*" est effectuée. L'index est obtenu en utilisant la combinaison des sous-titres, *scripts* et la détection d'activité labiale. Premièrement, les sous-titres sont alignés au script contenant le dialogue avec l'identité de chaque personne. L'hypothèse posée par les auteurs est que l'identité associée à chaque segment obtenu par alignement a une grande probabilité de contenir le visage associé à cette identité. Pour chaque segment aligné, les visages sont détectés et suivi puis un détecteur d'activité labiale permet d'associer le bon visage à l'identité. Cette méthode permet d'associer les séquences de visages à une identité avec une grande précision, car elle détecte les segments qui présentent très peu d'ambiguïtés. Malheureusement, cette méthode obtient un faible rappel causé en grande partie par les segments dans lesquels les personnes apparaissent, mais ne parlent pas. Pour retrouver ces segments, les auteurs proposent d'apprendre un modèle de visage et de costume pour chaque personne détectée automatiquement. Chaque personne détectée, et à laquelle aucune identité n'a été associée, est comparée à tous les modèles afin de déterminer l'identité la plus probable. Cette méthode obtient des résultats très prometteurs. L'inconvénient est qu'elle suppose que l'on possède un *script* du document audio-visuel ce qui n'est généralement pas le cas. Un autre inconvénient est qu'elle ne prend pas en compte les personnes qui ne parlent pas durant tout le document.

Dans [Cour et al., 2009], les auteurs proposent de regrouper l'apparence des personnes dans des épisodes de la série *Lost*. Un premier regroupement est effectué en utilisant un alignement entre le script contenant le dialogue avec les plans de la vidéo. À partir de cet alignement, chaque visage détecté est associé à un ou plusieurs labels (identité). Pour tous les visages ambiguës (associé à une liste de label), les auteurs proposent d'apprendre un *classifieur* qui incorpore des contraintes à partir des donnés annotés avec ambiguïtés. Les contraintes sont le degré d'ambiguïté, le mouvement des lèvres et le genre. Ensuite, le *classifieur* est utilisé pour sélectionner le label le plus probable.

Dans [Khoury et al., 2010], les auteurs proposent de regrouper les personnes dans une émission de $40mn$ de type *Talk show* en utilisant un algorithme hiérarchique as-

cendant. Premièrement, le visage est localisé par un détecteur *viola&Jones* et suivi sur tous les plans de l'émission. Sur chaque segment de visage suivi, 4 vecteurs de paramètres sont extraits (2 pour le visage et 2 pour le costume). Pour le visage, les paramètres utilisés sont le *SIFT (scale Invariant Feature Transform)* et l'histogramme de couleurs. Le costume est localisé par un rectangle sous le visage. Les paramètres extraits à partir du costume sont : la texture et l'histogramme de couleurs. Pour tous les segments de visages détectés, une matrice de similarité est obtenue pour chaque vecteur de paramètres. Un premier regroupement hiérarchique est effectué sur les segments où toutes les mesures de similarité s'accordent. Sur le reste des segments, un second regroupement est effectué en sélectionnant les vecteurs de paramètres dont la mesure de similarité s'accorde (un vecteur pour le costume et un vecteur pour le visage). Cette méthode est robuste lorsqu'un des vecteurs n'est pas fiable (exemple : pour le visage, les paramètres *SIFT* sont plus fiables que la couleur du visage lorsqu'il y a un changement d'éclairage). Enfin, un dernier regroupement est effectué sur les segments restants en utilisant uniquement les paramètres *SIFT* extraits sur le visage. Les résultats sur les $40mn$ de l'émissions montrent que fusionner les informations extraits du visage et le costume améliore le taux d'erreur de regroupement (environ 13% ce qui est considéré relativement acceptable contenu de la complexité du contenu).

1.4 Fusion

Afin de mieux comprendre notre environnement, nous utilisons souvent la fusion d'informations complémentaires. Par exemple : la synchronisation entre la parole et les lèvres afin de mieux comprendre un locuteur dans le cas d'un environnement bruyant [Silsbee and Bovik, 1996]. La fusion a pour objectif d'intégrer des sources d'informations complémentaires afin d'améliorer les résultats obtenus dans chaque modalité. Dans le cadre de notre étude, nous nous intéressons à l'indexation des personnes dans des contenus audiovisuels. Dans cette section, nous nous consacrons à la présentation de l'état de l'art de la fusion des modalités audio et visuelle.

1.4.1 Catégories de fusion

Dans un système multimodal, la fusion d'informations est souvent divisée en trois catégories : la fusion *précoce*, *intermédiaire* et *tardive* [Sanderson and Paliwal, 2004]. Dans les méthodes de fusion précoce, les informations sont combinées avant toute modélisation et classification. Dans la fusion intermédiaire, les informations sont combinées durant le processus de classification. Et enfin, dans la fusion tardive, les

informations sont combinées après le processus de classification de chaque modalité. La figure 1.9 présente les catégories de fusion dans un système multimodal.

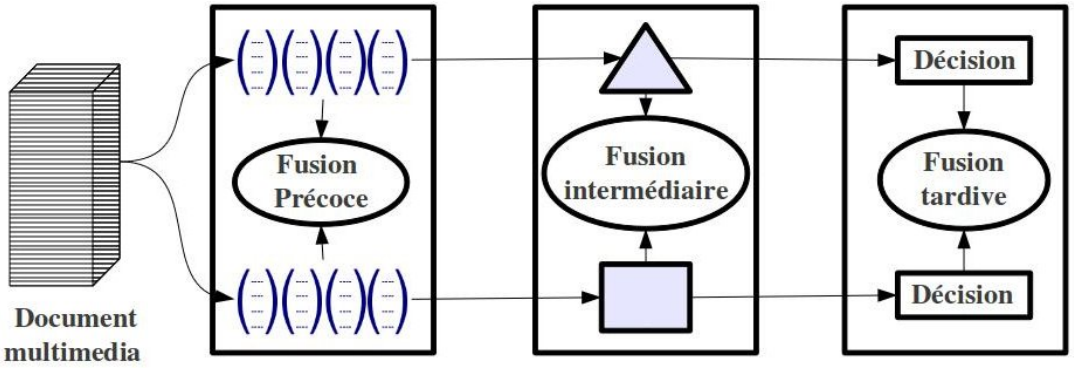


FIGURE 1.9 – Catégories de fusion dans un système multimodal

Fusion précoce

Dans la fusion précoce, il y a deux sous-catégories principales : fusion au niveau des capteurs et fusion au niveau des paramètres.

La fusion au niveau des capteurs : La fusion au niveau des capteurs suppose que l'on possède plusieurs capteurs d'une même modalité. Dans ce cas, la fusion consiste à associer des données brutes provenant de capteurs différents afin d'en extraire une information plus complète [Hall and Llinas, 1997]. Les méthodes de fusion utilisées diffèrent selon les applications. Par exemple, une reconstruction en *mosaïque* est utilisée pour créer une scène à partir des images provenant de plusieurs caméras fournissant chacune une représentation partielle de la scène [Zhu, 2005]. Dans certaines applications, la somme pondérée de capteurs ou une concaténation des caractéristiques peuvent être utilisés. Exemple : utilisation d'une somme pondérée pour combiner des données provenant de deux micros (pour réduire les effets du bruit) [Aarabi, 2003].

La fusion au niveau des paramètres : La fusion de paramètres consiste à combiner des vecteurs de caractéristiques extraits de différentes modalités afin de former un vecteur unique fourni entrée dans un système de classification (vérification d'identité, indexation, reconnaissance de la parole). Dans ces approches, les caractéristiques extraites dans chaque modalité sont souvent de nature différente. Dans ce cas, les

méthodes de classification possibles sont celles qui ne nécessite pas une normalisation des paramètres pour la fusion (exemple : réseaux de neurones, régression logistique). Dans [Chibelushi et al., 1997], afin de vérifier l'identité des personnes dans des séquences vidéos, les vecteurs de paramètres extraits de l'audio et du visuel sont concaténés pour en faire un vecteur unique à fournir comme entrée à un réseau de neurones.

Fusion intermédiaire

La fusion intermédiaire est utilisée généralement dans le but d'exploiter les corrélations des différentes modalités. Les modèles les plus répandus sont les Modèles de Markov cachés *HMM* qui sont adaptés pour la gestion de plusieurs flux de données. Par exemple, dans [Dupont and Luettin, 2000], des modèles *multi-stream HMM* sont utilisés pour la reconnaissance de la parole audiovisuelle. Des *HMM couplé* sont proposés dans [Nefian et al., 2003] pour la vérification de l'identité audio-visuelle.

Fusion tardive

Fusion de décisions : Pour chaque modalité, un processus de classification fournit une décision. Les décisions peuvent être combinées par un vote à la majorité [Genoud et al., 1996], une procédure de classement ([Radova and Psutka, 1997] *Condorcet*, *Borda*, etc), ou en utilisant les opérateurs *ET/OU*. L'inconvénient de ce type de fusion est qu'elle n'est robuste qu'à partir du moment où l'on possède beaucoup de décisions à fusionner.

La fusion de scores : Dans la fusion des scores, chaque modalité procure un score obtenu par un processus de classification. Ces scores sont souvent de nature différente. Dans ce cas, il est nécessaire de les normaliser afin de les représenter dans le même espace. Plusieurs fonctions de normalisation sont utilisées comme la *Z-Norm* ou la *Tanh Norm*. Une fois normalisés, les scores peuvent être combinés par une somme pondérée ou un processus *post-classificateur*. L'avantage de l'approche somme pondérée est que les poids peuvent être sélectionnés de manière à introduire des connaissances *a priori*, par exemple tenir compte de la fiabilité et la capacité de discrimination de chaque modalité. L'approche *post-classifieur* consiste à utiliser les scores de chaque modalité comme paramètres d'entrée pour apprendre à *post-classifieur* (réseau de neurones, *SVM*, régression logistique [Verlinde et al., 2000]). L'avantage est que selon certains *post-classifieurs*, il n'est pas nécessaire de normaliser les scores.

1.4.2 Fusion pour l'identification des personnes

Dans [Nefian et al., 2003], les auteurs proposent une méthode d'identification des personnes dépendant du texte dans des séquences vidéo en utilisant le visage, la voix et la synchronie entre la voix et les lèvres. Le système de vérification audiovisuel est appris de la manière suivante : pour chaque personne, la combinaison audio et mouvement des lèvres est modélisée par un *HMM couplé* pour chaque phonème. Pour chaque séquence de test, la vraisemblance de la séquence observée à être générée par le modèle de la personne k est obtenue par combinaison des vraisemblances du système de reconnaissance du visage ($L(O_f|k)$) et système de vérification audiovisuel ($L(O_a, O_v|k)$). La fusion est obtenue de la manière suivante :

$$L(O_f, O_a, O_v|k) = \lambda L(O_f|k) + (1 - \lambda)L(O_a, O_v|k) \tag{1.4}$$

Où λ est un poids de fusion fixé à priori. Les expériences sont effectuées sur des séquences issues de la base de donnée *XM2VTS* [Messer et al., 1999] sur lesquelles un bruit gaussien a été ajouté. Les résultats obtenus démontre une robustesse de la méthode à la qualité de l'audio.

Pour l'identification audio-visuelle des personnes, une autre approche est proposée dans [Li et al., 2005] basée sur les réseaux bayésiens dynamiques (*Dynamic Bayesian Network* ou *DBN*). À l'instant t, les observations sont les deux vecteurs de paramètres audio (*MFCC*) et visuels (*ACP* du visage). La structure du réseau bayésien est présentée de manière à ce que les observations audio et visuel soient conditionnellement dépendantes de l'état qui les génère (et donc conditionnellement indépendants). La probabilité jointe est donc définie comme produit des deux distributions de probabilités. Les résultats montrent que la fusion basée sur un modèle *DBN* améliore les performances du système de vérification de l'identité comparé aux systèmes mono-modaux.

Dans [Vallet et al., 2010], les auteurs utilisent des descripteurs visuels en plus des descripteurs acoustiques afin d'identifier les personnes dans des émissions de type *Talk shows*. Un *classifieur SVM* est appris pour chaque locuteur à partir de données annotées manuellement. Le *classifieur* prend en entrée des descripteurs audio (*MFCC*s) et visuels (signature de couleur du costume et du mouvement du locuteur). Cette méthode d'identification n'est pas très robuste aux tentatives d'impostures. Par contre, pour une application d'indexation de personnes, la méthode offre une bonne discrimination entre les locuteurs de la même émission.

1.4.3 Fusion pour la reconnaissance audiovisuelle de la parole

La reconnaissance de la parole audiovisuelle vise à utiliser la corrélation entre l'information audio (parole) et l'activité visuelle de la parole (exemple : mouvement des lèvres) afin d'améliorer les résultats des systèmes basés uniquement sur la modalité audio non robustes à la parole bruitée.

Dans [Dupont and Luettin, 2000], une méthode de combinaison des modèles acoustiques (*Perceptual linear predictive* ou *PLP*) et visuels pour améliorer la reconnaissance de la parole est proposée. Les paramètres visuels sont : le contour des lèvres et l'histogramme de niveaux de gris de la région de la bouche. La fusion est modélisée par un *multistream HMMs* qui modélise la synchronie entre les deux modalités. L'avantage de cette méthode est que la modalité visuelle apporte une information précieuse pour la reconnaissance de la parole dans un environnement sonore bruité.

Dans [Heracleous et al., 2010], les auteurs proposent de reconnaître la parole à partir de la fusion entre des paramètres acoustiques et de l'articulation. Les mouvements de la langue, des lèvres, de la mâchoire sont suivis par un dispositif *Articulographie Electro-Magnétique (EMA)*. La fusion est obtenue par une modélisation *HMM Multistream* qui combine les vraisemblances obtenues par des modèles *HMMs* sur chaque modalité (audio et visuelle). Dans le même article, une autre méthode de fusion tardive est proposée basée sur la classification des résultats obtenus par les modèles *HMMs* de chaque modalité. Pour chaque phonème, chaque modalité obtient une liste de vraisemblance. Soient $P(O_a, h)$ et $P(O_v, h)$ les score de vraisemblance de la séquence O au phonème h obtenues respectivement par les modalités audio et visuelle. La vraisemblance audio-visuelle est calculée de la manière suivante :

$$log(P(O_{av}, s)) = \lambda_a log(P(O_a, h)) + \lambda_e log(P(O_v, h)) \tag{1.5}$$

L'avantage de cette méthode de fusion tardive est qu'elle est robuste à l'asynchronie entre le mouvement visuel et la parole. Les résultats obtenus montrent une nette amélioration de la précision de reconnaissance des phonèmes dans le cas d'environnements bruités comparé au système utilisant uniquement l'information audio.

1.4.4 Fusion pour la détection de visages parlants

Dans la littérature, la plupart des méthodes de détection de visages parlants ont pour objectif d'améliorer les résultats de l'indexation en locuteurs.

Dans [Monaci et al., 2006], les auteurs proposent une méthode de détection des visages parlants basée sur la modélisation de la synchronie entre les caractéristiques audio et visuelles. La méthode est basée sur l'apprentissage d'un dictionnaire audio-visuel de la parole prononcée (une suite de chiffres). Les caractéristique audio et visuelles sont extraites toutes les 23 trames ($fps = 29.97$ trames/seconds). Les caractéristiques visuelles décrivent les mouvements typiques des différentes parties de la bouche pendant l'énoncé. Des fonctions traduisant la relation entre une prononciation et les lèvres sont apprises sur des séquences d'apprentissage. Ces fonctions sont apprises suivant 4 phases répétées jusqu'à convergence :

1. Localiser : pour chaque énoncé, trouver la position temporelle t_i qui maximise la corrélation entre ce mot et l'ensemble des caractéristiques audio extraites dans la base.
2. Apprendre : à l'instant t_i trouvé dans l'étape 1, trouver la structure visuelle qui représente le mieux la moyenne les caractéristiques visuels extraits.
3. Localiser : trouver la position temporelle t_j qui maximise la corrélation entre les caractéristiques visuelles trouvée à l'étape 2 les l'ensemble des caractéristiques visuels extraits.
4. Apprendre : à l'instant t_j trouvé à l'étape 3, trouver le mot audio qui représente le mieux, en moyenne, l'ensemble des caractéristiques audio extraites dans la base.

Afin de tester la capacité de cette méthode à modéliser la synchronie audio-visuelle et à déterminer le visage parlant, les tests sont effectués sur des séquences vidéo contenant deux personnes : un visage parlant (source de parole) et un visage silencieux imitant les mouvements des lèvres de la première personne. Les fonctions du dictionnaire obtiennent des résultats encourageants pour la détection de la synchronicité et la localisation du visage parlant.

Dans [Vajaria et al., 2008], les auteurs proposent d'utiliser le mouvement du corps pour détecter le visage parlant dans des réunions filmées par 4 caméras (une dans chaque coin de la pièce). D'abord, une méthode d'indexation en locuteurs basée sur la segmentation et le regroupement hiérarchique est utilisée afin d'effectuer un regroupement sommaire des locuteurs (les locuteurs sont regroupés dans plusieurs groupes). Ensuite, chaque groupe intermédiaire est associé à une région visuelle représentant la région du mouvement dominant. Pour chaque groupe intermédiaire, la signature audio est modélisés par des *GMMs* et sa région visuelle par un modèle basé sur l'analyse par composante principale (ACP). Les groupes intermédiaires sont progressivement affinés par la combinaison des plus proches paires

de segments audio-visuels. La mesure de distance entre deux segments est obtenue par une somme pondérée de la distance des segments dans chaque modalité (*KL* pour chaque modalité). Après chaque regroupement de deux segments, les modèles audio et visuels sont mis à jour. Le regroupement est réitéré jusqu'à satisfaction d'un critère d'arrêt. Les expériences effectuées sur le *NIST pilot meeting room corpus* [Garofolo et al., 2004] montrent une nette amélioration de l'indexation en locuteur. L'avantage de cette méthode est qu'elle est robuste aux variations de l'apparence du visage car elle ne nécessite pas une détection automatique des personnes.

Dans [Englebienne et al., 2009], les auteurs proposent une méthode de détection de visages parlants dans des vidéos de réunions filmes par plusieurs caméras. Cette méthode est basée sur l'information mutuelle. D'abord, le visage est détecté dans chaque trame. Ensuite, pour chaque visage détecté, les paramètres *SIFT* (*scale-invariant feature transform*) sont extraits. Le visage parlant est déterminé par la mesure de l'information mutuelle apportée par les paramètres acoustiques (énergie) et le vecteur de paramètres *SIFT*. Les améliorations apportée par cette méthode s'inscrivent parmi les meilleurs dans l'état de l'art de l'indexation en locuteurs basée uniquement sur l'audio dans ce contexte.

Dans [Knox and Friedland, 2010], les auteurs proposent d'utiliser le flux optique afin de déterminer le visage parlant dans une réunion contenant quatre participants filmés chacun par une caméra (*AMI corpus* [Carletta et al., 2005]). Les auteurs prennent l'hypothèse que le locuteur est celui qui effectue le plus de mouvement. D'abord, un index de locuteur est construit suivant la méthode *ICSI diarization system* [Wooters and Huijbregts, 2008]. Ensuite, le visage qui présente le plus de mouvement est associé au locuteur détecté. Les améliorations apportées par cette méthode s'inscrivent parmi les meilleures dans l'état de l'art de l'indexation en locuteurs basée uniquement sur l'audio dans ce contexte. L'inconvénient de cette méthode est qu'elle dépend du contexte d'étude. En effet, l'hypothèse que le locuteur est celui qui bouge le plus n'est pas souvent vérifiée dans d'autres types de contenus.

1.4.5 Fusion pour la structuration de documents audio-visuels

La fusion d'index de personnes obtenus de manière indépendante par l'audio et le visuel est un domaine nouveau. Dans [Jaffré et al., 2007], deux index de personnes sont construits de manière indépendante en utilisant les informations audio et visuelles. La recherche d'association suppose que plusieurs étiquettes audio peuvent être associées à plusieurs étiquettes visuelles. D'abord, une co-occurrence temporelle entre les étiquettes audio et visuelle est calculée. Puis de manière indépendante, chaque étiquette audio est associée à une étiquette visuelle et chaque étiquette vi-

suelle est associée à une étiquette audio par maximisation de la co-occurrence. La figure 1.10 présente un exemple d'index audio et visuel et les associations déterminées par cette méthode. Toutes les étiquettes qui sont associées de manière directe ou indirecte sont regroupées et considérées comme étant une seule personne. Dans ce cas, une étiquette audio peut être associée à une étiquette visuelle qui est associée à son tour à une autre étiquette audio différente de la première. Cette méthode d'association associe à chaque étiquette audio au moins une étiquette visuelle (et vice versa). L'avantage de cette méthode de fusion est qu'elle peut associer les personnes regroupées dans plusieurs *Clusters* par l'une des deux modalités ou les deux (cas de sur-segmentation). Mais, elle ne prend donc pas en compte les cas de *"voix off"* (personne qui parle mais qui n'apparaît jamais) ou les cas de personnes qui apparaissent mais ne parlent jamais.

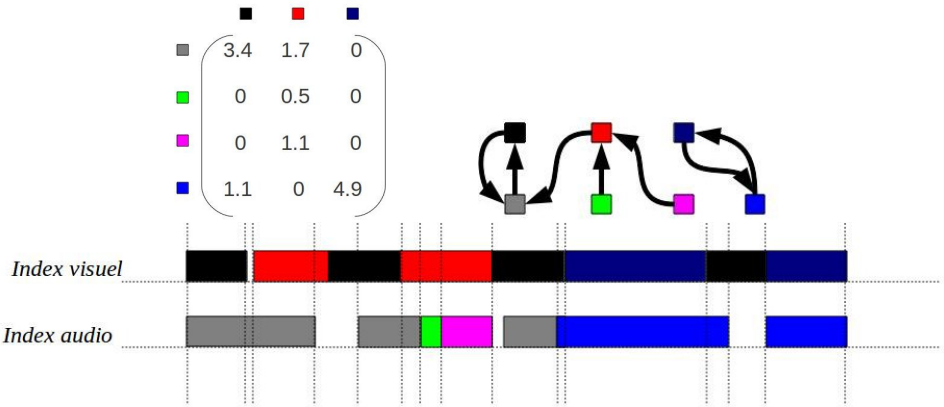


FIGURE 1.10 – Exemple de matrice de co-occurrence et des associations déterminées par la méthode décrite dans [Jaffré et al., 2007]. Chaque étiquette audio est associée à l'étiquette visuelle qui maximise sa co-occurrence (maximisation en ligne dans la matrice). De même, chaque étiquette visuelle est associée à l'étiquette audio qui maximise sa co-occurrence (maximisation en colonne dans la matrice). Toutes les étiquettes associées sont regroupées et considérée comme étant une seule personne. Les associations déterminent deux personnes : (gris + vert + magenta avec rouge + noir) et (bleue et bleue marine)

1.5 Conclusion

Dans cette section, nous avons présenté un tour d'horizon des méthodes utilisées pour l'indexation des personnes dans des contenus audio-visuels. Ce sujet est à l'intersection de plusieurs domaines de recherche et nécessite la combinaison de plusieurs technologies. Dans notre contexte d'étude, nous souhaitons indexer des personnes dans des émissions de type *Talk shows*. Dans ce contexte, il est très difficile de déterminer qui parle à quel moment en raison de la grande interactivité des dialogues caractéristique des *Talk shows*. La recherche dans ce contexte étant très récent, beaucoup d'efforts restent à faire en indexation de contenus télévisés, intrinsèquement multimédias. Les protocoles d'évaluation et bases de données restent encore aujourd'hui à définir. Un projet intitulé *REConnaissance de PERsonnes dans des Emissions audiovisuelles - REPERE* a été lancé en 2010 afin d'encourager la recherche dans ce domaine. Pour ces raisons, il est très difficile d'inscrire notre travail dans l'état de l'art. Nous avons tenté de comparer chaque technologie utilisée à son état de l'art quand cela nous a été possible de le faire.

Chapitre 2

Contexte d'étude et corpus

Nous nous intéressons à l'indexation des personnes dans des contenus audio-visuels de type émissions de plateaux (*Talk Shows*). Ce contexte présente de nombreuses difficultés en raison de la grande interactivité des dialogues. L'objet de ce chapitre est de présenter le contexte d'étude avec ses avantages et inconvénients. Nous présentons une analyse détaillée de la base de données, des annotations effectuées et des particularités de ces contenus audio-visuels.

2.1 Inventaire des corpus audio-visuels

Dans cette section, nous faisons un tour d'horizon des bases de données publiques et privées utilisées pour l'indexation audio, visuelle et audio-visuelle des personnes. La figure 2.1 présente des exemples de corpus audio, visuels et audio-visuels annotés par personne.

2.1.1 Corpus audio

Dans la communauté parole, plusieurs corpus publics annotés par personnes existent. En français, grâce aux campagnes d'évaluations *ESTER* [Geoffrois et al., 2006], plusieurs heures de parole annotées sont accessibles. Cette campagne a pour objectif de comparer les performances des systèmes d'analyse et d'indexation de documents audio contenant de la parole. Les évaluations s'organisent autour de trois grandes tâches : la segmentation (S), la transcription orthographique (T) l'extraction d'information (E). La tâche de segmentation comprend le suivi d'événements sonores (SES), segmentation en locuteurs (SRL) et suivi de locuteurs (SVL). Le corpus est constitué d'une centaine d'heures d'enregistrement d'émissions d'actualités radiophoniques françaises (*France Info, France Inter, RFI, France Culture et Radio Classique*) et francophones (Radio Télévision Marocaine *RTM*). Plusieurs annotations manuelles sont effectuées : transcription de la parole, entité nommés (pays, personnes, temps...). Dans le corpus, au total, 2172 locuteurs sont annotés (744 femmes, 1398 hommes et 20 enfants).

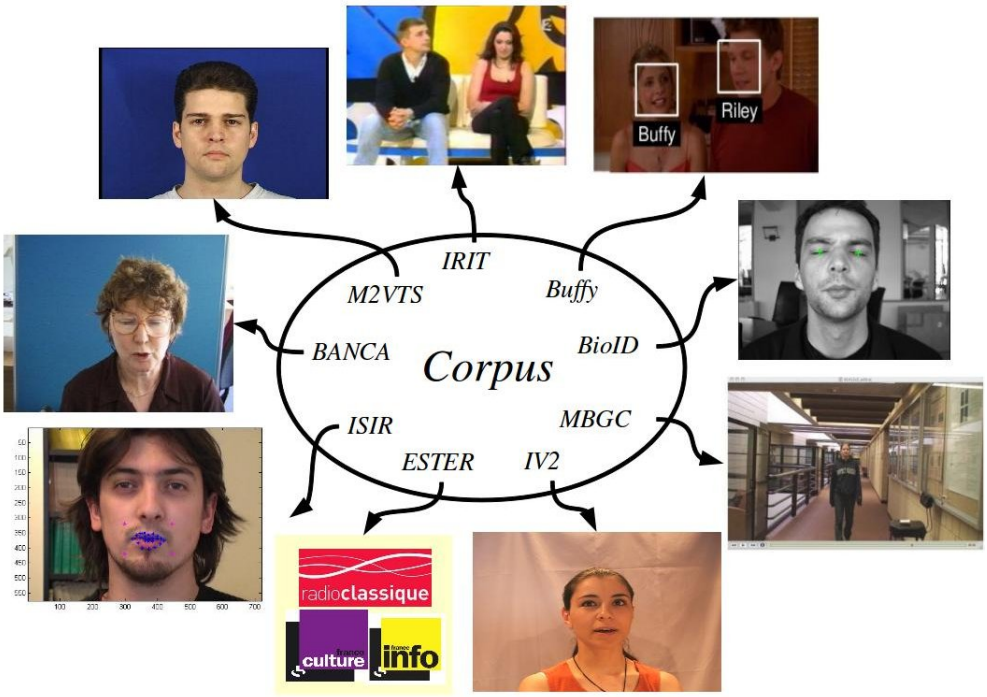


FIGURE 2.1 – Exemples de corpus audio-visuels de personnes

En 2008, la campagne d'évaluation *Ester2* [Galliano et al., 2009] enrichi le corpus de départ de 124h d'émissions d'information, des dossiers liés à l'actualité du moment et des émissions plus conversationnelles.

2.1.2 Corpus visuels

Il existe plusieurs corpus publics annotés par l'apparence des personnes dans des séquences vidéo. Parmi ces corpus, on trouve :

– Corpus *MBCG* : ce corpus est proposé par *NIST* dans le cadre de la campagne *Multiple Biometric Grand Challenge (MBGC)*. Cette campagne a pour objectif d'évaluer des systèmes de vérification biométrique en se basant sur les deux modalités *visage* et *iris* [Phillips et al., 2009]. Un des objectifs est la reconnaissance faciale des personnes qui marchent dans un couloir ou dans la rue. Bien que les conditions d'enregistrement représentent un challenge pour l'identification des personnes, chaque séquence contient un seul plan d'une seule personne

s'avançant dans un couloir.
- Corpus *ISIR* : le laboratoire *ISIR* (*Institut des Systèmes Intelligents et de Robotique*) a annoté des séquences de visages parlants de 6 individus apparaissant chacun durant une dizaine de secondes. Pour chaque personne, les coordonnées du visage et de 22 points sur le contour des lèvres sont annotées. Les conditions d'enregistrement dans ce corpus sont simples : une seule personne apparait face caméra.
- Corpus *Buffy* : Afin d'évaluer les méthodes d'indexation visuelles des personnes dans un contexte de télévision, le laboratoire *Visual Geometry Group* de l'université d'*Oxford* a annoté deux épisodes de la série américaine « *buffy contre les vampires* ». Ces annotations sont effectuées à titre privé, à partir des *DVDs* disponibles dans le commerce [Everingham et al., 2006].

2.1.3 Corpus audio-visuel

Il existe plusieurs corpus publiques pour la vérification de l'identité audio-visuelle des personnes. Parmi les plus connues : *Banca* [Bailly-Bailliére et al., 2003], *M2VTS* [Messer et al., 1999] et *IV2* (Identification par l'Iris et le Visage via la Vidéo). Ces bases de données se composent de séquences vidéos de personnes visibles, lisant un texte précis. Dans *Banca*, 3 scénarios d'enregistrement sont utilisés : *Controlled* (conditions contrôlées), *Degraded* (enregistrements dans un bureau avec un *webcam*) et Adverse (enregistrements dans un environnement sonore bruité). Dans le corpus *IV2*, les séquences sont enregistrées dans un studio disposant d'une caméra classique et d'un scanner *3D*. Malheureusement, tous ces corpus ne sont pas adaptés à notre contexte d'étude car, même si ils contiennent des séquences de différents visages parlants, le scénario d'enregistrement est très simple : un seul plan par séquence dans lequel apparaît un seul visage de face.

Très peu de travaux sont effectués sur des données de type émissions de plateaux. Afin d'évaluer les méthodes d'indexation audio-visuelles des personnes dans ce contexte, il est nécessaire d'avoir une base de données adaptée.
- Corpus *IRIT* : l'Institut de Recherche en Informatique de Toulouse [1] a collecté et annoté, à titre privé, deux émissions de plateaux : « *Pyramide*" »et « *Les amours* » [Jaffre and Joly, 2004]. En 2010, ces données ont été enrichies par 4 autres émissions de plateaux [Khoury, 2010].
- Projet *Quaero* [2] : un projet d'annotation des personnes dans 59 *JTs* de *France*

1. www.irit.fr
2. Quaero est un programme fédérateur de recherche et d'innovation industrielle sur les tech-

2 est en cours d'élaboration. Dans chaque trame, la région du visage, les coordonnées des yeux et de la bouche sont annotés.
– Corpus *Grand Échiquier* : Dans le cadre du projet *Infom@gic*, l'*INA* a mis à disposition de ces partenaire plusieurs épisodes de la célèbre émission *Le Grand Échiquier*. Certaines annotations sont disponibles comme les locuteurs, les événements sonores, etc.

Le projet *REPERE* lancé par l'*ANR* en 2010 a pour premier objectif la constitution d'une grande base de données de *JTs* et *Talk shows* annotées finement.

2.2 Présentation de la base de données *TSDB*

Afin de mesurer les performances d'un système d'indexation des personnes dans un contexte de télévision il est nécessaire d'avoir une base de données de programmes de télévision annotée par personne (JTs, *Talk shows*, reportages). Malheureusement, à notre connaissance, il n'existe pas de données *publiques* annotées avec les deux modalités : voix et apparence du visage. Il a été nécessaire de collecter et d'annoter une base de données de type émission de plateaux. Dans cette section, nous présentons la base de données *TSDB* (*Talk Show DataBase*).

2.2.1 Description du corpus

Cinq émissions du programme « *On n'a pas tout dit* »diffusé par la chaîne de télévision française *France 2* sont annotées. Ce programme, présentée par *Laurent Ruquier*, était diffusé du lundi au vendredi à partir de 19*h* en 2008. Le présentateur est entouré de ses chroniqueurs et invités placés autour d'une grande table ronde. Tous les jours *Laurent Ruquier* et ses chroniqueurs réagissaient à l'actualité du jour, phénomènes de société, événements culturels, etc. Un public, disposé autour des invités, réagit avec des applaudissements. La figure 2.2 montre des exemples de vues dans le *Talk show*.

Le choix de la base de données a été effectué de manière à anticiper de futures applications. En effet, dans ces contenus, le présentateur et tous ses chroniqueurs reviennent souvent dans plusieurs émissions ce qui permet d'avoir des données variées

nologies d'analyse automatique, de classification et d'utilisation de documents multimédias et multilingues www.quaero.org

pour chaque personne. Ces données peuvent être utilisées pour modéliser les personnes afin de les identifier dans d'autres épisodes.

FIGURE 2.2 – Exemples de plans collectés dans le corpus *TSDB*

Chaque épisode de l'émission est structuré en plusieurs parties correspondant à différents sujets de conversation. Ces parties sont séparées par des *jingles* sonores. Au début de l'émission, le présentateur commence par présenter les invités, les chroniqueurs, puis lance les sujets d'actualités à débattre. Durant une émission, les vues plateaux sont coupées de reportages extérieurs, clips de musique et génériques. Cinq épisodes (*S1, S2, S3, S4, S5*) correspondant à des diffusions du mois d'avril 2008 sont annotés (7, 8, 11, 14 et 15 Avril 2008). Dans ce corpus, les personnes apparaissent dans 4 types de plans :

- *Plan mono-visage* : vue concentré sur un seul visage en premier plan.
- *Plan multi-visages* : vue sur plusieurs visages en premier plan.
- *Plan général* : vue sur le public, les coupures de reportages et les vues générales autour de la table.
- *Plan de montage* : vue sur deux visages obtenue par montage. Apparaît souvent lorsqu'il y a un dialogue entre deux personnes éloignées.

2.2.2 Annotations

Afin d'évaluer nos méthodes d'indexation de personnes, nous avons annoté le corpus *TSDB* par personne de manière audio et visuelle. Les annotations sont effectuées

de la manière suivante :

Annotations audio

FIGURE 2.3 – Exemples d'annotations audio des personnes dans le corpus *TSDB* sur *transrciber*

Les séquences d'interventions sonores du présentateur, des chroniqueurs et des invités sont annotées par l'outil *transcriber*[3]. Pour chaque tour de parole, les informations annotées sont : début et fin de chaque séquence, l'identité du locuteur ainsi que le texte prononcé. D'autres informations sonores sont également annotées : applaudissements, musiques et séquences d'intervention de plusieurs personnes en même temps accompagnées de la transcription du texte dans le cas où celui-ci est compréhensible. La figure 2.3 montre un exemple du fichier *XML* de sortie avec les informations annotées.

3. http ://trans.sourceforge.net

Annotations visuelles

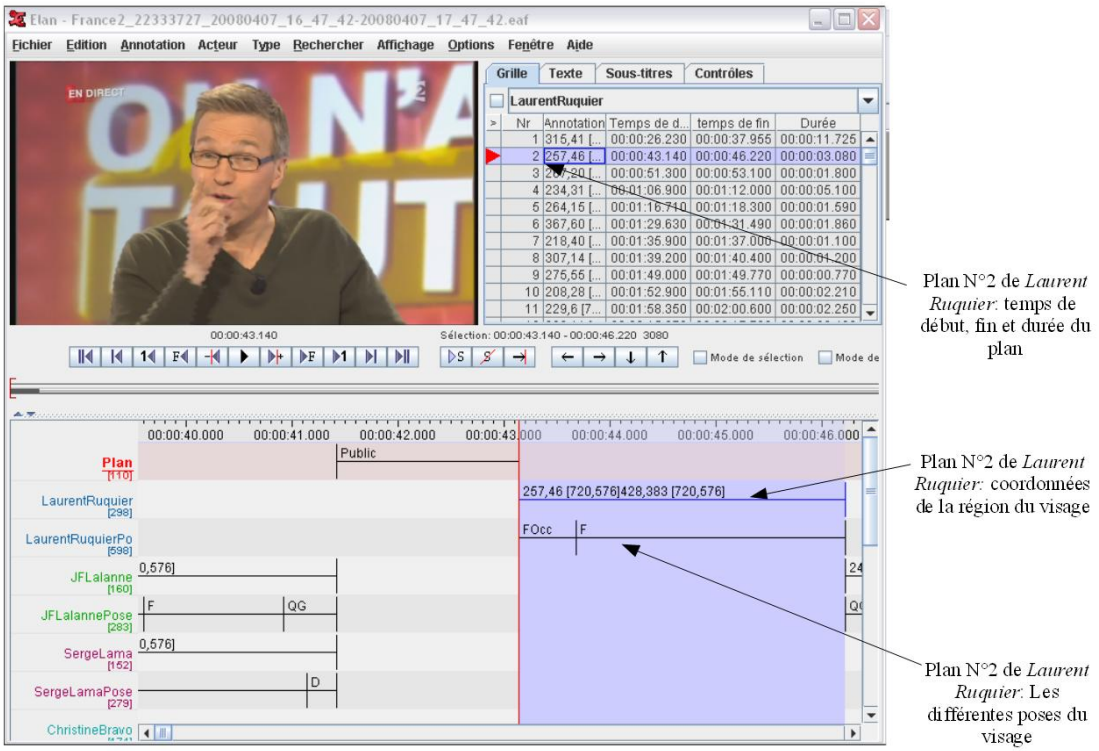


FIGURE 2.4 – Exemples d'annotations visuelles des personnes dans *TSDB* dans *Elan*

L'outil utilisé pour l'annotation visuelle des personnes est *Elan*[4]. Tous les plans d'apparence du présentateur, des chroniqueurs et invités sont annotés. La figure 2.4 montre un exemple de plans annotés et affichés par l'outil *Elan*. Pour chaque personne présente dans un plan, les informations annotées sont :

- L'identité de la personne.
- Les coordonnées de la région du visage dans le plan.
- La pose du visage : à droite (D), à gauche(G), frontal (F), quart droit (QD) ou gauche (QG), haut (H), bas (B)).
- La présence d'une occultation du visage (Occ).

4. http ://www.lat-mpi.eu/tools/elan/download

2.3 Analyse du corpus *TSDB*

Dans les émissions « *On n'a pas tout dit* », un épisode dure généralement 50*mn* durant lequel chaque personne intervient à des moments différents. Dans ce contexte, les interventions sont courtes, les dialogues sont interactifs et plusieurs personnes parlent et/ou apparaissent dans une même intervention. Le tableau 2.1 présente la structure générale de la base de données *TSDB* par émission. Dans chaque épisode, entre 7 et 9 personnes sont annotées (présentateur, chroniqueurs et invités). Dans ce tableau, la durée totale des personnes annotées de manière visuelle est supérieure à la durée totale de l'épisode. Cela s'explique par le fait que les plans multi-visages sont comptés autant de fois que le nombre de personnes qui y figurent.

Épisodes	*S1*	*S2*	*S3*	*S4*	*S5*
Nombre de personne	8	9	7	7	7
Durée totale de parole (s)	2347	2568	2014	2288	2844
Durée totale d'apparence (s)	3548	3720	3668	3049	3404
Durée totale des visages parlants (s)	1409	1505	981	1456	718

TABLE 2.1 – Statistiques générales des 5 épisodes du corpus *TSDB* (temps en secondes)

Dans ces épisodes, les personnes parlent et apparaissent de manière homogène à l'exception du présentateur qui intervient plus souvent. Le tableau 2.2 présente la répartition de la durée de la parole et de l'apparition de chaque personne dans l'épisode *S1*.

- En audio, durant un épisode, chaque personne parle approximativement 35 fois, chacune avec une durée moyenne de 6*s* (à l'exception du présentateur qui prend la parole plus souvent et plus longtemps). Dans ce type de contenus, les dialogues sont très spontanés, et les tours de parole sont très courts.
- En apparence, chaque personne (à part le présentateur) apparaît approximativement dans 140 plans, chacun avec une durée moyenne de 3*s*. Le présentateur quand à lui apparaît beaucoup plus que les autres (297 fois dans l'épisode *S1*). Les séquences d'apparition des personnes sont très courtes dans ce type de contenus. Ceci s'explique par le fait que dans ces plateaux, plusieurs caméras sont disposées autour des personnes afin de capturer les différentes réactions. Le dialogue étant très interactif, les plans changent très vite.
- En visage parlant, à l'exception du présentateur qui apparaît très souvent lorsqu'il parle (447 fois dans l'épisode *S1*), chaque personne parle et apparaît

simultanément environ 70 fois. La durée moyenne d'un segment de visage parlant est de 2.7s.

	Parole (#*tours parole*)	*Apparence* (#*plans visuels*)	*Visages parlants* (#*segments visage parlants*)	*Locuteur non visible* (% *temps parole*)	*visage silencieux* (% *temps d'apparence*)
P1	918 (175 tours)	812 (297 plans)	471 (188 Seg)	447 (48.7%)	341 (42.0%)
P2	193 (41 tours)	383 (126 plans)	133 (45 Seg)	59 (30.8%)	249 (65.1%)
P3	240 (56 tours)	545 (174 plans)	162 (64 Seg)	78 (32.4%)	382 (70.2%)
P4	304 (38 tours)	529 (160 plans)	190 (52 Seg)	114 (37.5%)	339 (64.1%)
P5	252 (38 tours)	240 (69 plans)	156 (62 Seg)	96 (38.1%)	84 (34.9%)
P6	119 (25 tours)	395 (152 plans)	80 (26 Seg)	39 (32.6%)	315 (79.6%)
P7	239 (28 tours)	483 (150 plans)	163 (49 Seg)	76 (31.7%)	320 (66.2%)
P8	81 (17 tours)	163 (53 plans)	52 (22 Seg)	29 (35.4%)	110 (67.6%)

TABLE 2.2 – Analyse de corpus - Durée (en seconde) des plans d'apparition et tours de parole pour chaque personne dans l'épisode *S1*

Pour chaque personne, le visage associé à une voix est visible plus de 60% de son temps de parole, alors que le temps de parole d'un visage n'est que de 35% de la durée totale de l'apparence de ce visage dans l'épisode. Ainsi, pour ces émissions de télévision, la probabilité qu'un locuteur soit visible est beaucoup plus élevée (presque deux fois plus) que la probabilité qu'un visage soit parlant. Ceci introduit la problématique de l'indexation des personnes dans ce type de contenus dans lesquels aucune synchronisation entre les séquences de parole et l'apparence n'est garantie.

2.4 Problématiques d'indexation des personnes dans les *Talk shows*

L'indexation des personnes dans un contexte de *Talk shows* est un problème très difficile à cause des différentes ambiguïtés que présentent l'information audio, visuelle et leurs associations. Dans ces contenus, les interventions sont très interactives, spontanées et expressives rendant difficile la détection et l'identification des personnes par un processus automatique.

2.4.1 Difficultés de l'information audio

Dans les émissions de télévision de type *Talk show*, l'audio est caractérisé par des dialogues très interactifs :

- La parole est spontanée et très expressive : les personnes s'expriment de plusieurs façons différentes (changement de ton : rire, colère, étonnement, etc) et le dialogue est non fluide (hésitations, silences).
- Les tours de parole sont très courts. Le tour de parole de chaque personne est estimé en moyenne à 6 secondes.
- Plusieurs personnes peuvent parler en même temps.
- Présence de séquences d'applaudissements, de musique, jingles, etc.

2.4.2 Difficultés de l'information visuelle

FIGURE 2.5 – Exemples de visages dans la base de données *TSDB*

Dans les émissions de télévision de type *Talk show*, les personnes se présentent autour d'une table et discutent de manière très interactive. Les visages apparaissent avec beaucoup de variations de pose, de conditions d'éclairage et d'expressions faciales (rire, colère, étonnement, grimaces). Parfois, les visages apparaissent avec des occultations (lunettes, main, verre à eau, etc). La figure 2.5 présente des exemples de visages apparaissant dans la base de données *TSDB*. Il est très difficile de détecter et reconnaître les personnes par leurs visages.

FIGURE 2.6 – Les ambiguïtés dans l'association des informations audio et visuelle dans la base de données *TSDB* (résultat du détecteur de visage *Viola&Jones* [Viola and Jones, 2001]). Situation (a) : *Jérémy Michalak* (à gauche avec un cercle rouge) est le visage parlant. Situation (b) : le locuteur *Laurent Ruquier* n'est pas filmé. Situation (c) : le locuteur *Jérémy Michalak* (à gauche) n'est pas détecté automatiquement

2.4.3 Difficultés dans l'association des informations audio et visuelle

Dans un contexte de télévision, la synchronie entre les séquences de parole des personnes et leurs séquences d'apparence n'est pas garantie (voir section 2.3). Dans ce cas, l'association des modalités audio et visuelle est un exercice qui consiste à répondre aux trois questions suivantes : « *qui parle* »« *qui apparaît ?* »et « *qui est le visage parlant ?* ». Trois types de situations ambiguës rendent difficile la détermination de quel visage associer à quelle séquence audio :

- *Situations de multi-visage* : lorsque plusieurs personnes, parmi lesquelles le locuteur, apparaissent dans l'image (exemple (a) dans la figure 2.6). Dans ce cas, comment déterminer quel visage correspond au locuteur ?
- *Situations où le locuteur n'est pas filmé :* lorsque la caméra filme un plan contenant d'autres personnes (exemple (b) dans la figure 2.6). Dans ce cas, comment déterminer qu'aucun visage ne correspond au locuteur ?
- *Situations où le locuteur est non détecté en visuel :* lorsque le visage parlant est non détecté mais bien présent, alors que l'on détecte d'autre personnes (exemple (c) dans la figure 2.6).

2.4.4 Avantages des *Talk shows*

Dans la tâche de structuration audio-visuelle des contenus par personne, les contenus de type *Talk shows* présentent tout de même certains avantages.

Avantages dans les interventions audio

Étant donné que l'indexation des personnes est effectuée dans un même épisode, certaines variabilités temporelles dans la voix sont à exclure. Par exemple, les changements de la voix pour cause de maladie (grippe, angines, fumeurs, etc) ou vieillesse (variation de la voix entre jeune âge, adulte et vieux). Ces variations introduisent beaucoup de difficultés et constituent des *challenges* pour l'identification des personnes. En intra-épisode, les variabilités de prise de son sont rares dans notre base de données (cas de reportages de type micro-trottoir). Un changement de microphone introduit des ambiguïtés dans le processus d'indexation automatique.

Avantages dans les interventions visuelles

De la même manière que pour l'audio, certaines variations temporelles de l'apparence des personnes sont à exclure. Par exemple les changements dus à la croissance, rides, variations du poids, variations de la couleur de cheveux, changement de vêtements, etc. Un autre avantage présenté par ces contenus est que les caméras disposées autour de la table sont fixes. Un mouvement de caméra rend plus difficile l'analyse de l'apparence des personnes qui sont déjà en mouvement.

2.5 Conclusion

Indexer automatiquement des personnes dans un contexte de télévision est un véritable *challenge*. Plusieurs ambiguïtés dans l'audio, l'image et leur association rendent les technologies existantes de détection, regroupement et identification des personnes peu fiables. Dans ce chapitre, nous avons présenté et analysé la base de données *TSDB* sur laquelle nous avons évalué nos méthodes. Pour plusieurs raisons, la collecte, l'annotation et la diffusion de bases de données issues de la télévision est très difficile même par un organisme scientifique. La première raison est la notion du droit à l'image qui consiste à permettre à chaque personne de s'opposer à l'utilisation « *commerciale ou non* »de son image. Aussi, les procédures de négociations avec les droits de production des chaînes de télévision sont très compliqués. C'est pourquoi La base de données *TSDB* est collectée et annotée par *France télécom* à titre *privé*.

Chapitre 3

Protocole d'évaluation

Notre objectif est la structuration des contenus audio-visuels par personne en utilisant les modalités audio et visuelle. En indexation mono-modale, plusieurs protocoles d'évaluation sont utilisés pour mesurer les erreurs. En indexation audio-visuelle, l'origine des erreurs est plus complexe car issue de la combinaison de plusieurs types d'erreurs. Il est donc nécessaire d'analyser les nouvelles erreurs et de définir une façon de les mesurer. Dans ce chapitre, nous présentons les mesures de performance généralement utilisées pour évaluer les systèmes d'indexation des personnes et le protocole expérimental que nous avons utilisé pour évaluer les performances de nos méthodes.

3.1 Tour d'horizon

Plusieurs campagnes d'évaluation sont lancées dans le but d'encourager la recherche dans le domaine de l'indexation audio-visuelle. Ces campagnes fournissent des bases de données ainsi que des protocoles d'évaluation afin de comparer les performances des différents systèmes. Dans cette section, nous présentons quelques protocoles d'évaluation des systèmes d'indexation proposés lors de campagnes d'évaluations.

3.1.1 Mesure de pureté

La mesure de pureté a été introduite dans les systèmes d'indexation en locuteur dans [Solomonoff et al., 1998]. Un système d'indexation en locuteur construit un index de personnes détectées et regroupés en *Clusters*. Dans un *Cluster*, la pureté traduit le taux d'éléments étrangers au *Cluster*. Deux mesures de pureté peuvent être calculées : puretés du locuteur P_{Loc} et du groupe P_{Cl} :

$$P_{Cl} = \frac{1}{N_0} \sum_{i=1}^{N} p_i n_i \qquad avec \qquad p_i = \sum_{j=1}^{S} \frac{n_{ij}^2}{n_i^2} \tag{3.1}$$

$$P_{Loc} = \frac{1}{N_0} \sum_{j=1}^{S} p_j n_j \qquad avec \qquad p_j = \sum_{i=1}^{N} \frac{n_{ij}^2}{n_j^2} \tag{3.2}$$

où :

- N : nombre de classes du document audio.
- S : nombre de locuteurs du document audio.
- N_0 : nombre de trames du document audio.
- n_i : nombre de trames de la classe i.
- n_j : nombre de trames du locuteur j.
- n_{ij} : nombre de trames dans la classe i prononcées par le locuteur j.

P_{Loc} traduit le fait qu'un locuteur soit dispersé sur plusieurs groupe (*Clusters*), tandis que P_{Cl} traduit le fait qu'un *Cluster* contienne des données provenant de plusieurs locuteurs. Par exemple, si $P_{Loc} = 1$ et $P_{Cl} = 0.5$, cela signifie que toutes les données pour un même locuteur sont regroupées dans un même *Cluster*, mais que dans chaque *Cluster*, il y a en moyenne 2 locuteurs.

Ces mesures de pureté sont principalement utilisées en indexation en locuteurs. Elles peuvent également être utilisées dans d'autres systèmes d'indexation basés sur un processus de détection et regroupement (notamment en indexation de visages parlants).

3.1.2 Protocole d'évaluation *TRECVID*

La campagne d'évaluation *TRECVID* (*TREC Video Retrieval Evaluation*) lancée depuis 2001 par l'organisme *NIST* a pour objectif d'encourager la recherche d'information dans des contenus audio-visuels. Les tâches évaluées dans cette campagne sont des tâches de détection de concept prédéfinis (personne, voiture, animal, etc). Le protocole d'évaluation de la campagne consiste à détecter automatiquement les concepts, puis à déterminer la pertinence de la réponse du système en calculant les mesures *Précision* et *Rappel*. Dans la tâche de détection de personnes dans des séquences vidéos, un plan considéré pertinent est un plan dans lequel une personne est détectée et correctement associée à la bonne identité. Les mesures de Précision et Rappel sont calculées, par rapport aux plans, pour chaque personne de la manière

suivante :

$$Precision(P_i) = \frac{N^{Correct}(P_i)}{N^{Det}(P_i)} \qquad Rappel(P_i) = \frac{N^{Correct}(P_i)}{N^{Ref}(P_i)} \tag{3.3}$$

- $N^{Correct}(P_i)$: nombre de plans où la personne P_i est correctement détectée.
- $N^{Ref}(P_i)$: nombre de plans où la personne P_i est annotée en référence.
- $N^{Det}(P_i)$: nombre de plans où la personne P_i est détectée automatiquement.

3.1.3 Protocole d'évaluation *ESTER*

La campagne d'évaluation *ESTER* [Galliano et al., 2009] a pour objectif d'évaluer les systèmes d'analyse et d'indexation de documents audio en français. Les tâches sont organisées autour de l'évaluation de la segmentation et regroupement de locuteurs *(S)*, la transcription de la parole *(T)*, et l'extraction d'information *(E)*. Dans notre cas, nous nous intéressons au protocole d'évaluation de l'indexation des personnes (tâche *S*). L'outil proposé par *ESTER* pour l'évaluation des performances des systèmes d'indexation en locuteurs est *SpkrSegEval-v23.pl*. Cet outil permet de trouver la meilleure correspondance entre l'ensemble des étiquettes de référence et étiquettes obtenues par une segmentation et regroupement automatique. À partir de ces associations, plusieurs métriques sont calculées :

- *Reference Time* : correspond à la durée totale de personnes annotées manuellement.
- *Cluster Time* : correspond à la durée totale de personnes détectées par le système d'indexation automatique.
- *Correct Time* : correspond à la durée totale des segments détectés automatiquement et correctement associés à la bonne identité.
- *Error Time* : correspond aux erreurs sur les identités (arbitraires) des locuteurs.
- *False Alarm Time* : correspond à la durée totale de personnes automatiquement détectées mais non référencées.
- *Missed Time* : correspond à la durée totale de personnes référencées mais non détectées automatiquement.

La figure 3.1 schématise les différentes métriques pour l'indexation. À partir de ces métriques, plusieurs taux d'erreurs peuvent être calculés. Ces erreurs diffèrent selon qu'on se positionne du coté de la référence ou de la réponse automatique. Se

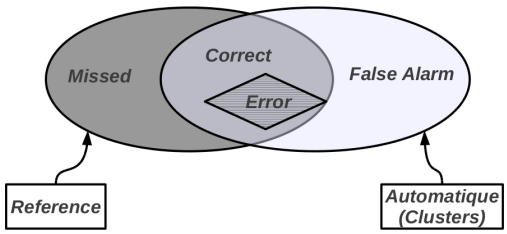


FIGURE 3.1 – Métriques pour l'évaluation des systèmes d'indexation. La référence correspond aux plans annotés manuellement. L'automatique correspond à l'ensemble des plans détectés automatiquement. On distingue 3 types de segments dans l'ensemble des plans détecté automatiquement : segment ne contenant pas de personne (*False Alarm Time*), contenant une personne associé à la mauvaise personne (*Error Time*) et les segments contenant une personne correctement identifiée (*Correct time*). On distingue deux types de segments dans l'ensemble de référence : segments dans lesquels on a détecté une personne automatiquement et segments dans lesquels aucune personne n'est détectée automatiquement (*Missed Time*)

positionner côté référence signifie que les métriques calculées par l'outil sont comparées au temps total de référence. Dans ce cas, on privilégie la capacité du système à retrouver le plus de possible de segments référencés. Dans le cas ou l'on se positionne côté réponse du système, les métriques sont comparés au temps total du *Cluster*. Les taux calculés montrent la composition de la réponse en terme bonnes et mauvaises réponse. Souvent, les évaluations se positionnent côté référence par le taux d'erreur *Diarization Error Rate* qui est calculé de la manière suivante :

$$DiarizationErrorRate = \frac{ErrorTime + FalseAlarmTime + MissedTime}{ReferenceTime} \tag{3.4}$$

Afin d'évaluer la réponse du système, il est intéressant de mesurer le taux de perte (calculé par rapport au temps de référence) et la composition de la réponse automatique (calculé par rapport au temps de réponse automatique). Le taux de perte est calculé de la manière suivante :

$$MissedDurationRate(MDR) = \frac{MissedTime}{ReferenceTime} \tag{3.5}$$

Pour la composition de la réponse automatique, il y a 3 possibilités de réponse du système automatique : segments bien détectés et correctement identifiés (*Correct Duration Rate*), segments détectés et faussement identifiés (*Error Duration Rate*)

ou segments qui ne sont pas dans la référence (*False Alarm Rate*, cas de personnes détectées automatiquement alors qu'il y en a pas). La composition de la réponse automatique est calculée de la manière suivante :

$$\begin{aligned} CorrectDurationRate(CDR) &= \frac{CorrectTime}{ClusterTime} \\ ErrorDurationRate(EDR) &= \frac{ErrorTime}{ClusterTime} \\ FalseAlarmRate(FAR) &= \frac{FalseAlarmTime}{ClusterTime} \end{aligned} \tag{3.6}$$

En comparaison avec les évaluations *TRECVID*, le *Correct Duration Rate (CDR)* correspond à la mesure de précision définie précédemment.

3.2 Proposition d'un protocole d'évaluation

3.2.1 Notations

Soit x une séquence audio-visuelle dont on extrait un vecteur de paramètres acoustiques noté x_A et un vecteur de paramètres visuels noté x_V. On note $A(x_A)$ la fonction d'annotation audio qui associe à vecteur de paramètres acoustiques x_A une étiquette audio a_i avec $i \in \{a_1, .., a_K\}$. On note $V(x_V)$ la fonction d'annotation visuelle qui associe à un vecteur de paramètres visuels x_V à une étiquette audio v_i avec $i \in \{v_1, .., v_L\}$.

Un segment de visage parlant x est associé à deux étiquettes : $A(x_A) = a_i$ provenant de l'index audio et $V(x_V) = v_j$ provenant de l'index visuel. On note l'association des étiquettes audio et visuel par $a_i \Leftrightarrow v_j$.

3.2.2 Origines des erreurs

En indexation audio-visuelle de personnes, après avoir identifié les associations entre les personnes détectées de manière audio et visuelle, les segments de visages parlants sont obtenus par intersection des index audio et visuel. On distingue trois types d'erreurs : *Error Time*, *False Alarm Time* et *Missed Time* définis précédemment. Ces erreurs sont issues de propagation des erreurs d'une des deux modalité audio ou visuelle (ou des deux).

Dans la réponse du système de structuration audio-visuel, un segment de visage parlant est considéré comme une erreur lorsque les deux systèmes d'indexation détectent le visage parlant mais l'associent tous les deux à la mauvaise personne. La

figure 3.2 présente la répercussion des erreurs issues des deux systèmes se structuration sur le système de fusion.

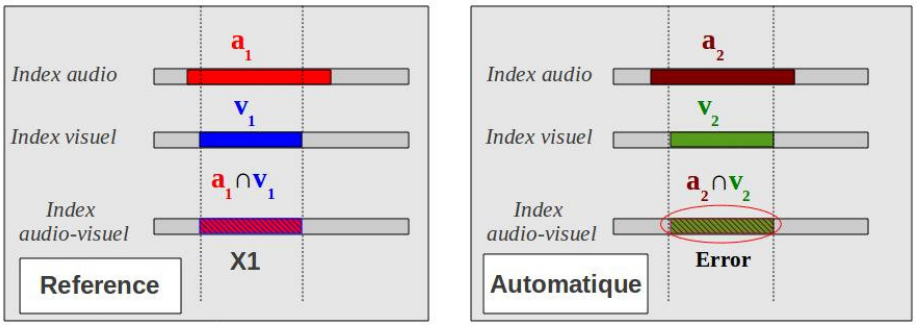


FIGURE 3.2 – *Error Time* : exemple de segment considéré comme une erreur en structuration audio-visuelle. Le segment de visage parlant $X1$ annoté manuellement par (a_1, v_1) est correctement détecté mais associé à la mauvaise identité par les deux modalités (a_2,v_2)

Un segment de visage parlant $x(x_A, x_V)$ est considéré comme une fausse alarme dans deux situations :

- lorsque l'une des deux modalités (ou les deux) commet une fausse alarme (voir exemple *False Alarm1* dans la figure 3.3).
- Dans le cas d'une erreur d'identification d'un des deux systèmes (ou les deux) qui favorise une association $A(x_A) \Leftrightarrow V(x_V)$ (voir l'exemple *False Alarm2* dans la figure 3.3.

Un segment de visage parlant $x(x_A, x_V)$ n'est pas retrouvé automatiquement par le système de structuration audio-visuel dans deux situations :

- Lorsque que la personne n'est pas détectée par l'une (ou les deux) modalité(s) (voir l'exemple *Missed Error1* dans la figure 3.4).
- Dans le cas d'une erreur d'identification d'un des deux systèmes (ou les deux) qui ne favorise pas une association $A(x_A) \not\Leftrightarrow V(x_V)$ (voir l'exemple *Missed Error2* dans la figure 3.4.

3.2.3 Mesures de performances

Une méthode de structuration de documents audio-visuels par personne permet d'obtenir un index audio-visuel des personnes. Afin dévaluer objectivement cet index, il est comparé à l'index de référence annoté manuellement. L'évaluation de nos expériences est effectuée grâce à l'outils *SpkrSegEval-v23.pl*. La mesures *précision* Pr,

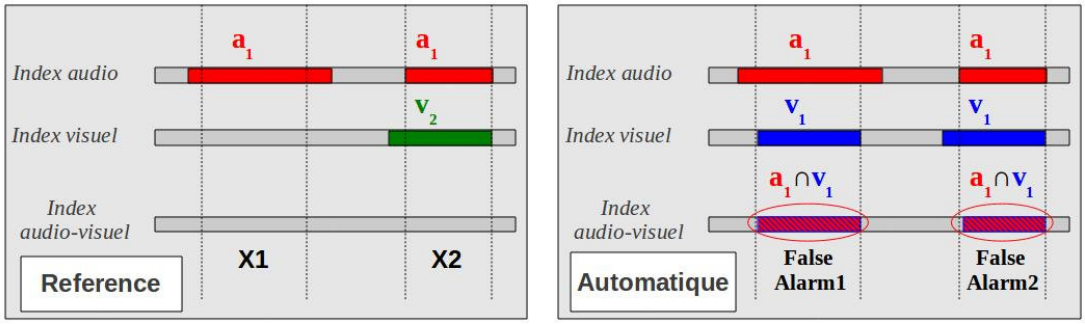


FIGURE 3.3 – *False Alarm Time* : exemples de fausses alarmes. Dans le cas *False Alarm1*, le segment $X1$ est correctement détecté et identifié par le système audio comme étant a_1. Le système visuel détecte une personne par erreur et l'associé à v_1. Cette erreur conduit à la détection d'un visage parlant alors qu'il n'y en a pas. Dans le cas *False Alarm2*, le segment $X2$ est correctement détecté et identifié par l'audio comme étant a_1. Le système visuel détecte une personne correctement mais l'associe à la personne v_1 par erreur. Cette erreur favorise l'association $a_1 \Leftrightarrow v1$ et conduit par erreur à la détection d'un visage parlant

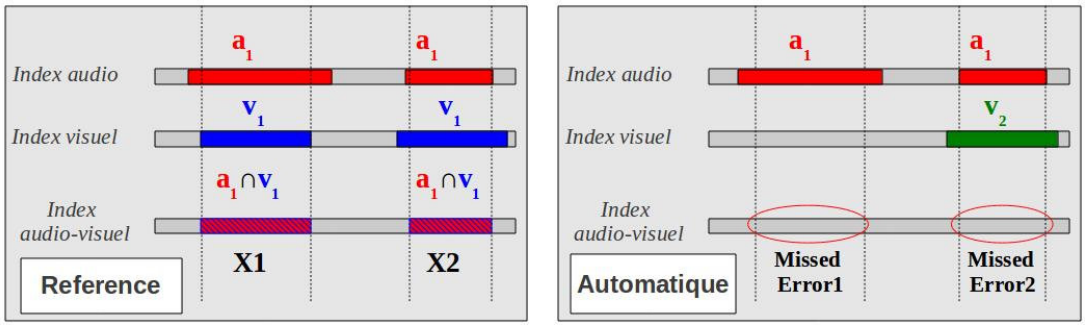


FIGURE 3.4 – *Missed Time* : exemples de segments de visages parlants non détectés. Dans le cas *Missed Error1*, le segment $X1$ est correctement détecté et identifié par le système audio comme étant a_1 mais le système visuel ne le détecte pas. Cette erreur conduit à la non détection du visage parlant. Dans le cas *Missed Error2*, le segment $X2$ est correctement détecté et identifié par l'audio comme étant a_1. Le système visuel détecte le visage existant mais l'associe à la personne v_2 par erreur. Cette erreur ne favorisant pas une association $a_1 \not\Leftrightarrow v_2$ conduit à la non détection du visage parlant

le *rappel* Ra et la *F-mesure* F_m qui combine la précision et le rappel sont calculées de la manière suivante :

$$Pr = \frac{CorrectTime}{ClusterTime}, \qquad Ra = \frac{CorrectTime}{ReferenceTime} \tag{3.7}$$

$$F_m = \frac{2 \times Pr \times Ra}{Pr + Ra} \tag{3.8}$$

Compte-tenu de la difficulté à traiter des documents multimédias, on considère 2 types d'évaluation : *l'évaluation complète* et *l'évaluation restreinte* :

L'évaluation complète

Dans *l'évaluation complète*, les mesures de performances sont calculées sur toute la durée du document multimédia.

L'évaluation restreinte

Dans *l'évaluation restreinte*, les mesures de performances sont calculées sur une partie sélectionnée du document à structurer. Cette restriction est une convention souvent utilisée en analyse audio [Galliano et al., 2009] afin de se concentrer spécifiquement sur l'évaluation du regroupement en locuteur sans être masqué par d'autres problèmes telle qu'une détection erronée de segment de parole. Les index de références ainsi que ceux obtenus de manière automatique subissent deux prétraitements :

- Une fenêtre de 0.25 secondes est supprimée sur les frontières des segments (dans la référence et les segments détectés automatiquement). L'intérêt de faire ça est de tolérer des décalages ($< 0.25s$) entre les frontières de référence et les frontières automatiques.
- Les segments référencés en tant que séquences de non parole ne sont pas pris en compte dans l'évaluation. Ces segments sont annotés comme des séquences de non-parole, paroles qui se chevauchent (multi-intervenants), applaudissements, musique, reportages, etc.

Chapitre 4

Système de structuration de documents audio-visuels

Rappelons que notre objectif est de structurer des programmes de télévision en termes de visage parlants sans aucun dictionnaire prédéfini d'identités. Dans la littérature, la plupart des méthodes d'indexation des personnes dans des contenus audio-visuels proposent de construire l'index à partir de l'intersection des index de locuteur et visage obtenus de manière indépendante [Everingham et al., 2006, Jaffré et al., 2007].

Dans ce chapitre, nous présentons en première partie les systèmes de structuration basés sur l'information audio et visuelle de manière indépendante. Ensuite, nous présentons la méthode de fusion des deux index afin d'obtenir l'index de visages parlants. Le choix d'effectuer l'indexation de manière indépendante par l'audio et le visuel se justifie par le fait que nous voulons offrir aux utilisateurs le choix de naviguer dans un document en sélectionnant les séquences d'interventions en parole d'une personne spécifique, ou les séquences de sa présence à l'écran ou encore les séquences de ses interventions en parole et en présence à l'écran.

4.1 Système basé sur l'information audio

Dans cette partie, le but est de construire un index de locuteur à partir d'un document audio en détectant et regroupant les interventions sonores des personnes. Nous avons utilisé un système développé en interne à *OrangeLabs* qui s'inspire de la méthode présenté dans [Deléglise et al., 2005]. Cette méthode de structuration est divisée en deux phases. Une première phase de segmentation et regroupement disjoints est effectuée. Nous avons utilisé une méthode basée sur un algorithme de regroupement hiérarchique ascendant avec un critère *BIC* (Bayesian Information Criterion) [Barras et al., 2006]. Ensuite, de manière itérative, une nouvelle segmentation et regroupement est effectuée basée sur un modèle *HMM* évolutif.

4.1.1 Traitements préliminaires

Premièrement, les coefficients *MFCC* (*Mel-frequency cepstral coefficients*) sont extraits sur une fenêtre de $32ms$ chaque $16ms$. Le vecteur des paramètres de dimension $Dim = 36$ est composé des 12 premiers coefficients *MFCCs* et des dérivés premières et secondes. Ensuite, une première segmentation du signal en parole/non parole est effectuée afin de ne conserver que les séquences des interventions sonores des personnes. La méthode utilisée est basée sur une classification à deux classes parole et non parole. Les classes sont modélisées par une mixture de 64 gaussiennes (*GMMs*).

4.1.2 Première phase : structuration et segmentation disjoints

Segmentation

Cette phase consiste à découper chaque segment classé comme étant de la parole en petits segments contenant un seul locuteur. La méthode de segmentation est basée sur la mesure de similarité selon le critère *BIC* [Barras et al., 2006] entre toutes les deux fenêtres consécutives de taille fixe afin de détecter un changement de locuteur.

Regroupement

Une fois que la séquence audio est découpée en segments supposés contenir chacun un seul locuteur, la phase de regroupement consiste à rassembler tous les segments d'un même locuteur. Nous avons utilisé une méthode de regroupement hiérarchique ascendante basée sur le critère *BIC* (voir la description dans le chapitre 1). À chaque itération, les deux segments qui présentent un minimum de distance selon le critère *BIC* sont regroupés. L'algorithme s'arrête lorsque toutes les variations Δ_{BIC} entre les groupes dépassent un seuil théorique S fixé à 0.

4.1.3 Seconde phase : structuration et segmentation conjoints

Les résultats de la première phase de segmentation et regroupement sont utilisés pour initialiser un modèle *HMM* dans lequel chaque état représente un locuteur et les transitions représentent le passage d'un locuteur à un autre. Chaque *Cluster* obtenu dans la première phase est utilisé pour apprendre un modèle de voix *GMM* (64 gaussiennes). Un décodage *Viterbi* permet d'obtenir une nouvelle segmentation (détection

de changements de locuteurs par les transitions du *HMM*). À partir de cette nouvelle segmentation, un nouveau regroupement des *Clusters* est effectué. Ce regroupement est basé sur le critère *CLR* (*Cross Liklihood Ratio*) [Deléglise et al., 2005]. Les nouveaux *Clusters* permettent de faire évoluer le nombre d'états du *HMM*. les *GMMs* sont ré-estimés à partir de la nouvelle segmentation et regroupement. De manière itérative, le *HMM* est réinitialisé afin de re-segmenter la séquence audio. Ce processus de segmentation et regroupement est réitéré jusqu'à stabilisation du découpage.

À la fin du processus de regroupement, chaque segment audio-visuel x dont on a extrait le vecteur x_A de paramètres acoustiques est associé à une étiquette audio notée $A(x_A) = a_i$ parmi les K étiquettes audio $a_{i=1,..,K}$ détectées automatiquement. Le nombre d'étiquettes audio peut être supérieur au nombre exact de locuteurs intervenant dans le document sonore.

4.1.4 Résultats et discussion

À la fin du processus de segmentation et regroupement, chaque segment détecté automatiquement est attribué à un groupe audio. Le système de structuration basé sur l'information audio est évalué sur les cinq épisodes $(S1, .., S5)$ de la base de données *TSDB* présenté dans le chapitre 2. Afin de mieux analyser la réponse du système, nous présentons les résultats par plusieurs mesures : taux de perte (*MDR*) et composition de la réponse automatique *CDR + EDR + FAR*, et par les mesures *Précision + Rappel + F-mesure* (voir chapitre 3).

Le nombre de *Clusters* détectés automatiquement

En analysant les groupes audio détectés, on constate que le système de regroupement produit au moins un *Cluster* contenant les séquences ambiguës de brouhaha, d'échanges très rapides et de double parole. De plus, certaines personnes présentant une variabilité importante de la voix selon l'expressivité (voix « normale »et voix « énervée »ou sur bruit de fond important) peuvent donner lieu à deux *Clusters* différents. En particulier, la voix du présentateur *Laurent Ruquier* qui parle sur les rires ou les applaudissements, pour reprendre le fil de l'épisode, présente une voix assez différente de sa voix normale. Dans ce cas, le système a créé 2 groupes de la même personne.

Le tableau 4.1 présente le nombre de personnes audio détectées automatiquement par le système basé sur le critère *BIC* ainsi que le nombre de personnes annotées dans la référence.

Épisodes	*S1*	*S2*	*S3*	*S4*	*S5*
# personnes détectées automatiquement	11	10	8	12	8
# de personnes annotées	8	9	7	7	7

TABLE 4.1 – Regroupement de locuteur par le critère *BIC* - Nombre de *Clusters* audio détectés automatiquement dans les épisodes de la base de données *TSDB*

Résultats du *Clustering* audio

Épisodes	*CDR + EDR + FAR*	*MDR*	*Pr*	*Ra*	F_m
S1	68.7 + 17.5 + 13.8	3.5	68.7	75.4	71.9
S2	70.6 + 18.2 + 11.2	6.6	70.6	70.2	72.3
S3	**46.0 + 25.2 + 28.8**	**13.1**	**46.0**	**56.1**	**50.6**
S4	75.4 + 16.9 + 7.7	8.6	75.4	74.7	75.0
S5	72.0 + 22.4 + 5.6	10.4	72.0	68.4	70.1
All	66.5 + 20.0 + 13.4	8.4	66.5	69.0	68.0

TABLE 4.2 – *Évaluation complète* des résultats de l'indexation en locuteurs sur les 5 épisodes de la base de données *TSDB*. L'épisode $S3$ obtient des taux particulièrement bas

Résultats Le tableau 4.2 présente les performances du système de structuration basée sur l'audio des épisodes de la base de données *TSDB* selon la procédure d'évaluation *complète*. Le taux *CDR* (*Correct Duration Rate*) correspondant au temps correctement associé à la bonne personne audio et varie de 46.0 à 72.0% selon les épisodes. Le taux de perte *MDR* varie également beaucoup (de 3.5 à 13.1%) et le système détecte entre 5.6 et 28.8% de fausses alarmes (segments dans lesquels une personne est détectée automatiquement alors qu'il n'y a personne qui parle).

Discussion Les variations des résultats d'un épisode à un autre s'expliquent par la forte interactivité des dialogues dans ce contexte. En particulier, l'épisode $S3$ obtient une précision de $Pr = 46\%$ avec un rappel $Ra = 56.1\%$. Ces taux particulièrement bas s'expliquent par le fait que dans cet épisode, souvent, les tours de paroles de différentes personnes se chevauchent.

Résultats Le tableau 4.3 présente les performances du système de structuration basée sur l'audio selon la procédure d'évaluation *restreinte*. Le *CDR* varie de 71.8

Épisodes	*CDR + EDR + FAR*	*MDR*	*Pr*	*Ra*	F_m
S1	$84.1 + 15.8 + 0$	4.9	84.1	80.1	82.1
S2	$82.3 + 17.7 + 0$	5.1	82.3	78.1	80.2
S3	$71.8 + 28.2 + 0$	11.6	71.8	63.4	67.4
S4	$87.7 + 11.3 + 0$	7.3	87.7	82.1	84.8
S5	$82.5 + 17.5 + 0$	9.1	82.5	75.2	78.7
All	$81.7 + 18.1 + 0$	7.6	81.7	75.8	78.6

TABLE 4.3 – *Évaluation restreinte* des résultats de l'indexation en locuteurs sur les 5 épisodes de la base de données *TSDB*

à 87.7% selon les épisodes. Le taux d'erreur *EDR* varie de 11.3 à 28.2%. Dans l'évaluation *restreinte*, les taux d'erreurs sont calculés sur les segments qui ont été annoté comme étant de la parole. Ce qui explique que le système ne détecte pas de fausses alarmes.

Discussion Comparé à l'évaluation *complète*, les performances sont améliorées ($+15,1\%$ en précision et $+6,7\%$ en rappel). Ceci s'explique par la suppression des segments annotés manuellement comme étant des séquences de non-parole ou parties ambiguës (double parole). Notre système n'est pas robuste à la parole dans un contexte très interactif, mais quand le son est clair, la précision de la réponse du système est d'environ 80%. Jusqu'à présent, peu de travaux ont été effectués sur des contenus aussi interactifs. Beaucoup d'efforts méritent d'être fait dans ce domaine afin d'améliorer les performances.

4.2 Système basé sur l'information visuelle

Dans cette section, de la même façon que pour le locuteur, l'objectif est de construire automatiquement un index de personnes dans un contenu de télévision basé uniquement sur l'apparence, sans aucune liste prédéfinie des identités des participants. Dans ce contexte, en raison de la forte variabilité de l'apparence du visage d'une même personne, il est très difficile de détecter et d'identifier les visages avec une grande fiabilité. En effet, même s'il existe des méthodes de détection de visages avec différentes poses, la reconnaissance du visage reste très difficile dans le cas de visage non frontal, expressif ou avec des occultations. Le costume des personnes étant sujets à moins de variations que le visage (voir le chapitre 1), nous avons choisi de détecter et regrouper les plans dans lesquels apparait chaque personne en utilisant

la signature des couleurs de leurs costumes.

4.2.1 Détection des costumes

Afin de détecter les costumes à partir d'une image, nous nous sommes inspirés de la méthode présentée dans [Jaffre and Joly, 2004]. Cette méthode est basée sur la recherche d'un rectangle sous le visage détecté automatiquement. D'abord, le visage est détecté en utilisant l'implémentation *OpenCV* de l'algorithme *Viola&Jones* [Viola and Jones, 2001]. Ensuite, un rectangle de la région du costume est déterminé sous le visage. Ce rectangle est proportionnel à la taille du visage détecté : $\times 3.6$ la largeur du visage pour la largeur du costume et $\times 1.5$ la hauteur du visage. La figure 4.1 présente la détection des costumes sur une image extraite de la base de données *TSDB*. La zone du costume est restreinte à une petite zone afin d'éviter de prendre des pixels de l'arrière plan.

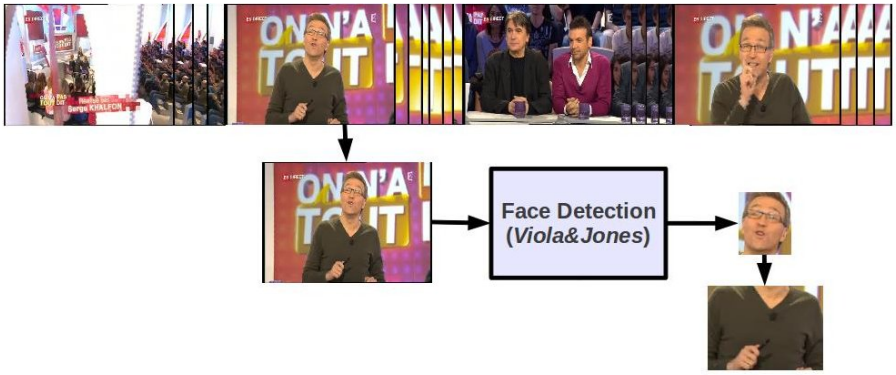


FIGURE 4.1 – Exemple de détection du costume à l'aide du visage

Sur un plan donné, un détecteur de visage est lancé sur chaque trame, puis le costume est localisé pour chaque visage détecté. Ainsi, chaque personne détectée est associée à une série de costumes détectés sur toutes les trames du plan.

4.2.2 Représentation du costume

Chaque costume détecté est représenté par sa signature de couleurs. Les couleurs sont codé dans le système *RGB* (Rouge, Vert, Bleu). La figure 4.2 montre un exemple de représentation de la distribution des couleurs d'un costume détecté.

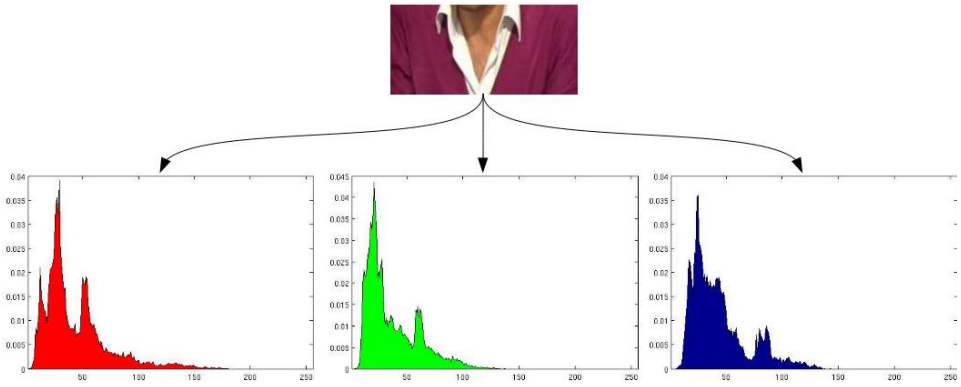


FIGURE 4.2 – Représentation du costume par la distribution des couleurs (RGB)

4.2.3 Sélection du meilleur costume

Afin de représenter la personne détectée dans un plan à partir de la série de costumes détectés, le costume le plus représentatif de la série de costumes est sélectionné. Ce costume « centroïde »est déterminé par la sélection du costume le plus proche des autres costumes de la série en termes de corrélation des histogrammes de couleurs. La figure 4.3 illustre un exemple du costume sélectionné à partir d'une série de costumes.

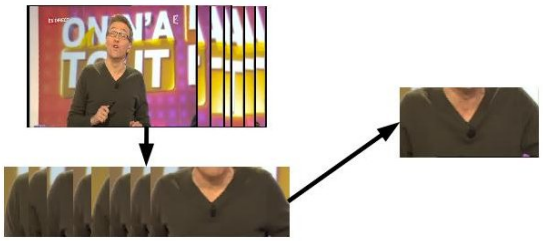


FIGURE 4.3 – Exemple de sélection du meilleur costume

4.2.4 Regroupement des costumes

La plupart des approches de regroupement de personnes sont basées sur des algorithmes hiérarchiques ascendants où les groupes de personnes (appelés *Clusters*) sont construits de manière itérative. Dans nos expériences, trois mesures de similarités ont été testées : euclidienne, corrélation, *Ward* [Ward, 1963]. La méthode de *Ward* a été

sélectionnée pour avoir présenté le meilleur regroupement au sens pureté des *Clusters*.

Au départ de l'algorithme de regroupement *Ward*, chaque costume détecté dans l'épisode est associé à un *Cluster* unique. À chaque itération, toutes les combinaisons des groupes sont étudiées et les *Clusters* qui présentent le minimum de perte d'information sont regroupés. La perte d'information entre deux groupes A et B est calculée de la manière suivante :

$$\Delta(A,B) = \sum_{i \in A \cup B} ||X_i - \bar{X}_{A \cup B}||^2 - \sum_{i \in A} ||X_i - \bar{X}_A||^2 - \sum_{i \in B} ||X_i - \bar{X}_B||^2 \qquad (4.1)$$

où $\bar{X}_{A \cup B}$ représente le centroïde du groupe $A \cup B$. La figure 4.4 illustre l

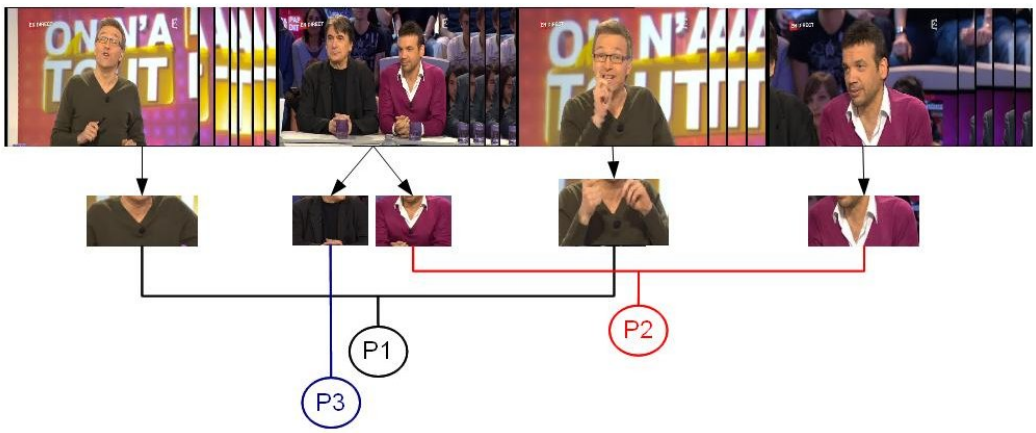


FIGURE 4.4 – Regroupement de Costumes

Dans les méthodes de regroupements hiérarchiques, les éléments sont regroupés sans tenir compte de l'information temporelle. Toutefois, les costumes détectés dans le même plan ne devraient impérativement pas être associés à la même personne. Dans certains cas, il arrive que les personnes apparaissant dans le même plan aient des costumes similaires (en terme de signature de couleurs). Dans l'algorithme de regroupement *Ward*, nous avons ajouté une contrainte temporelle afin de rendre impossible le regroupement des costumes détectés dans le même plan. Pour cela, nous avons intégré la liste des costumes détectés dans le même plan afin de leur affecter une très grande distance. L'algorithme s'arrête lorsque toutes les variations $\Delta(A,B)$

entre les groupes dépassent un seuil S fixé.

À la fin du processus de regroupement, chaque segment audio-visuel x dont on a extrait le vecteur de paramètres du costume x_C est associé à une étiquette visuelle $V(x_C)$ parmi les L étiquettes visuelles $v_{i=1,..,L}$ déterminées automatiquement.

4.2.5 Résultats et discussion

Étant donné la complexité du contexte, nous avons souhaité limiter la détection de personne aux plans annotés contenant les personnes à indexer afin d'éviter les plans généraux, plans publiques et des vues hors plateau. Cette évaluation est équivalente à l'évaluation *restreinte* en indexation en locuteur.

Le tableau 4.4 présente le nombre de personnes visuelles détectées automatiquement par le système ainsi que le nombre de personnes annotées dans la référence. Afin de simplifier le problème, nous avons considéré que le nombre de personnes est connu a priori.

Épisodes	*S1*	*S2*	*S3*	*S4*	*S5*
# personnes détectées automatiquement	9	9	7	7	10
# de personnes annotées	8	9	7	7	7

TABLE 4.4 – Regroupement des costumes - Nombre de *Clusters* visuels détectés dans les épisodes de la base de données *TSDB*

Épisodes	*CDR + EDR + FAR*	*MDR*	*Pr*	*Ra*	F_m
S1	69.9 + 30.1 + 0	9.1	69.9	63.6	66.6
S2	81.7 + 18.7 + 0	7.7	81.7	75.3	78.4
S3	90.2 + 9.8 + 0	11.0	90.2	80.3	85.0
S4	67.5 + 32.5 + 0	7.7	67.5	62.3	64.8
S5	83.9 + 16.1 + 0	13.5	83.9	72.3	77.8
All	78.6 + 21.4 + 0	9.8	78.6	70.8	74.5

TABLE 4.5 – Évaluation des résultats de la structuration par le costume sur les 5 épisodes de la base de données *TSDB*

Résultats Le tableau 4.5 résume les performances du système de structuration de personne basée sur les costumes sur la base de données *TSDB*. Le taux de *EDR*

correspondant à la durée des plans associés à la mauvaise personne. Ce taux varie de 9.8 à 32.5% selon les épisodes. Ces erreurs de regroupement s'expliquent par le fait que dans ces émissions de télévision, il arrive que deux personnes soient habillées avec des costumes semblables (voir la figure 4.5). Le taux de perte *MDR* (variant de 7.7 à 13.5%) correspond aux plans dans lesquels la détection de costume a échoué. Quand au *FAR*, les valeurs nulles s'expliquent par l'hypothèse de simplification de l'étude en restreignant la recherche de personne uniquement dans les plans où l'on a annoté la présence des personnes.

Discussion La méthode de structuration basée uniquement sur le costume peut introduire des erreurs dans le regroupement en cas de costumes semblables en termes de signature de couleurs. La figure 4.5 montre un exemple de costume type pour chaque personne dans l'épisode *S2* avec la matrice de similarité (distance euclidienne) entre tous les costumes détectés dans chaque plan durant l'épisode. On remarque que les costumes des personnes 1, 2, 3 et 5 sont proches. Un moyen d'améliorer les performances du système de regroupement basé sur les costumes est de modéliser ceux-ci en utilisant la forme ou la texture afin de minimiser les erreurs de regroupement de costumes appartenant à différentes personnes [Jaffré, 2005]. Une autre façon de palier au problème des costumes similaires est d'inclure le visage pour mieux discriminer les costumes. Mais l'amélioration apportée peut dans ce cas être masquée par des erreurs de discrimination par le visage vue sa variabilité dans ce contexte.

4.3 Appariement audio-visuel

L'objectif du système est d'indexer automatiquement des visages parlants dans des contenus de télévision. Dans ce but, deux index de personnes $a_{i=1,..,K}$ et $v_{i=1,..,L}$ sont construits de manière indépendante en utilisant les informations audio et visuelle. Une fois ces index obtenus, il est nécessaire d'apparier automatiquement chaque personne détectée de manière audio à chaque personne détectée de manière visuelle.

4.3.1 Recherche d'associations

La fonction de liaison entre les étiquettes audio et les étiquettes visuelles est obtenue comme suit : explorer toutes les combinaisons possibles et sélectionner celle qui maximise la durée « totale »de l'intersection de paires associées. Cette association est formalisée par la fonction $f_{Association}(v_i) = a_k$. La fonction de liaison maximise

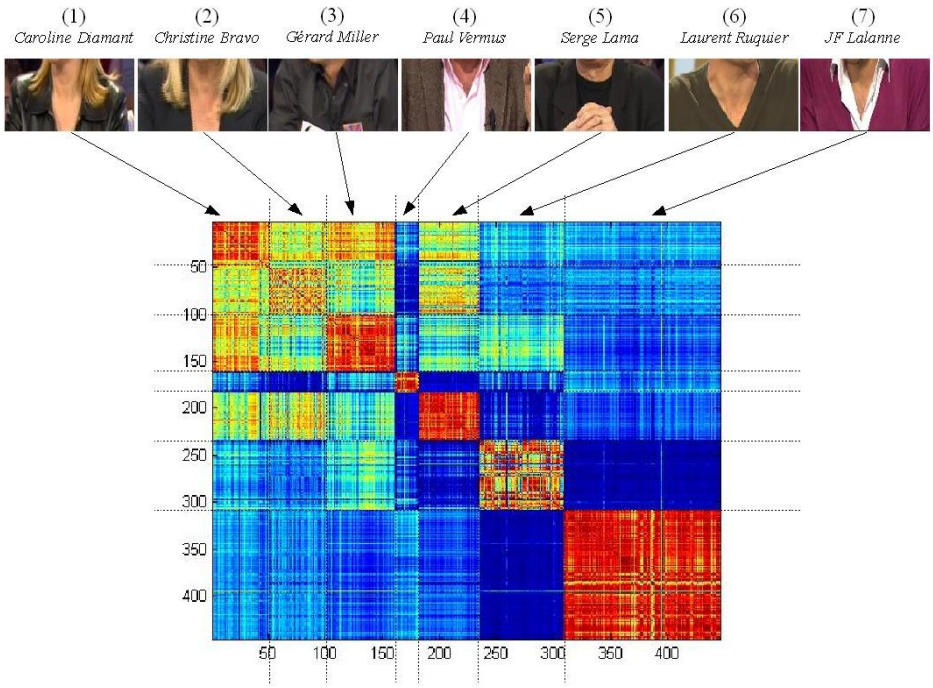


FIGURE 4.5 – Exemples de costumes dans l'épisode *S2* - matrice de similarité entre les costumes détectés

la couverture globale des visages parlants et donc suppose que, globalement sur l'ensemble de contenu, le visage qui apparaît le plus longtemps lorsque quelqu'un parle est bien le visage du locuteur. Cette hypothèse est vérifiée dans ce type de contenus (voir l'analyse du contexte présenté dans le chapitre 2). L'avantage de cette méthode est qu'elle permet de traiter les cas de voix off ou de visages muets ($f_{Association}(v_i)$ peut être nulle). La figure 4.6 présente un exemple d'associations entre 3 étiquettes audio (grise, rose et bleue) et 3 étiquettes visuelles (noire, rouge et bleue marine) par maximisation de la couverture globale.

4.3.2 Fusion d'index

A l'issue de l'étape de recherche d'association, chaque segment audio-visuel x dont on extrait le vecteur de paramètres acoustiques x_A et le vecteur de paramètres du costume x_C est associé à une paire d'étiquettes $(A(x_A), V(x_C))$. Dans l'index issu de l'intersection, seuls les segments audio-visuel x où $A(x_A) = f_{association}(V(x_C))$ sont sélectionnés comme segments de visages parlants.

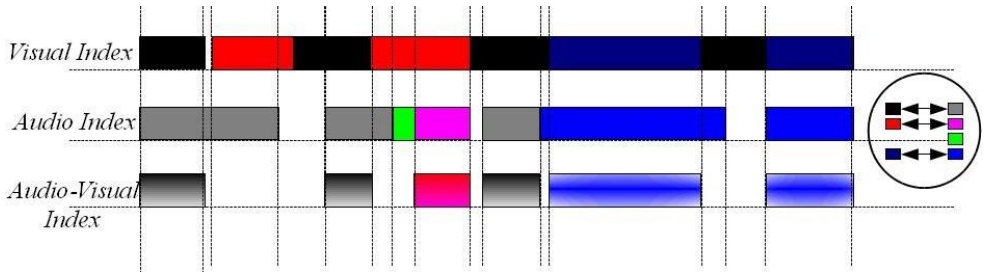


FIGURE 4.6 – Exemple de fusion d'index par maximisation de la couverture globale. Les étiquettes visuelles : noire, rouge et bleue marine sont associées respectivement aux étiquettes audio grise, rose et bleue

4.3.3 Résultats et discussion

Dans le système de structuration audio-visuel, l'origine des erreurs provient de la propagation des erreurs de détection et de regroupements des deux modalités audio ou visuelle (voir la présentation des types d'erreurs dans 3).

Résultats

Épisodes	*CDR + EDR + FAR*	*MDR*	*Pr*	*Ra*	F_m
S1	86.2 + 2.3 + 11.5	41.6	86.1	56.8	68.4
S2	88.4 + 1.8 + 9.8	35.1	88.4	63.6	74.0
S3	65.0 + 8.9 + 26.0	41.4	65.0	51.6	57.5
S4	91.8 + 0.6 + 7.6	37.9	91.8	61.7	73.9
S5	87.8 + 4.1 + 8.1	53.4	87.8	44.6	59.2
All	83.8 + 3.5 + 12.6	41.9	83.8	55.7	66.7

TABLE 4.6 – *Évaluation complète* en audio des résultats de la structuration audio-visuelle sur les 5 épisodes de la base de données *TSDB*

Le tableau 4.6 présente les résultats du système de structuration audio-visuelle des épisodes de l'émission de télévision par personne pour l'évaluation *complète*. Dans cette procédure d'évaluation, une grande partie des segments de visages parlants n'est pas détectée (entre 35.1 et 53.4%). Ces erreurs peuvent provenir de deux sources : par la non-détection automatique de la personne par l'une des deux modalités ou une erreur d'étiquetage automatique de l'un des deux systèmes (voir la description du protocole d'évaluation dans la section 3.2 du chapitre 3). Le taux de fausses alarmes

FAR varie entre 7.6 et 26%. Ce taux reflète l'impact des erreurs de détection et regroupement (qui favorisent une association) produites par les deux systèmes de structuration audio et visuelle.

Le tableau 4.7 présente les performances du système de structuration audio-visuelle suivant la procédure d'évaluation *restreinte*. Dans la fusion, l'*évaluation restreinte* consiste à restreindre l'évaluation aux segments de visages parlants qui sont inclus dans l'ensemble des segments annotés comme parole.

Épisodes	*CDR + EDR + FAR*	*MDR*	*Pr*	*Ra*	F_m
S1	$94.8 + 1.9 + 3.3$	39.2	94.8	61.4	74.5
S2	$93.5 + 1.8 + 4.7$	33.2	93.5	65.7	77.2
S3	$78.9 + 9.3 + 11.8$	36.2	78.9	55.4	65.1
S4	$95.3 + 1.7 + 3.0$	35.3	95.3	63.7	76.3
S5	$92.0 + 3.8 + 4.2$	51.5	92.0	56.2	69.8
All	$90.9 + 3.7 + 5.4$	39.1	90.9	60.5	72.6

TABLE 4.7 – *Évaluation restreinte* en audio des résultats de la structuration audio-visuelle sur les 5 épisodes de la base de données *TSDB*

Dans cette évaluation, grâce à la suppression des segments audio ambigus, l'index de personnes obtenu par le système de structuration basé sur l'audio commet moins de fausses alarmes (voir les résultats dans le tableau 4.3). Cette amélioration se reflète dans le taux *FAR* du système audio-visuel qui diminue significativement pour chaque épisode comparé au *FAR* de l'évaluation *complète*. La totalité des fausses alarmes dans l'évaluation *restreinte* vient donc des erreurs de regroupement qui favorise une association. Par contre, le taux de perte *MDR* est franchement dégradé. Ces taux rassemblent les erreurs commises par l'étiquetage automatique d'un des deux systèmes de structuration audio et visuels (ou les deux). En résumé, lors de l'évaluation *restreinte*, le taux *FAR* est réduit, la *CDR* (*Correct Duration Rate*) augmente, mais le taux *MDR* ne s'améliore pas.

Par souci de comparaison, nous avons souhaité comparer notre méthode de recherche d'associations avec la méthode proposée dans [Khoury et al., 2010] dans le même contexte (voir la description de la méthode dans la section 1.4.5 du chapitre 1). Le tableau 4.8 résume les performances du système de structuration audio-visuel en utilisant la méthode d'association proposée dans [Khoury et al., 2010].

Comparé aux résultats de obtenues avec notre méthode d'association, le taux de perte *MDR* est amélioré ($-21,5\%$ en moyenne). Cette amélioration s'explique par le

Épisodes	*CDR + EDR + FAR*	*MDR*	*Pr*	*Ra*	F_m
S1	63.4 + 28.4 + 8.3	15.3	63.4	58.5	60.8
S2	63.4 + 13.2 + 10.6	14.9	76.2	72.6	74.3
S3	52.9 + 29.2 + 17.9	15.5	52.9	54.4	53.6
S4	59.8 + 23.0 + 17.2	19.9	59.8	57.9	58.9
S5	70.8 + 13.3 + 15.9	22.3	70.8	65.3	68.0
All	62.1 + 21.4 + 14.0	17.6	62.1	61.7	53.1

TABLE 4.8 – *Évaluation restreinte* en audio des résultats de la structuration audio-visuelle par appariement d'index selon la méthode [Khoury et al., 2010] sur les 5 épisodes de la base de données *TSDB*

fait que la méthode de fusion autorise l'association de plusieurs étiquettes audio à une étiquette visuelle. Par conséquent, les séquences très ambiguës de « brouhaha »sont systématiquement associes à des étiquettes visuelles. Par contre, la précision du système diminue significativement (-26% en moyenne). Cette diminution s'explique par le principe de maximisation de la co-occurrence locale. Par ce principe, certaines étiquettes visuelles se sont associées formant une seule personne.

Analyse des erreurs du système de structuration audio-visuelle

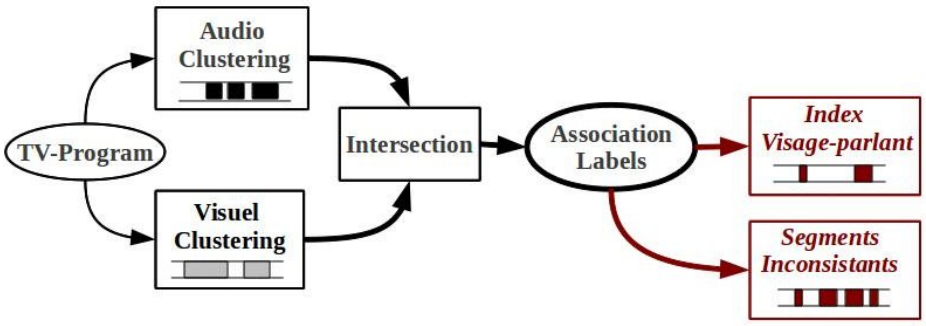


FIGURE 4.7 – Système de structuration audio-visuelle

Dans notre système de structuration, après intersection des index audio et visuel, seuls les segments où les étiquettes audio et visuelle s'associent sont sélectionnés comme étant visages parlants (segments consistants). Les autres segments sont considérés comme inconsistants ($A(x_A) \neq f_{assoc}(V(x_C))$ et annotés en tant que visages

non parlants. La figure 4.7 présente les segments obtenus par le système de structuration audio-visuel. Les segments inconsistants présentent deux cas :

- Cas A : pas de visages parlants dans le segment (le locuteur n'est pas la personne qui apparaît).
- Cas B : il y a effectivement un visage parlant dans le segment, mais non détecté en raison d'une erreur d'étiquetage dans le regroupement audio ou visuelle (ou les deux).

Les résultats sur la base de données *TSDB* montre que notre méthode arrive à détecter seulement 60% du temps totale de visages parlants de référence, mais avec une précision d'environ 90%. Nous avons souhaité analyser les sources d'erreurs qui engendrent une perte de segment de visages parlants dans notre système. Un segment de visage parlant n'est pas retrouvé pour deux raisons : erreur de détection ou erreur d'étiquetage automatique (séquences détectées, mais attribuées à la mauvaise personne). Chaque source d'erreur peut être due à l'échec du système de structuration basée sur l'audio, le visuel ou les deux à la fois. La décomposition du taux de segments non détecté par le système MDR en fonction des différentes causes est la suivante :

$$\begin{aligned} E_{Detection} &= E^{Audio}_{Detection} + E^{Visuel}_{Detection} - E^{Audio-visuel}_{Detection} \\ E_{Clustering} &= E^{Audio}_{Clustering} + E^{Visuel}_{Clustering} - E^{Audio-visuel}_{Clustering} \\ MDR &= E_{Detection} + E_{Clustering} = 100\% \end{aligned}$$

Le tableau 4.9 présente la composition du taux d'erreur MDR pour chaque épisode de la base de données *TSDB*. Les causes principales des erreurs diffèrent selon les épisodes : dans l'épisode $S1$, le MDR provient en majorité des erreurs de non-détection ; dans l'épisode $S5$ les origines des erreurs sont équilibrées, tandis que dans les épisodes $S2$,$S3$ et $S4$, le MDR est très majoritairement provoqué par des erreurs de *Clustering*, en audio ou en visuel. Les erreurs de regroupements s'expliquent par la qualité de l'audio dans le contexte de télévision qui rend difficile le regroupement des personnes. Les erreurs de regroupement visuel s'expliquent par le fait que certains costumes de personnes différentes sont similaires en terme de signature de couleurs. Les erreurs de détection visuelle correspondent aux séquences où le visage n'a pas été détecté. En particulier dans l'émission $S1$, beaucoup de segments de visages parlants ont été perdus à cause de la non détection des visages due à la présence de beaucoup de plans généraux de personnes.

Type d'erreur →	$E_{Detection}$				$E_{Clustering}$			
Épisodes ↓	*Audio*	*Visuel*	*deux*	*Total*	*Audio*	*Visuel*	*deux*	*Total*
S1	7.96	60.05	3.28	64.73	22.11	20.61	7.46	35.24
S2	14.62	14.20	4.37	24.45	27.08	58.19	9.72	75.55
S3	14.44	10.36	3.46	21.19	53.06	45.89	20.30	78.65
S4	13.60	11.50	4.05	21.05	16.62	67.25	4.93	78.94
S5	14.86	37.77	5.02	47.49	27.50	29.48	4.59	52.39
All	13.09	26.77	4.04	35.78	29.27	44.28	9.40	64.15

TABLE 4.9 – Composition du taux d'erreur *MDR* sur les 5 épisodes de la base de données *TSDB*

4.4 Conclusion

Dans ce chapitre, nous avons présenté une méthode de regroupement des personnes basée sur les informations audio et visuelle de manière indépendante et combinant les index résultants pour obtenir les séquences de visages parlants. Les résultats sur la base de données *TSDB* montrent que seulement 60% du temps total des visages parlants est détecté par notre méthode, mais avec une précision de l'ordre de 90%. Par conséquent, lorsque les deux systèmes de structuration de personne audio et visuel sont d'accord, la réponse a une forte probabilité d'être correcte.

L'analyse des erreurs a montré qu'une partie significative des visages parlants non détectées sont causés par des erreurs de *Clustering*. Dans ce qui suit, nous souhaitons détecter et corriger ces erreurs de *Clustering* par un processus automatique. Afin de détecter une erreur de regroupement, nous avons développé des indicateurs de présence de visages parlants dans les segments inconsistants. Dans le chapitre suivant, nous présentons une méthode de détection de présence de visages parlant basée sur la détection de l'activité visuelle de la parole.

Chapitre 5

Activité visuelle de la parole

Dans le chapitre précédent, nous avons présenté un *système initial* de structuration audio-visuelle. La réponse du système obtient une grande précision pour un taux de perte important. Ces pertes sont dues majoritairement à des erreurs de regroupement audio et/ou visuel. Notre objectif est de détecter ces erreurs de regroupement afin de les corriger automatiquement. Des indicateurs de présence de visages parlants peuvent être utilisés pour détecter des erreurs de regroupement. L'activité visuelle de la parole est un indicateur de présence de visage parlant.

Dans ce chapitre, une méthode de détection d'activité visuelle de la parole ainsi que son intégration dans le système d'indexation de visage parlant sont proposées.

5.1 Tour horizon

Dans la littérature, il existe plusieurs méthodes utilisant l'information visuelle de l'activité de la parole afin d'améliorer des systèmes de reconnaissance de la parole [Dupont and Luettin, 2000, Heckmann et al., 2001], systèmes de lecture sur les lèvres et audio/vidéo synchronie [Rúa et al., 2008]. La plupart de ces méthodes nécessitent un haut niveau de représentation des lèvres (contours des lèvres, ouverture, surface, largeur, ...). Dans [Dupont and Luettin, 2000], afin d'améliorer la reconnaissance de la parole, les auteurs proposent de combiner les paramètres acoustiques avec les paramètres visuels représentés par le contour des lèvres et l'histogramme des niveaux de gris de la région autour de la bouche. Pour le même objectif, dans [Heckmann et al., 2001], les paramètres visuels de la parole sont représentés par la hauteur extérieure et intérieure des lèvres ainsi que la surface des lèvres. Dans [Rúa et al., 2008], des coefficients *(DCT)* (*Discrete Cosine Transform Coefficients*) sont extraits de la zone des lèvres et combinés avec les coefficients *MFCCs* afin de mesurer la synchronie entre l'audio et le mouvement des lèvres.

Très peu d'auteurs se sont concentrés uniquement sur la détection d'une activité de la bouche afin de localiser le visage parlant. Dans [Everingham et al., 2006, Saenko et al., 2005] les auteurs utilisent une différence entre pixels de deux régions

consécutives de la bouche pour détecter une activité de la parole. Notre contribution à ces travaux est de développer une mesure de détection de l'activité de la bouche basée sur le désordre des directions de pixel de la zone de la bouche.

5.2 Système de détection de l'activité des lèvres

L'objectif est de détecter un mouvement des lèvres dans une séquence vidéo d'un visage. Dans notre contexte d'étude, en raison de la faible résolution des images vidéo, il est très difficile d'extraire la forme des lèvres avec une grande fiabilité. Nous avons choisi de représenter la région de la bouche comme un rectangle de pixels.

5.2.1 Description

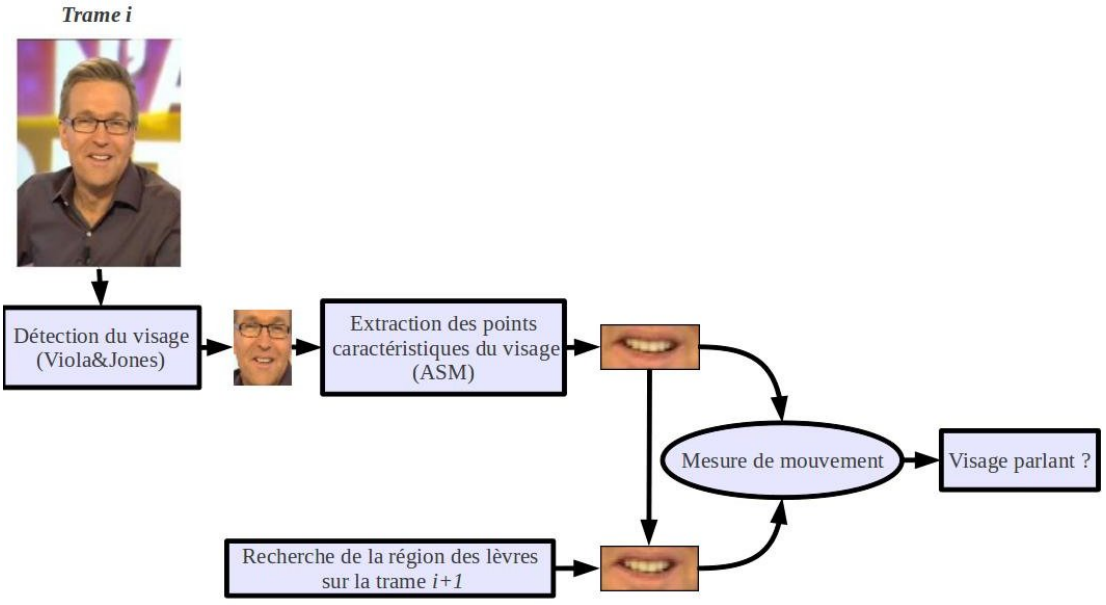


FIGURE 5.1 – Détail du système de détection de l'activité des lèvres

La figure 5.1 représente le détail des étapes de notre méthode pour la détection d'activité des lèvres afin de détecter la présence de visages parlants. À partir d'une image contenant un visage, le mouvement des lèvres est calculé de la manière suivante :

1. Un détecteur *Viola&Jones* [Viola and Jones, 2001] est appliqué pour localiser le visage.
2. La région de la bouche est localisée en utilisant un détecteur de caractéristiques du visage [Milborrow and Nicolls, 2008].
3. Sur la trame suivante, la région des lèvres alignée avec la trame précédente est recherchée.
4. Le mouvement est calculé entres les deux régions des lèvres alignées.

5.2.2 Extraction des caractéristiques du visage

La détection des points caractéristiques du visage consiste à localiser automatiquement les régions spécifiques du visage telles que les lèvres, le nez, les sourcils et les yeux. Pour extraire les caractéristiques d'un visage, nous avons utilisé le logiciel *Stasm* [Milborrow and Nicolls, 2008]. Ce système est basé sur des modèles actifs de forme (*Active Shape Model - ASMs*) décrit dans [Cootes et al., 1995]. D'abord, le visage est détecté en utilisant le détecteur *Viola&Jones*[Viola and Jones, 2001]. Ensuite, le modèle de visage appris est positionné sur le visage détecté et itérativement déformé jusqu'à ce qu'il corresponde au visage détecté.

FIGURE 5.2 – Exemples de détection de caractéristiques du visage - 68 points. Chaque œil est représenté par 5 points (coins à gauche et à droite, ouverture sur le dessus et le dessous des yeux et le centre de l'iris) et chaque sourcil est représenté par 6 points. Le contour des lèvres est décrit par 19 points (18 points pour les lèvres supérieures et inférieure et un point représentant le centre de la bouche). Le nez est représenté par 12 points et le contour du visage par 15 points

La figure 5.2 présente des exemples de détection des caractéristiques du visage à l'aide de l'outil *Stasm*[1]. Les caractéristiques du visage sont localisées avec une grande fiabilité dans des conditions d'éclairage standards, visage frontal et expression du visage classique (une évaluation du logiciel sur la base de données *BioID* est présentée dans l'annexe 8.5). Néanmoins, le système n'est pas suffisamment fiable avec les changements de pose tels que des visages regardant vers le haut ou vers le bas et des expressions telles qu'une bouche grande ouverte. Ceci s'explique par le fait que le modèle de forme est appris sur une base de données de visage de face, sans expression et donc ne modélise pas toutes les distorsions du visage.

FIGURE 5.3 – Exemples d'erreurs dans la détection de caractéristiques du visage

La figure 5.3 présente des exemples d'erreurs de détection des caractéristiques du visage avec la méthode *ASM*. Dans les cas **(a)** et **(c)**, les lèvres ne sont pas détectées, car le modèle de forme est appris sur des lèvres de face et fermées. Souvent lorsque la bouche est ouverte, le système détecte les narines ou la moustache comme étant des lèvres. Dans le cas **(b)**, les lèvres et le nez sont correctement localisés malgré la présence de l'occultation du doigt mais l'œil droit n'est pas correctement localisé à cause de la pose du visage et de rides sous l'œil dues à l'expression du visage.

5.2.3 Sélection de la région des lèvres

Afin de minimiser les variations dues à la localisation des caractéristiques du visage d'une trame à une autre, les régions de la bouche doivent être alignées. Comme illustré dans la figure 5.4, pour chaque trame, la région des lèvres est localisée en utilisant le détecteur de caractéristiques du visage. Dans la trame suivante, le rectangle autour des lèvres est élargi afin de rechercher la meilleure région alignée avec

1. http://www.milbo.users.sonic.net/stasm/

la précédente. Cet alignement est déterminé en sélectionnant la région qui minimise la moyenne de la différence au carré (*Mean Squared Difference - MSD*).

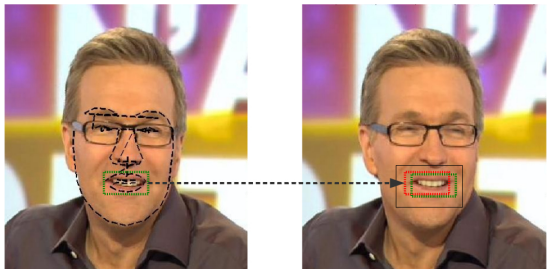


FIGURE 5.4 – Recherche de région autour des lèvres. Dans l'image à gauche, la région des lèvres (en vert) est détectée par l'outil *Stasm*. Ensuite, cette même région est représentée dans l'image à droite (trame suivante) en rouge. Cette région n'est pas alignée avec la bouche de la trame précédente. Le rectangle est élargi (en noir) afin de rechercher la région de la bouche alignée (en vert) avec celle de la trame précédente

5.2.4 Mesure de l'activité visuelle de la parole

Notre objectif est de mesurer l'activité de la bouche afin de localiser le visage qui correspond au locuteur. L'activité visuelle de la parole se caractérise par une déformation des lèvres que l'on peut assimiler à une dynamique de mouvement désordonné autour de la région des lèvres (par opposition à un mouvement des lèvres qui serait issu d'un mouvement général du visage). Afin de mesurer un mouvement correspondant à une activité de la parole, nous proposons de mesurer le degré de désordre des directions des pixels autour des lèvres obtenu par le flux optique [Ogale and Aloimonos, 2005].

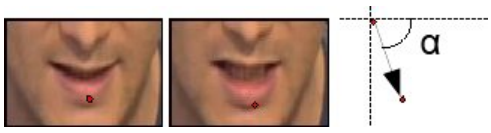


FIGURE 5.5 – Directions des mouvements des pixels estimés avec le flux optique

Le flux optique mesure la projection des objets en mouvement calculée à partir des variations d'intensité des pixels entre 2 images consécutives. Chaque pixel est associé à un *vecteur 2D* représentant la direction estimée de mouvement (voir la

figure 5.5). Dans nos expériences, nous avons découpé l'espace des directions des pixels en 24 axes.

Soit x_i vecteur de paramètres visuels extraits de la trame i d'une séquence vidéo contenant n trames $x_{1..n}$. Afin de mesurer le désordre des directions des pixels, l'entropie des angles de mouvements des pixels est calculée autour de la région des lèvres. Entre deux visages détectés dans deux trame consécutives x_i et x_{i+1}, l'entropie de la direction des pixels autour de la région des lèvres est calculée de la manière suivante :

$$\text{Entropy}(\text{x}_\text{i}, \text{x}_\text{i+1}) = -\sum_{j=1} P(\alpha_j) * log(P(\alpha_j)) \tag{5.1}$$

où $P(\alpha_j)$ est la probabilité qu'un pixel choisi au hasard dans la région des lèvres bouge avec un angle α_j. La figure 5.5 montre un exemple de direction du mouvement estimé par flux optique pour un pixel.

Pour chaque visage détecté dans la séquence $x_{1..n}$, la mesure du mouvement des lèvres est calculée de la manière suivante :

$$\text{Mv(x)} = \frac{1}{n-1} \sum_{j=1}^{n-1} Entropy(x_{j-1}, x_j) \tag{5.2}$$

5.3 Évaluation

5.3.1 Expérience

Afin de comparer la méthode proposée, nous avons également expérimenté la méthode basée sur le *MSD* proposée dans [Everingham et al., 2006] pour détecter une activité visuelle de la parole dans un contexte de télévision. La différence de la moyenne quadratique entre les pixels est calculée entre deux régions consécutives de la bouche après alignement. Un seuil fixe est appliqué pour déterminer si les lèvres bougent.

5.3.2 Protocole

Pour évaluer la capacité du détecteur de mouvement des lèvres à séparer les visages parlants des visages non parlants, nous nous sélectionné, à partir de la base *TSDB*, les *plans de visages parlants* où le visage prend la parole sur toute la durée du plan et les *plans visages non-parlants* où le visage ne parle pas pendant toute

la durée du plan. Afin de comparer nos résultats avec la méthode proposée dans [Everingham et al., 2006], nous avons sélectionné les plans où le visage est visible (pose frontale ou quart droit/gauche).

Au total, la base sur laquelle nous avons expérimenté notre méthode est composée de 581 plans de visages parlants (durée moyenne des plans est 3.45s) et 667 plans de visages non-parlants (durée moyenne 2.17s). Dans le contexte de *Talk shows*, les plans de visages parlants sont relativement plus longs que les plans de visages non-parlants (voir le tableau 5.6).

Durée	$< 1s$	de 1 à 4s	$> 4s$	*Total*
# *Visages parlants*	6	425	150	581
# *Visages non parlants*	7	613	47	667

FIGURE 5.6 – Taille des séquences de visages parlants et non parlants

5.3.3 Résultats des alignements

Dans des contenus de télévision, les visages peuvent se déplacer très rapidement. Afin de calculer une variation de la bouche sans tenir compte des mouvements du visage, il est nécessaire d'aligner les régions de la bouche. La figure 5.7 montre trois exemples d'alignement des régions des lèvres provenant de deux trames consécutives. Pour chaque exemple, la première image est le rectangle de la région des lèvres obtenu dans une trame donnée par extraction des paramètres du visage. La seconde image est le même rectangle de la région des lèvres dans la trame suivante sans alignement. La troisième image est le rectangle de la région des lèvres dans la trame suivante après alignement par minimisation l'erreur quadratique moyenne (*mean squared difference*).

5.3.4 Résultats et discussion

Résultats

La figure 5.8 résume les performances du système de détection d'activité des lèvres pour la classification des visages parlants en utilisant les méthodes basée sur le *MSD* et le désordre des directions des pixels (Mv). Les performances obtenues en mesurant le désordre des directions de mouvement sont nettement meilleures, selon le taux d'égale erreur *EER* (voir le chapitre 3) que celles obtenues par une moyenne

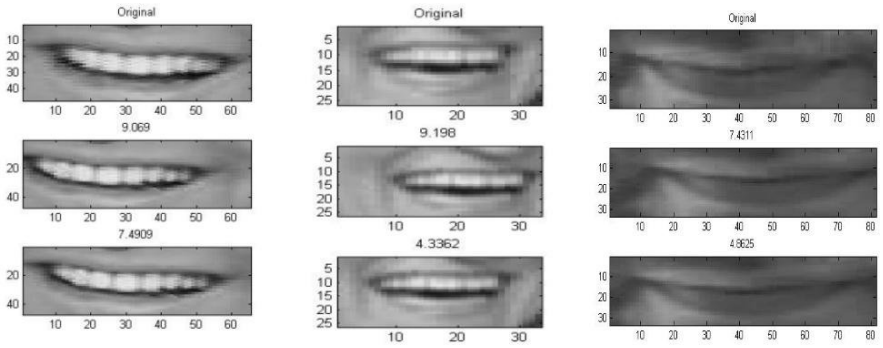
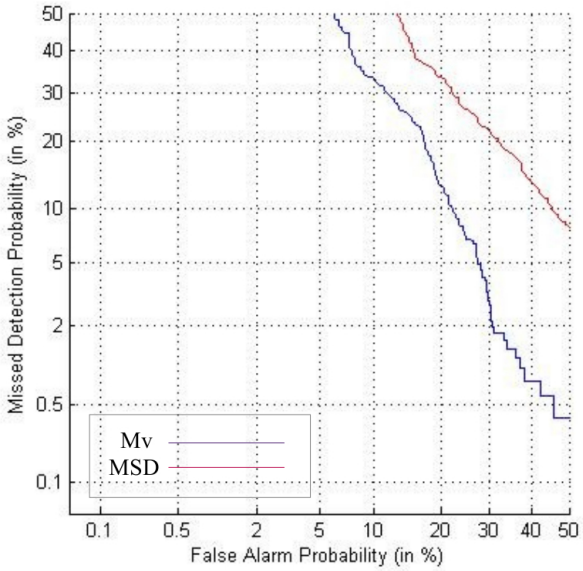


FIGURE 5.7 – Trois exemples d'alignements des régions des lèvres. L'image originale représente la région des lèvres obtenue par *Stasm*. La seconde image représente la même région des lèvres dans la trame suivante (sans alignement). La troisième image représente la région des lèvres dans la trame suivante après alignement



	EER
Mv	17.52%
MSD	25.64%

FIGURE 5.8 – Courbe DET - Performance du système de détection d'activité de la bouche pour la classification de visages parlants

des différences au carré des pixels.

Les figures 5.9 et 5.10 présentent des exemples de bonne classification de séquences de visages. La méthode de détection de mouvement est fiable dans le cas de visage non parlant avec l'apparition d'un doigt dans la région des lèvres (cas (a) dans la figure 5.9)

FIGURE 5.9 – Exemples de bonnes classifications des segments non parole. Cas (a) $Mv = 0.65$. Cas (b) $Mv = 0.79$. Cas (c) $Mv = 0.63$

FIGURE 5.10 – Exemple de bonnes détections d'activité de la parole. Cas (a) $Mv = 1.73$. Cas (b) $Mv = 1.11$. Cas (c) $Mv = 1.63$

Notre méthode d'évaluation fait l'hypothèse qu'une activité de bouche signifie activité de la parole. Après une analyse des erreurs, nous avons remarqué que le

système n'est pas fiable dans le cas de grimaces, de gestes brusques de bouche ou de fous rires. La figure 5.11 montre des exemples d'erreurs de séquences de visages non parlants classés par notre méthode en visages parlants. La valeur du mouvement est nettement plus élevée que le seuil correspondant, ce qui signifie que la personne est classée comme ayant une activité de bouche par notre méthode, hors celle-ci ne parle pas. Il est important de noter également que la détection des lèvres n'est pas fiable pour les bouches grandes ouvertes.

FIGURE 5.11 – Exemples d'erreurs de détection d'activité de la parole. Cas (a) $Mv = 1.62$. Cas (b) $Mv = 1.59$

Conclusion

L'objectif est d'utiliser l'information visuelle pour détecter la présence de visage parlant. Dans cette section, nous avons proposé une nouvelle méthode de détection de l'activité de la parole basée sur le désordre de direction des pixels autour de la région des lèvres. Une amélioration significative est observée comparé à la différence entre les pixels. Les résultats montrent également que dans les données réelles, l'activité de la bouche indique dans la plupart des cas, une activité de la parole. Une façon d'améliorer la classification des visages parlants est d'étudier des méthodes pour améliorer la fiabilité de la mesure proposée dans les cas de rires ou grimace. Premièrement, en utilisant un détecteur de caractéristique de visage qui prend en compte tous les types de distorsions de la bouche afin de corriger des erreurs de détection de la région des lèvres qui faussent le calcul du degré de désordre des directions de pixels. En plus, un mouvement des lèvres peut également correspondre à une séquence de rire (considéré comme non parole). L'information audio peut être utilisée en plus du mouvement des lèvres afin de déterminer si le mouvement détecté

correspond bien à une activité visuelle de la parole ou à un rire.

5.4 Mesure de l'activité visuelle de la parole dans le système de structuration

À l'intersection des index audio et visuels, deux types de segments sont obtenus : *segments consistants* dans lesquels les étiquettes audio et visuelle sont associées (annotés en visage-parlants) ou *segments inconsistants* dans lesquels les étiquettes audio et visuelle ne s'associent pas. Ces segments inconsistants peuvent être des segments ne contenant pas de visage-parlant (le locuteur n'est pas celui qui est visible) ou des segments de visages parlant mal étiquetés en raison d'une erreur d'étiquetage dans la modalité audio ou visuelle. D'après l'analyse des erreurs dans les segments inconsistants (voir la section 4.3.3 du chapitre 4), la part des erreurs due au regroupement est significative. Ceci s'explique par le fait que dans un contexte de télévision, il est très difficile de regrouper les segments de la même personne avec beaucoup de fiabilité.

Dans cette partie, l'idée est de remettre en question, dans les segments inconsistants, les modalités audio et visuelle afin de détecter une erreur d'étiquetage et de la corriger. En effet, une erreur d'étiquetage automatique conduisant à un segment inconsistant entraîne la non-détection d'un visage parlant. Afin de déterminer une erreur d'étiquetage, nous tentons de suspecter la présence d'un visage parlant en utilisant le détecteur d'activité des lèvres présenté précédemment. Ensuite, l'erreur suspectée est corrigée selon un processus de modification de manière à ce que les étiquettes audio et visuelle s'associent. La figure 5.12 présente le processus de récupération de segments de visages parlants annotés comme inconsistants en corrigeant les erreurs d'étiquetage.

5.4.1 Système de structuration audio-visuel modifié

Soit x un segment audio-visuel annoté inconsistant par le *système initial*. Pour intégrer la mesure d'activité des lèvres $Mv(x)$ dans le système d'indexation, la taille du segment x est utilisée comme une mesure de confiance. Cette mesure de confiance suppose que plus le segment est long, plus la confiance au détecteur du mouvement

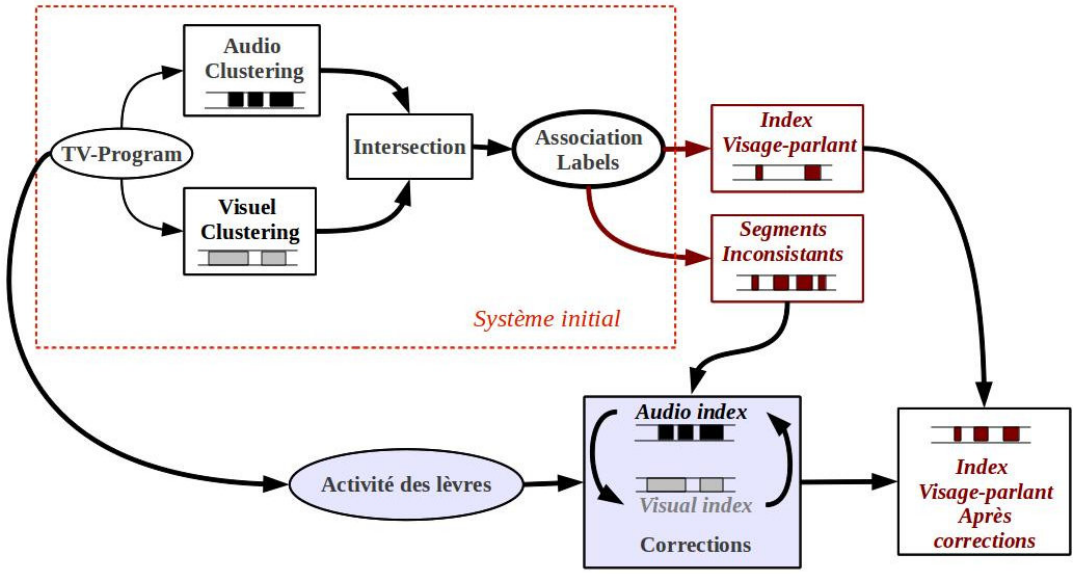


FIGURE 5.12 – Système de structuration audio-visuel modifié. Processus de correction des segments inconsistants obtenus par erreur d'étiquetage automatique par le *système initial*

des lèvres est grande. La décision $Lip(x)$ est obtenue comme suite :

$$Lip(x) = \begin{cases} 1 & \text{si } Mv(x) > \theta_1 \text{ and } length(x) > \theta_2 \\ 0 & \text{sinon} \end{cases} \tag{5.3}$$

où les seuils θ_1, θ_2 sont optimisés dans l'ensemble de développement.

Pour chaque segment *inconsistant* x dont on a extrait les vecteurs de paramètres acoustiques x_A et le vecteur de paramètres du costume x_C, les étiquettes audio $A(x_A)$ et visuelle $V(x_C)$ ne s'associent pas. Si un mouvement des lèvres est détecté, une modification systématique de l'étiquette de la modalité considérée comme la moins fiable est effectuée afin de la rendre compatible avec l'étiquette de la modalité considérée comme la plus fiable. Le processus de modification systématique est formalisé dans l'algorithme 1.

Selon la modalité considérée la moins fiable, le processus de modification $modification(A(x_A), V($ est la suivante :

– *Modification audio* $\rightarrow$ modifier l'étiquette audio de manière à la rendre compatible à celle du visuelle : $A(x_A) = f_{association}(V(x_C))$.

Algorithm 1 Systematic modification

for chaque segment $x(x_A, x_C)$ où $f_{association}(V(x_C)) \neq A(x_A)$ **do**

  if *présence de visage parlant* **then**

    $modification(A(x_A), V(x_C))$

  end if

end for

– *Modification visuelle* → modifier l'étiquette visuelle de manière à la rendre compatible à celle de l'audio : $V(x_C) = f^{-1}_{association}(A(x_A))$.

5.4.2 Résultats et discussion

Résultats

Le tableau 8.6 résume les performances du système de structuration des visages parlants en utilisant la mesure de mouvement des lèvres afin de corriger des erreurs d'étiquetage automatique. Pour chaque schéma de modification, nous présentons les taux de *précision*, *rappel* et *F-mesure* obtenus pour chaque émission ainsi que la moyenne des émissions.

Épisodes	Système initial			Modification visuelle			Modification audio		
	Pr	Ra	F_m	Pr	Ra	F_m	Pr	Ra	F_m
S1	94.8	61.4	74.5	92.9	74.3	82.5	77.4	62.4	69.1
S2	93.5	65.7	77.2	90.7	73.7	81.3	82.6	66.8	73.9
S3	78.9	55.4	65.1	72.2	59.8	65.5	66.1	57.1	61.3
S4	95.3	63.7	76.3	92.9	76.1	83.6	77.6	66.9	71.8
S5	92.0	56.2	69.8	88.6	61.2	72.4	84.4	58.5	69.1
All	90.3	60.8	72.7	86.8	69.2	76.9	77.2	62.1	68.8

TABLE 5.1 – Performances du système de structuration basé sur la mesure d'activité des lèvres dans la base de données *TVSDB*

Dans le schéma de modification visuelle, la précision diminue comparé au *système initial* (−3.1% en moyenne sur les 5 épisodes) alors que le rappel augmente (+8.4% en moyenne). L'amélioration du rappel s'explique par le fait que le processus de modification a permis de récupérer des segments de visages parlants perdus pour cause

d'erreur de *clustering* visuel. La diminution de la précision s'explique par l'introduction d'erreurs lors de la modification systématique des étiquettes visuelles. Ces erreurs peuvent avoir deux origines : modification erronée (c.à.d seulement l'étiquette audio est erronée) ou modification insuffisante (c.à.d la modification s'effectue seulement sur l'étiquette visuelle erronée alors que l'étiquette audio et également erronée).

Dans le schéma de modification audio, la précision diminue significativement comparé au *système initial* (-13.1% en moyenne) alors que le rappel augmente très peu ($+1,3\%$ en moyenne). De la même manière que pour le schéma de modification visuelle, la diminution de la précision s'explique par l'introduction d'erreurs causées par une modification erronée ou insuffisante.

Discussion

Le processus de modification des étiquettes visuelles améliore sensiblement les résultats du *système initial*, alors que le schéma de modification des étiquettes audio échoue toujours. Bien que les systèmes d'indexation basés sur l'audio et le visuel puissent être considérés comme tout aussi fiables l'un que l'autre en raison de leurs performances de base (voir les résultats du *système initial* dans le chapitre 4), ils jouent un rôle très asymétrique dans les schémas de modification. En effet, comme nous l'avons présenté dans le chapitre 2, sur ce type de contenu, le visage associé à une voix est visible plus de 60% de son temps de parole, alors que le temps de parole d'un visage n'est que de 35% de la durée totale de l'apparence de ce visage dans l'épisode. Cette différence de probabilités a priori explique pourquoi dans le cas où l'étiquette audio et visuelle ne s'associent pas, la modification du label visuel est plus efficace que la modification d'étiquette audio.

5.4.3 Conclusion

Dans cette partie du chapitre, nous avons proposé une méthode d'intégration du détecteur de l'activité visuelle de la parole dans le *système initial* dans le but de détecter automatiquement des erreurs de regroupement afin de les corriger. Le processus de correction est basé sur une modification systématique d'une modalité considéré *a priori* la moins fiable par l'utilisateur. Le schéma de correction basé sur la modification de l'étiquette visuelle a permis de récupérer plusieurs segments de visages-parlants perdus dans le *système initial*.

Conclusion

Notre objectif est d'annoter automatiquement les séquences de visages parlants dans un contexte de *Talk shows*. Dans ce contexte, déterminer qui parle à quel moment est difficile en raison des nombreuses ambiguïtés que présente ce contexte. Dans cette première partie du rapport, nous avons proposé une méthode de structuration des personnes qui utilise l'information audio et visuelle de manière indépendante, puis qui combine les deux index obtenus pour construire un index de visages parlants. La méthode proposée basée sur l'intersection des index audio et visuel parvient à détecter avec une grande précision les séquences de visages parlants, au prix d'un taux de rejet élevé.

Dans cette partie du rapport, nous avons également proposé d'intégrer au système de structuration audio-visuelle un détecteur d'activité visuelle de la parole basé sur la mesure du désordre des directions de pixels dans la région des lèvres. La prise en compte de l'activité labiale permet de corriger des erreurs du *Clustering* visuel et permet ainsi de retrouver plusieurs séquences de visages parlants perdus dans le système initial. La méthode proposée améliore significativement le rappel en dégradant légèrement la précision du système d'indexation.

Perspectives Plusieurs améliorations peuvent être apportées. D'abord, concernant le système d'indexation basé sur l'information visuelle, même si le regroupement basé sur la signature des couleurs du costume obtient des résultats intéressants, la méthode est très sensible au fait que les gens peuvent porter des costumes similaires pendant la même émission. Une façon d'améliorer le système est d'introduire des informations sur la forme des costumes, ou d'inclure un regroupement basé sur le visage.

Concernant le système d'indexation basé sur l'information audio, à notre connaissance, rares sont les études menées sur des contenus aussi interactifs. La méthode mérite des améliorations afin de tenir compte des caractéristiques acoustiques de type : rires, chevauchement de la parole.

Compte tenu de la bonne précision du système d'indexation audio-visuel modifié, dans la seconde partie du rapport, nous voulons utiliser les séquences de visages parlants détectés pour apprendre des modèles non supervisés afin de récupérer les segments manqués pour cause d'erreur de regroupement. Ce modèle peut être un modèle acoustique, modèle de visage, ou même un modèle combiné voix-visage.

Deuxième partie

Identification audio-visuelle des personnes

Introduction

Notre objectif est l'indexation des contenus de type *Talk shows* par personne dans le but d'offrir à un utilisateur la possibilité de naviguer dans un contenu audio-visuel par personne et de retrouver des interventions spécifiques. Dans cette partie, nous nous intéressons à l'indexation des personnes basées sur des systèmes de vérification de l'identité.

Un système de vérification de l'identité nécessite une modélisation d'identité d'une personne à partir d'une collection d'exemples. Selon les applications, la collection de données peut être fournie et annotée manuellement (apprentissage supervisé) ou issue d'un processus automatique (apprentissage non supervisé). Dans notre étude, nous nous sommes intéressés à deux applications des modèles de vérification de l'identité : application dans la structuration et dans l'identification.

Application dans la structuration

Dans la première partie du rapport, nous avons présenté des méthodes de structuration d'émissions de type *Talk shows* par personne. Ces méthodes sont basées sur la détection automatique des séquences d'interventions des personnes (tours de parole et plans d'apparence) et le regroupement des interventions de la même personne. Les résultats obtenus montrent que nos méthodes possèdent une grande précision pour un faible rappel. Ce taux s'explique en grande partie par des erreurs de regroupement d'une des modalités. À partir des interventions détectées automatiquement, nous souhaitons apprendre des modèles afin d'identifier les personnes détectées automatiquement et dont le processus de regroupement est erroné. Ces modèles de personnes sont donc appris à partir d'exemples issus d'un processus automatique.

Application dans l'identification

Une fois qu'un épisode est structuré par personnes, une collection d'interventions audio, visuelle et audio-visuelle est obtenue pour chaque personne présente dans l'épisode. Le dictionnaire des identités des personnes étant ouvert, on ne pourra pas identifier toutes les personnes. Par contre, il est possible de vérifier l'identité de certaines interventions (présentateur, invités, personnalités populaires, etc) à partir du moment où l'on possède des échantillons annoté manuellement de la personne.

Dans ce cas, il est nécessaire d'apprendre des modèles audio, visuels et audio-visuels de personnes de manière supervisée.

Organisation

La deuxième partie du rapport est organisée de la manière suivante :

D'abord, les modèles de vérification de l'identité utilisés dans nos expériences sont présenté dans le chapitre 6. Le chapitre 7 est dédié à l'application des modèles de vérification de l'identité dans le système de structuration. Cette application est basée sur un apprentissage non supervisé des modèles des personnes à partir des *Clusters* obtenus par le *Système initial*. Dans le chapitre 8, nous présentons l'application des modèles de vérification de l'identité dans les systèmes biométriques pour l'identification des personnes. Cette application est basée sur un apprentissage supervisé des modèles des personnes à partir de collections annotées manuellement.

Chapitre 6

Modèles de vérification de l'identité

Dans un processus de vérification de l'identité, deux étapes se distinguent : une étape d'apprentissage (consiste à apprendre un modèle à partir d'échantillons) et une étape de vérification de l'identité d'un exemple test (qui consiste à comparer le test au modèle appris). L'étape d'apprentissage consiste à modéliser une personne à partir d'une collection de données lui appartenant. L'étape de vérification de l'identité consiste à vérifier si un échantillon donnée correspond bien à la personne que l'on souhaite identifier. Pour cela, l'échantillon est comparé au modèle de la personne obtenu dans l'étape d'apprentissage. On appelle *accès client* lorsque la personne possède réellement l'identité qu'elle prétend avoir et *accès imposteur* dans le cas contraire. Dans ce chapitre, nous décrivons les différents modèles de vérification de l'identité audio-visuelle que nous avons utilisés dans nos expériences.

6.1 Présentation des systèmes de vérification de l'identité

6.1.1 Système de vérification du locuteur

Chaque individu possède une signature vocale propre. Cette signature dépend des caractéristiques anatomiques et comportementales. La modélisation d'un locuteur consiste à apprendre une empreinte vocale qui permet de le différencier des autres locuteurs. L'une des grandes difficultés de la modélisation d'un locuteur est la variabilité de la voix aux conditions physiques (maladie, croissance) et environnementales (bruit de fond, type de capteur, etc). Une autre difficulté est la variabilité de la voix selon le scénario d'enregistrement (rire, colère, etc).

Dans nos expériences, nous avons utilisé un système de vérification de l'identité du locuteur développé par *OrangeLabs*. Cet outil est basé sur une modélisation par mixture de gaussiennes *GMMs* avec un modèle du monde (*UBM*) [Reynolds et al., 2000]. La figure 6.1 présente les étapes de modélisation du locuteur en *GMMs*.

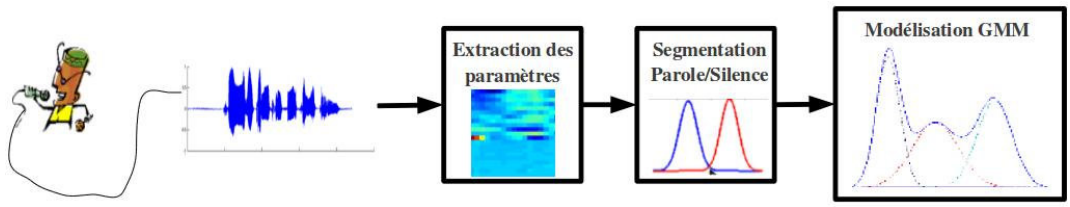


FIGURE 6.1 – Détail de la modélisation du locuteur en modèle de mixture de gaussiennes *GMMs*

Modélisation du locuteur

Extraction des paramètres

Premièrement, les coefficients *MFCCs* sont extraits toutes les $16ms$ sur des fenêtres glissantes de $32ms$. Chaque trame est associée à un vecteur de paramètres composé de l'énergie et des 13 premiers *MFCC* auxquels on ajoute les dérivées premières et secondes. Le vecteur paramètres extraits est de dimension $dim = 42$.

Segmentation Parole/Silence

Afin de ne conserver que les zones de parole, les trames de silence et de parole son séparées. Un bref état de l'art des méthodes de segmentation est présenté dans la partie 1.2.2 du chapitre 1. La méthode de segmentation en silence/parole utilisée est basée sur une modélisation multi-gaussienne bi-classes des coefficients *MFCCs*.

Modèle du monde (*UBM*)

En raison de la grande variabilité de la voix, les échantillons qui servent à la modélisation d'un locuteur ne sont pas suffisant pour apprendre un modèle qui permette de le différencier des autres locuteurs. Un modèle du monde (*Universal Background Model*) est un modèle *GMM* global entraîné sur une grande base de données de parole de plusieurs personnes différentes de manière à couvrir le plus de variabilité dans la voix. Un mélange de gaussiennes est une somme pondérée de n densités gaussiennes. Les paramètres du modèles *UBM* sont appris en maximisant la vraisemblance selon l'algorithme *EM* [Dempster et al., 1977]. Dans nos expériences, la dimension des *GMM* est de $n = 256$.

Modèle du locuteur

Le modèle *UBM* couvre un maximum de variabilité. Il est possible d'adapter le modèle UBM qui couvre un maximum de variabilité à chaque locuteur à partir de ses données d'apprentissage. Cette adaptation est obtenue l'approche *MAP* (*Maximum A Posteriori*) [Reynolds et al., 2000].

Vérification du locuteur

Au moment de vérifier si le locuteur d'une séquence de test x est la personne λ, les paramètres extraits x_A de la séquence de test sont comparé au modèle du locuteur λ et au modèle du monde (UBM). Le score de vérification de l'identité noté S_A est le rapport entre la vraisemblances du test au modèle du locuteur λ et la vraisemblance du test au modèle du monde UBM. Le score $S_A(x_A|\lambda)$ se présente comme suit :

$$S_A(x_A|\lambda) = \frac{1}{n} \log \frac{p(x_A|\lambda)}{p(x_A|UBM)} \tag{6.1}$$

avec n la longueur de la séquence x. Le locuteur de la séquence de test est identifié comme étant la personne λ si le score S_A est supérieur à un seuil fixe.

6.1.2 Système de vérification de l'identité visuelle

Les informations visuelles que nous avons étudiées au cours de nos expériences sont le visage et le costume.

Système de vérification du visage

Nous nous sommes basés sur le système développé dans l'équipe *TSI* de *Telecom ParisTech* par *Hervé Bredin* qui a été intégré dans l'outil de référence de l'identité des visages parlants présenté dans [Bredin et al., 2006]. Ce système est décrit en détails dans [Bredin, 2007]. L'approche utilisée pour la vérification de l'identité du visage est basée sur les *EigenFaces* [Turk and Pentland, 1991]. Ce modèle utilise la redondance de l'information apportée par la vidéo de manière à extraire les paramètres du visage dans chaque trame.

Apprentissage du modèle de visage

Premièrement, un détecteur de *Viola&Jones* [Viola and Jones, 2001] est utilisé afin de détecter le visage dans chaque trame de la séquence. Ensuite chaque visage

détecté est normalisé : alignement des yeux horizontalement (méthode de détection des yeux est décrite dans [Fasel et al., 2005]), redimensionnement et suppression des pixels de l'arrière plan. Chaque visage détecté est projeté sur l'espace des visages appris par *ACP* à partir d'une grande base de données de visage suivant la méthode *EigenFaces*. La figure 6.2 montre un exemple de projection d'un vecteur de paramètres (visage candidat) dans l'espace des visages. Pour une séquence de visages, le vecteur de paramètres est la concaténation des visages projetés. Afin de bien modéliser la séquence, les N meilleurs visages sont sélectionnés selon le critère *DFFS* (*Distance From Face Space*). On notera $x_V^\lambda = \{x_1^\lambda, .., x_N^\lambda\}$ l'ensemble des vecteurs de paramètres du visage x_i^λ extraits de la séquence d'apprentissage de la personne λ.

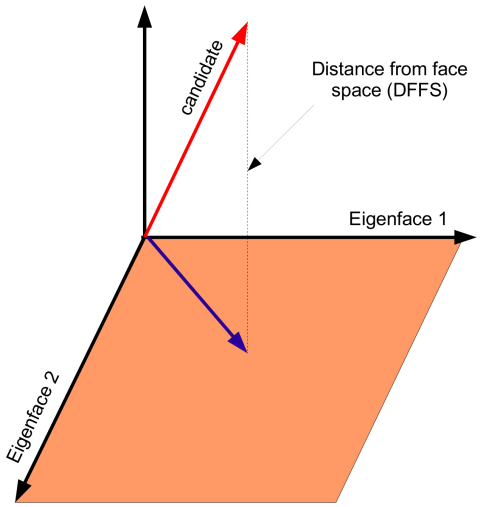


FIGURE 6.2 – Projection d'un visage candidat dans l'espace de visage

Vérification du visage

Au moment d'une vérification de l'identité d'une séquence test x, les paramètres du visage sont extraits pour chaque visage de la séquence de test de la même manière que lors de l'apprentissage du modèle. Soit $x_V = \{x_1, .., x_M\}$ l'ensemble de vecteurs des paramètres extraits de la séquence test. Le score de vérification de l'identité du visage (noté S_V) est obtenu en mesurant la distance *Mahalanobis* entre chaque vecteur de paramètres de test avec chaque vecteurs de paramètres de la séquence d'apprentissage. La distance *Mahalanobis* entre les deux vecteurs x_i et x_j^λ est calculé

de la manière suivante :

$$Mahalanobis(x_i, x_j^\lambda) = \sqrt{(x_i - x_j^\lambda)' \Sigma_\lambda^{-1} (x_i - x_j^\lambda)} \tag{6.2}$$

avec Σ_λ la matrice de covariance de vecteur $x_V^\lambda = \{x_1^\lambda, .., x_N^\lambda\}$. Une fois que les $M \times N$ distances sont calculées, le score de vérification de l'identité du visage $S_V(x_V|\lambda)$ est obtenu en moyennant les distances calculées.

Système de vérification du costume

Le costume ne constitue pas une information très robuste pour l'identification des personnes. Ceci dit, en l'absence de notion d'imposteur (personne qui tente de prendre l'identité de quelqu'un d'autre), il est possible d'utiliser le costume comme information de discriminante à condition que les personnes ne changent pas de vêtement entre la séquence d'apprentissage et de test.

Modèle de costume

Soit x une séquence audio-visuelle d'apprentissage de la personne de la personne λ contenant L trames. Le costume est détecté dans chaque trame de la séquence d'apprentissage selon le principe présenté dans la sous-section 4.2.1 du chapitre 4. Chaque costume détecté est représenté par l'histogramme des couleurs (codage RGB). Le modèle de costume est la concaténation des histogrammes de couleurs des costumes détectés noté $x_C^\lambda\{x_1^\lambda, .., x_L^\lambda\}$.

Vérification par le costume

Au moment d'une vérification de l'identité d'une séquence test x de taille K, les paramètres du costume sont extraits pour chaque costume de la séquence de test de la même manière que lors de l'apprentissage du modèle. Soit $x_C = \{x_1, .., x_K\}$ l'ensemble de vecteurs des paramètres extraits de la séquence test. Une distance *Euclidienne* est calculée entre chaque vecteur de paramètres de test et chaque vecteur de paramètres de la séquence d'apprentissage. Le score de vérification de l'identité par le costume $S_C(x_C|\lambda)$ est la moyenne des distances calculées

$$S_C(x_C|\lambda) = \frac{1}{K \times L} \sum_{i=1}^{K} \sum_{j=1}^{L} ||x_i - x_j^\lambda|| \tag{6.3}$$

6.1.3 Fusion des scores

La fusion consiste à combiner les scores de manière à obtenir une performance globale meilleure que celle de chaque système pris tout seul. Un bref état de l'art des techniques de fusion pour l'identification des personnes est présenté dans la sous section 1.4.2 du chapitre 1. Plusieurs travaux ont démontré que la fusion des scores par combinaison linaire permet d'obtenir de très bons résultats. Ces méthodes de fusion sont très simples à mettre en œuvre, mais nécessitent une phase de normalisation des scores afin de les représenter sur la même échelle.

Normalisation des scores

Les scores obtenus à partir des différents systèmes de vérification de l'identité ne sont pas représentés dans la même échelle. En effet, les scores du modèle de vérification de l'identité basés sur le visage et le costume représentent des distances (*Mahalanobis*, *euclidienne*) alors que les scores du modèle de vérification du locuteur représentent un rapport entre deux vraisemblances. La fonction de normalisation la plus utilisée est la *Z-Norm*. Soit x_i un vecteur de paramètres extrait dans la modalité i et $\tilde{S}_i(x|\lambda)$ son score de vérification de l'identité. La fonction de normalisation *Z-Norm* est calculée de la manière suivante :

$$\tilde{S}_i(x_i|\lambda) = f(S_i(x_i|\lambda), \mu^{imp}, \sigma^{imp}|\lambda) = \frac{(S_i(x_i|\lambda) - \mu^{imp})}{\sigma^{imp}} \tag{6.4}$$

avec μ^{imp} et σ^{imp} la moyenne et écart type des scores *Imposteurs* (appris à partir d'un ensemble de développement).

Dans [Li et al., 2008], les auteurs démontrent que la fonction de normalisation la plus robuste est la *Tanh Norm*. Cette méthode exploite la propriété de la fonction tangente hyperbolique pour replacer les scores dans le même espace de représentation. La *Tanh Norm* normalisation se présente comme suit :

$$\begin{aligned} \tilde{S}_i(x_i|\lambda) = f(S_i(x_i|\lambda), \mu^c, \sigma^c|\lambda) &= 0.5 + 0.5 * tanh(0.01 \times Z) \\ avec \quad \mathrm{Z} &= \frac{(S_i(x_i|\lambda) - \mu^c)}{\sigma^c} \end{aligned} \tag{6.5}$$

avec μ^c et σ^c la moyenne et écart type des scores *Clients* (appris à partir d'un ensemble de développement). L'intérêt de centrer et réduire les scores (Z dans l'équation 6.5) avant de les normaliser est de ramener la distribution des scores *Clients* aux alentours 0 avec une variance de 1. Les valeurs des scores *Imposteurs* vont s'étaler. La fonction de normalisation *tanh* permet de représenter les scores dans l'intervalle $[0, 1]$.

Méthode de fusion

Au final, différents scores sont obtenus par différents systèmes de vérification de l'identité de la personne. Afin de vérifier si une séquence audio-visuelle de test x correspond à la personne λ, les vecteurs de paramètres extraits dans chaque modalité x_i sont comparés au modèle de la personne λ dans cette modalité. Après normalisation des scores obtenus par les systèmes de vérification du visage et du locuteur, le score final $S_{final}(x|\lambda)$ est obtenus par combinaison linaire de la manière suivante :

$$\begin{cases} S_{final}(x|\lambda) = \sum_{i=1}^{n} w_i \tilde{S}_i(x_i|\lambda) \\ avec \quad \tilde{S}_i(x_i|\lambda) = f_i(S_i(x_i|\lambda), \mu_i^c, \sigma_i^c|\lambda) \end{cases} \tag{6.6}$$

avec les poids $\sum_{i=1}^{n} w_i = 1$ et f_i définie dans l'équation 6.5. L'estimation des poids est effectuée par optimisation des performances dans l'ensemble de développement. En résumé, les paramètres de fusion sont : les fonctions normalisation des scores f_i ainsi que leurs poids de fusion w_i dans chaque modalité ($i \in 1, .., n$).

6.2 Introduction de mesures de qualité dans la fusion

Les systèmes de vérification de l'identité sont très sensibles à la qualité du signal. L'intérêt majeur d'introduire des mesures de qualité est de pouvoir adapter la confiance accordé aux différentes modalités. Ainsi, accorder plus d'importance au système de reconnaissance basé sur l'image dans le cas où la séquence de parole est bruitée ou plus d'importance au système de de reconnaissance du locuteur lorsque le visage n'est pas.

Dans cette partie du chapitre, nous proposons une méthode d'introduction de mesures de qualité dans la fusion pour la vérification de l'identité. D'abord, un bref état de l'art des systèmes biométriques basés sur des mesures de qualité est présenté. Ensuite, nous exposons notre contribution dans la fusion basée sur les mesures de qualité.

6.2.1 Tour d'horizon

Dans la littérature, plusieurs travaux ont démontré l'apport des mesures de qualité dans les systèmes biométriques.

Dans [Fierrez-Aguilar et al., 2005], les auteurs proposent une méthode de fusion des modalités empreinte digitale et signatures basée sur des mesures de qualité subjectives. Un expert humain attribue des mesures de qualité subjectives sur l'empreinte digitale. Cette mesure de qualité est concaténée aux scores de vérification de l'identité de la signature et l'empreinte digitale afin de former un vecteur d'entrée à une méthode de classification (SVM). Les résultats montrent que le système de vérification de l'identité est plus robuste à la qualité des deux modalités.

Dans [Richiardi et al., 2006], les auteurs proposent de vérifier l'identité des locuteurs en introduisant plusieurs mesures de confiances calculées sur la qualité du signal audio ainsi que les distributions *Clients/Imposteurs*. Ces mesures de confiances sont utilisées pour apprendre un modèle graphique pour prédire une mesure de confiance globale qui va déterminer si l'on peut croire au score de vérification de l'identité locuteur. Les résultats démontrent une amélioration des taux de bonnes classifications des tentatives d'accès clients et imposteurs mais ne mesure pas le pourcentage de tentatives auquel aucune décision n'est prise. Les auteurs proposent de renvoyer à un opérateur humain les tentatives dans lesquelles la mesure de confiance globale n'est pas suffisante.

Dans [Poh et al., 2007], les auteurs proposent de choisir une des deux règles de fusion : somme ou produit selon des mesures de qualité dans un système biométrique audio-visuel. Les mesures de qualités sont : la pose du visage, l'illumination et le SNR audio. Pour chaque enregistrement, les scores sont regroupés en fonction des mesures de qualité. Ensuite, dans chaque groupe de qualité, les scores sont combinés par somme pondérée. Les scores de groupes résultants sont fusionnés par produit afin d'obtenir le score audio-visuel de l'enregistrement. Les résultats obtenus sur la base de données $XM2VTS$ [Messer et al., 1999] montrent une amélioration significative des performances comparé à une fusion classique. Dans [Kryszczuk et al., 2005], une décision binaire est prise en fonction de mesures de qualité afin de choisir la modalité à considérer pour la vérification de l'identité (audio ou visage). Dans [Poh and Bengio, 2005], les auteurs proposent une méthode de fusion dépendante des mesures de qualité du signal basée sur une somme pondérée des scores de classificateurs mono-modaux mais sans aucun résultat expérimental.

6.2.2 Notre méthode de fusion

Les dépendances

Les systèmes de vérification audio-visuelle de l'identité basés sur une seule modalité sont très sensibles à la qualité des données. Les performances diminuent lorsque l'au-

dio et/ou l'image sont en mauvaise qualité. Il serait intéressant d'adapter la confiance associée à chaque modalité en fonction des conditions d'enregistrement.

Selon le contexte d'étude et le type de données, différentes mesures de qualités peuvent être calculées (le rapport signal à bruit, la netteté de l'image, la pause du visage, etc). Soient $Q = \{q_1, ..q_n\}$ des mesures de qualité des n modalités utilisées pour la vérification de l'identité. Dans ce cas, le score de fusion final de l'équation 6.6 devient :

$$\begin{cases} S_{final}(x|\lambda) = \sum_{i=1}^{n} w_i(Q) \times \tilde{S}_i(x_i|\lambda) \\ avec \quad \tilde{S}_i(x_i|\lambda) = f_i(S_i(x_i|\lambda), \mu_i^c, \sigma_i^c|\lambda) \end{cases} \tag{6.7}$$

De plus, dans chaque modalité, la distribution des scores dépend de la qualité du signal d'entrée (voir les résultats expérimentaux du chapitre 8). Notre contribution consiste à adapter les paramètres de normalisation des scores à la qualité du signal. En effet, les paramètres de normalisation des scores qui sont appris sur les distributions des scores *Clients* sont également dépendants de la qualité du signal d'entrée. L'intérêt de cette normalisation basée sur les mesures de qualité est démontrée dans le chapitre 8. Avec ces considérations, l'équation 6.7 devient :

$$\begin{cases} S_{final}(x|\lambda) = \sum_{i=1}^{n} w_i(Q) \times \tilde{S}_i(x_i|\lambda) \\ avec \quad \tilde{S}_i(x_i|\lambda) = f_i(S_i(x_i|\lambda), \mu_i^c(q_i), \sigma_i^c(q_i)|\lambda) \end{cases} \tag{6.8}$$

Les classes de dégradation du signal

Afin d'apprendre les paramètres de fusion avec leurs dépendances à la qualité du signal, nous avons procédé de la manière suivante :

1. Définir des Classes de dégradation du signal en fonction des mesures de qualité $Q(q_1, .., q_n)$. Soit M le nombre de classes de dégradation du signal. On note $C_{i=1}^{M}$ les classes de dégradation du signal.
2. Estimer les paramètres de fusion (poids et paramètres de normalisation) dans chaque classe.
3. Apprendre une fonction de prédiction des classes de dégradation $\zeta(Q) = C_i$ avec $i \in 1, .., M$ afin de prédire automatiquement le niveau de dégradation en fonction des mesures de dégradation du signal dans chaque modalité. La fonction $\zeta(Q)$ est apprise en utilisant des méthodes de classification standards comme les machines à vecteurs support (SVM), la régression logistique, K-*moyenne*, etc.

Fusion

Le score final de l'équation 6.8 devient :

$$\begin{cases} S_{final}(x|\lambda) = \sum_{i=1}^{n} w_i(\zeta(Q)) \times \tilde{S}_i(x_i|\lambda) \\ avec \quad \tilde{S}_i(x_i|\lambda) = f_i(S_i(x_i|\lambda), \mu_i^c(\zeta(Q)), \sigma_i^c(\zeta(Q))|\lambda) \end{cases} \tag{6.9}$$

6.3 Conclusion

Dans ce chapitre, nous avons décrit les principes des systèmes de vérification de l'identité que nous avons utilisés dans nos expériences ainsi que des méthodes de fusion. Notre contribution consiste dans la définition d'une méthode de fusion basée sur des mesures de qualité calculables automatiquement. Cette méthode est basée sur la définition de classes de dégradation du signal dans lesquelles les paramètres de fusion sont optimisés. Ces classes de dégradation doivent traduire la confiance accordée à chaque modalité. Dans le chapitre 8, La méthode de fusion est appliquée dans un système de vérification de l'identité audio-visuelle.

Chapitre 7

Application des modèles non-supervisés pour la structuration

Notre objectif est de construire un index audio-visuel des interventions de chaque personne dans des contenus de type *Talk shows*. Le *système initial* de structuration proposé dans la première partie du rapport est basé sur la combinaison des index audio et visuel obtenus par segmentation et regroupement des personnes. Malheureusement, cette approche cumule les erreurs obtenues dans chaque modalité. Dans ce chapitre, nous proposons une méthode de correction mutuelle des erreurs de *Clustering* dans laquelle chaque modalité pourra pallier aux faiblesses de l'autre afin d'améliorer le système global.

7.1 Principe

Le système initial présenté dans le chapitre 4 est basé sur la construction de deux index de personnes, l'un basé sur la parole et l'autre sur les costumes. Ensuite, une fonction d'appariement est recherchée afin d'associer les séquences d'interventions audio et visuelles provenant de la même personne. Enfin, les séquences de visages parlants sont obtenues par intersection des segments audio et visuels dont les étiquettes sont appariées. À l'intersection des index audio et visuels, chaque segments audio-visuels x est associé à une paire d'étiquettes audio et visuelle $(A(x_A), V(x_C))$. Seuls les segments où les étiquettes audio et visuelle s'associent $(A(x_A)) = f_{assoc}(V(x_C)))$ sont considérés comme des segments de visages parlant (segments *consistants*). Les autres segments sont considérés comme *inconsistants* $(A(x_A)) \neq f_{assoc}(V(x_C)))$. Cette inconsistance s'explique par deux raisons :

1. *Cause A* : pas de présence de visage parlant (le locuteur n'est pas la personne qui apparaît).
2. *Cause B* : il y a bien une présence d'un visage parlant, mais non détectée en raison d'une erreur d'étiquetage dans le regroupement audio ou visuel (ou les

deux).

Le principe d'amélioration est basé sur la remise en question, dans ces segments inconsistants, l'une des deux modalités audio ou visuelle considérée la moins fiable afin de la corriger en cas de détection d'erreur d'étiquetage. La figure 7.1 présente le processus de correction mutuelle des erreurs de chaque modalité basé sur des modèles de vérification de l'identité.

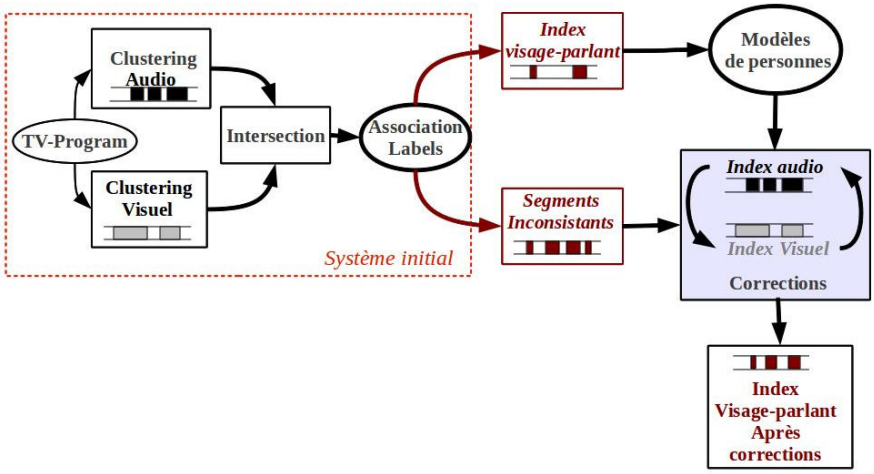


FIGURE 7.1 – Processus de correction mutuelle des segments inconsistants basé sur les modèles de vérification de l'identité

D'après les résultats du système initial présenté dans le chapitre 4, la précision des *Clusters* de visages-parlants est élevée alors que le rappel est faible. Dans ce chapitre, l'idée est d'utiliser ces groupes homogènes obtenus automatiquement pour apprendre des modèles audio et visuel de chaque groupe. Le processus de correction détermine pour chaque segment la modalité qui a échoué en fonction des scores de vérification de l'identité audio-visuelle.

7.2 Les modèles de vérification audio-visuelle de l'identité

Les séquences de visages-parlant détectés automatiquement par le *système initial* ont une forte probabilité d'être correctement regroupés dans des ensembles homogènes (*Clusters*). Nous utilisons ces groupes homogènes afin d'apprendre un

modèle de locuteur et un modèle de costume pour chaque groupe détecté automatiquement. Étant donné que ces modèles audio et visuels sont disponibles pour chaque groupe, il est alors possible d'identifier les personnes dans chaque segment *inconsistant*, puis de déterminer non seulement une présence de visage parlant, mais aussi son groupe audio-visuel.

7.2.1 Vérification de l'identité audio

Nous avons utilisé le modèle de vérification du locuteur présenté dans le chapitre 6. Ce modèle est basé sur des modèles *GMM* avec un modèle du monde. Le modèle du monde *UBM* est entraîné à partir d'émissions non annotées du programme « *On n'a pas tout dit* ». Soit x un segment audio-visuel annoté *inconsistant* par le système initial. Le vecteur de paramètres acoustiques $x_A = x_A^1, .., x_A^n$ extrait de la séquence est comparé au modèle du groupe λ et au modèle du monde (UBM). Le score de vérification de l'identité du locuteur $S_a(x_A|\lambda)$ est calculé de la manière suivante :

$$\begin{aligned} S_a(x_A|\lambda) &= \frac{1}{n} \log \frac{p(x_A|\lambda)}{p(x_A|UBM)} \\ p(x_A|\lambda) &= \prod_i p(x_A^1|\lambda) \end{aligned} \tag{7.1}$$

avec n la taille de la séquence audio-visuelle x.

7.2.2 Vérification de l'identité visuelle

Le modèle de vérification de l'identité visuelle est basé sur la signature des couleurs des costumes. Ce modèle est présenté dans la section 6.1.2 du chapitre 6. Soit x une séquence audio-visuelle d'apprentissage de la personne de la personne λ contenant L trames. Le costume est détecté dans chaque trame de la séquence d'apprentissage selon le principe présenté dans la sous-section 4.2.1 du chapitre 4. Chaque costume détecté est représenté par l'histogramme des couleurs (codage RGB). Pour chaque segment audio-visuel x, le vecteur de paramètres du costume x_C est extrait de la manière suivante :

- Dans chaque trame, le costume est détecté selon le principe présenté dans le chapitre 4.
- Ensuite, le costume centroïde de la collection de costumes détectées est sélectionné.
- x_C est l'histogramme de couleurs du costume centroïde (codage RGB).

Afin d'appendre un modèle de costume à partir d'une collection de L segments audio-visuels, l'histogramme de couleurs du costume centroïde est extrait pour chaque segment. Le modèle de costume est représenté par le vecteur de concaténation des histogrammes noté $x_C^\lambda\{x_1^\lambda, .., x_L^\lambda\}$.

Soit x un segment audio-visuel annoté *inconsistant* par le système initial. Le vecteur de paramètres du costume x_C extrait de la séquence x est comparé au modèle du groupe λ. Le score de vérification de l'identité du costume $S_c(x_C|\lambda)$ est calculé de la manière suivante :

$$S_c(x_C|\lambda) = \frac{1}{L}\sum_{i=1}^{L} ||x_C - x_i^\lambda|| \tag{7.2}$$

7.2.3 Vérification de l'identité audio-visuelle

Pour un segment audio-visuel x dont on extrait les vecteurs de paramètres acoustiques x_A et de costume x_C, le score de vérification de l'identité audio-visuelle est une combinaison linéaire des scores normalisés $\tilde{S}_a(x_A|\lambda)$ et $\tilde{S}_c(x_C|\lambda)$. La fonction de normalisation utilisée est $Z-Norm$ (voir le chapitre 6). Pour un segment x, le score de vérification audio-visuelle $S_{final}(x|\lambda)$ est calculé de la manière suivante :

$$S_{final}(x|\lambda) = \alpha\tilde{S}_a(x_A|\lambda) + (1-\alpha)\tilde{S}_c(x_C|\lambda) \tag{7.3}$$

avec le poids de fusion $\alpha \in [0, 1]$.

7.3 Les hypothèses de modification

Le processus de modification est basé sur la détermination automatique, pour chaque segment inconsistant (obtenu par le système initial), la modalité à corriger. Soit x un segment audio-visuel inconsistant. On définit trois hypothèses de modification : la première considère qu'aucune modalité n'a échoué, la seconde considère que c'est la modalité audio qui a échoué et le troisième considère que c'est la modalité visuelle. L'hypothèse que les deux modalités ont échoué n'est pas traitée. La figure 7.2 présente les 3 hypothèses de modification. Chaque hypothèse est associée à un score de vérification de l'identité audio-visuelle calculé de la manière suivante :

1. **Hypothèse** H_1 : pas d'erreur de regroupement audio ou visuel. Dans ce cas, le segment est réellement un segment de visage non parlant. Le score de vérification audio-visuelle du segment est dans ce cas calculé comme suite :

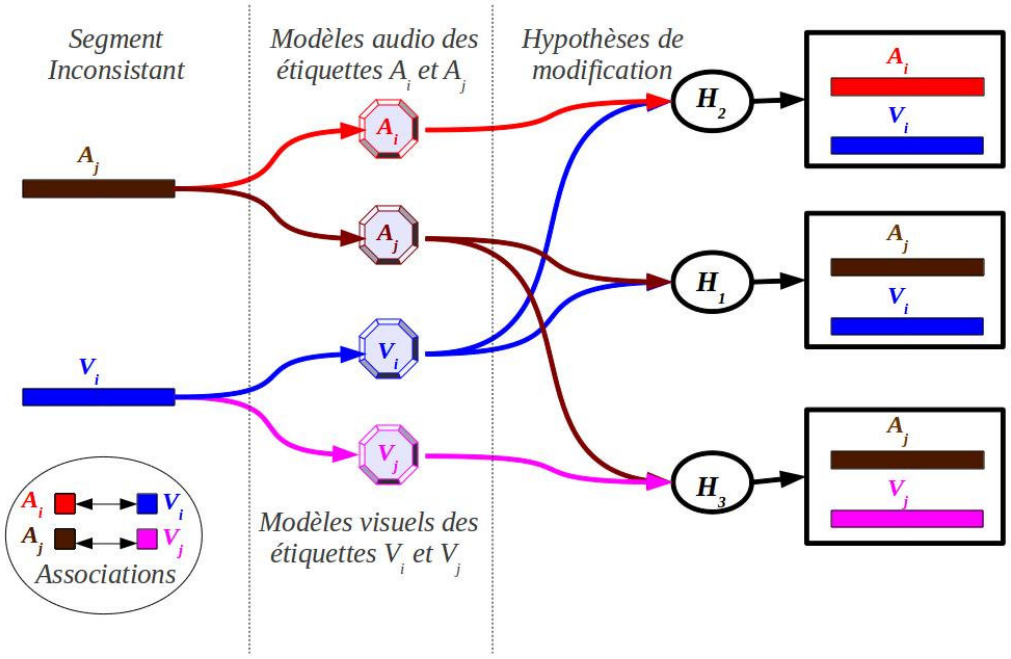


FIGURE 7.2 – Exemples d'hypothèses de modification. Dans H_1, aucune modalité n'a échoué. Dans H_2, la modalité audio a échoué. Dans H_3, la modalité visuelle a échoué

$$S^{H_1}(x) = \alpha \tilde{S}_a(x_A|A(x_A)) + (1-\alpha)\tilde{S}_c(x_C|V(x_C))$$

2. **Hypothèse** H_2 : le système de regroupement basé sur le costume a commis une erreur. Seule l'étiquette audio est considérée pour le calcul du score de vérification de l'identité audio-visuelle. La séquence est comparée au modèle de locuteur et de costume de la personne portant l'étiquette du label audio. Le score de vérification audio-visuelle du segment, basé sur l'identité audio, est dans ce cas calculé comme suite :

$$S^{H_2}(x) = \alpha \tilde{S}_a(x_A|A(x_A)) + (1-\alpha)\tilde{S}_c(x_C|f^{-1}_{assoc}(A(x_A)))$$

3. **Hypothèse** H_3 : le système de regroupement basé sur le locuteur a commis une erreur. Seule l'étiquette visuelle est considérée pour le calcul du score de vérification de l'identité audio-visuelle. La séquence est comparée au modèle audio et de costume de la personne portant l'étiquette du label visuel. Le score de vérification audio-visuelle du segment, basé sur l'identité visuelle, est dans

ce cas calculé comme suite :

$$S^{H_3}(x) = \alpha \tilde{S}_a(x_A | f_{assoc}(V(x_C)) + (1-\alpha)\tilde{S}_c(x_C|V(x_C))$$

7.4 Améliorations basées sur la vérification d'identité de l'audio-visuelle

Notre objectif est de récupérer des segments de visages parlants perdus dans le système initial à cause d'une erreur de regorgement. Ce principe de la méthode basée sur les modèles de vérification de l'identité est divisé en deux phases. La première phase consiste à détecter automatiquement les erreurs de regroupement. Dans la seconde phase, un schéma de correction est appliqué afin de corriger l'erreur détectée.

7.4.1 Détection des erreurs de regroupement

Afin de détecter une erreur de regroupement, deux indicateurs sont utilisés :

Détection basée sur le mouvement des lèvres

Pour chaque segment inconsistant, l'erreur de regroupement est suspectée si un mouvement des lèvres est détecté. Nous avons utilisé le détecteur de mouvement des lèvre est présenté dans le chapitre 5.

Détection basée sur les hypothèses de modification

Pour chaque segment inconsistant, l'erreur de regroupement est suspectée par l'invalidation de la première hypothèse ($\neg H_1$). l'hypothèse H_1 est invalidée par le fait que le score $S^{H_1}(X)$ qui préconise de ne rien faire est moins bon que l'un des scores des deux autres hypothèses qui préconisent une correction d'un label.

7.4.2 Schéma de modification basé sur la vérification d'identité de l'audio-visuelle

Dans le cas où la présence d'un visage parlant et suspectée, l'identité du segment est déterminée par la modalité la plus fiable. Cette modalité est déterminée en utilisant le meilleur score de vérification de l'identité calculée pour les deux hypothèses

Algorithm 2 Modification basée sur la vérification d'identité de l'audio-visuelle

for chaque segment *inconsistant* x **do**

if *Détection d'erreur de regroupement* **then**

if $S^{H_2}(x) > S^{H_3}(x)$ **then**

$V(x_C) = f_{assoc}^{-1}(A(x_A))$

else

$A(x_A) = f_{assoc}(V(x_C))$

end if

end if

end for

H_2 et H_3. Si le meilleur score est obtenu par l'hypothèse H_2, cela signifie que le segment est identifié comme étant la personne déterminée par le regroupement audio. Dans ce cas, le label visuel est corrigé de manière à le rendre compatible avec le label audio par la fonction d'association. Si le meilleur score est obtenu par l'hypothèse H_3, le segment est identifié comme étant la personne déjà déterminée par le regroupement basé sur le costume. Dans ce cas, c'est le label audio qui est corrigé afin de le rendre compatible avec le label visuel déterminé par la fonction d'association. Ce processus de modification est formalisé dans l'algorithme 2.

7.5 Expériences

Afin de comparer les performances du schéma de modification basée sur les modèles de vérification de l'identité avec les performances des schémas de modification basé sur une modification systématique (présentés dans le chapitre 5), trois expériences sont effectuées :

Expérience 1 Pour chaque segment *inconsistant* x, la détection d'erreur de regroupement est basée sur l'invalidation de l'hypothèse H_1. Le schéma de modification est basé sur la correction systématique de l'étiquette de la modalité considérée a priori la moins fiable. Ce processus de modification systématique est détaillé dans

le chapitre 5. Deux types de modification systématique :

- Modification systématique de l'audio : l'étiquette audio est systématiquement modifiée de manière à la rendre compatible à l'étiquette visuelle ($A(x_A) = assoc(V(x_C))$). Les résultats sont présentés dans $\neg H_1$ + *modification systématique de l'audio.*
- Modification systématique du visuel : l'étiquette visuelle est systématiquement modifiée de manière à la rendre compatible à l'étiquette audio ($V(x_A) = f^{-1}_{assoc}(A(x_A))$). Les résultats sont présentés dans $\neg H_1$ + *modification systématique du visuel*

Expérience 2 Pour chaque segment *inconsistant* x, la détection d'erreur de regroupement est basée est basée sur l'invalidation de l'hypothèse H_1. Le schéma de modification est basé sur la correction par l'hypothèse qui obtient le meilleur score. Les résultats sont présentés dans $\neg H_1$ + *modification basée sur la vérification d'identité de l'audio-visuelle.*

Expérience 3 Pour chaque segment *inconsistant* x, la détection d'erreur de regroupement est basée sur le détecteur de mouvement des lèvres présenté dans le chapitre 5. Le schéma de modification est basé sur les scores de vérification de l'identité. Les résultats sont présentés dans *Lip + modification basée sur la vérification d'identité de l'audio-visuelle.*

7.6 Résultats et discussion

Le tableau 7.1 résume les performances du système d'indexation des visages-parlants dans les différents schémas de correction des erreurs de regroupement et avec les deux détecteur d'erreurs de regroupement. Pour chaque schéma de modification, nous présentons la moyenne des taux de *Précision*, *Rappel* et *F-mesure* obtenus pour les 5 épisodes de la base de données *TSDB*. Dans le schéma de modification basée sur les scores de vérification de l'identité audio-visuelle, le paramètre de fusion α est fixé à 0.8 après optimisation de la *F-mesure*.

Dans le schéma de modification systématique, l'utilisation du détecteur de mouvement des lèvres pour suspecter la présence de visage parlant donne de meilleurs résultats que $\neg H_1$ pour le processus de modification visuelle.

Dans le schéma de modification basé sur les scores de vérification audio-visuelle, lorsque le détecteur de mouvement des lèvres est utilisé afin de suspecter une présence de visage parlant, les améliorations obtenues ne sont pas très significatives comparé

Schémas de modification/moyenne sur les 5 épisodes	*Pr*	*Ra*	F_m
Système initial	90.3	60.8	72.7
Lip + modification systématique de l'audio	77.2	62.1	68.8
Lip + modification systématique du visuel	86.8	69.2	76.9
$\neg H_1$ *+ modification systématique de l'audio*	78.5	62.2	69.4
$\neg H_1$ *+ modification systématique du visuel*	83.9	66.3	74.1
$\neg H_1$ *+ modification basée sur la vérification d'identité*	85.2	70.0	76.1
Lip + modification basée sur la vérification d'identité	88.5	68.5	77.3

TABLE 7.1 – Performances du système d'indexation des visages-parlants dans les différents schémas de modification sur la base de données *TSDB*

à un schéma de modification systématique du visuel. Cela signifie que les modèles de vérification utilisés ne sont pas assez précis pour obtenir une amélioration significative. Cependant, ces modèles sont efficaces parce qu'ils permettent de détecter une présence de visage-parlant (en utilisant $\neg H_1$). Certes, cette détection est un peu moins efficace que celle basée sur le détecteur de mouvement de lèvres, mais elle permet néanmoins une amélioration significative du *système initial* (en moyenne +4.6 dans la *F-mesure*).

7.7 Conclusion et perspectives

Dans ce chapitre, nous avons présenté une méthode de récupération des segments de visages parlants perdus par le *système initial* pour cause d'erreurs de regroupement. La majorité de ces erreurs sont dues à la complexité des données dans des contenus de télévision. Le principe de la méthode est d'apprendre des modèles non supervisés à partir des groupes de personnes obtenus par le *système initial* afin d'identifier les segments ambigus. À partir de ces modèles, plusieurs processus de récupération sont évalués.

Toutes les tentatives de correction proposées améliore le taux $F - mesure$. Ces améliorations sont dues principalement à une augmentation du taux de rappel au détriment d'une légère diminution de la précision.

Travaux futurs Les modèles non supervisés peuvent être utilisés afin de détecter des personnes dans d'autres épisodes de l'émission. De plus, dans nos schémas de modification, l'hypothèse que les deux modalités ont échoué n'est pas traitée. Il serait intéressant d'étudier un schéma de correction dans ce cas.

Chapitre 8

Application des modèles dans les systèmes biométriques

Dans ce chapitre, nous présentons l'application des modèles de vérification de l'identité dans les systèmes biométriques. Une application de notre méthode de fusion basée sur des mesures de qualités calculables automatiquement est expérimenté sur la base de données biométrique *Banca*.

8.1 Les systèmes Biométriques

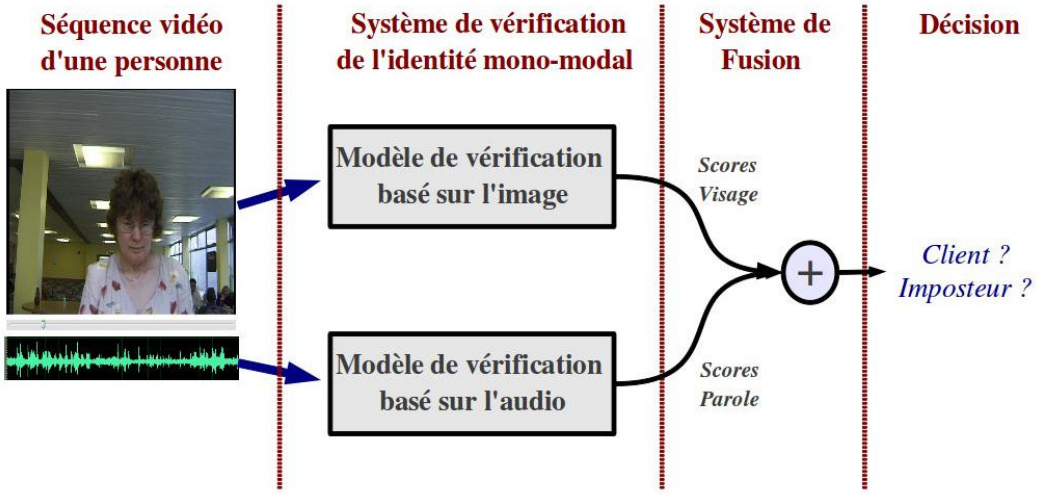


Figure 8.1 – Architecture du système de vérification de l'identité audio-visuelle

La vérification audiovisuelle de l'identité intègre les deux informations complémentaires audio et visuelle afin d'améliorer les performances de chaque système pris séparément. En effet, les systèmes de vérification de l'identité basés uniquement sur la modalité

audio sont très sensibles au type de micros utilisés pour l'enregistrement, l'environnement acoustique, le contexte d'enregistrement (expressivité, maladie, etc). Symétriquement, les systèmes de vérification de l'identité basés sur la modalité visuelle sont sensibles aux conditions d'éclairage, la caméra, pose, expression faciale, etc.

Architecture générale

L'architecture générale d'un système de vérification de l'identité audio-visuelle est divisé en trois composantes :

- Un module de vérification de l'identité basé sur l'information audio : prend en entrée une séquence audio à comparer au modèle de la personne prétendue.
- Un module de vérification de l'identité basé sur l'information visuelle : prend en entrée une image ou une séquence d'image à comparer au modèle visuel de la personne prétendue.
- Un module de fusion qui combine les deux scores obtenus afin de prendre la décision finale à savoir si la personne est bien celle qu'elle prétend être.

Évaluation

Dans les systèmes biométriques, deux erreurs peuvent survenir : fausses acceptations et faux rejets. La figure 8.2 montre ces erreurs par rapport aux distribution des scores *Clients* et *Imposteurs*.

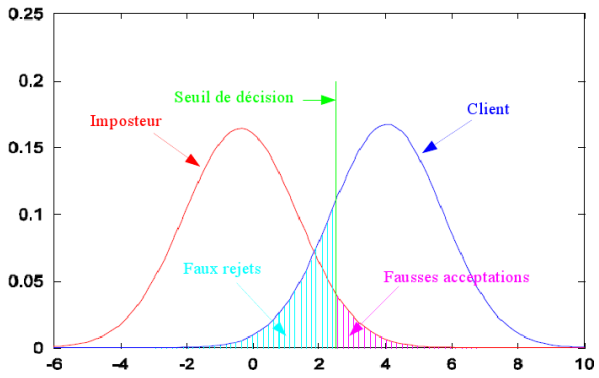


FIGURE 8.2 – Distribution des scores *Client/Imposteur*

À partir de ces erreurs, deux mesures de performances sont calculées [Petrovska-Delacrétaz et al.
le taux de fausses acceptations (FAR) et le taux de faux rejets (FRR) sont calculées de la manière suivante :

$$\text{FAR} = \frac{\textbf{Nombre de fausses acceptations}}{\textbf{Nombre d'accès Imposteurs}}$$
$$\text{FRR} = \frac{\textbf{Nombre de faux rejets}}{\textbf{Nombre d'accès Clients}}$$

La courbe *DET* (Detection Error Tradeoff) décrit le taux de faux rejet (*FRR*) en fonction du taux de fausse acceptation (*FAR*) en faisant varier le seuil de décision. Cette courbe permet de visualiser différents points de fonctionnement et de trouver un compromis entre les différentes erreurs grâce au réglage du seuil de décision. Deux points de fonctionnements sont souvent utilisés pour mesurer les performances dans un système biométrique :

- Le taux d'égale erreur (*EER*) : le point où on a autant de faux rejets que de fausses acceptations, autrement dit le point où la courbe coupe la diagonale.
- La fonction de coût de détection (*DCF*) : somme pondérée des mesures *FAR* et *FRR*. Lorsque l'objectif est d'être robuste aux accès *Imposteurs* (le cas des systèmes biométriques), un grand poids est accordé aux fausses acceptations (*FAR*). Dans les évaluations *NIST*, le taux *DCF* se calcule de la manière suivante : $\text{DCF} = 0.99 * \text{FAR} + 0.1 * \text{FRR}$

8.2 Base de données *Banca*

La base de données biométrique audio-visuelle *Banca*[1] a été acquise en quatre langues différentes (anglais, français, italien et espagnol) et dans les deux modalités visage et voix.

Description

Les séquences sont enregistrées en 12 sessions (durant 3 mois) auprès de différents capteurs (2 caméras et 2 micros) et de plusieurs scénarios (contrôlé, dégradés et les effets indésirables). Au total, 208 personnes sont capturés, 52 dans chaque langue (26 hommes et 26 femmes). Dans notre étude, nous nous somme intéressés à la base constituée en anglais. La population a été séparée en deux groupes noté *G1* et *G2* contenant chacun 26 individus (13 hommes et 13 femmes).

Chaque individu enregistre 12 sessions. Pour chaque session, l'individu fait deux enregistrements : un accès *Client* (déclarant être lui même) et un accès *Imposteur*

1. http ://www.ee.surrey.ac.uk/CVSSP/banca/

FIGURE 8.3 – Exemples de plans dans la base de données *Banca*. Les conditions : *Controlled*, *Degraded et Adverse*

(déclarant être une autre personne). Au cours de chaque enregistrement, le sujet est invité à lire un texte contenant 12 chiffres au hasard, son nom, son adresse et date de naissance. La durée moyenne d'un enregistrement est de vingt secondes. Les 12 sessions sont séparées en 3 différents scénarios :

- *Controlled* pour les sessions $1-4$, enregistrées dans un environnement contrôlé avec un arrière plan neutre et une bonne caméra.
- *Degraded* pour les sessions $5-8$, enregistrées grâce à une webcam dans un environnement assez peu bruyant (bureaux).
- *Adverse* pour les sessions $9-12$, enregistrées dans un réfectoire avec beaucoup de bruit de fond mais une bonne caméra.

Protocole d'évaluation dans *Banca*

Le protocole utilisé dans nos expériences est le protocole appelé *Pooled* distribué avec la base de donnée *Banca*. Dans ce protocole, l'apprentissage des modèles de chaque personne est effectué sur un seul enregistrement obtenu dans un environnement contrôlé (*Controlled*). Le reste de la base de données est utilisé pour le test. Plus précisément, lors des phases d'apprentissage et de test, nous suivons le protocole suivant :

Phase d'apprentissage Pour chaque individu λ, utiliser l'enregistrement de l'accès *Client* de la première session de l'individu pour apprendre les modèles audio et visuels.

Phase de test Pour chaque individu λ, utiliser les enregistrements de la manière suivante :

- En accès *Client* : utiliser les sessions 2 à 4 de la condition *Controlled*, 6 à 8 de la condition *Degraded* et 10 à 12 de la condition *Adverse*. Les sessions 5 et 9 ne sont pas utilisées afin de conserver les mêmes proportions pour chaque condition. En résumé, le protocole P prévoit 234 tests *Clients* (78 dans chaque condition).
- En accès *Imposteur* : utiliser les 12 sessions (toutes conditions confondues). Le protocole P prévoit 312 tests *Imposteurs* (104 dans chaque condition).

8.3 Performances des modèles de personnes dans le système biométrique

Ensemble de développement

Malheureusement, *Banca* ne possède pas une grande collection et pas d'ensemble de développement. Nous avons exploité le fait que la base est divisée en deux groupes distincts pour apprendre les poids de fusion ainsi que les paramètres de normalisation de chaque groupe à partir de l'autre groupe.

Système de vérification audio

Nous avons utilisé un système de vérification de l'identité du locuteur développé par *OrangeLabs*. Cet outil basé sur des *GMMs* avec un modèle du monde est présenté dans le chapitre 6. Le modèle du monde *UBM* est entraîné à partir d'une très grande base de données *NIST* (anglais américain).

Système de vérification du visage

Nous avons utilisé le système de vérification de l'identité du visage décrit dans [Bredin, 2007]. L'espace des visages de dimension 80 est appris à partir de 2200 visages provenant de plusieurs base de données de visages (*Biomet, ATT, Banca*). Et pour chaque séquence vidéo, les cents meilleurs visages selon le critère *DFFS* sont sélectionnés.

Fusion

Pour un segment audio-visuel x, les scores de vérification de l'identité du locuteur et

du visage obtenus sont normalisés par la fonction *Tanh Norm*. La fusion est réalisée par combinaison linéaire selon la méthode présentée dans la section subsec :SumFusion du chapitre 6. Les poids de fusion (w_A, w_V) sont estimés par minimisation du *DCF* sur l'ensemble du développement. Le score de fusion est calculé de la manière suivante :

$$\begin{cases} S_{final}(x|\lambda) = w_a \tilde{S_A}(x_A|\lambda) + w_v \tilde{S_V}(x_V|\lambda) \\ avec \quad w_A + w_V = 1 \end{cases} \tag{8.1}$$

La figure 8.4 montre les variations du *DCF* en fonction du poids accordé à la modalité audio w_A pour chaque groupe dans *Banca*. En optimisant le *DCF*, on accorde environ 70% du poids aux scores issu du système de vérification de l'identité locuteur et 30% scores issu du système de vérification de l'identité visage.

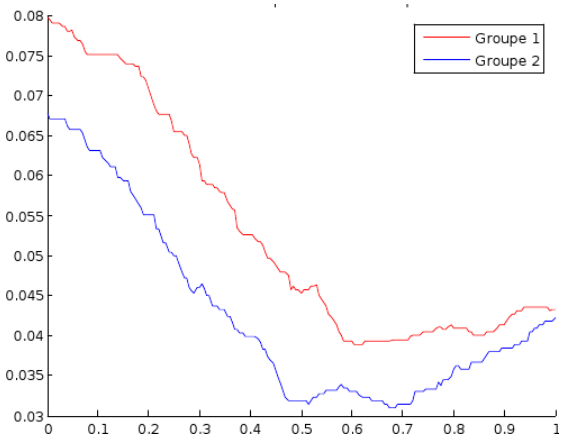


FIGURE 8.4 – Valeurs du *DCF* en fonction du poids du système de vérification du locuteur (w_A) dans *Banca*. Groupe *G1* en rouge et groupe *G2* en bleu

Système de référence *Biosecure*

Un système référence de vérification de l'identité des visages parlants est présenté dans [Bredin et al., 2006] dans le cadre de la campagne d'évaluation *The BioSecure Multimodal Evaluation Campaign (BMEC)*. Le système de vérification du visage est le même que celui utilisé dans nos expériences tandis que le système de vérification de l'identité locuteur est basé sur l'outil *BECARS* [Blouet et al., 2004]. La fusion est basée sur une classification *SVM* des scores audio et du visage. Ce système

de référence évalué sur *Banca* obtient un taux $EER = 7.7\%$ (voir détails dans [Bredin et al., 2006]).

Résultats

	Speaker	*SpeakerBecars*	*Face*	*SumFusion*
G1 (EER)	13.1	19.9	20.5	10.1
G2 (EER)	9.7	17.0	20.5	7.7
Moyenne (EER)	11.4	18.4	20.5	8.9

TABLE 8.1 – Performances des systèmes de vérification de l'identité dans *Banca*. Les taux *EER* obtenus par notre systèmes de vérification du locuteur (*Speaker*), le système *BECARS* (*SpeakerBecars*), le visage (*Face*) et la fusion par somme pondérée (*SumFusion*)

Le tableau 8.1 présente les taux *EER* obtenus par notre système de vérification du locuteur (*Speaker*), le système *BECARS* (*SpeakerBecars*), le visage (*Face*),et la fusion par somme pondérée (*SumFusion*) pour chaque groupe.

Dans la vérification de l'identité du locuteur, les résultats obtenus par notre système sont meilleurs que ceux du système *BECARS*. Ceci s'explique par l'adaptation du modèle du monde déjà existant pour d'autres données pour les données *Banca*. Dans la vérification de l'identité du visage, le taux *EER* est de 20.5% en moyenne. Ce taux d'erreur élevé s'explique par le choix du modèle de vérification du visage qui est assez simple. Concernant la fusion, la combinaison par somme pondérée des scores (*SumFusion*) améliore considérablement les performances du système biométrique comparé aux résultats dans chaque modalité. En moyenne, le taux *EER* est de 8.9%. Néanmoins, notre système est moins bon que celui du système de référence *Biosecure*.

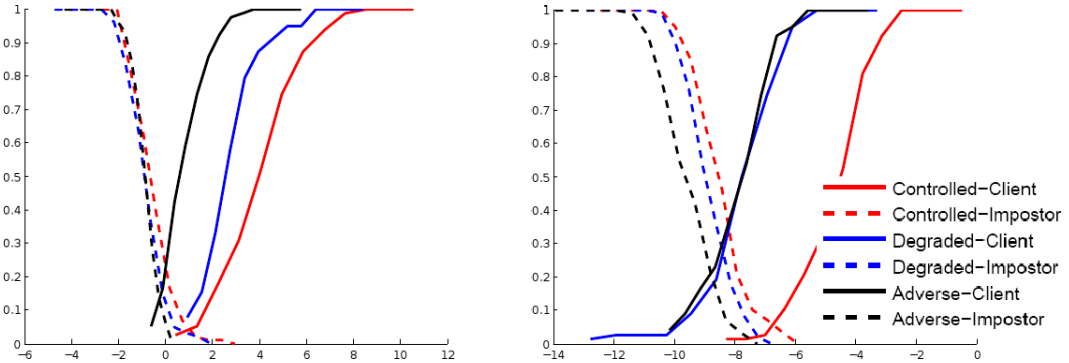
8.4 Introduction de mesures de qualité dans la fusion

Les systèmes de vérification de l'identité du locuteur et du visage sont sensibles aux conditions d'enregistrements. Sur *Banca*, la séparabilité des scores *Clients/Imposteurs* est nettement plus grande dans les conditions d'enregistrements contrôlées (*Controlled*). Les performances du modèle de vérification de l'identité du visage n'est pas

assez performant lorsque les enregistrements sont effectués par une webcam (condition *Degraded*). Les performances du modèle de vérification de l'identité du locuteur diminuent lorsque les enregistrements audio sont bruités (*Adverse*). Il est donc intéressant d'inclure des mesures de qualité audio-visuelles, de manière à ce que la fusion des scores prenne en compte les performances de chaque module de reconnaissance automatique.

Dans cette section, nous présentons l'utilisation de la méthode d'introduction des mesures de qualité dans la fusion par somme pondérée (décrite dans la section 6.2 du chapitre 6) dans la base de donnée *Banca*. Dans la fusion basée sur les mesures de qualité, les paramètres de normalisation des scores et les poids de fusion sont estimés dans chaque condition dans *Banca* à partir de l'ensemble de développement. Ces paramètres sont appliqués pour chaque classe de dégradation du signal prédite de manière non supervisée.

8.4.1 Les dépendances



(a) (b)

FIGURE 8.5 – Histogrammes cumulés des scores *Client/Imposteurs* dans chaque condition du groupe *G2* dans *Banca*. (a) Scores obtenus par le système de vérification de l'identité locuteur. (b) Scores obtenus par le système de vérification de l'identité du visage

La figure 8.5 présente les histogrammes cumulés des scores *Clients/Imposteurs* obtenus dans chaque condition dans *Banca*. Pour chaque accès *Client*, les résultats expérimentaux montrent que la sortie du système est très dépendante de la qualité du signal concernés. Ainsi, les paramètres de normalisation estimés sur les scores *Clients* varient en fonction des conditions d'enregistrement. Il est donc nécessaire

d'adapter les paramètres de normalisation aux conditions d'enregistrement. Aussi, Le seuil de décision est diffèrent dune condition à l'autre. Cette différence de seuil engendrée par la normalisation globale des scores implique une perte de performance lors de la fusion. Pour remédier à ce problème, il faut procéder à une normalisation conditionnelle dépendante des qualités des deux modalités.

8.4.2 Les mesures de qualité dans *Banca*

Il est souvent très difficile de définir des mesures de qualité du signal car celles-ci sont très subjectives. Dans le cas de notre étude, les mesures doivent traduire des mesures de confiance sur le système de vérifications de l'identité du locuteur et du visage.

Mesure de la qualité de la parole

Pour mesurer la qualité audio q_A d'une séquence audio x, l'information généralement utilisée est le rapport signal à bruit (*Signal to Noise Ratio*). Ce rapport compare la force du signal audio à celle du bruit de fond. Un rapport signal à bruit faible signifie que le signal est très bruité, tandis qu'un rapport élevé indique un son clair. Sur une séquence audio x, le *SNR* est estimé de la manière suivant :

$$q_A(x) = \text{SNR}(x) = 10 \times \log_{10}\left(\frac{E_{Parole}}{E_{Bruit}}\right) \tag{8.2}$$

avec E_{Bruit} et E_{Parole} la moyenne de l'énergie des trames détectées comme étant respectivement du bruit et parole de la séquence audio x (énergie est exprimée en *db*).

La figure 8.6 présente les histogrammes du *SNR* par condition dans *Banca* pour les groupes *G1* et *G2*. Le *SNR* apporte une information significative sur la qualité de l'enregistrement audio (plus la valeur est grande, meilleure est la qualité du signal audio). Sur la figure, on arrive à distinguer les trois conditions dans *Banca* avec de grandes valeurs pour les enregistrements contrôlés (*Controlled*), la condition dégradée (*Degraded*) en deuxième position et les plus faibles valeurs pour la condition *Adverse* (environnement sonore très bruité). Les valeurs du *SNR* peuvent être utilisées comme indicateur de qualité audio q_A dans *Banca*.

Mesure de la qualité de l'image

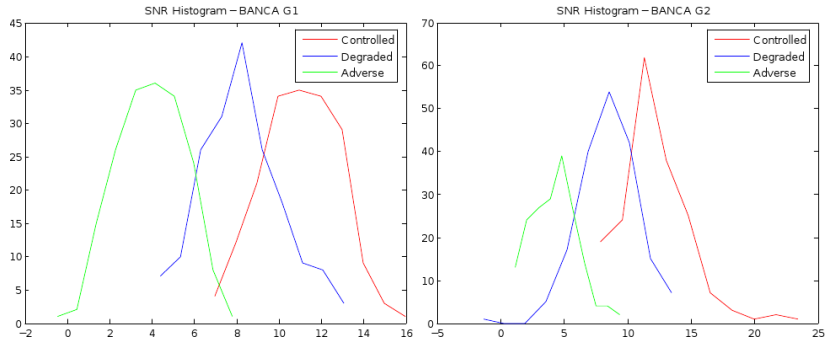


FIGURE 8.6 – Histogramme de la qualité audio q_A (SNR) en fonction des conditions d'enregistrement dans *Banca* pour les groupes *G1* et *G2*

Les performances du système de vérification de l'identité du visage dépendent de deux types de qualité de l'image : qualité au sens netteté de l'image et la qualité de la pose du visage. Sur *Banca*, étant donné que les individus sont face à la caméra, la position du visage ne peut pas être un indicateur de qualité. De plus, dans le modèle de vérification de l'identité du visage utilisé, les meilleurs visages son sélectionnés dans chaque séquence par minimisation de *DFFS* (voir présentation du système dans le chapitre 6). Cette sélection élimine d'avance les visages qui se présentent avec une mauvaise pose. Donc, nous nous intéresserons par conséquent à la détection de la qualité d'image au sens netteté. L'entropie qui est une mesure du désordre indique l'uniformisation de l'image. Plus la valeur de l'entropie est petite, plus l'image est uniforme. Sur une collection d'images $x = \{x^1, .., x^n\}$ de taille n, la qualité $q_V(x)$ est estimée de la manière suivante :

$$\begin{cases} q_V(x) = \frac{1}{N} \sum_{i=1}^{N} Entropy(x^i) \\ avec \quad Entropy(x^i) = -\sum_{j=1}^{256} P_j * log(P_j) \end{cases} \tag{8.3}$$

P_j représente la probabilité de l'intensité j, estimé à partir de l'histogramme des niveaux de gris (256 valeurs).

La figure 8.7 présente l'histogramme de la qualité de l'image q_V par conditions dans *Banca*. Les conditions sont relativement séparables et triées par ordre de qualité d'enregistrement. En conclusion, q_V peut être utilisé comme indicateur de qualité visage dans *Banca*.

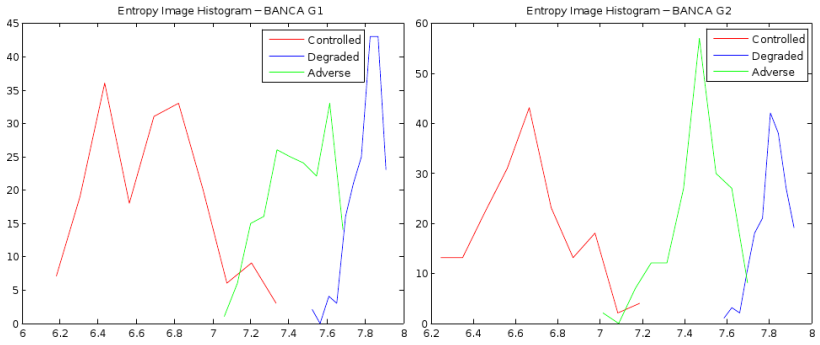


FIGURE 8.7 – Histogramme de la qualité de l'image q_V (*Entropie*) en fonction des conditions d'enregistrement sur *Banca* pour les groupes *G1* et *G2*

8.4.3 Détection des classes de dégradation du signal dans *Banca*

Dans *Banca*, les classes de dégradation du signal sont étiquetés en *Controlled, Degraded et Adverse* ($M = 3$ conditions). La fonction $\zeta(q_A, q_V) = C_i$ de prédiction des classes de dégradation du signal $C_{i=1}^{3}$ est apprise à partir des mesures de qualité q_A et q_V (voir le chapitre 6). Afin de déterminer les conditions dans Banca, nous avons appris un *classifieur* de type (SVM) [2] dans l'ensemble de développement.

G1 - Prédictions/Réels	*Controlled*	*Degraded*	*Adverse*
Controlled	157	27	0
Degraded	25	135	23
Adverse	0	20	159
G2 - Prédictions/Réels	*Controlled*	*Degraded*	*Adverse*
Controlled	157	19	0
Degraded	25	127	26
Adverse	0	36	156

FIGURE 8.8 – Résultats de la détection des conditions par *SVM* dans *Banca* pour les groupes *G1* et *G2* (tableau de contingence)

Le tableau 8.8 présente les résultats de prédiction par *SVM* des trois conditions dans *Banca* pour les groupes *G1* et *G2* (tableau de contingence). Dans *Banca*, les

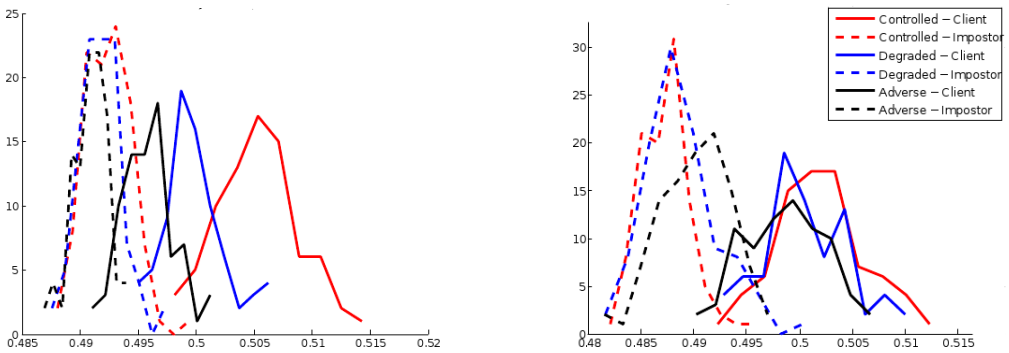
2. http ://svmlight.joachims.org/

mesures de qualité q_A et q_V permettent d'apprendre un *classifieur* textitSVM qui obtient un taux de bonne classification de 82.9% pour le groupe *G1* et de 80.7% pour le groupe *G2*.

8.4.4 Estimation des paramètres de fusion

Paramètres de normalisation des scores

Pour chaque modalité i, les paramètres de normalisation *Tanh Norm* sont μ_i^c et σ_i^c (moyenne et écart type des scores *Clients*. Dans *Banca*, ces paramètres de normalisation sont optimisés pour chaque condition dans l'ensemble de développement. Dans la fusion, ces paramètres sont appliqués dans la classe de dégradation lui correspondant obtenus par de manière non supervisée par l'algorithme *SVM*.



(a) (b)

FIGURE 8.9 – Histogrammes des scores *Clients/Imposteurs* dans chaque condition du groupe *G2* dans *Banca*. (a) scores de fusion basée sur une normalisation indépendante des conditions d'enregistrement. (b) scores de fusion basée sur une normalisation dépendante des conditions d'enregistrement

La figure 8.9 présente les scores obtenus par la fusion basée sur une somme pondérée avec une normalisation indépendante des mesures de qualité (a) et après normalisation dépendante des mesures de qualité (b). Dans cette expérience, les poids de fusion w_A et w_V sont appris de manière globale par optimisation du *DCF*.

Estimation des poids de fusion

Les poids de fusion sont estimés par optimisation du *DCF* pour chaque condition dans *Banca* à partir de l'ensemble de développement et sont appliqués pour chaque classe de dégradation prédite de manière non supervisée par l'algorithme *SVM*.

Fusion

Pour un segment audio-visuel x, une fois la classe de dégradation du signal $\zeta(q_A, q_V) = C_j$ déterminée, le score de vérification de l'identité audio-visuel est calculé en utilisant les paramètres de la optimisé pour la classe prédite dans l'ensemble de développement. Le score final devient :

$$S_{final}(x|\lambda) = w_A(C_j) \times \tilde{S}_A(x_A|\lambda) + w_V(C_j) \times \tilde{S}_V(x_V|\lambda) \tag{8.4}$$

$$\begin{cases} \tilde{S}_A(x_A|\lambda) = f_A(S_A(x_A|\lambda), \mu_A^c(C_j), \sigma_A^c(C_j)|\lambda) \\ \tilde{S}_V(x_V|\lambda) = f_V(S_V(x_V|\lambda), \mu_V^c(C_j), \sigma_V^c(C_j)|\lambda) \end{cases}$$

Apport de la normalisation dépendante des mesures de qualité

Afin de mesurer l'apport de la normalisation dépendante des mesures de qualité, une expérience est menée dans laquelle seulement les poids de fusion w_A et w_A sont optimisés par classe de dégradation du signal tandis que les paramètres de normalisation sont optimisés de manière indépendante (sans prise en compte des mesures de qualité. Les résultats de cette expérience sont présentés par *GN_QualityFusionSum*.

8.4.5 Résultats

La figure 8.10 présente la courbe *DET* pour chaque groupe dans *Banca* des scores obtenus par les systèmes de vérification de l'identité visage (*Face*), du locuteur (*Speaker*) et des fusions par somme pondérée (*SumFusion*) et de fusion dépendante des mesures de qualité (*QualityFusionSum*) pour chaque groupe dans *Banca*.

Le système de fusion basée sur les mesures de qualité (*QualityFusionSum*) apporte une amélioration significative des résultats de vérification de l'identité audio-visuelle dans *Banca* (en moyenne $EER = 6.9$). Cette amélioration s'explique par l'adaptation des paramètres de normalisation et poids de fusion à la qualité du signal.

Dans l'expérience *GN_QualityFusionSum*, les poids de fusion sont dépendants des mesure de qualité alors que les paramètres de normalisation sont indépendants des mesures de qualité. Les performances obtenues sont moins bonnes que ceux de

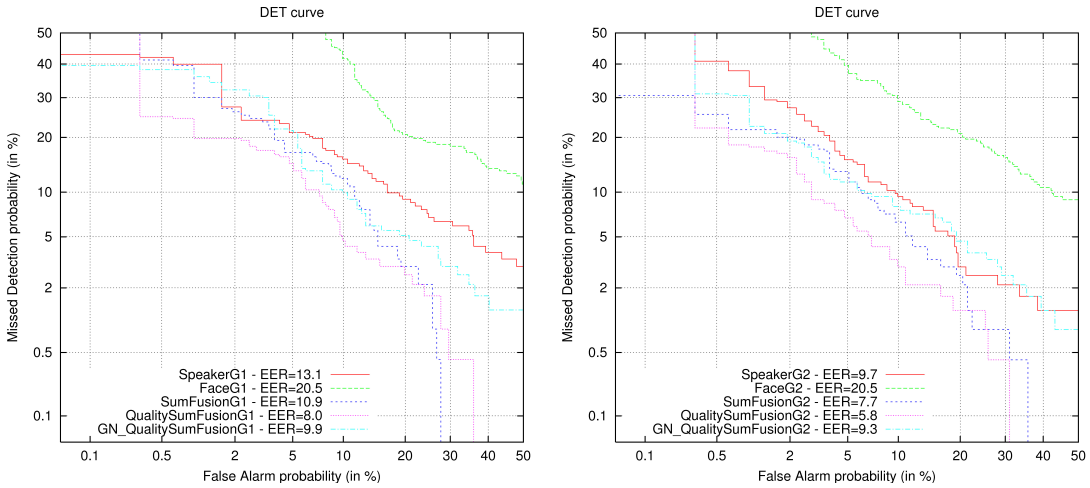


FIGURE 8.10 – Performances des systèmes de fusion pour chaque groupe dans *Banca*

QualityFusionSum dans laquelle toutes les dépendances à la qualité du signal sont prise en compte.

Dans *Banca*, l'estimation des paramètres de fusion (poids de fusion et paramètres de normalisation) dépendants des mesures de qualité s'effectue sur de très petites collections. Ces résultats peuvent être améliorés dans le cas où les paramètres sont estimés sur une plus grande base de données.

8.5 Conclusion

Les systèmes de vérification de l'identité basés sur les modalités audio et visage sont très sensibles à la qualité du signal. Dans ce chapitre, nous avons présenté l'application de notre méthode de fusion basée sur des mesures de qualité dans un système biométrique. Au cours de nos expériences dans *Banca*, nous avons soulevé le problème de la sensibilité du seuil de décision et des poids de fusion à la qualité du signal. Nous avons montré l'intérêt de notre contribution qui consiste à intégrer la dépendance des paramètres de normalisation à la qualité du signal. Notre méthode a permis une amélioration significative des résultats de la vérification de l'identité audio-visuelle (résultats meilleurs que ceux du système référence *Biosecure* [Bredin et al., 2006]).

Perspectives Dans un contexte de télévision, d'autres mesures de qualités pourrait êtres exploitées. Par exemple, en audio, une mesure automatique de l'expressivité de

la parole (émotions, ..) pourrait être utilisée comme indicateur de performance du système de vérification en locuteur. En visuel, certaines poses du visage peuvent être détectées automatiquement ([Bailly and Milgram, 2009]) et pourraient donc être utilisées comme indicateurs de performance du système de vérification du visage.

Conclusions générales et perspectives

Les systèmes d'indexation audio-visuelle des personnes ont pour objectif de faciliter la recherche d'interventions audio-visuelles de certaines personnalités publiques. Cette problématique implique la détection audio-visuelle des personnes, le regroupement des séquences détectées automatiquement et l'association d'une identité à chaque groupe. Chacun de ces domaines représente à lui seul un thème de recherche à part entière. Dans notre étude, nous souhaitons indexer des interventions de personnes dans des contenus de télévision. Dans ce type de contenus, la détection et l'identification des personnes par un processus automatique est très difficile en raison de nombreuses ambiguïtés dans l'audio, l'image et dans leur association (interactivité entre les dialogues, variations de pose du visage, asynchronie la parole et l'apparence, etc). L'objectif principal de la thèse est de proposer des solutions robustes pour lever toutes ces ambiguïtés.

Contributions

Première contribution Nous avons proposé une méthode de construction d'index de personnes à partir d'épisodes de *Talk Shows* en utilisant l'information audio et visuelle de manière indépendante, puis en combinant ces deux index afin d'obtenir les séquences d'intervention audio-visuelle des personnes. Le système est basé sur la détection et le regroupement par la signature vocale et le costume des personnes. Nous avons proposé une méthode de combinaison des index basée sur la maximisation de la couverture globale des groupes de personnes détectées. Un très bon taux de précision est obtenu sur les interventions audio-visuelles des personnes, mais avec un taux de rappel plus faible que dans chacune des modalités.

Nous avons travaillé sur l'amélioration du système d'indexation en remettant en cause la fiabilité des étiquettes audio et visuelles de manière à ce que chaque modalité corrige les erreurs d'annotations automatiques de l'autre modalité. La procédure d'amélioration du système initial est divisée en deux étapes : détection automatique de l'erreur d'indexation puis correction de l'erreur.

Deuxième contribution Afin de détecter une erreur d'indexation, nous avons travaillé sur la détection automatique de l'activité visuelle de la parole dans le but de confirmer la présence d'un visage parlant dans les séquences ambiguës. Nous avons proposé une nouvelle méthode de détection de l'activité visuelle de la parole basée sur la mesure du degré de désordre de la direction des pixels autour de la région des lèvres. Comparée aux approches classiques utilisées dans de telles vidéos, notre mesure permet une amélioration significative du taux de bonne classification. En particulier, notre méthode s'avère être plus robuste à un mouvement global du visage, mais n'est pas fiable pour les rires ou les grimaces où l'on enregistre bien un mouvement des lèvres sans activité de la parole.

Troisième contribution Afin de corriger l'erreur détectée automatiquement, nous avons travaillé sur des procédures de modification de la modalité qui a échoué. Considérant la grande précision du *système initial*, nous avons appris des modèles non supervisés des personnes à partir des séquences détectées automatiquement par le *système initial*. Ces modèles sont utilisés afin de vérifier l'identité de la personne dans les segments ambigus. Les résultats montrent une amélioration significative du rappel expliquée par la récupération d'une grande partie des visages parlants non détectés par le *système initial*.

Quatrième contribution Dans les systèmes biométriques audio-visuels, nous avons concentré notre étude sur les techniques de fusion robustes pour la vérification d'identité audio-visuelle. Au cours de notre étude, nous avons soulevé le problème de la sensibilité des modèles de vérification à la qualité du signal. Nous avons proposé une méthode d'introduction de mesures de qualité dans le but d'adapter la confiance accordée aux différentes modalités. Nous avons également démontré la dépendance des paramètres de normalisation des scores à la qualité du signal. Afin de résoudre ce problème, nous avons proposé de définir des classes de dégradation du signal dans lesquelles les paramètres de normalisation sont optimisés. Ces classes de dégradations qui traduisent la confiance accordée à chaque modalité sont déterminées par des mesures de qualité calculables automatiquement. Cette méthode de fusion a permis une amélioration significative des performances du système biométrique audio-visuel.

Cinquième contribution Au cours de la thèse, nous avons été heurtés au problème de l'inexistence de bases de données publiques et communes pour l'indexation des personnes dans un contexte de télévision. Nous avons collecté et annoté finement cinq épisodes de l'émission de télévision "*On n'a pas tout dit*". Une analyse détaillée de la base de données est effectuée afin de mieux comprendre le contexte d'étude.

Ces contenus sont mis à disposition de tous les partenaires du projet *QUAERO*.

L'origine des erreurs d'un système d'indexation audio-visuel de personne est complexe car issue de la combinaison d'erreurs commises dans chaque modalité. Nous avons présenté une analyse détaillée afin de déterminer le type des erreurs et définir un protocole d'évaluation des méthodes de structuration de contenus de télévision par personne.

Perspectives

L'information audio

Il existe une grande communauté scientifique dédiée à la recherche en indexation en locuteurs. En particulier en France, grâce à la Campagne d'évaluation *ESTER* qui vise à évaluer les systèmes de transcription et d'indexation d'émissions radiophoniques. Le challenge aujourd'hui est de travailler sur des contenus très interactifs comme les *Talks shows* télévisés. Dans ce contexte, les séquences sont souvent courtes, le dialogue est interactif et la parole est très expressive rendant les méthodes existantes d'indexation en locuteur et la vérification de l'identité peu fiables. Ces domaines de recherche méritent des efforts afin d'améliorer les performances.

Une piste d'amélioration de l'indexation en locuteur dans ce contexte est d'utiliser d'autres informations disponibles et complémentaires à l'audio. Outre le visage, la signature gestuelle peut être utilisée pour l'identification du locuteur. Partant du principe que chaque personne possède une manière visuelle d'expression, dans [Bregler et al., 2009] des méthodes sont développées afin d'améliorer l'identification du locuteur par la signature de la gestuelle du haut du corps et du visage. L'image peut également apporter une information supplémentaire sur la séquence de parole prononcée par le locuteur grâce à l'analyse du mouvement des lèvres et de la synchronie audio-visuelle.

L'information visuelle

Le système d'indexation des personnes basé sur la détection et le regroupement des costumes nécessite une amélioration afin de le rendre robuste aux regroupements de costumes similaires appartenant à différentes personnes. L'intégration du visage et de l'arrière plan en plus du costume peut minimiser les erreurs de regroupement. Des résultats préliminaires montrent que l'utilisation du visage en plus du costume améliore les résultats de regroupement de personnes portant des costumes similaires

mais introduit certaines erreurs de regroupement à cause de la variabilité d'apparence du visage.

La méthode de détection de costume utilisée est dépendante de la détection du visage. Les détecteurs de visages frontaux et de profils ne sont pas fiables dans le cas où les visages se présentent avec beaucoup de variabilité (mouvements brusques ou rapides, expressions faciales, etc). Il serait intéressant d'étudier une nouvelle méthode de détection des personnes indépendante du détecteur de visage. Dans un contexte de télévision, des modèles de plans peuvent être appris à partir des plans où la personne est bien détectée afin de retrouver les personnes dont le visage n'est pas visible. Une autre piste serait de développer des méthodes de détection automatique des costumes ou de silhouettes.

Enfin, notre méthode d'indexation des personnes basée sur les costumes fait l'hypothèse que chaque personne conserve les mêmes habits durant un même épisode de l'émission. Cette hypothèse est souvent vérifiée dans le cas des *Talk Shows* mais pas dans d'autres applications comme les films ou des séries télévisées. Dans ce cas, le costume est sujet à beaucoup de variabilités.

Concernant les modèles de vérification de l'identité visuelle, on peut envisager d'avoir plusieurs modèles du visage de manière à balayer ses principales variabilités dans ce contexte. Par exemple, apprendre pour chaque personne un modèle de visage de profil gauche/droit, un modèle frontal et un modèle pour chaque expression, etc. Malheureusement, cela nécessiterait d'avoir beaucoup de données.

La fusion

La technique de détection de présence de visages parlants que nous avons proposés est basée sur la détection de l'activité des lèvres. Cette mesure pourrait être améliorée par l'analyse de la synchronie entre la parole et le mouvement des lèvres. Dans un contexte de télévision, il est très difficile de détecter les lèvres avec une grande précision. Dans [Rúa et al., 2008], une méthode de mesure de synchronie qui ne nécessite pas une modélisation fine de la région des lèvres est proposée.

Concernant le processus de corrections des erreurs d'indexation basé sur les scores de vérification de l'identité audio-visuelle, un processus itératif d'apprentissage des modèles pourrait améliorer le taux de rappel du système d'indexation. Il s'agit de réapprendre de manière itérative les modèles audio et visuels de vérification de l'identité de chaque personne en intégrant les segments que le processus de correction

a permis de retrouver. Le schéma de correction peut être appliqué à nouveau afin de corriger d'éventuels segments.

Les modèles non supervisés des personnes

Notre système de structuration de contenus audio-visuels par personne nous a permis d'apprendre de manière non supervisée des modèles audio et visuels de personnes à partir des groupes homogènes de personnes détectées et regroupés automatiquement dans un même épisode de l'émission. Nous pouvons envisager d'utiliser ces modèles non supervisés afin de retrouver des personnes dans d'autres épisodes de l'émission.

Association d'une identité à un modèle de personne

Les personnes détectées et regroupées automatiquement par le système de structuration peuvent être associées à une identité de trois façons :

- **L'identification supervisée** consiste à associer une identité à chaque personne détectée en utilisant des données annotées manuellement. À partir de ces données, des modèles audio-visuels des personnes sont appris constituant ainsi un dictionnaire des participants à l'épisode de l'émission. Dans ce cas, une vérification de l'identité de chaque personne détectée automatiquement est effectuée afin de lui associer une identité.
- **L'identification semi-automatique** consiste à faire intervenir un annotateur humain afin d'associer une identité aux modèles des personnes obtenus de manière non supervisée. Cela revient à avoir un dictionnaire semi-automatique des participants à l'épisode de l'émission. Ce dictionnaire pourra ensuite être utilisé pour identifier des personnes dans le même épisode ou dans d'autres épisodes de l'émission.
- **L'identification automatique** consiste à associer une identité à chaque personne détectée sans utiliser de données annotées manuellement. Des informations supplémentaires susceptibles de contenir l'identité d'une personne peuvent être extraites automatiquement telles que le texte qui apparaît sur l'image (extraction par *OCR*[3]) ou la transcription de la parole. Souvent, dans ce type de contenus, les personnes participant à l'émission sont présentées en début de l'émission de manière orale (par le présentateur) et visuelle (par une incrustation textuelle).

3. Optical character recognition

Les applications futures

L'indexation audio-visuelle des personnes est un domaine relativement nouveau. De nombreuses applications apparaissent principalement sur Internet. Les chaînes de télévision proposent de revoir les émissions sur leurs sites. L'Institut National de l'Audio-visuel (*INA*) possède une très importante base de données de contenus télévisuels accessible. En plus, de nouveaux services privés proposent d'indexation et de mettre à disposition à des utilisateurs des contenus télévisuels tels qu'*EXALEAD*[4] ou *2424actu* par *Orange*[5] qui indexent et offrent plusieurs possibilités de navigation dans les journaux télévisés de plusieurs chaînes de télévision. Toutes ces nouvelles applications nécessitent de développer de nouvelles techniques d'indexation adaptées à ce contexte. S'il existe beaucoup de travaux académiques dans le domaine de l'indexation des personnes dans des contenus audio ou visuels de manière indépendante, beaucoup d'efforts restent à faire en indexation de contenus télévisés et multimédias en général. En 2010, l'Agence Nationale de la Recherche (*ANR*) a lancé un appel à projet intitulé REconnaissance de PERsonnes dans des Emissions audiovisuelles (*REPERE*) afin d'encourager la recherche dans ce domaine, et en particulier pour proposer une base de données publique commune à tous les chercheurs.

D'autres nouvelles applications peuvent être imaginées dans le domaine de la vidéos à la demande (*Video On Demand - VOD*) comme l'indexation des personnes dans les films ou séries télévisées. Dans ce contexte, plusieurs travaux de recherches existent tels que le projet *PittPatt*[6] ou *VisRec*[7] de l'université d'*Oxford*. Des efforts restent à faire dans le cadre des films d'animation où le challenge est de détecter et d'identifier des personnages déformables.

4. http://voxaleadnews.labs.exalead.com/
5. www.2424actu.fr
6. http://facemining.pittpatt.com
7. http://www.robots.ox.ac.uk/~vgg/projects/visrec

Annexe A

Le système de structuration sur le *Grand Échiquier*

Le *Grand Échiquier* est une émission diffusée par la chaîne de télévision française *Antenne 2*. Ce programme, présentée par *Jaques Chancel*, était diffusé de 1972 à 1986. Chaque épisode de l'émission durait environ $3h30$ durant lesquelles le présentateur recevait un invité spécial entouré d'invités secondaires et d'un public disposé autour des invités. La figure 8.11 montre des exemples de plans dans l'émission.

FIGURE 8.11 – Exemples de plans collectés dans le corpus *Grand Échiquier*

Un épisode de l'émission est structuré en plusieurs parties correspondant à des interviews de l'invité principal. Ces interviews sont entrecoupées de passages musicaux, d'extrait de film et d'interviews hors plateaux.

annotations

Nous avons annoté $30mn$ de l'épisode correspondant à *CPB84052346* (de $672s$ à $2385s$). Cinq personnes sont annotées en audio et en visuel à l'aide des outils

Transcriber et *Elan*. Le tableau 8.2 présente les statistiques des annotations du corpus *Grand Échiquier*.

	Durée totale (s)	*Durée moyenne (s)*	*# segments*
Parole	579	5.6	100
Apparence	1401	6.6	211
Visages parlants	510	5.6	91

TABLE 8.2 – Statistiques des annotations du corpus *Grand Échiquier*

La durée moyenne d'une intervention audio est du même ordre que celle des *Talks shows* dans le corpus *TSDB*. Par contre, la durée moyenne d'un segment de visage parlant est beaucoup plus grande dans le corpus *Grand Échiquier*. Cela s'explique par le fait que dans ces émissions, lorsqu'une personne prenait la parole, elle avait plus d'espace pour intervenir. Pour cette émission, lorsqu'une personne intervient, la probabilité qu'elle soit visible est de 90%. Le phénomène d'asynchronie entre les interventions audio avec les interventions visuelles est beaucoup moins important que dans le corpus *TSDB*. Par contre, le temps de parole d'un visage visible n'est que de 35% de la durée totale de l'apparence de ce visage dans l'épisode. Cela s'explique par tous les plans généreux ou une personne est filmée alors qu'elle ne parle pas.

Expériences

Nous avons testé nos méthodes de structurations présentées dans le chapitre 4 sur le corpus *Grand Échiquier*. De la même manière que pour la base de données *TSDB*, les résultats sont présenté par les mesures : taux de perte (*MDR*) et composition de la réponse automatique *CDR + EDR + FAR*, et par les mesures *Précision + Rappel + F-mesure* (voir chapitre 3).

Structuration par l'audio

Le système d'indexation en locuteur détecte 5 *Clusters* ce qui correspond au nombre de personnes intervenant dans la séquence évaluée. Dans le corpus *TSDB*, des *Clusters* supplémentaires sont détectés correspondants aux séquences de brouhaha, de double parole, ou de variabilité dans la façon de s'exprimer d'une même personne. Dans le corpus *Grand Échiquier*, la parole est beaucoup moins expressive et il y a très peu de séquence de double parole ou de brouhaha.

Clustering Audio	*CDR + EDR + FAR*	*MDR*	*Pr*	*Ra*	F_m
Évaluation complète	68.5 + 1.5 + 30.0	16.6	68.5	81.8	74.6
Évaluation restreinte	95.0 + 1.6 + 3.5	15.3	95.0	83.2	88.6

TABLE 8.3 – Évaluation des résultats de l'indexation en locuteurs sur le *Grand Échiquier*

Le tableau 8.3 présente les performances du système d'indexation en locuteurs selon la procédure d'évaluation *complète* et *restreinte*. Le taux *CDR* correspondant au temps correctement associé à la bonne personne audio. Ce taux est de 68.5% lorsque l'évaluation est effectuée sur toute la séquence audio annotée alors qu'il est de 95.0% lorsque l'évaluation est effectuée sur les séquences correspondant à de la parole. Le système détecte 30.0% de fausses alarme dans l'évaluation complète (segments dans lesquels une personne est détectée automatiquement alors qu'il n'y a personne qui parle). Ce taux s'améliore significativement lorsque l'évaluation est restreinte aux séquences de parole. Le taux de perte *MDR* est d'environ 16% pour les deux évaluations.

Comparé aux résultats obtenus dans le corpus *TSDB*, les performances du système d'indexation en locuteur sont nettement meilleures pour le grand Échiquier (+13% en précision, +7.4% en rappel).

Structuration par le costume

Clustering Video	*CDR + EDR + FAR*	*MDR*	*Pr*	*Ra*	F_m
Évaluation	85.2 + 14.8 + 0	31.5	85.2	58.3	69.3

TABLE 8.4 – Évaluation des résultats de la structuration par le costume sur le *Grand Échiquier*

Le tableau 8.4 présente les performances du système d'indexation basé sur le costume. Le taux *CDR* est de 85.2%. Le taux de perte *MDR* est d'environ 31%. Ce taux s'explique principalement par des segments dans lesquelles le visage n'a pas pu être détecté automatiquement. Le taux d'erreur est de 14.8%.

Comparé aux résultats obtenus dans le corpus *TSDB*, la précision du système d'indexation basé sur le costume est nettement meilleure dans le *Grand Échiquier* (+6.6%). Cette amélioration est due au fait que le costume est plus discriminant dans le *Grand Échiquier* (pas de costumes similaires). Par contre, le taux de rappel diminue

considérablement comparé au rappel obtenu dans le corpus *TSDB* (−12.5%). Cette détérioration est due aux segments d'apparence des personnes perdus pour cause de non-détection du visage.

Dans ce contexte, on a plus de mal à détecter les visages des personnes lorsqu'elles ne parlent pas. Cela s'explique par le fait que dans cette émission, il y a très peu de caméras. Les personnes sont disposées en rond. Lorsqu'une personne intervient, la caméra se focalise généralement sur elle. Les autres personnes autour sont visibles, mais ne sont pas face caméra.

Structuration audio-visuelle

Clustering Audio-visuel	*CDR + EDR + FAR*	*MDR*	*Pr*	*Ra*	F_m
Évaluation complète	68.9 + 1.8 + 29.3	30.0	68.9	68.2	68.6
Évaluation restreinte	95.7 + 1.8 + 2.5	28.0	95.6	70.5	81.2

TABLE 8.5 – Évaluation des résultats de la structuration audio-visuelle sur le *Grand Échiquier*

Le tableau 8.5 présente les résultats du système de structuration audio-visuelle selon la procédure d'évaluation *complète* et *restreinte*. Dans l'évaluation *complète*, une grande partie des segments de visages parlants n'est pas détectée (*MDR*= 30.0%). Ces erreurs peuvent provenir de deux sources : par la non-détection automatique de la personne par l'une des deux modalités ou une erreur d'étiquetage automatique de l'un des deux systèmes. Le taux de fausses alarmes *FAR* est très élevé (FAR = 29.3%). Ce taux s'explique principalement par l'impact des fausses alarmes produites par le système d'indexation en locuteur dans cette évaluation.

Dans l'évaluation *restreinte*, grâce à la suppression des segments audio ambigus, l'index de personnes obtenu par le système de structuration basé sur l'audio commet moins de fausses alarmes (voir les résultats de l'évaluation *restreinte* dans le tableau 8.3). Cette amélioration se reflète dans le taux *FAR* du système audio-visuel qui diminue significativement comparé au *FAR* de l'évaluation *complète* (−26.8%). Par contre, le taux de perte *MDR* reste très important. Ce taux s'explique par les erreurs de regroupement d'un des deux systèmes de structuration audio et visuels (ou les deux).

Comparé aux résultats de l'indexation des visages parlants dans le corpus *TSDB*, la précision du nettement meilleures pour le *Grand Échiquier* (+4.8%). Le taux de rappel est amélioré considérablement comparé au rappel obtenu dans le corpus *TSDB*

(+10%). Ces résultats s'expliquent principalement par le fait que notre système de regorgement audio-visuel commet moins d'erreurs.

Conclusion

Le *système initial* développé pour les contenus de type *Talk Shows*, marche convenablement pour le *Grand Échiquier*. Cela démontre la robustesse de notre approche dans un contexte de télévision. Les résultats de la structuration obtenus sont nettement meilleurs que ceux obtenus par le *système initial* dans le corpus *TSDB*. Dans le *Grand Échiquier*, ces améliorations s'expliquent principalement par les dialogues moins interactifs.

Annexe B

Évaluation de l'outil *Stasm* sur *BioID*

Dans cette partie, nous souhaitons évaluer le logiciel de détection de caractéristiques du visage *Stasm* sur des visages frontaux et non expressifs.

Base de données *BioID*

BioID est une base de données publique de visages[8] . Les images (384 × 288 pixel) sont capturés en noir et blanc sur 23 personnes. Au total, 1521 visages sont enregistrés. Les personnes se présentent face caméra avec une expression neutre. Pour chaque visage, des points clés des caractéristiques de visages sont annotés manuellement :

- 4 points sur les lèvres.
- 3 points sur chaque œil.
- 2 points sur chaque sourcil.
- 3 points sur le nez.
- 3 point sur le contour du visage.

Au total, 20 points sont placés manuellement sur chacun des 1521 visages. La figure 8.12 montre un exemple des annotations effectuée sur un visage de la base de données *BioID*.

L'outil *Stasm*

Nous avons évalué le détecteur de caractéristiques du visage *Stasm*[9] basé sur des modèles actifs de forme (Active Shape Model - ASMs) décrit dans [Cootes et al., 1995]. La figure 8.13 montre un exemple d'annotations automatiques obtenues par l'outil *Stasm* sur un visage de la base de données *BioID*. Cet outils permet de détecter 68 points du visage :

- 5 points sur chaque œil (coins gauche et droit, ouverture sur le dessus et le dessous des yeux et le centre de l'iris).

8. www.bioid.com/research/index.html
9. http://www.milbo.users.sonic.net/stasm/

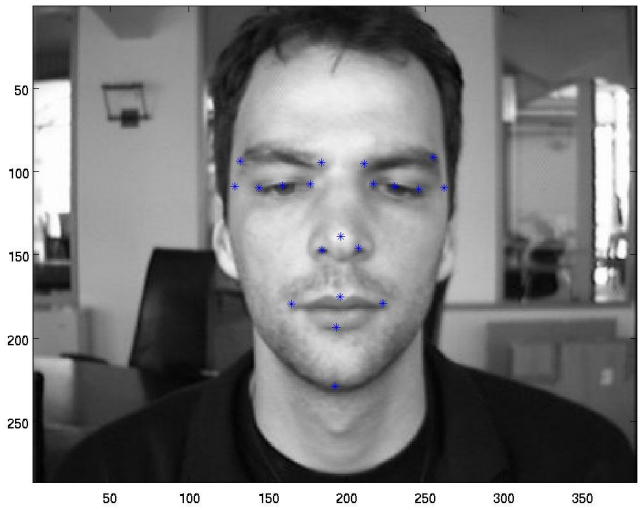


FIGURE 8.12 – Exemples de visage de la base de données *BioID* avec les 20 points annotés manuellement

- 6 points sur chaque sourcil.
- 19 points sur le contour des lèvres (18 points pour les lèvres supérieures et inférieure et un point représentant le centre de la bouche).
- 12 points sur le nez.
- 15 point sur le contour du visage.

Évaluation

Pour un annotateur humain, il est difficile de positionner les caractéristiques du visage avec beaucoup de précision. Dans ce cas, l'évaluation objective des résultats de la détection automatique de caractéristiques du visage devient compliquée. Nous avons défini la notion de bonne détection de point basée sur sa distance par rapport au point annoté. Dans la base de données *BioID*, un point est considéré bien détecté lorsqu'il est à une distance de 4 pixels de l'annotation manuelle, la taille d'une image étant de 384×288 pixel. Donc un point mal localisé se trouve à plus de 4 pixels du point annoté manuellement. Pour chaque caractéristique du visage, on définit l'erreur E_i qui représente le pourcentage d'images dans lesquelles le caractéristique est détecté avec i points mal localisés.

Soit N le nombre de points annotés sur le caractéristique du visage. L'erreur générale sur ce caractéristique (notée E) est calculée par combinaison linéaire des erreurs E_i de la manière suivante :

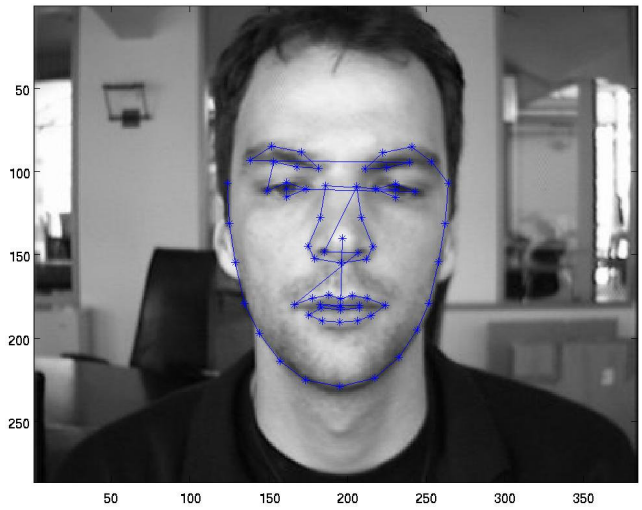


FIGURE 8.13 – Exemples de visage de la base de données *BioID* avec les 68 points détectés par *Stasm*

$$E = \sum_{i=1}^{N} i \times E_i$$

Dans l'évaluation, nous avons sélectionné les points détectées par l'outil *Stasm* qui correspondent aux points annotés manuellement : 4 points pour les lèvres, 3 points pour chaque œil, points pour chaque sourcil, 3 points pour le nez et les 1 points pour le menton.

Expérience

Discussion

À part pour le menton, le pourcentage des caractéristiques dans lesquels tous les points sont mal détectés est très faibles.

Les yeux En moyenne, 95.6% des yeux sont correctement détectés sans aucune erreur et seulement 0.15% des yeux sont mal détectés. De manière subjective, ces résultats sont très acceptables.

Caractéristiques	E_0	E_1	E_2	E_3	E_4	E
Lèvres	55.2	27.3	10.8	4.9	1.9	7.1
Œil gauche	96.5	2.9	0.3	0.3	–	7.1
Œil droit	94.7	4.0	1.3	0	–	1.1
Nez	85.4	11.0	2.4	1.2	–	3.2
Sourcil gauche	72.6	23.9	3.5	–	–	10.3
Sourcil droit	70.2	24.9	4.9	–	–	11.5
Menton	82.9	17.1	–	–	–	17.1

TABLE 8.6 – Performances du système d'indexation avec la mesure d'activité des lèvres dans la base de données *TVSDB*

Les sourcils Pour les sourcils, le point intérieur annoté manuellement est diffèrent du point détecté automatiquement. Dans l'annotation manuelle, l'annotateur a localisé l'intérieur haut du sourcil alors que l'annotation automatique détecte le point intérieur bas. Cela explique pourquoi seulement en moyenne 71.2% des sourcils sont détectés sans aucune erreur. Par contre, seulement $4,2\%$ des sourcils sont mal détectés par les deux points.

Le menton Le menton est considéré mal détecté dans 17.1% des visages. Le détecteur *Stasm* détecte un point extrême sud du visage. Le menton étant un peu carré, il est difficile de comparer objectivement l'annotation manuelle et automatique. En moyenne, 82% du menton est correctement détecté sans aucune erreur.

Les lèvres Pour les lèvres, seulement 55.2% sont correctement détectés sans aucune erreur. Ce taux relativement bas monte à 82.5% dès que l'on autorise une seule erreur dans la détection automatique (sur les 4 points) et 93.3% pour deux erreurs. Ces résultats peuvent s'expliquer par deux raisons : d'abord, l'annotation manuelle des lèvres n'est pas très précise (en particulier sur le point de la lèvre inférieure). Aussi, comparé aux autres caractéristiques du visage, les lèvres sont sujettes à beaucoup de distorsions même sur des visages neutres. Le modèle de visage *ASM* appris dans *Stasm* ne prend pas en compte toutes les distorsions des lèvres. Ceci pourrait expliquer que les lèvres ne soient pas détectées avec beaucoup de précision. Ceci dit, seulement 1.9% des lèvres sont mal détectés par les 4 points. Donc la localisation des lèvres est souvent correcte, mais le détecteur n'est pas tout le temps précis sur la détection du contour des lèvres.

Liste des publications

Liste des publications

- M. Bendris, D. Charlet, and G. Chollet. Introduction of quality measures in audio-visual identity verification. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Taipei, Taiwan. 2009.
- M. Bendris. Introduction of indexing people problematic in TV-Content. *Seminar on Information, Signal, Images et Vision : Indexation scalable et Cross Media*. Paris, France. 2009.
- M. Bendris, D. Charlet, and G. Chollet. Talking Faces indexing in TV-Content. *International Workshop on Content-Based Multimedia Indexing (CBMI)*. Grenoble, France. 2010.
- M. Bendris, D. Charlet, and G. Chollet. Lip activity detection for talking faces classification in TV-Content. *International Conference in Machine Vision (ICMV)*. Hong Kong, China. 2010.
- T. Baerecke, M. Bendris, M. Campedel, M. Detyniecki, D. Marraud. Traitement des modalités « image » et « vidéo ». *Chapter Book : Sémantique et Multimodalité en analyse de l'information. Infom@gic* Project. 2011.
- J. Carrive, J. Razik, M. Bendris, S. Vanni, L. Rigouste, D. Marraud, G. Chollet, C. Brun. Technologies d'indexation pour la valorisation du patrimoine audiovisuel. *Chapitre du livre : Sémantique et Multimodalité en analyse de l'information. Infom@gic* Project. 2011.
- M. Bendris, D. Charlet, and G. Chollet. People indexing in TV-Content using lip-activity and unsupervised audio-visual identity verification. *International Workshop on Content-Based Multimedia Indexing (CBMI)*. Madrid, Spain. 2011.

Liste des tableaux

2.1 Statistiques générales des 5 épisodes du corpus *TSDB* (temps en secondes) . . . 50
2.2 Analyse de corpus - Durée (en seconde) des plans d'apparition et tours de parole pour chaque personne dans l'épisode *S1* . . . 51

4.1 Regroupement de locuteur par le critère *BIC* - Nombre de *Clusters* audio détectés automatiquement dans les épisodes de la base de données *TSDB* . . . 70
4.2 *Évaluation complète* des résultats de l'indexation en locuteurs sur les 5 épisodes de la base de données *TSDB*. L'épisode *S*3 obtient des taux particulièrement bas . . . 70
4.3 *Évaluation restreinte* des résultats de l'indexation en locuteurs sur les 5 épisodes de la base de données *TSDB* . . . 71
4.4 Regroupement des costumes - Nombre de *Clusters* visuels détectés dans les épisodes de la base de données *TSDB* . . . 75
4.5 Évaluation des résultats de la structuration par le costume sur les 5 épisodes de la base de données *TSDB* . . . 75
4.6 *Évaluation complète* en audio des résultats de la structuration audio-visuelle sur les 5 épisodes de la base de données *TSDB* . . . 78
4.7 *Évaluation restreinte* en audio des résultats de la structuration audio-visuelle sur les 5 épisodes de la base de données *TSDB* . . . 79
4.8 *Évaluation restreinte* en audio des résultats de la structuration audio-visuelle par appariement d'index selon la méthode [Khoury et al., 2010] sur les 5 épisodes de la base de données *TSDB* . . . 80
4.9 Composition du taux d'erreur *MDR* sur les 5 épisodes de la base de données *TSDB* . . . 82

5.1 Performances du système de structuration basé sur la mesure d'activité des lèvres dans la base de données *TVSDB* . . . 97

7.1 Performances du système d'indexation des visages-parlants dans les différents schémas de modification sur la base de données *TSDB* . . . 129

8.1 Performances des systèmes de vérification de l'identité dans *Banca*. Les taux *EER* obtenus par notre systèmes de vérification du locuteur (*Speaker*), le système *BECARS* (*SpeakerBecars*), le visage (*Face*) et la fusion par somme pondérée (*SumFusion*) 137
8.2 Statistiques des annotations du corpus *Grand Échiquier* 156
8.3 Évaluation des résultats de l'indexation en locuteurs sur le *Grand Échiquier* 157
8.4 Évaluation des résultats de la structuration par le costume sur le *Grand Échiquier* 157
8.5 Évaluation des résultats de la structuration audio-visuelle sur le *Grand Échiquier* 158
8.6 Performances du système d'indexation avec la mesure d'activité des lèvres dans la base de données *TVSDB* 164

Liste des acronymes

AAM	Modèles actifs d'apparence - *Active Appearance Models*
ACP	Analyse en Composantes Principales - *Principal Component Analysis*
ANR	Agence Nationale de la Recherche
ASM	Modèles actifs de forme - *Active Shape Models*
BIC	*Bayesian Information Criterion*
CLR	Rapport de vraisemblance croisé - *Cross Likelihood Ratio*
DBN	Réseaux bayésiens dynamiques - *Dynamic Bayesian Network*
DCF	Fonction de coût de détection - *Detection Cost Function*
DET	Courbe de détection - Detection Error Tradeoff
DFFS	Distance à l'espace de visage - Distance From Face Space
EER	Taux d'égale erreur - *Equal Error Rate*
EM	Espérance et maximisation - *Expectation Maximization*
FAR	Taux de fausses acceptations - *False Acceptation Rate*
FRR	Taux de faux rejets - *False Rejection Rate*
GLR	Rapport de vraisemblance généralisé - *Generalized Likelihood Ratio*
GMM	Modèle de mélange de gaussiennes - *Gaussian Mixture Model*
HMM	Modèle de Markov caché - *Hidden Markov Model*
INA	Institut National de l'Audiovisuel
JT	Journal Télévisé
KL	Distance *Kullback-Leibler*
MAP	Maximum à postériori - *Maximum A Posteriori*
MBGC	*Multiple Biometric Grand Challenge*
MFCC	*Mel-Frequency Cepstral Coefficients*
MSD	*Mean Squared Distance*
REPERE	REconnaissance de PERsonnes dans des Émissions audiovisuelles
SIFT	Transformation de caractéristiques visuelles invariante à l'échelle - *Scale-Invariant Feature Transform*
SVM	Machine à vecteurs de support - *Support Vector Machine*
TSDB	Base de données d'émissions de plateaux - *Talk Shows DataBase*
UBM	Modèle du monde - *Universal Background Model*

Bibliographie

[Aarabi, 2003] Aarabi, P. (2003). The fusion of distributed microphone arrays for sound localization. *EURASIP Journal on Applied Signal Processing*, 4 :338–347.

[Acosta et al., 2002] Acosta, E., Torres, L., Albiol, A., and Delp, E. (2002). An automatic face detection and recognition system for video indexing applications. In *Internationl conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages 3644–3647.

[Arandjelovic and Zisserman, 2005] Arandjelovic, O. and Zisserman, A. (2005). Automatic face recognition for film character retrieval in feature-length films. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 860–867. IEEE Computer Society.

[Bailly and Milgram, 2009] Bailly, K. and Milgram, M. (2009). Head pan angle estimation by a nonlinear regression on selected features. *International Conference on Computer Analysis of Images and Patterns (CAIP)*.

[Bailly-Bailliére et al., 2003] Bailly-Bailliére, E., Bengio, S., Bimbot, F., Hamouz, M., Kittler, J., Mariéthoz, J., Matas, J., Messer, K., Popovici, V., and Porée, F. (2003). The BANCA Database and Evaluation Protocol. In *4th International Conference on Audio- and Video-Based Biometric Person Authentication, Guildford*, volume 2688, pages 625–638. Springer.

[Barras et al., 2006] Barras, C., Zhu, X., Meignier, S., and Gauvain, J.-L. (2006). Multistage speaker diarization of broadcast news. *IEEE Transactions On Audio Speech And Language Processing*, 14 :1505–1512.

[Blouet et al., 2004] Blouet, R., Mokbel, C., and Chollet, G. (2004). Becars : a free software for speaker verification. *ODYSSEY - The Speaker and Language Recongnition Workshop. Toledo, Spain*, pages 145–148.

[Boccignone et al., 2005] Boccignone, G., Chianese, A., Moscato, V., and Picariello, A. (2005). Foveated shot detection for video segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 15 :365–377.

[Boreczky and Rowe, 1996] Boreczky, J. S. and Rowe, L. A. (1996). Comparison of video shot boundary detection techniques. *Journal of Electronic Imaging*, 5 :122.

[Bourel et al., 2000] Bourel, F., Chibelushi, C. C., and Low, A. A. (2000). Robust facial feature tracking. In *11th British Machine Vision Conference (BMVC). Bristol, UK*, pages 232–241.

[Bowen and Hansen, 2005] Bowen, Z. and Hansen, J. (2005). Efficient audio stream segmentation via the combined t2 statistic and bayesian information criterion. *IEEE Transactions on Audio, Speech and Language Processing*, 13 :467–474.

[Bredin, 2007] Bredin, H. (2007). *Verification de l'identite d'un visage parlant. Apport de la mesure de synchronie audiovisuelle fac aux tentatives deliberees d'imposture*. PhD thesis, Telecom ParisTech. Paris, France.

[Bredin et al., 2006] Bredin, H., Aversano, G., Mokbel, C., and Chollet, G. (2006). The biosecure talking-face reference system. In *2nd Workshop on Multimodal User Authentication*.

[Bregler et al., 2009] Bregler, C., Williams, G., Rosenthal, S., and McDowall, I. (2009). Improving acoustic speaker verification with visual body-language features. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1909–1912. IEEE Computer Society.

[Carletta et al., 2005] Carletta, J., Ashby, S., Bourban, S., Flynn, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., and Reidsma, M. W. P. D. (2005). The ami meeting corpus : A pre-announcement. *Lecture Notes in Computer Science*, 3869 :28–39.

[Cernekova et al., 2003] Cernekova, Z., Kotropoulos, C., and Pitas, I. (2003). Video shot segmentation using singular value decomposition. *International Conference on Multimedia and Expo*, 1 :301–304.

[Chen and Gopalakrishnan,] Chen, S. S. and Gopalakrishnan, P. S. Speaker, environment and channel change detection and clustering via the bayesian information criterion. *Proc DARPA Broadcast News Transcription and Understanding Workshop*, pages 127–132.

[Chiang et al., 2003] Chiang, C.-C., Tai, W.-K., Yang, M.-T., Huang, Y.-T., and Huang, C.-J. (2003). A novel method for detecting lips, eyes and faces in real time. *Real-Time Imaging*, 9 :277–287.

[Chibelushi et al., 1997] Chibelushi, C. C., Mason, J., and Deravi, F. (1997). Feature-level data fusion for bimodal person recognition. *6th International Conference on Image Processing and its Applications*, 1 :399–403.

[Comaniciu et al., 2003] Comaniciu, D., Ramesh, V., and Meer, P. (2003). Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25 :564–577.

[Cootes et al., 2001] Cootes, T. F., Edwards, G. J., and Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23 :681–685.

[Cootes et al., 1995] Cootes, T. F., Taylor, C. J., Cooper, D. H., and Graham, J. (1995). Active shape models-their training and application. *Computer Vision and Image Understanding. New York, USA*, 61 :38–59.

[Cour et al., 2009] Cour, T., Sapp, B., Jordan, C., and Taskar, B. (2009). Learning from ambiguously labeled images. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 919–926.

[Dai and Nakano, 1996] Dai, Y. and Nakano, Y. (1996). Face-texture model based on sgld and its application in face detection in a color scene. *Pattern Recognition*, 29 :1007–1017.

[Dalal and Triggs, 2004] Dalal, N. and Triggs, W. (2004). Histograms of oriented gradients for human detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). San Diego, USA*, 1 :886–893.

[Deléglise et al., 2005] Deléglise, P., Estève, Y., Meignier, S., and Merlin, T. (2005). The lium speech transcription system : a cmu sphinx iii-based system for french broadcast news. *International Speech Communication Association (Interspeech). Lisbon, Protugal*.

[Dempster et al., 1977] Dempster, A. P., Laird, N., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *The Royal Statistical Society Series B Methodological*, 39 :1–38.

[Duffner and Garcia, 2005] Duffner, S. and Garcia, C. (2005). A connexionist approach for robust and precise facial feature detection in complex scenes. *4th International Symposium on Image and Signal Processing and Analysis (ISPA). Zagreb, Croatia*.

[Dupont and Luettin, 2000] Dupont, S. and Luettin, J. (2000). Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*.

[Eickeler et al., 2001] Eickeler, S., Wallhoff, F., Iurgel, U., and Rigoll, G. (2001). Content-based indexing of images and video using face detection and recognition methods. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 3 :1505–1508.

[Englebienne et al., 2009] Englebienne, G., Kröse, B. J., and Noulas, T. K. (2009). Multimodal speaker diarization. *Computer Vision and Image Understanding*, pages 1–34.

[Everingham et al., 2006] Everingham, M., Sivic, J., and Zisserman, A. (2006). Hello ! my name is ... buffy " – automatic naming of characters in tv video. *The British Machine Vision Conference (BMVC)*, 3 :1–10.

[Fasel et al., 2005] Fasel, I., Fortenberry, B., and Movellan, J. (2005). A generative framework for real time object detection and classification. *Computer Vision and Image Understanding*, 98 :182–210.

[Féraud et al., 2001] Féraud, R., Bernier, O. J., Viallet, J.-E., and Collobert, M. (2001). A fast and accurate face detector based on neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23 :42–53.

[Ferrari et al., 2008] Ferrari, V., Marin-Jimenez, M., and Zisserman, A. (2008). Progressive search space reduction for human pose estimation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Alaska, USA*, 2 :1–8.

[Fierrez-Aguilar et al., 2005] Fierrez-Aguilar, J., Ortega-garcia, J., Gonzalez-rodriguez, J., and Bigun, J. (2005). Discriminative multimodal biometric authentication based on quality measures. *Pattern Recognition*, 38 :777–779.

[Galliano et al., 2009] Galliano, S., Gravier, G., and Chaubard, L. (2009). The ESTER2 evaluation campaign for the rich transcription of French radio broadcast. *10th Annual Conference of the International Speech Communication Association (Interspeech). Brighton, United Kingdom.*

[Garcia and Delakis, 2002] Garcia, C. and Delakis, M. (2002). A neural architecture for fast and robust face detection. *Object recognition supported by user interaction for service robots*, 0 :44–47.

[Garofolo et al., 2004] Garofolo, J. S., Laprun, C. D., Michel, M., Stanford, V. M., and Tabassi, E. (2004). The nist meeting room pilot corpus. *4th Conference on Language Resources and Evaluation (LREC). Lisbon, Portugal.*

[Genoud et al., 1996] Genoud, D., Gravier, G., Bimbot, F., and Chollet, G. (1996). Combining methods to improve speaker verification decision. In *Proceedings of International Conference on Spoken Language Processing*, volume 3, pages 1756–1759.

[Geoffrois et al., 2006] Geoffrois, E., Gravier, G., Bonastre, J.-F., Mostefa, D., Choukri, K., and Galliano, S. (2006). Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. *5th international conference on Language Resources and Evaluation (LREC). Genoa, Italy.*

[Goldberger and Roweis, 2004] Goldberger, J. and Roweis, S. (2004). Hierarchical clustering of a mixture model. *Advances in Neural Information Processing Systems*, pages 505–512.

[Hall and Llinas, 1997] Hall, D. L. and Llinas, J. (1997). An introduction to multisensor data fusion. *Proceedings of the IEEE*, 85 :6–23.

[Heckmann et al., 2001] Heckmann, M., Berthommier, F., and Kroschel, K. (2001). A hybrid ann/hmm audio-visual speech recognition system. *International Conference on AuditoryVisual Speech Processing Proceedings (AVSP).*

[Heracleous et al., 2010] Heracleous, P., Badin, P., Bailly, G., and Hagita, N. (2010). Exploiting multimodal data fusion in robust speech recognition. *International Conference on Multimedia and Expo (ICME). Singapour.*

[Jaffré, 2005] Jaffré, G. (2005). *Indexation de la vidéo par le costume.* PhD thesis, University of Paul Sabatier France.

[Jaffre and Joly, 2004] Jaffre, G. and Joly, P. (2004). Costume : A new feature for automatic video content indexing. In *Coupling approaches, coupling media and coupling languages for information retrieval (RIAO 2004). Avignon*, pages 314–325.

[Jaffré et al., 2007] Jaffré, G., Pinquier, J., Senac, C., and Khoury, E. E. (2007). Association of audio and video segmentations for automatic person indexing. *International Workshop on Content-Based Multimedia Indexing. Bordeaux, France.*

[Johnson, 1967] Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32 :241–254.

[Khoury, 2010] Khoury, E. E. (2010). *Unsupervised Video Indexing based on Audiovisual Characterization of Persons.* PhD thesis, University of Paul Sabatier. Toulouse, France.

[Khoury et al., 2010] Khoury, E. E., Senac, C., and Joly, P. (2010). Face-and-clothing based people clustering in video content. In *International conference on Multimedia information retrieval (MIR). Philadelphia, USA*, pages 295–304.

[Knox and Friedland, 2010] Knox, M. T. and Friedland, G. (2010). Multimodal speaker diarization using oriented optical flow histograms. *International Conference of the International Speech Communication Association (Interspeech)*, pages 290–293.

[Kryszczuk et al., 2005] Kryszczuk, K., Richiardi, J., Prodanov, P., and Drygajlo, A. (2005). Error handling in multimodal biometric systems using reliability measures. *13th European Signal Processing (EUSIPCO). Istanbul, Turkey*, pages 4–8.

[Li et al., 2005] Li, D., Sang, L., Yang, Y., and Wu, Z. (2005). Bimodal speaker identification using dynamic bayesian network. *Advances in Biometric Person Authentication*, 3338 :1–24.

[Li et al., 2008] Li, Y., Zhang, H., and Wang, L. (2008). Multi-modal biometric verification based on far-score normalization. *International Journal of Computer Science and Network Security (IJCSNS)*, 8 :250–254.

[Luhong et al., 2000] Luhong, L., Haizhou, A., and Xu, G. (2000). Face detection based on template matching and neural network verification. *International Conference on Image*.

[Matthews and Baker, 2004] Matthews, I. and Baker, S. (2004). Active appearance models revisited. *International Journal of Computer Vision*, 60 :135–164.

[McKenna et al., 1998] McKenna, S. J., Gong, S., and Raja, Y. (1998). Modelling facial colour and identity with gaussian mixtures. *Pattern Recognition*, 31 :1883–1892.

[Meignier et al., 2001] Meignier, S., françois Bonastre, J., and Igounet, S. (2001). E-hmm approach for learning and adapting sound models for speaker indexing. *A Speaker Odyssey The Speaker Recognition Workshop*, pages 175–180.

[Messer et al., 1999] Messer, K., Matas, J., Kittler, J., uergen Luettin, and Maître, G. (1999). XM2VTSDB : The Extended M2VTS Database. In *Second International Conference on Audio and Video-based Biometric Person Authentication*, pages 72–77.

[Milborrow and Nicolls, 2008] Milborrow, S. and Nicolls, F. (2008). Locating facial features with an extended active shape model. In *10th European Conference on Computer Vision : Part IV. Marseille, France*, pages 504–513.

[Monaci et al., 2006] Monaci, G., Pierre, V., Mailhe, B., Lesage, S., and Gribonval, R. (2006). Learning multimodal dictionaries : Application to audiovisual data. *International Workshop on Multimedia Content Representation, Classification and Security (MRCS). Istanbul, Turkey*, pages 538–545.

[Nefian et al., 2003] Nefian, A. V., Liang, L. H., Fu, T., and Liu, X. X. (2003). A bayesian approach to audio-visual speaker identification. *4th international conference on Audio- and video-based biometric person authentication (AVBPA). Guildford, UK*, pages 761–769.

[Ogale and Aloimonos, 2005] Ogale, A. S. and Aloimonos, Y. (2005). Shape and the stereo correspondence problem. *International Journal of Computer Vision*, 65 :147–162.

[Petrovska-Delacrétaz et al., 2009] Petrovska-Delacrétaz, D., Chollet, G., and Dorizzi, B. (2009). Guide to biometric reference systems and performance evaluation. *Springer*.

[Phillips et al., 2009] Phillips, J. P., Flynn, P. J., Beveridge, R. J., Scruggs, T. W., O'Toole, A. J., Bolme, D. S., Bowyer, K. W., Draper, B. A., Givens, G. H., Lui, Y. M., Sahibzada, H., Scallan, J. A., and Weimer, S. (2009). Overview of the multiple biometrics grand challenge. In *Third International Conference on Advances in Biometrics (ICB). Alghero, Italy*, volume 5558, pages 705–714. Springer.

[Poh and Bengio, 2005] Poh, N. and Bengio, S. (2005). Improving fusion with margin-derived confidence in biometric authentication tasks. *Audio and video-based biometric person authentication (AVBPA). NY, USA*, 3546 :474–483.

[Poh et al., 2007] Poh, N., Kittler, J., and Fatukasi, O. (2007). Quality controlled multimodal fusion of biometric experts. *12th Iberoamerican Congress on Pattern Recognition (CIARP). Viña del Mar-Valparaiso, Chile*, pages 881–890.

[Radova and Psutka, 1997] Radova, V. and Psutka, J. (1997). An approach to speaker identification using multiple classifiers. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Washington DC, USA*, volume 2, pages 1135–1138.

[Reynolds et al., 2000] Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10 :19–41.

[Reynolds et al., 1998] Reynolds, D. A., Singer, E., ans Gerald C. O'Leary ans Jack J. McLaughlin, B. A. C., and Zissman, M. A. (1998). Blind clustering of speech utterances based on speaker and lan- guages characteristics. *5th International Conference on Spoken Language Processing (ICSLP).*

[Reynolds and Torres-Carrasquillo, 2005] Reynolds, D. A. and Torres-Carrasquillo, P. A. (2005). Approaches and applications of audio diarization. *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP).*, 5 :953–956.

[Richiardi et al., 2006] Richiardi, J., Prodanov, P., and Drygajlo, A. (2006). Confidence and reliability measures in speaker verification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toulouse, France*, volume 343, pages 574–595. Elsevier.

[Rúa et al., 2008] Rúa, E. A., Bredin, H., Mateo, C. G., Chollet, G., and Jiménez, D. G. (2008). Audio-visual speech asynchrony detection using co-inertia analysis and coupled hidden markov models. *Pattern Analysis and Applications*, 12 :271–284.

[Saenko et al., 2005] Saenko, K., Livescu, K., Siracusa, M., Wilson, K., and Darrell, J. G. T. (2005). Visual speech recognition with loosely synchronized feature streams. *10th International Conference on Computer Vision (ICCV). Beijing, China*, 1 :1424–1431.

[Sanderson and Paliwal, 2004] Sanderson, C. and Paliwal, K. K. (2004). Identity verification using speech and face information. *Digital Signal Processing*, 14 :449–480.

[Shahraray, 1995] Shahraray, B. (1995). Scene change detection and content-based sampling of video sequences. *Digital video compression : algorithms and technologies*, 2419 :2–13.

[Siegler et al., 1997] Siegler, M. A., Jain, U., Raj, B., and Stern, R. M. (1997). Automatic segmentation, classification and clustering of broadcast news audio. *DARPA Speech Recognition Workshop. Virginia, USA*, 2 :97–99.

[Silsbee and Bovik, 1996] Silsbee, P. L. and Bovik, A. C. (1996). Computer lipreading for improved accuracy in automatic speech recognition. *IEEE Transactions On Speech And Audio Processing*, 4 :337–351.

[Smeaton et al., 2010] Smeaton, A. F., Over, P., and Doherty, A. R. (2010). Video shot boundary detection : Seven years of trecvid activity. *Computer Vision and Image Understanding*, 114 :411–418.

[Solomonoff et al., 1998] Solomonoff, A., Mielke, A., Schmidt, M., and Gish, H. (1998). Clustering speakers by their voices. *International Conference on Acoustics Speech and Signal Processing (ICASSP). New Jersey, USA*, 2 :757–760.

[Spors and Rabenstein, 2001] Spors, S. and Rabenstein, R. (2001). A real-time face tracker for color video. *International Conference on Acoustics Speech and Signal Processing (ICASSP). Utah, USA*, 3 :1493–1496.

[Sue E. Johnson, 1998] Sue E. Johnson, P. C. W. (1998). Speaker clustering using direct maximisation of the mllr-adapted likelihood. *5th International Conference on Spoken Language Processing (ICSLP). Sydney, Australia*, 5 :1775–1779.

[Turk and Pentland, 1991] Turk, M. and Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3 :71–86.

[Vajaria et al., 2008] Vajaria, H., Sarkar, S., and Kasturi, R. (2008). Exploring cooccurence between speech and body movement for audio-guided video localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 18 :1608–1617.

[Vallet et al., 2010] Vallet, F., Essid, S., Carrive, J., and Richard, G. (2010). Robust visual features for the multimodal identification of unregistered speakers in tv talk-shows. *International Conference on Image Processing (ICIP). Hong Kong, China.*

[Verlinde et al., 2000] Verlinde, P., Chollet, G., and Acheroy, M. (2000). Multi-modal identity verification using expert fusion. *Information Fusion*, 1 :17–33.

[Viola and Jones, 2001] Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 1 :511–518.

[Ward, 1963] Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58 :236–244.

[Wooters and Huijbregts, 2008] Wooters, C. and Huijbregts, M. (2008). Multimodal technologies for perception of humans. pages 509–519. Springer-Verlag.

[Yang et al., 2002] Yang, M.-H., Kriegman, D. J., and Ahuja, N. (2002). Detecting faces in images : A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24 :34–58.

[Zhang et al., 1993] Zhang, H., Kankanhalli, A., and Smoliar, S. W. (1993). Automatic partitioning of full-motion video. *Multimedia Systems*, 1 :10–28.

[Zhu et al., 2005] Zhu, X., Barras, C., Meignier, S., and Gauvain, J.-L. (2005). Combining speaker identification and bic for speaker diarization. *International Speech Communication Association (Interspeech). Lisbon, Portugal*.

[Zhu, 2005] Zhu, Z. (2005). Mosaic-based 3d scene representation and rendering. In *11th InternationalConference on Image Processing*, volume 21, pages 739–754.

Résumé

Le développement et l'amélioration du réseau Internet a permis de mettre un grand nombre de contenus télévisuels à disposition des utilisateurs. Afin de faciliter la navigation parmi ces vidéos, il est intéressant de développer des technologies pour indexer les personnes automatiquement. Les solutions actuelles proposent de construire l'index audio-visuel des personnes par combinaison des index audio et visuel obtenus de manière indépendante. Malheureusement, pour les émissions de télévision, il est difficile de détecter et de regrouper les personnes automatiquement à cause des nombreuses ambiguïtés dans l'audio, le visuel et leur association (interactivité des dialogues, variations de pose du visage, asynchronie entre la parole et l'apparence, etc). Les approches basées sur la fusion des index audio et visuel combinent les erreurs d'indexation issues de chaque modalité.

Les travaux présentés dans ce rapport exploitent la complémentarité entre les informations audio et visuelle afin de palier aux faiblesses de chaque modalité. Ainsi, une modalité peut appuyer l'indexation d'une personne lorsque l'autre est jugée peu fiable. Nous proposons une procédure de correction mutuelle des erreurs d'indexation de chaque modalité. D'abord, les erreurs sont détectées automatiquement à l'aide d'indicateurs de présence de visage parlant. Puis, la modalité qui a échoué est corrigée grâce à un schéma automatique.

Nous avons proposé en premier lieu un *système initial* d'indexation de visages parlants basé sur la détection et le regroupement du locuteur et du costume. Nous proposons une méthode de combinaison d'index basée sur la maximisation de la couverture globale des groupes de personnes. Ce système, évalué sur des émissions de plateaux, obtient une grande précision ($\sim 90\%$), mais un faible rappel (seulement 55% des visages parlants sont détectés).

Afin de détecter automatiquement la présence d'un visage parlant dans le processus de correction mutuelle, nous avons développé une nouvelle méthode de détection de mouvement des lèvres basée sur la mesure du degré de désordre de la direction des pixels autour de la région des lèvres. L'évaluation, réalisée sur le corpus de d'émission de plateaux, montre une amélioration significative de la détection des visages parlants comparé à l'état de l'art dans ce contexte. En particulier, notre méthode s'avère être plus robuste à un mouvement global du visage.

Enfin, nous avons proposé deux schémas de correction. Le premier est basé sur une modification systématique de la modalité considérée *a priori* la moins fiable. Le second compare des scores de vérification de l'identité non supervisée afin de déterminer quelle modalité a échoué et la corriger. Les modèles non supervisés des personnes sont appris à partir des ensembles homogènes de visages parlants obtenus automatiquement par le *système initial*. Les deux méthodes de correction conduisent à une amélioration significative des performances (+2 à 5% de la *F-mesure*).

Nous nous sommes également intéressé aux systèmes biométriques audio-visuels et particulièrement sur les techniques de fusion tardives pour la vérification d'identité. Nous avons proposé une méthode de fusion dépendante de la qualité du signal dans chaque modalité.