Modélisation non-supervisée de signaux sociaux

Stéphane Michelet

THÈSE DE DOCTORAT DE
l'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité

Informatique

École doctorale
Sciences mécaniques, acoustique, électronique & robotique de Paris

Présentée par

Stéphane MICHELET

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

Modélisation non-supervisée de signaux sociaux

soutenue le 10 Mars 2016 devant un jury composé de :

Rapporteurs	Pr Samia BOUCHAFA	IBISC Université d'Evry
	Dr Slim ESSID	TSI TELECOM PariTech
Examinateurs	Pr Jean Claude MARTIN	LIMSI Université Paris Sud
	Pr Jean Luc ZARADER	ISIR UPMC
Directeurs	Dr Catherine ACHARD	ISIR UPMC
	Pr Mohamed CHETOUANI	ISIR UPMC

Résumé

Dans une interaction sociale, nous adaptons notre comportement à celui de nos interlocuteurs. L'étude et la compréhension des mécanismes sous-jacents à cette adaptation constituent le coeur du Traitement du Signal Social. Le but de cette thèse est de proposer des méthodes d'étude et des modèles pour l'analyse des signaux sociaux dans un contexte d'interaction en exploitant à la fois des techniques issues du traitement du signal et de la reconnaissance des formes.

Tout d'abord, une méthode non supervisée permettant de mesurer l'imitation entre deux partenaires en termes de délai et de degré est proposée en étudiant uniquement des données gestuelles. Dans un premier temps, des points d'intérêts spatio-temporels sont détectés afin de sélectionner les régions les plus importantes des vidéos. Ils sont ensuite décrits à l'aide d'histogrammes pour permettre la construction de modèles sac-de-mots dans lesquels l'information spatiale est réintroduite. Le degré d'imitation et le délai entre les partenaires sont alors estimés de manière continue grâce à une corrélation-croisée entre les deux modèles sac-de-mots.

La deuxième partie de cette thèse porte sur l'extraction automatique d'indices permettant de caractériser des interactions de groupe. Après avoir regroupé tous les indices couramment employés dans la littérature, nous avons proposé l'utilisation d'une factorisation en matrice non négative. En plus d'extraire les indices les plus pertinents, celle-ci a permis de regrouper automatiquement et de manière non supervisée des meetings en 3 classes correspondant aux trois types de leadership tels que définis par les psychologues.

Enfin, la dernière partie se focalise sur l'extraction non supervisée d'indices permettant de caractériser des groupes. La pertinence de ces indices, par rapport à des indices ad-hoc provenant de l'état de l'art, est ensuite validée dans une tâche de reconnaissance des rôles.

Mots clés Apprentissage non supervisé, Interaction, Traitement du Signal Social, Imitation, Reconnaissance de rôles, Extraction d'indices, Sac-de-Mots, Factorisation en Matrices Non-négatives

---

Abstract

In an social interaction, we adapt our behavior to our interlocutors. Studying and understanding the underlying mecanisms of this adaptation is the center of Social Signal Processing. The goal of this thesis is to propose methods of study and models for the analysis of social signals in the context of interaction, by exploiting both social processing and pattern recognition techniques

First, an unsupervised method allowing the measurement of imitation between two partners in terms of delay and degree is proposed, only using gestual data. Spatio-temporal interest point are first detected in order to select the most important regions of videos. Then they are described by histograms in order to construct bag-of-words models in which spatial information is reintroduced. Imitation degree and delay between partners are estimated in a continuous way thanks to cross-correlation between the two bag-of-words models.

The second part of this thesis focus on the automatic extraction of features permitting to characterizing group interactions. After regrouping all features commonly used in literature, we proposed the utilization of non-negative factorization. More than only extracting the most pertinent features, it also allowed to automatically regroup, and in an unsupervised manner, meetings in three classes corresponding to three types of leadership defined by psychologists.

Finally, the last part focus on unsupervised extraction of features permitting to characterize groups. The relevance of these features, compared to ad-hoc features from state of the art, is then validated in a role recognition task.

Keywords Unsupervised learning, Interaction, Social Signal Processing, Imitation, Role recognition, Feature extraction, Bag-of-words, Non-negative Matrix Factorization

---

Table des matières

Table des figures

Liste des tableaux

Introduction générale

L'être humain est un être complexe, qualifié d'être intelligent. Mais que signifie être intelligent ? Est ce qu'une mesure du quotient intellectuel est la meilleure prédiction du succès de quelqu'un dans la vie ? Des recherches en sciences cognitives avancent le fait que cette vision de l'intelligence est trop restreinte, car elle ignore un grand nombre de capacités ayant une importance dans la vie de tous les jours. En effet, l'aptitude de chacun à réagir au monde qui l'entoure et à interagir avec lui est capitale dans notre vie [82]. Les compétences qui sont associées à cette capacité sont appelées l'intelligence sociale [4], et incluent des éléments tels que la faculté à exprimer et reconnaître des signaux sociaux et des comportements sociaux comme les tours de parole, la politesse, l'empathie. L'intelligence sociale permet de gérer ces différents éléments afin de s'adapter de manière efficace aux autres personnes et ainsi gagner leur coopération.

L'étude des interactions sociales, bien qu'assez ancienne en psychologie et sociologie [107], est très récente dans le domaine du traitement automatique. Nous présenterons donc tout d'abord son histoire, définirons ses signaux et fonctions, avant de présenter la méthode générique utilisée en traitement automatique pour l'étudier. Nous présenterons enfin quelques unes des applications découlant des travaux de ce domaine, avant de conclure sur les challenges à résoudre.

Le traitement du signal social

En 2007, Alex Pentland introduit l'expression "Traitement du Signal Social" ("Social Signal Processing" [137]) afin de décrire les méthodes du domaine des Sciences de l'Ingénieur tentant de déduire des informations liées aux interactions sociales à partir d'indices de comportements non verbaux. Depuis, le domaine s'est élargi et traite maintenant des scénarios plus larges, tout en restant dans l'étude de signaux produits lors d'interactions sociales. Dans ceux-ci, la formation et l'ajustement des relations et interactions entre plusieurs agents (humains ou artificiels) sont étudiés. Ce domaine se situe à la croisée de trois grandes disciplines de recherche, à savoir la compréhension des comportements humains (Psychologie du comportement), la compréhension de l'influence des autres sur les pensées, sentiments et comportements (Psychologie sociale) et les Sciences de l'Ingénieur.

Un des buts du Traitement du Signal Social est de fournir aux machines ou robots une intelligence sociale. Il se concentre en particulier sur la modélisation, l'analyse et la synthèse de comportements non verbaux dans les interactions sociales [183]. L'idée clé est qu'une machine peut participer à une interaction sociale si elle arrive à automatiquement comprendre ou synthétiser les nombreux indices, verbaux et non verbaux (attitudes corporelles, regards, ...), que les personnes utilisent.

Signaux sociaux

Les signaux sociaux sont l'ensemble des signaux non verbaux qui permettent de transmettre un message, complémentaire du message verbal. Ces signaux sont variés, et apparaissent dans toutes les modalités. Les signaux qui vont être présentés ci dessous sont ceux qui ont été identifiés par la recherche en psychologie comme étant les plus importants dans les jugements humains des comportements sociaux. Plus de détails peuvent être trouvés dans [107, 67]. Ekman [67] lie les signaux non verbaux à cinq catégories : états affectifs/attitudes/cognitifs (joie, peur, stress...), emblèmes (des signaux spécifiques à une culture), manipulateurs (actions utilisées pour agir sur les objets de l'environnement ou sur soi même comme se mordiller les lèvres), illustrateurs (actions qui accompagnent le discours comme le pointage du doigt) et régulateurs (médiateurs de la conversation, comme l'échange de regard, hochement de tête...).

Les signaux sociaux existent dans toutes les modalités. L'apparence physique par exemple, est primordiale dans l'interaction sociale. Elle est constituée de la taille, la forme du corps, la physionomie, la couleur des cheveux et de la peau, ainsi que d'éléments plus volatiles tels que les vêtements, des bijoux ou du maquillage. Toutes ces informations ont un fort lien avec "l'attractivité". Ainsi, l'effet halo, aussi appelé "ce qui est beau est bon" [57], montre que l'apparence physique est associée inconsciemment à de la compétence et de la gentillesse. La forme du corps et la physionomie sont souvent attribuées à des traits de personnalité [47]. La posture ou les gestes sont révélateurs d'un état émotionnel, de la personnalité ou même d'un statut [121]. Il est aussi possible de citer les expressions faciales, le regard, le centre d'attention en ce qui concerne le visage [188], la prosodie, les tours de parole et les silences pour la voix [131], et la distance inter-personnelle ou le placement du siège pour l'aspect spatial et environnemental.

Tous les comportements non verbaux permettent d'envoyer de puissants messages qui influencent plusieurs aspects de l'interaction sociale. De manière générale, les psychologues ont identifié six grandes fonctions de la communication non verbale [184].

Tout d'abord, la communication non verbale permet de se forger une impression. Avant même de commencer une interaction, les gens se font une opinion sur les autres, en se basant sur des indices comme l'apparence ou les mouvements qui sont les premières choses que l'on remarque et forment ainsi une première impression [4]. Bien que l'on dise souvent que l'apparence n'est pas importante, des études de psychologie montrent le contraire. Le phénomène de "ce qui est beau est bon" [57] est un exemple concret que l'apparence joue sur la perception de désirabilité, de compétence ou même de gentillesse [47].

Ensuite, la communication non verbale permet également de gérer les interactions. En effet, elle permet par exemple à chacun de savoir quel est le bon moment pour intervenir. Elle permet d'améliorer la qualité de l'interaction en rendant celle-ci plus harmonieuse, en réduisant les interruptions et en améliorant les transitions entre les tours de parole.

La troisième fonction est l'expression des émotions. Celles-ci sont principalement portées par les expressions faciales et les gestes du corps qui apparaissent comme les éléments les plus importants dans le jugement humain des comportements affectifs [4].

La quatrième fonction, et non des moindres, est l'envoi de messages relationnels. Ces messages permettent de traduire son propre ressenti ou le type de relation que l'on souhaite établir (coopérative, agressive, ...). Le visage, à travers les expressions faciales, permet ainsi d'exprimer

beaucoup d'informations concernant des états cognitifs (intérêt, étonnement, ...), des traits de personnalité (extraversion, tempérament, ...), des comportements sociaux (accord, rapport, ...) ou des signaux sociaux (statut, fiabilité, ...). Les gestes et la posture permettent eux aussi d'exprimer des choses telles que l'ennui (manipulation de petits objets environnant), une attitude négative (posture fermée), l'engagement (posture faisant face à l'interlocuteur)...

La cinquième fonction est celle de la tromperie, et de sa détection.. Les indices non verbaux se répartissent alors en deux catégories. Alors que la première catégorie est utilisée de manière consciente afin de rendre des mensonges plus crédibles, la seconde est hors de contrôle et mène à un ressenti négatif qui peut être utilisé pour détecter les mensonges. Par exemple les menteurs font beaucoup d'efforts pour contrôler leurs expressions faciales, mais considèrent souvent le langage corporel comme non important et le négligent. Or celui-ci envoie alors des signaux contradictoires, qui peuvent donc être utilisés pour détecter le mensonge.

Enfin, les messages non verbaux sont de forts signaux de persuasion. Ainsi, des personnes dominantes ont tendance à plus toucher qu'à être touchées, regardent les autres moins qu'elles ne sont regardées, et contrôlent le temps et l'espace des interactions [5].

Toutes ces fonctions sont analysées et utilisées dans trois buts principaux. Le premier concerne l'analyse des émotions sociales. Les émotions vont des émotions intra-personnelles, telle que la joie ou la peur qui ont été longuement étudiées, à des émotions inter-personnelles comme l'empathie, l'envie, l'admiration. Le second but concerne l'analyse d'attitudes sociales, telles que la dominance, la personnalité et leurs effets. Enfin le dernier concerne l'analyse des relations sociales dans un groupe, notamment par la reconnaissance de rôles.

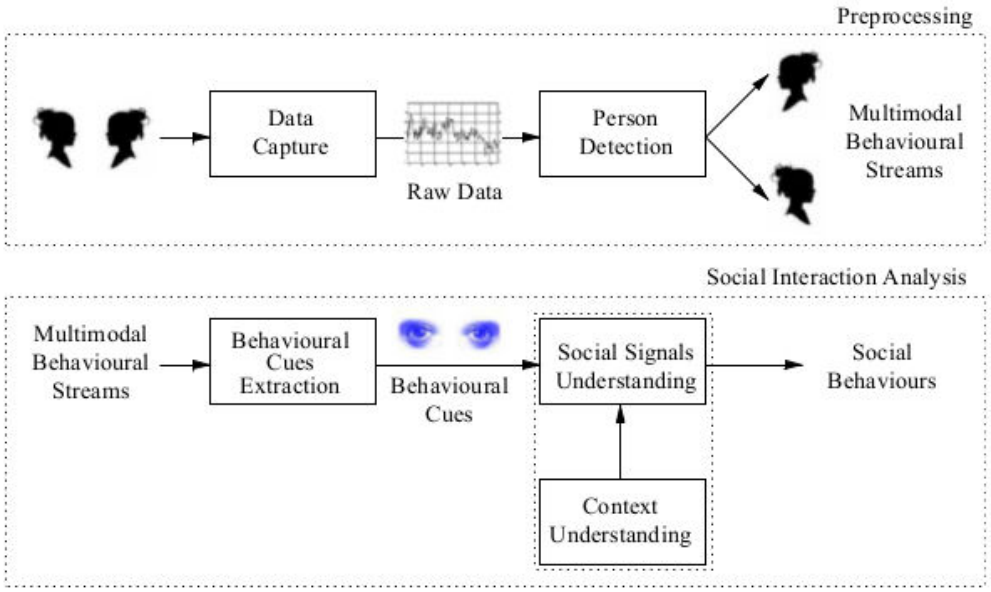
Méthodologie

L'analyse automatique des signaux sociaux comporte quatre grands sous problèmes, et est illustrée sur la figure ci dessous (extraite de [183]).

1. l'enregistrement des données ;
2. la détection des personnes ;
3. l'extraction depuis les données audios, vidéos ou physiologiques d'indices non verbaux et l'interprétation de ces indices en terme de signaux sociaux ;
4. la classification de ces signaux en tant que donnée sociale ou comportementale, en tenant compte du contexte.

L'avènement des capteurs peu chers a permis d'améliorer et systématiser le premier problème, notamment à travers l'utilisation de salles multimodales "intelligentes" capturant à la fois l'audio, la vidéo, et des éléments supplémentaires (tableau, présentations,...) de manière synchronisée [128, 186].

La seconde étape a elle aussi été simplifiée, puisqu'il n'est pas rare que les bases, lors de leur création, prévoient la séparation des données pour chaque personne à l'aide de capteurs différents (cas des microphones cravates par exemple). Cependant, malgré tout, les émissions sonores liées aux autres participants sont souvent présentes et nécessitent tout de même une étape de détection des personnes afin de nettoyer les données.



Méthodologie générale du traitement du signal social (extrait de Vinciarelli [183]).

La troisième étape extrait des indices des différentes modalités que nous avions évoqué précédemment, comme l'apparence, les gestes, la posture, le regard, les expressions faciales, la prosodie ou la proximité spatiale. Cette étape est complexe, et un grand nombre de travaux ont porté sur l'amélioration des techniques d'extraction et de reconnaissance [183, 78, 136].

La dernière étape joue elle aussi un rôle crucial dans le domaine du traitement du signal social. Vinciarelli et al. résument par exemple dans [183] les liens qui unissent les indices non verbaux et les comportement sociaux tels que l'émotion, la personnalité, la dominance, le statut, la persuasion, la régulation ou le rapport.

Applications

Un grand nombre d'applications s'ouvrent à l'analyse automatique des signaux sociaux. Nous ne citerons ici que quelques unes d'entre elles, mais des études plus complètes décrivent avec précision d'autres applications [180, 137, 136, 78]. Ces applications touchent un grand nombre de domaines, et par exemple les travaux de Pentland [138] sont reconnus comme une découverte capitale qui va faire avancer et changer les pratiques de management.

L'indexation multimédia [193] est la première application à laquelle on pense. De nos jours, de plus en plus de données sont enregistrées, échangées, et classées chaque jour. L'avantage apporté par l'analyse automatique des signaux sociaux est crucial, car cette information est actuellement manquante dans la plupart des systèmes, et seule l'annotation manuelle (des rôles, personnalités, contexte...) permet actuellement de la remplir.

Une seconde application concerne les téléphones mobiles. De nos jours, la plupart des personnes emmènent avec elles toute la journée leur téléphone mobile. Les informations de position spatiale, et de communication sont riches et peuvent être analysées [147], par exemple pour obtenir des informations sur le type de vie sociale (est ce que les amis proches sont liés au lieu de

travail, à des clubs extérieurs, ... est ce qu'une personne mange de manière régulière sur son lieu de travail ou a tendance à utiliser ce temps pour voir d'autres amis) que mène une personne [58].

Un troisième et dernier exemple est l'interaction homme-machine. Des études indiquent que non seulement nous effectuons les même indices non verbaux que nous interagissions avec un humain ou une machine, mais en plus nous avons tendance à attribuer des caractéristiques humaines (personnalité, émotions, ...) aux machines avec lesquelles nous interagissons [131].

Challenges

Le domaine fait face à de très nombreux challenges techniques et scientifiques [182]. Concernant les challenges techniques, la plupart des travaux fonctionnent sur des données de laboratoires, très contraintes et plus faciles à traiter que des données naturelles. Dans le futur, les méthodes devront être robustes au bruit, et capables de s'adapter dans divers contextes. De plus, ces méthodes sont souvent lentes, et ne sont donc pas applicables en temps réel dans des systèmes intégrés.

D'autre part, des challenges scientifiques existent. Le premier est le passage de l'étude de l'individu à l'étude des dyades et des groupes. En effet, un groupe de personnes en interaction est bien plus que la simple somme de ses membres. Cependant, pour l'instant, la plupart des études se concentrent sur l'analyse des individus. Ainsi, le développement de méthodes considérant le groupe dans son ensemble va devenir clé dans l'étude de ces groupes. Le second challenge scientifique repose sur le fait qu'un grand nombre d'études se basent sur une seule modalité pour analyser les comportement sociaux. Comment intégrer plusieurs modalités afin d'avoir une information plus complète ? La fusion doit-elle se faire au niveau des indices ou au niveau des comportements ? Doit on intégrer toutes les modalités ? Enfin, le plus grand challenge consiste en la compréhension, la modélisation et la représentation du contexte. Comment intégrer le contexte dans cette fusion afin d'avoir le meilleur résultat possible ? Comment faire pour que les algorithmes, au lieu d'être entraînés, soient capables d'apprendre avec de nouveaux utilisateurs dans de nouveaux contextes ?

Contributions

Dans la première partie de cette thèse, nous avons introduit une méthode de mesure d'imitation gestuelle entre deux partenaires interactifs. Contrairement à la plupart des approches de la littérature, aussi bien la temporalité que la forme des gestes sont pris en compte lors de cette évaluation. De plus, la méthode proposée permet d'estimer les trois grandeurs caractéristiques de l'imitation telles que définies par les psychologues : le degré d'imitation entre les partenaires, le délai entre les partenaires, et l'orientation de l'imitation. Ces trois grandeurs, mesurées de manière continue, à chaque instant de l'interaction, seront fort utile pour développer de nouveaux modèles d'interaction qui permettront aux robots de mieux interagir avec les humains.

Dans un second temps, nous nous sommes intéressés aux descripteurs permettant de caractériser les interactions sociales et avons mis en place une factorisation en matrices non-négatives

pour classer de manière non supervisée des meetings en trois grands types d'interactions de groupes. En plus de trouver automatiquement des types d'interactions qui correspondent à des types de leardership définis pas des psychologues, cette factorisation fait ressortir les signaux sociaux les plus déterminants dans la caractérisation des interactions.

Enfin la dernière partie se focalise sur la recherche non supervisée de signaux sociaux pertinents (motifs) pour la caractérisation des interactions de groupe. Une méthode spectrale permet de sélectionner les motifs les plus importants parmi tous les motifs possibles. Une tâche de reconnaissance de rôles utilisant les motifs trouvés permet de valider leur pertinence.

Plan de la thèse

Cette thèse se penche sur plusieurs des challenges identifiés dans la section précédente.

Ainsi, la **première partie** considère une modélisation interpersonnelle : l'imitation dans une interaction dyadique. L'imitation est un signal fort du maintient de l'interaction [53], et son évaluation est souvent globale (quantité de mouvement). Dans cette partie, nous proposons un modèle non supervisé d'évaluation automatique de l'imitation, sans a priori sur les gestes effectués. Nous essaierons de répondre aux deux questions suivantes : Comment évaluer de manière objective et automatique l'imitation ? Quelles sont les grandeurs temporelles à considérer lors de son évaluation ?

Dans une **seconde partie**, nous nous intéressons tout d'abord au problème de la caractérisation d'une interaction, à travers l'extraction d'indices. Contrairement aux approches courantes, les indices sont considérés selon plusieurs points de vue : intra-personnel, inter-personnel et de groupe. Une classification non supervisée des interactions est ensuite effectuée afin d'obtenir une information globale sur la structure du groupe.

Enfin, la **troisième partie** considère la place de l'individu dans un groupe, à travers son rôle. Pour cela de nouveaux indices sont obtenus de manière non supervisée. Leur pertinence est analysée à travers l'évaluation de la reconnaissance automatique des rôles.

Première partie

Imitation

Introduction

Dès que l'on interagit avec une autre personne, une coordination apparaît à plusieurs niveaux [19]. La plus évidente est celle de la parole, où une adaptation mutuelle des tours de parole des partenaires est nécessaire afin de maintenir un échange. Mais nous coordonnons aussi d'autres comportements non verbaux afin qu'ils s'harmonisent avec ceux de nos partenaires interactifs.

Comme le montrent les recherches en psychologie [149], énormément d'information passe par les comportements non verbaux comme les gestes des mains ou de la tête, le regard, les sourires, la prosodie,... Cette harmonisation ou synchronisation de nos comportements non verbaux prend une place importante dans bien des aspects de nos vies sociales. Ainsi, la psychologie développementale a longtemps étudié cette synchronisation, par exemple dans la relation mère-enfant [72, 73], patient-psychothérapeute [149] ou entre des amis [74]. Les résultats montrent qu'elle facilite et permet d'améliorer la qualité de l'interaction, améliorant ainsi les résultats de ces interactions, comme le développement de l'enfant, le succès de la thérapie, ou l'importance de l'amitié.

Une autre façon d'adapter son comportement à l'autre réside dans l'imitation. Ainsi, face à une personne qui parle doucement, on aura tendance à ne pas parler fort et face à une personne triste, on ne sera pas joyeux. D'un point de vue neurosciences cognitives, cette imitation a pu être démontrée dans les années 1990 par Giacomo Rizzolatti [153], puisque des neurones miroirs ont été trouvés et leurs signaux enregistrés dans l'aire prémotrice du cortex de singes, permettant de démontrer que ces neurones s'activent indifféremment lors de la réalisation d'une tâche ou de la simple observation de cette tâche. Ces neurones sont ainsi supposés jouer un rôle important dans l'apprentissage par imitation, dans la vie sociale mais aussi dans des comportements tels que l'empathie [120]. Merleau-Ponty disait en 2002 [125] "c'est à travers mon corps que je comprends les autres, comme c'est à travers mon corps que je perçois les choses". Cette phrase illustre un des buts de l'imitation : comprendre l'autre, par l'accession sensorimotrice à l'état intérieur du partenaire. Si l'imitation est nécessaire au maintien d'une bonne interaction, elle doit également être justement dosée. Ainsi, lorsqu'il y a trop d'imitation, le partenaire peut sembler se moquer tandis que si elle n'est pas assez marquée, le partenaire ne montrant pas d'imitation peut sembler inintéressé. L'imitation sert de glu sociale [8], en permettant d'augmenter le sentiment de similarité, d'affiliation et de sympathie. Elle est mise en place de manière non consciente lorsqu'une personne souhaite s'intégrer dans un groupe, ou lorsqu'elle est en accord avec ce qui est dit. L'imitation est donc une donnée utile pour estimer la qualité d'une interaction. Les approches utilisées en psychologie pour mesurer cette qualité d'interaction consistent en un visionnage intensif de vidéos de sessions, en une annotation manuelle de paramètres et en questionnaires donnés à remplir aux sujets d'étude. Bien que des psychologues entraînés puissent

prédire avec une grande précision les résultats de comportements sociaux, ce genre d'approche est très coûteux en temps et le suivi de nombreux groupes n'est alors pas possible. La mesure automatique de l'imitation est donc nécessaire et va au-delà du cadre de la psychothérapie, puisqu'elle peut par exemple servir à mesurer la qualité des échanges dans des scénarios divers comme les relations interpersonnelles d'un manager et de ses employés. Dans cette thèse, nous nous intéressons uniquement à l'imitation gestuelle et ne traitons donc pas de l'adaptation audio ou émotionnelle. L'imitation, aussi appelée mimétisme, concerne la coordination de mouvements à la fois dans le temps (synchronie) et dans la forme (même geste). Elle est présente de manière naturelle dans les interactions comme le montre la figure I.0.0.1.

FIGURE I.0.0.1 Deux amis adoptent une posture congruente et démontrent des gestes d'imitation lors de la discussion.

Tout d'abord un état de l'art présente les travaux effectués sur l'imitation en psychologie ainsi que les méthodes computationnelles apparues dans le domaine du traitement du signal social. Une méthode automatique d'évaluation de l'imitation est ensuite présentée, ainsi que la création d'une base de données adaptée à sa mesure. Enfin les résultats de la méthode seront analysés avant de conclure.

Chapitre I.1

État de l'art

Sommaire

I.1.1 Histoire et premières recherches

Les premières définitions de l'imitation (aussi appelée mimicry, mirroring, congruence, ou effet caméléon), comme celle de Charles Darwin [52], voyaient l'imitation comme un synonyme de l'empathie (la capacité à comprendre les sentiments et émotions d'un autre individu). Les premières recherches menées sur ce sujet allaient en ce sens, les personnes imitant plus étant perçues comme plus empathiques. En effet, selon la théorie de l'esprit, plus une personne imite, plus elle est capable de comprendre la personne en face [40] .

Dès 1890 des psychologues s'intéressent à l'imitation [93]. Ils différencient alors imitation consciente et inconsciente, chacune poursuivant des buts différents. L'imitation est étudiée en fonction de ce qui est imité (le verbal, facial, émotionnel ou comportemental), et est démontrée comme capitale dans l'interaction (communication, liant social, ostracisme, objectif d'affiliation [110]) ainsi que dans des cadres non sociaux (self-monitoring [41, 39], pouvoir de persuasion [178]). Elle est impactée par de nombreux facteurs comme l'humeur, le désir de convaincre, la capacité d'auto-régulation ou bien encore par la quantité de croyances partagées [39], et il a été établi qu'il existe un lien entre le degré de mimétisme, la perception de la fluidité de l'interaction et le degré d'appréciation de l'interaction entre les partenaires. De plus, l'imitation est considérée comme un processus de communication permettant de maintenir l'interaction [53].

La théorie de l'imitation commença par l'étude approfondie sur ses fonctions sociales, puis elle se tourna vers ses racines cognitives [108, 142]. Cependant, ces deux aspects sont en réalité inter-connectés [40].

Dans ce document, nous suivrons la définition de l'imitation de Bernieri et al. [18], qui définit l'effet caméléon comme la coordination d'actions ou événements entre des partenaires, à la fois en forme et en temps.

Dans la suite de ce chapitre, nous allons présenter divers travaux du domaine du traitement du signal social portant sur l'imitation, sur les méthodes d'évaluation de l'imitation, ainsi que sur les bases de données la concernant. Ensuite, nous décrirons comment se place notre méthode par rapport à ces travaux, et nous finirons par présenter les challenges qui s'offrent à nous.

I.1.2 Méthodes computationnelles

La plupart des méthodes de l'état de l'art pour mesurer l'imitation s'appuient sur deux grandes étapes : l'extraction de caractéristiques et une mesure de similarité [173]. Elles doivent ensuite (mais toutes ne le font pas), valider le score obtenu, comme illustré sur la figure I.1.2.1.

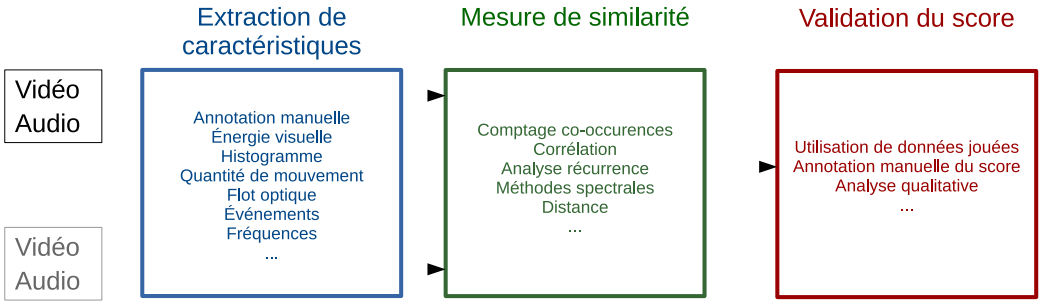


FIGURE I.1.2.1 Grandes étapes de la mesure de l'imitation.

L'extraction de caractéristiques peut être faite de plusieurs manières : alors que certaines études se concentrent sur l'étude des mouvements d'une partie du corps en particulier, d'autres capturent le mouvement global de chaque personne. Il est intéressant de noter que l'étude de ces mouvements globaux donne ensuite non pas lieu à une mesure d'imitation comme il est souvent clamé, mais seulement à une mesure de synchronie : la similarité entre les événements (du mouvement global) s'appuie en effet seulement sur une similarité temporelle, et non sur la forme précise des mouvements.

Par exemple, un grand nombre d'études se basent sur la tête, puisqu'elle est à la fois un lieu d'expression des émotions, d'une participation active à travers le mouvement des lèvres et de la tête, ou de l'accord à travers les hochements. Plusieurs méthodes existent pour la suivre, allant des algorithmes de suivi basés sur les vidéos [35, 179], aux dispositifs de capture du mouvement [7]. Quand au mouvement global de chaque personne, il est souvent capturé grâce aux Images d'Énergie de Mouvement (MEI) [2, 149] ou à leurs dérivées [55, 176].

Une fois les caractéristiques extraites, plusieurs mesures entre les deux séries temporelles peuvent être utilisées pour calculer la similarité (même forme au même instant), la synchronie (coordination temporelle) ou l'imitation (coordination en temps et en forme). Il existe trois grandes techniques de calcul de l'imitation : la corrélation, l'analyse de récurrence ou les mé-

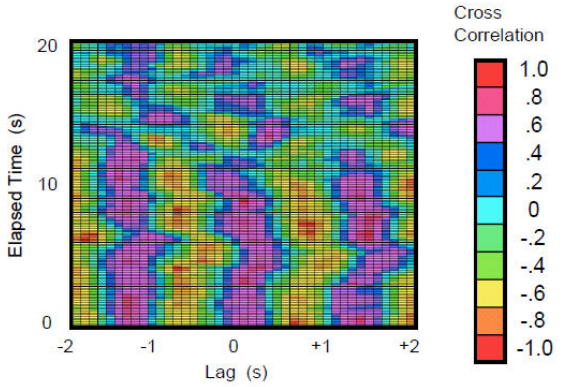


FIGURE I.1.2.2 Illustration d'un calcul de corrélation croisée, extrait de Boker et al. [26].

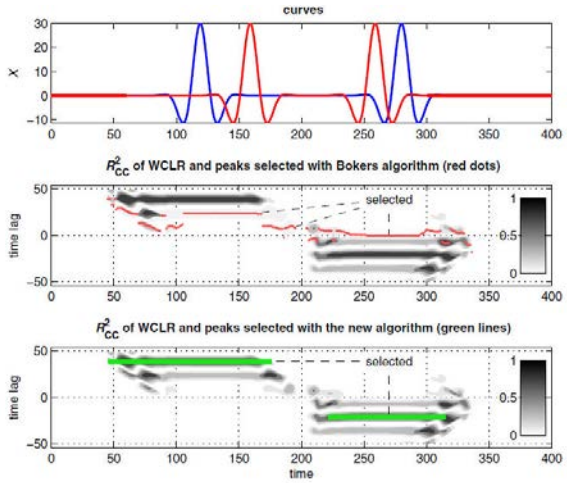


FIGURE I.1.2.3 Résultats de deux algorithmes de sélection de pic, extrait de Altman et al. [3].

thodes spectrales. La corrélation [7, 26, 2] est souvent calculée sur des fenêtres temporelles d'interaction courtes, pour plusieurs délais temporels, comme illustré sur la figure I.1.2.2. Ensuite, un algorithme de détection de pic est utilisé pour obtenir une mesure unique à chaque instant (voir figure I.1.2.3). L'analyse de récurrence [179] est parfois utilisée comme une substitution à la corrélation. Basée sur la théorie des systèmes dynamiques couplés, elle permet d'avoir une représentation graphique de la dynamique des systèmes couplés. Les points de récurrence sont les points temporels où les deux systèmes présentent une coordination temporelle, comme illustré sur la figure I.1.2.4. Cependant, ces méthodes donnent de bien maigres résultats pour la détection d'imitation, car c'est souvent l'aspect temporel qui est mis en avant, en reléguant au second plan la mesure de la ressemblance dans la forme des mouvements (ou en les considérant de manière globale à l'aide de quantité de mouvement). La mesure de l'imitation nécessite une description plus fine des gestes, qui ne se base pas juste sur une quantité de mouvement, à travers notamment des techniques de reconnaissance de gestes.

On notera aussi différentes approches dans l'appréhension de la temporalité. Ainsi certains considèrent des événements discrets et calculent leur co-occurrence [71], font une évaluation continue en se basant sur des mouvements globaux [175] (voir figure I.1.2.5), évaluent des mouvements locaux sur des périodes globales [196] ou comparent des modèles [171, 135]. Un

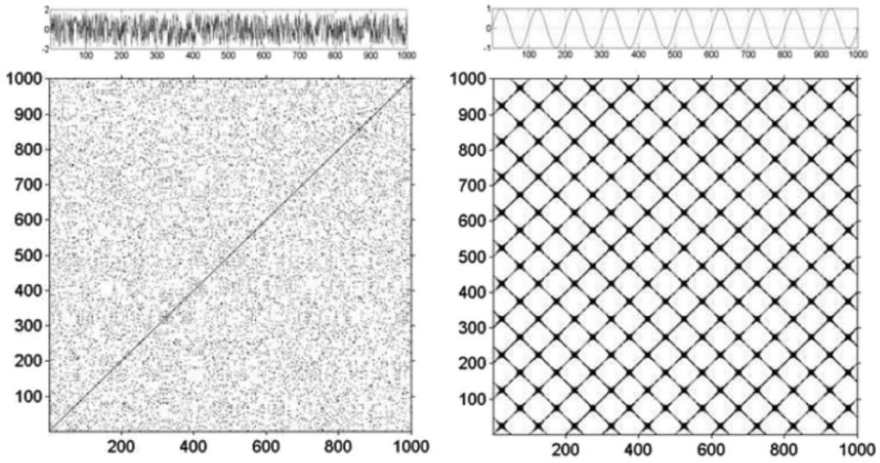


FIGURE I.1.2.4 Illustration d'un tracé d'analyse de récurrence pour deux types de signaux : un signal bruité (gauche) et un signal sinusoïdal (droite). Extrait de Varni et al. [179].

certain nombre non exaustifs de publications sont présentées ci-dessous, et résumées dans la table I.1.1.

Dans [143], Quiroga et al. souhaitent détecter des crises d'épilepsie en analysant les signaux EEG. Pour cela, ils développent une méthode pour mesurer la synchronisation ainsi que le délai temporel entre des événements dans des séries temporelles. Le degré de synchronisation est obtenu à partir du comptage du nombre d'apparences quasi-simultanées d'événements, et le délai est calculé à partir de la préséance des événements d'un signal par rapport à l'autre signal. Cette méthode est ensuite appliquée à des signaux électroencéphalogrammes (EEG) de rat ainsi qu'à des enregistrements intra-crâniens humains, dans lesquels apparaissent des crises d'épilepsie. Les résultats sont interprétés afin de mettre en évidence les phénomènes d'épilepsie. L'utilisation de données de substitution permet de montrer que cette mise en évidence n'est pas aléatoire. Cependant, aucune quantification ou validation des résultats n'est effectuée, seule l'interprétation est présentée.

Dans [190], Watanabe et al. prouvent l'importance de l'utilisation de l'information de respiration lors de conférence utilisant des agents virtuels. Pour cela, des agents virtuels reproduisent les mouvements de tête, bras et corps des sujets. On y ajoute (ou non), la respiration. Les sujets sont placés dans des pièces différentes, communiquant entre eux à travers leurs avatars respectifs. Plusieurs situations sont proposées, dans lesquelles soit les sujets voient un seul agent virtuel (celui du partenaire), ou bien à la fois l'agent virtuel les représentant et celui de leur partenaire. Watanabe et al. utilisent tout d'abord des questionnaires remplis par les participants après l'expérience pour montrer l'intérêt d'ajouter la respiration dans les informations affichées ainsi que de voir son propre avatar afin de faciliter l'interaction. Ils calculent ensuite une corrélation croisée entre les mouvements de tête verticaux des participants (calculée sur une fenêtre glissante de 30 secondes, avec un délai maximal de 2 secondes). Ils montrent ainsi que cette corrélation est plus forte lorsque les deux avatars sont présents, ou lorsque la respiration est montrée. Ceci est validé par des T-test (pour $p<0,1$).

Dans [148], Ramseyer et al. souhaitent démontrer l'intérêt d'une thérapie. Pour cela, ils vont se baser sur une mesure de la coordination entre patient et thérapeute. Tout d'abord, la quantité

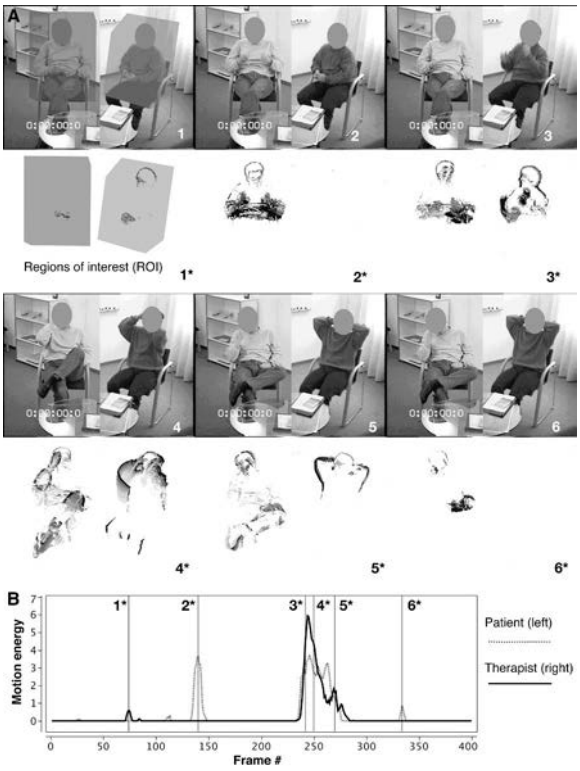


FIGURE I.1.2.5 Quantité de mouvement évaluée de manière globale dans des régions d'intérêt, extrait de Ramseyer et al. [149].

de mouvement de chaque personne est calculée grâce aux Images d'Énergie de Mouvement (MEI). Une corrélation croisée entre les quantités de mouvements des deux participants (sur une fenêtre de 1 minute et un délai maximal de 4 secondes) permet d'obtenir une mesure générale de la coordination. Pour être sûr que la coordination mesurée entre le patient et son thérapeute est pertinente, la même mesure est appliquée sur des données mélangées (à l'échelle des fenêtres temporelles) et un test statistique est effectué. Enfin, les scores sont calculés sur des séances de début de thérapie et de fin de thérapie et permettent de montrer grâce à un test statistique que la thérapie améliore la coordination entre le patient et son thérapeute.

Dumas et al. souhaitent mettre en avant dans [66] l'émergence d'une synchronisation des ondes cérébrales lors d'une interaction sociale. Pour cela, deux sujets sont placés face à un écran projetant les mains du partenaire. Plusieurs configurations existent : un des deux partenaires imite l'autre de manière continue, ou les deux s'imitent l'un l'autre de manière libre (ce qui donne lieu à des changements d'orientation dans l'interaction). Ils annotent à la fois les épisodes de synchronie (alignement temporel du début et fin de gestes) et d'imitation (ressemblance des gestes en forme et en direction spatiale). Des casques EEG permettent de mesurer l'activité cérébrale des partenaires. Des tests statistiques montrent qu'il existe des couplages au niveau neuronal entre le cerveau du leader et celui de l'imitateur, et que ceux-ci sont différents en fonction des bandes de fréquences étudiées.

Altman et al. [3] montrent que pour des signaux cycliques, la corrélation croisée est biaisée et proposent un modèle de régression croisée afin de s'affranchir de l'auto-corrélation. Ils comparent ces deux modèles sur un exemple joué afin de montrer que leur modèle a de meilleures

performances dans le cas de signaux répétitifs. Ensuite, ils utilisent de vraies données d'interaction, sur lesquelles sont calculées des Analyses de l'Énergie du Mouvement (MEA), qui contiennent une information globalisée sur les mouvements des participants. Ils appliquent ensuite leur modèle ainsi qu'un nouvel algorithme de sélection de maxima sur des dyades d'enfants. Deux conditions sont testées, suivant si les enfants sont amis ou non, et ils mettent en avant les différences entre ces deux conditions. Aucune évaluation de la qualité et des performances de leur algorithme n'est présentée, si ce n'est sur un exemple joué.

Sun et al. montrent dans [176] qu'il existe une adaptation des partenaires lors d'une interaction sociale. Pour cela, ils extraient tout d'abord automatiquement des caractéristiques audio (pitch, énergie et taux de parole). Ensuite, ils utilisent une corrélation croisée entre les signaux de deux partenaires interactifs, ainsi qu'entre un présentateur et un des partenaires. C'est ensuite l'étude qualitative et l'interprétation des variations de la corrélation qui leur permet d'affirmer qu'il existe bien une adaptation du comportement des partenaires lors de l'étude de l'interaction. Un article complémentaire [174] se base quand à lui sur des caractéristiques visuelles (Images de Mouvement Accumulées, AMI) pour calculer les corrélations croisées. Enfin, dans un autre article [175], Sun et al. détectent tout d'abord des régions d'intérêt en utilisant des quadtree, à partir desquelles sont extraites des caractéristiques de flot optique. Ensuite, ils calculent la corrélation entre des séries de caractéristiques visuelles et en déduisent un score d'imitation global sur toute l'interaction. La dynamique de l'imitation à travers le temps est étudiée, montrant que l'imitation facilite l'interaction et les échanges d'idées entre les participants. Cependant, les caractéristiques encodées représentent des quantités de mouvement, ce qui mène plus à une mesure de synchronie qu'à une mesure d'imitation.

Feese et al. [71] étudient plusieurs types de leadership. Pour cela, ils extraient automatiquement divers indices (hochements de tête, gesticulations, bougeotte, changements de posture, main qui touche le visage et croisement de bras). Ils comparent ensuite les types de leadership en comptant le nombre d'événements du même type dans des fenêtres temporelles limitées. Ceci donne une mesure d'imitation pour chaque type de leadership pour chaque type d'indice, dans les deux sens d'imitation. Un test statistique permet ensuite de voir quels sont les indices pertinents qui différencient les types de leadership.

Bilakhia et al. veulent dans [23] créer des modèles afin de détecter l'imitation. Pour cela, ils extraient des caractéristiques cepstrales, et les changements dans les expressions faciales (largeur et hauteur de la bouche, pose des sourcils, ...). Ils créent ensuite deux modèles, un pour l'imitation, et l'autre pour la non-imitation. Les modèles sont prédictifs et utilisent les caractéristiques d'une personne pour reconstruire celles du partenaire. Une séquence inconnue est ensuite passée dans les deux modèles (imitation et non imitation), et celui avec l'erreur de reconstruction la plus faible permet de libeller la séquence en imitation ou non imitation. Les résultats sont évalués grâce à une annotation manuelle, ce qui mène à un taux moyen de vrai positif pour l'imitation de 77,5% et de 60% pour la non-imitation, différence due à deux classes très déséquilibrées.

Dans [196], Xiao et al. souhaitent montrer que la synchronie augmente au cours d'une interaction. Pour cela, ils modélisent les mouvements de la tête en utilisant un Modèle de Mixtures de Gaussiennes (GMM) sur les fréquences spectrales tirées des vecteurs de mouvement

de la tête. Ensuite ils quantifient la similarité entre les partenaires grâce à la divergence de Kullback-Leibler calculée à partir de probabilités a posteriori des GMM. Les scores du début et de la fin de l'interaction sont comparés, et un test statistique permet de valider l'évolution.

Dans [54], Delaherche et al. souhaitent séparer la question du timing (rythmes) et de la ressemblance. Pour cela, ils détectent des pics dans les mouvements, et envoient les séquences à deux modules distincts. Le premier ne considère que les temps entre les événements des partenaires, pour en sortir deux indices, un de rythme interpersonnel et un de rythme intrapersonnel. Ces deux indices sont ensuite testés sur des données, et un test statistique montre leur pertinence à dissocier des séquences synchrones et non synchrones. Le second module quand à lui se base sur des points d'intérêts décrits par des histogrammes de flot optique ou de gradients orientés, qui sont ensuite concaténés dans des sacs-de-mots (BOW). Ces BOW sont ensuite comparés grâce à un SVM 1-classe qui permet de mesurer une distance entre eux. Une courbe ROC montre l'efficacité de la méthode. Cette méthode est d'ailleurs reprise dans [56] où elle est appliquée à une base de mouvements de mains, et à laquelle une analyse par récurrence est ajoutée.

I.1.3 Bases de données

A part les travaux récents de Sun et al. [173], nous n'avons pas trouvé dans la littérature de bases de données publiques et annotées permettant l'étude de l'imitation dans un contexte dyadique et gestuel. Dans cette base de données [173], deux participants sont assis en face à face, et débattent sur un sujet prédéfini, comme illustré sur la figure I.1.3.6. Les auteurs mettent en avant la relation entre les occurrences d'imitation et les émotions. Plusieurs annotations sont utilisées (main, mouvement de tête et de corps, expressions faciales, tours de paroles, ...) afin de montrer cette relation, à travers deux expériences : un débat sur un sujet politique et un jeu de rôles. Cependant cette base, plus dédiée à l'analyse d'expressions faciales, présente très peu de gestes et de mouvements.

FIGURE I.1.3.6 Image illustrant la base de données de Sun et al. [173].

Sanchez-Cortes et al. ont présenté le corpus ELEA dans [160], lequel consiste en 10 heures d'enregistrement de réunions. Il est demandé aux participants de résoudre une épreuve de

survie en condition de températures extrêmes sans assignation préalable de rôles. Les vues vidéos présentées par le corpus mettent en avant le haut du corps de quatre participants dans une configuration portable, et la tête dans la configuration fixe, comme illustré sur la figure I.1.3.7. Cependant, la complexité introduit par les réunions crée une difficulté supplémentaire dans l'analyse de l'imitation : à chaque instant, il est nécessaire de déterminer quel participant est en train d'imiter quel autre participant.

FIGURE I.1.3.7 Image illustrant la base de données de Sun et al. [160]. Une des webcams est entourée en rouge, et le micro omnidirectionnel en bleu.

Delaherche et al. introduisent dans [54] une base de données où deux participants s'imitent l'un l'autre de manière synchronisée (avec des gestes prédéfinis), comme illustré sur la figure I.1.3.8. Le but de cette base était de séparer l'aspect temporel de l'aspect forme dans l'imitation, et la base a été construite de telle sorte à ce que les participants fassent le même geste au même instant (à l'aide d'un métronome). Cependant, dans les interactions naturelles, deux partenaires exécutent rarement le même geste simultanément, puisque l'imitation nécessite un délai entre les partenaires (un participant en suit un autre). Ainsi, une parfaite synchronisation temporelle dans les gestes provient plutôt d'une intention partagée que d'imitation.

FIGURE I.1.3.8 Séquence des gestes de la base de Delaherche et al. [54].

Ainsi, ce manque de bases de données nous a mené à créer une nouvelle base de données respectant les conditions suivantes : l'imitation gestuelle doit être effectuée dans un cadre dyadique, sans gestes prédéfinis. La liberté laissée aux participants, à la fois dans la forme et dans le temps, mène à une base de données d'imitation gestuelle plus naturelle (voir section I.2.5 , et la figure I.1.3.9), et qui pourra être utile aux futurs chercheurs ayant l'intention de développer de nouveaux algorithmes en leur évitant l'étape de collection de données.

(a) (b) (c) (d)

FIGURE I.1.3.9 Quatre exemples d'images extraites de la base créée.

I.1.4 Évaluation

Beaucoup d'études mettent en relation la similarité (ou les co-occurrences) avec une autre variable (comme le succès d'un traitement thérapeutique par exemple) ou une hypothèse (par exemple "est-ce que le fait de retirer une modalité dans l'interaction baisse sa qualité ?"), en supposant que la-dite similarité est un indice fiable d'imitation.

Cependant, une des premières questions qui se pose lorsqu'on utilise la corrélation ou la corrélation croisée est la suivante : le score obtenu est-il significatif ? Pour répondre à cette question, Bernieri et al. [18] proposent de créer des données artificielles ("surrogate data") à partir de données réelles, en n'appareillant pas correctement les dyades de la base de donnée. Ainsi, au lieu d'avoir les participants A_1 (resp. A_2) interagissant avec son partenaire B_1 (resp. B_2), les données de A_1 sont appareillées avec les données de B_2. Ainsi, une interaction artificielle (n'ayant jamais eu lieu) est obtenue, qui permet alors de déterminer le score moyen obtenu lors de données pseudo-aléatoire d'interaction. Un test statistique permet ensuite de s'assurer que les scores d'imitation obtenus contiennent bien de l'information, ou si certaines fenêtres temporelles contiennent un haut taux d'imitation.

I.1.5 Cadre

Dans ce chapitre, nous nous concentrerons sur l'imitation gestuelle qui permet de mieux comprendre la coordination des actions entre les partenaires, à la fois temporellement et en terme de forme [18], tâche qui constitue un verrou pour le développement de robots socialement adaptés.

Alors que les travaux actuels mènent seulement à la présence d'imitation, et parfois au délai entre les partenaires, il est nécessaire d'évaluer à la fois son degré (à quel point les deux actions ou gestes se ressemblent), le délai entre les partenaires et l'orientation entre les partenaires (qui mène et qui suit). De plus, ces mesures doivent être estimées de manière continue, à chaque instant, afin de permettre une étude fine de l'interaction. Celle-ci pourrait être exploitée pour construire des robots ou agents virtuels hautement réactifs et socialement adaptés.

De plus, les nouvelles méthodes devraient tout d'abord être validées sur des bases de données d'interactions dyadique dans lesquelles l'imitation est présente et connue, puis dans des bases naturelles, où l'imitation est plus diffuse (en intensité) et moins présente (en fréquence d'apparition).

TABLE I.1.1 Études portant sur l'imitation.

Étude	Caractéristique(s)	Modèle Métrique	Performances Résultats
Quirogua et al. [143]	événement de signaux EEG	comptage de co-occurences	Détection de crises épileptiques. Utilisation de données jouées pour montrer que la détection ne détecte rien dans ces cas là. Interprétation des résultats (sans quantification).
Wanabe et al. [190]	mouvement verticaux de la tête	corrélation croisée	Des questionnaires et des T-test prouvent que l'affichage de la respiration améliore la qualité de l'interaction
Ramseyer et al. [148]	Quantité de mouvement (MEI)	corrélation croisée	Utilisation de données jouées pour montrer la validité de la mesure. Puis application des mesures pour démontrer l'amélioration de l'interaction au cours d'une thérapie.
Dumas et al. [66]	signaux EEG de mouvements de main	différences statistiques	Prouvent qu'il existe une synchronisation des ondes cérébrales lors de l'interaction sociale (calculs de différences statistiques entre cas d'imitation synchronisée et cas de non imitation).
Altman et al. [3]	Quantité de mouvement (MEA)	regression croisée	Cette méthode alternative à la corrélation croisée permet de s'affranchir de l'auto-correlation, forte dans les signaux cycliques. Et ainsi d'avoir ensuite une meilleure estimation du délai.
Sun et al. [176]	Hauteur, énergie et taux de parole	corrélation croisée	L'étude des variations de la corrélation au cours du temps montre qu'il existe une adaptation des partenaires lors d'une interaction. Aucune évaluation ni validation n'est présenté par l'article.
Sun et al. [174]	Quantité de mouvement (AMI)		
Sun et al. [175]	Quadtree puis flot optique		
Feese et al. [71]	événements visuels	comptage de co-occurences	Un test statistique sur les mesures d'imitation permet de voir quels sont les indices pertinents qui différencient les types de leadership.
Bilakhia et al. [23]	Cepstre (MFCC) et changements d'expressions faciales	Deux modèles de regression : imitation et non-imitation	Les modèles permettent de libeller chaque sequence en imitation (resp. non-imitation). Taux de vrai positif 77,5% (resp. 60%).
Xiao et al. [196]	fréquences des mouvement de la tête modélisées par GMM	Synchronie mesurée par divergence KL	La Synchronie est calculée pour le début et la fin de l'interaction. Des tests statistiques prouvent l'augmentation de synchronie.
Delaherche et al. [54, 56]	Points d'intérêts concaténés dans un BOW	SVM 1-classe	Comptage du nombre de points pour lesquels la distance entre les gestes est faible, puis classification en imitation/non-imitation. Score F1 = 0.7848.

Chapitre I.2

Évaluation automatique de l'imitation entre deux partenaires interactifs

Sommaire

I.2.1 Introduction

Contrairement à la plupart des méthodes présentées précédemment, la méthode que nous introduisons dans ce chapitre utilise à la fois la coordination temporelle des gestes et leur apparence afin de mesurer l'imitation. Elle diffère aussi des approches usuelles d'étude des interactions homme-homme qui considèrent des événements discrets et calculent leur co-occurrence [71], font une évaluation continue en se basant sur des mouvements globaux [175], évaluent des mouvements locaux sur des périodes globales [196] ou comparent des modèles [171, 135].

De plus, elle tire parti des études menées en psychologie et permet de mesurer l'imitation entre deux partenaires en termes de délai et de degré et ce, en étudiant uniquement des données gestuelles. Elle diffère également de beaucoup de travaux dans le sens où elle conduit à une mesure continue de l'interaction et ce dans le but de modéliser la communication sociale de manière fine. Comme nous allons le voir ensuite, la méthode s'appuie sur des modèles sac-de-mots (BOW pour Bag Of Words) pour caractériser chaque vidéo puis sur une mesure de similarité pour estimer les paramètres de l'imitation.

I.2.2 Vue d'ensemble de la méthode

La méthode se concentre sur l'évaluation automatique de l'imitation entre deux partenaires dans des vidéos synchronisées utilisant seulement des données gestuelles. Des études dans le domaine de la psychologie [73] sur les interactions parent-enfant ont montré que la synchronie est basée sur trois paramètres majeurs : l'orientation dans la relation (quelle est la personne qui mène), le délai entre les partenaires (le temps entre un stimuli et la réponse associée) et le degré de synchronie (quantification de la ressemblance entre le stimuli et la réponse). Cette définition est ici étendue à un cas particulier de la synchronie : l'imitation. Notre méthode se propose d'évaluer les différents paramètres sus-cités en utilisant une caractérisation des gestes. La plupart des méthodes de l'état de l'art se base sur de la reconnaissance de gestes parmi un ensemble de gestes prédéfinis. Le but de notre méthode n'est pas de reconnaître des gestes mais d'identifier si deux personnes s'imitent ou non. Ainsi, notre méthode va permettre d'évaluer la similarité des actions effectuées par les partenaires, sans reconnaître l'action elle même et donc sans contrainte sur l'action (qui peut sortir de l'ensemble des gestes déjà vus dans la base d'apprentissage).

La méthode est représentée figure I.2.2.1. La première étape de l'approche consiste à détecter les parties de l'image contenant une information d'intérêt. Pour cela, nous extrayons les Points d'Intérêt Spatio-Temporels (STIPs pour Spatio-Temporal Interest Points) qui révèlent de l'information tant dans l'espace spatial que temporel (voir section I.2.3.1).

La seconde étape consiste à décrire les STIPs, en utilisant des histogrammes locaux basés sur le gradient, le flot optique, ou une combinaison de ces deux descripteurs (section I.2.3.2). Une technique de clustering appliquée aux points détectés mène à un dictionnaire de mots visuels qui permet de limiter la description des points à un vocabulaire prédéfini. Les STIPs, décrits à l'aide de ces mots, sont ensuite accumulés dans un modèle sac-de-mots afin de caractériser les vidéos (section I.2.3.3). Comme nous le verrons dans la section I.2.3.4 , un inconvénient majeur du modèle sac-de-mots est la perte de l'information spatiale, qui sera alors réintroduite.

Une fois que le modèle sac-de-mots a été obtenu pour toutes les vidéos, une corrélation-croisée est utilisée sur les sac-de-mots pour mesurer la similarité et évaluer l'imitation entre deux vidéos (section I.2.4). La section I.2.5 quand à elle présentera la base de données qui a été créée pour évaluer la méthode. La corrélation croisée, mesurée pour des vidéos d'imitation et de non-imitation, permet de valider l'ensemble de la méthode, comme expliqué dans la section I.2.6 .

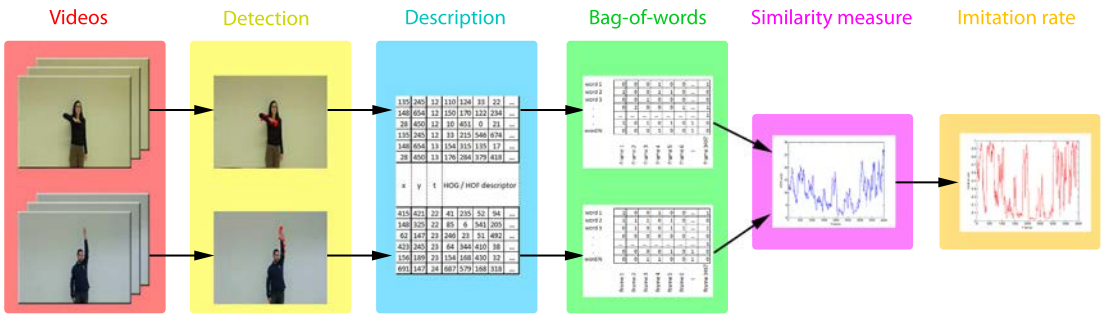


FIGURE I.2.2.1 Processus de la méthode d'évaluation de l'imitation.

I.2.3 Modélisation des actions de partenaires interactifs par sac-de-mots

Le modèle sac-de-mots a été largement utilisé par la communauté de traitement des images afin de réduire le coût computationnel pour l'analyse d'images et de vidéos. En effet, au lieu de comparer des images entières, seulement quelques régions d'intérêt sont étudiées. Cependant, comme nous le montrerons dans la section I.2.3.4 , ce modèle possède des inconvénients majeurs, comme par exemple la perte d'information spatiale. La construction du modèle de sac-de-mots est basée principalement sur deux étapes présentées ci-après : la détection de points d'intérêts et leur description.

I.2.3.1 Détection

La détection de STIPs peut se faire à l'aide de multiples approches. Des études complètes et des comparaisons de détecteurs peuvent être trouvées dans la littérature, comme par exemple [187], [166], ou [79]. Nous ne présenterons ici que les deux approches les plus populaires : une extension du détecteur de Harris, et un détecteur utilisant les filtres de Gabor.

I.2.3.1.a Harris 3D

L'approche proposée par Laptev dans [111] est basée sur le détecteur de Harris [88], généralisé à des données spatio-temporelles. Elle a été largement utilisée dans le domaine de la

reconnaissance d'action [187].

Le détecteur de Harris est un détecteur de coins, et se base sur la formulation de Moravec. Dans cette formulation, un coin a une grande valeur si des changements existent dans toutes les directions.

Le détecteur de Harris a tout d'abord été étendu dans une première version multi-échelle. Dans celle-ci, les coins sont détectés à plusieurs échelles, chaque échelle correspondant à une dégradation progressive de l'image originale en étant filtrée par une gaussienne. Ainsi, un coin sera présent à certaines échelles et pas à d'autres, et l'échelle où s'opère la disparition du coin est appelée échelle caractéristique.

Pour finir, le détecteur de Laptev, aussi appelé Harris 3D est seulement une extension de ce processus multi-échelle, en considérant le temps comme une nouvelle dimension. Ainsi les points sont détectés à de multiples échelles spatiales et temporelles. Ceci permet de détecter des points d'intérêts présentant des caractéristiques fortes dans les vidéos (à la fois en temps et dans l'espace), tout en gardant un niveau de détail, qui peut être évalué grâce à l'échelle de chaque point, comme illustré sur la figure I.2.3.2.

Ainsi, pour l'imitation, certains gestes peuvent être caractérisés par un mouvement global d'un bras. Ils sont alors présents à une échelle temporelle faible (mouvement "lent"), et une échelle spatiale faible aussi (un bras n'est pas juste un détail). Un hochement de tête quand à lui se caractérise par une échelle temporelle grande (mouvement rapide) et une échelle spatiale grande aussi (les contours de la tête ne bougent pas, mais le visage bouge).

FIGURE I.2.3.2 Détection de points d'intérêts spatio-temporels (extrait de [106]).

I.2.3.1.b Filtres de Gabor

La seconde approche, proposée par Dollár dans [59], se base sur les filtres de Gabor, qui sont des fonctions harmoniques multipliées par une gaussienne. La fonction de réponse est donnée par l'équation I.2.1, où $I(x, y, t)$ représente la série temporelle d'images, $g(x, y; \sigma)$ représente le noyau de lissage Gaussien appliqué seulement aux deux dimensions spatiales, $h_{ev}(t; tau, omega)$ et $h_{od}(t; tau, omega)$ sont une paire quadratique de filtres de Gabor 1D appliqués uniquement à la dimension temporelle, et définis par les équations I.2.2 et I.2.3 (omega est souvent fixé à 4/tau). Les paramètres σ et τ correspondent approximativement aux échelles spatiale et temporelle de détection.

$$R(x, y, t) = (I(x, y, t) \star g(x, y; \sigma) \star h_{ev}(t; \tau, \omega))^2 + (I(x, y, t) \star g(x, y; \sigma) \star h_{od}(t; \tau, \omega))^2 \quad \text{(I.2.1)}$$

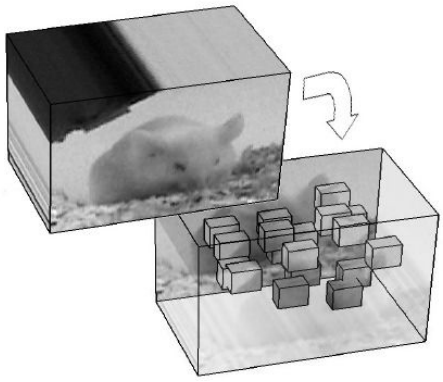


FIGURE I.2.3.3 Illustration des cuboids. Extrait de [59].

$$h_{ev}(t;\tau,\omega) = -cos(2\pi t\omega)e^{-t^2/\tau^2} \tag{I.2.2}$$

$$h_{od}(t;\tau,\omega) = -sin(2\pi t\omega)e^{-t^2/\tau^2} \tag{I.2.3}$$

Les points d'intérêt sont les maximum locaux de la fonction de réponse $R(x,y,t)$. Ce détecteur était à l'origine conçu pour détecter des mouvements périodiques rapides, comme les battements d'aile d'un oiseau, mais s'est révélé être très efficace et robuste dans la détection d'éléments non périodiques comme des coins spatio-temporels, ce qui est important dans l'étude de l'imitation. Bien qu'il soit vrai qu'un certain nombre de gestes humains possèdent une certaine périodicité (faire "coucou" de la main, marcher, courir, applaudir, etc), un grand nombre de gestes ne le sont pas, et nécessitent donc un détecteur efficace sur les événements non-périodiques. Ce détecteur a été choisi plutôt que celui de Laptev pour son efficacité, comme démontré dans [106] et surtout parce qu'il permet une représentation plus dense des vidéos en détectant plus de points significatifs. Nous utilisons le code fourni par l'auteur.

I.2.3.2 Description

Après la détection des Points d'Intérêts Spatio-Temporels (STIPs), chacun d'entre eux est décrit par un vecteur caractéristique aussi appelé descripteur. L'objectif principal de ce processus est de trouver des caractéristiques qui décrivent le motif. De multiples descripteurs ont été développés en traitement d'images et de vidéos au fil des années. Nous allons seulement en présenter deux, qui ont été introduits par Dollár dans [59] et Laptev dans [112].

I.2.3.2.a Cuboïdes

Le descripteur présenté par Dollár dans [59] considère des cuboïdes de 19x19x11 pixels (qui correspondent à 9 pixels de chaque côté, et 5 images avant et après le point d'intérêt détecté), comme illustré sur la figure I.2.3.3. Les niveaux de gris de ce cuboïde sont conservés intégralement, ce qui mène à un vecteur de description d'une longueur 3971 pour chaque point. Ce descripteur est de très grande dimension (donnée par le voisinage brut de chaque point), et est sensible aux changements d'intensité.

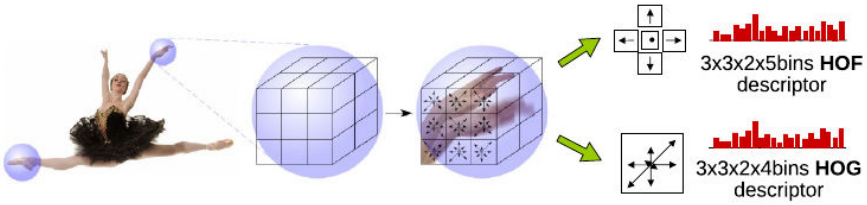


FIGURE I.2.3.4 Illustration des histogrammes (de gradient ou de flot optique), qui sont concaténés pour obtenir un vecteur de description (à droite). Extrait de [106].

I.2.3.2.b Histogrammes de Gradients Orientés et Histogrammes de Flot Optique

Ces descripteurs sont calculés sur un volume spatio-temporel autour de chaque point composé d'un voisinage spatial 3x3 et 2 images dans le temps, comme décrit dans l'article originel [112].

Les Histogrammes de Gradients Orientés (HOG) sont très similaires au très célèbre descripteur SIFT. Pour chaque cellule du volume spatio-temporel, les HOG sont calculés en utilisant plusieurs orientations (ici 4, mais 8 sur la figure I.2.3.4). Ils sont alors concaténés en un seul vecteur caractéristique, menant alors à un descripteur de taille 72 (3x3x2x4 éléments).

Les Histogrammes de Flot Optique (HOF) sont basés sur le même principe que les HOG mais en utilisant les vecteurs du flot optique. Les HOF sont calculés sur chaque cellule du volume, mais en utilisant 5 catégories (4 pour l'orientation et une pour l'absence de mouvement). Les HOF sont alors concaténés dans un descripteur de taille 90 (3x3x2x5 éléments).

Le descripteur HOG-HOF est la concaténation des descripteurs HOG et HOF, et est donc de dimension 162. Une fois les points décrits à l'aide d'un des descripteurs présentés, les modèles sac-de-mots sont construits.

I.2.3.3 Modèle sac-de-mots

L'algorithme k-moyennes est utilisé sur les descripteurs d'une base de données afin d'obtenir un dictionnaire de k mots visuels. Cette étape est faite hors ligne sur une base externe introduite dans [54] et décrite dans la section I.2.5.3 .

Pour de nouvelles vidéos, les points d'intérêt spatio-temporels (STIPs) sont détectés et caractérisés grâce aux descripteurs évoqués dans le paragraphe précédent (HOG, HOF ou HOG-HOF). Ils sont alors ensuite assignés à des mots du dictionnaire de manière souple, proportionnellement à l'inverse de la distance entre le point et le centre des k-moyennes. Tous les mots sont alors rassemblés au cours du temps dans une variable BOW où BOW(k) est un vecteur dont la composante m spécifie le nombre de fois où le mot m apparaît dans l'intervalle d'images [k,k+25]. Ce modèle décrit la structure temporelle de la vidéo, et résume son contenu visuel.

L'information spatiale semble essentielle à l'étude de l'imitation gestuelle. Cependant, comme il peut être vu sur la figure I.2.3.5, le modèle BOW perd la structure spatiale. En effet, quelle que soit la position de deux points d'intérêt détectés, ils apparaissent de la même façon dans le BOW, ce qui peut être préjudiciable à la reconnaissance de l'imitation. De ce fait, l'information spatiale va être réintroduite dans le modèle BOW.

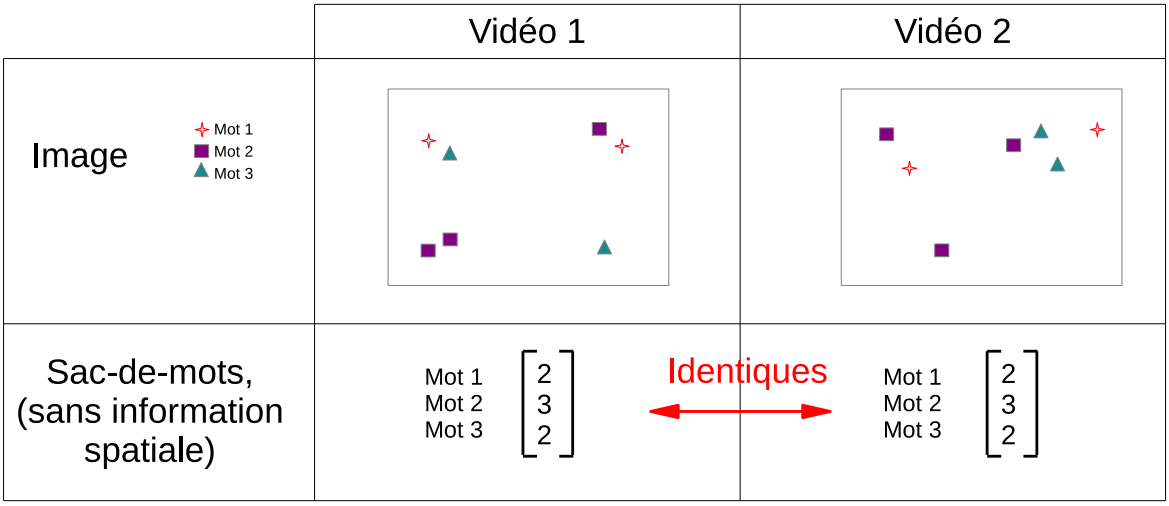


FIGURE I.2.3.5 Sans information spatiale, les deux BOW sont les mêmes, alors même que les images ne se ressemblent pas.

I.2.3.4 Réintroduction de l'information spatiale

Nous venons de voir que le modèle classique de BOW perd l'information spatiale. Cependant, pour de la reconnaissance d'imitation, il est impensable de ne pas avoir cette information, car un geste est à la fois défini par son mouvement, sa forme, mais aussi par sa position par rapport au reste du corps. Afin de réintroduire l'information spatiale, la première idée consiste à étendre le vecteur caractéristique en ajoutant les coordonnées spatiales au descripteur. Cependant, ceci mène souvent à de mauvais résultats, puisque les coordonnées comptent seulement pour deux composantes (x et y) du vecteur, alors que le descripteur possède soit 72 (HOG), 90 (HOF) ou même 162 (HOG-HOF) composantes.

Une seconde solution, plus complexe, consiste à séparer le contenu visuel et les coordonnées spatiales pour chaque point. Deux dictionnaires sont alors appris pour tous les points de la base d'apprentissage, que nous allons détailler ci-dessous.

I.2.3.4.a Dictionnaire spatial

Le premier dictionnaire ne prend en compte que la position des points. Afin de pouvoir détecter l'imitation en s'affranchissant de la position des participants dans les vidéos, les coordonnées des points sont normalisées dans un espace centré sur le visage des participants, comme illustré dans la figure I.2.3.6. Tout d'abord, les visages sont détectés dans chaque image en utilisant un détecteur de visage appelé IntraFace [197]. Un filtre de Kalman est ensuite utilisé afin de stabiliser les résultats et de gérer les quelques détections manquantes. Les points détectés sont alors recentrés dans un espace centré sur le visage (invariance en translation du participant), et normalisé par la taille de la boite enveloppante (invariance en distance participant-caméra). Un algorithme k-moyennes est enfin appliqué. Ses k-centres ($Cluster_1$ à $Cluster_4$ représentés sur la figure I.2.3.7) correspondent à des zones spatiales, et forment le dictionnaire spatial, qui sera présenté dans la section I.2.6.2 .

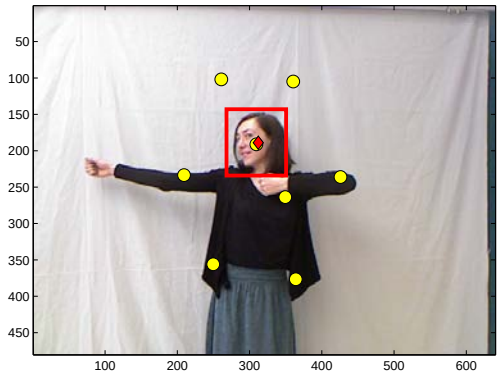


FIGURE I.2.3.6 Visage détecté par IntraFace, et filtré par un filtre de Kalman. Les clusters spatiaux sont montrés en jaune.

I.2.3.4.b Dictionnaire visuel

Le second dictionnaire est appris seulement avec les descripteurs visuels (comme dans le modèle BOW classique présenté dans la section I.2.3.3), et mène à des mots visuels (Mot_1 à Mot_3 dans la figure I.2.3.7).

I.2.3.4.c Modèle BOW avec information spatiale

Une fois les deux dictionnaires appris, le modèle BOW avec information spatiale est obtenu par la concaténation des descripteurs (Mot) estimés pour chaque cluster spatial ($Cluster$), comme illustré dans la figure I.2.3.7. Chaque point participe à chaque composante du descripteur final de manière inversement proportionnelle à sa distance avec les clusters spatiaux et les mots visuels (codage souple).

I.2.4 Mesure de la similarité

Lorsque toutes les vidéos sont décrites à l'aide d'un modèle BOW, une mesure de similarité est utilisée afin de mesurer le délai, le degré et l'orientation de l'imitation entre les partenaires. La corrélation croisée est l'une des méthodes les plus simples et les plus utilisées afin de comparer deux séries temporelles. Dans les interactions naturelles, le délai d'imitation entre les partenaires varie tout le temps, et les partenaires changent de rôle constamment. Ainsi, une mesure de similarité comme la corrélation, si elle est évaluée globalement, ne permet pas de prendre en compte les variations de délai entre partenaires. Il existe deux possibilités pour répondre à cela : considérer l'interaction sur des fenêtres temporelles plus courtes (dans lesquelles le délai sera supposé constant), ou utiliser une mesure dynamique. Ainsi, nous proposons aussi d'évaluer une version modifiée de Dynamic Time Warping dans laquelle la similarité est mesurée.

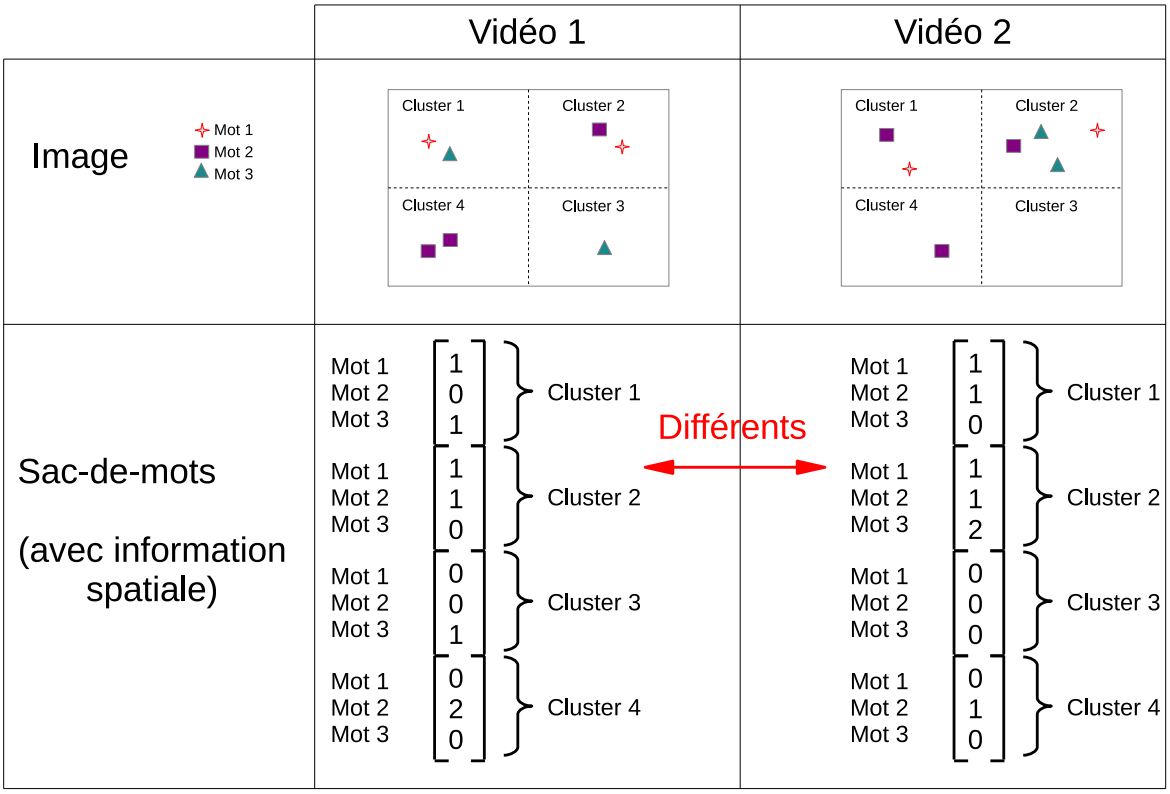


FIGURE I.2.3.7 Introduction de l'information spatiale dans le BOW qui permet d'avoir une représentation disctinctes des deux images.

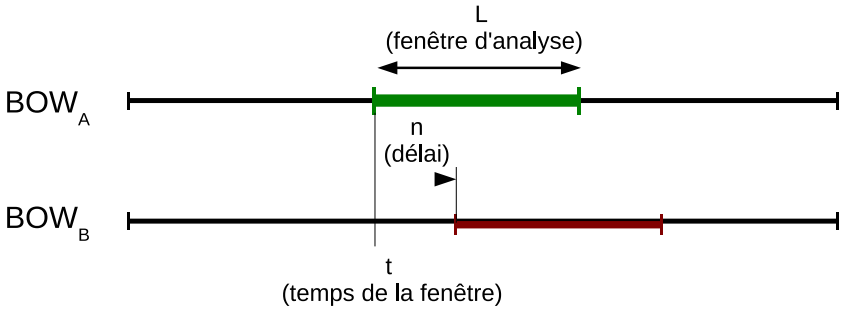


FIGURE I.2.4.8 Illustration des variables utilisées.

I.2.4.1 Corrélation croisée

Pour estimer localement, au temps t, la corrélation entre deux signaux, deux paramètres doivent être considérés : la durée L de la fenêtre temporelle pendant laquelle les signaux sont observés et le décalage temporel (délai) n entre ces signaux, comme illustré sur la figure I.2.4.8. Si $BOW_A(t)$ et $BOW_B(t)$ représentent les BOW caractérisant les séquences A et B, la corrélation estimée avec un délai n et une fenêtre d'observation de longueur L est donnée par :

$$[BOW_A \star BOW_B]_n(t) = \sum_{k=t}^{t+L} \frac{BOW_A(k)^T . BOW_B(k+n)}{||BOW_A(k)||_2 * ||BOW_B(k+n)||_2} \tag{I.2.4}$$

où $||x||_2$ est la norme L2 de x.

I.2.4.2 Dynamic Time Warping

Le Dynamic Time Warping (DTW) est une mesure qui permet de comparer deux séries temporelles non alignées. Cependant, comme nous allons le présenter ici, nous n'allons pas utiliser le DTW pour mesurer une distance, mais pour mesurer une similarité.

Alors que la distance mesure le degré de "différence" entre deux éléments, un indice de similarité mesure le degré de "ressemblance". Dans le cas d'une distance, on cherche habituellement les éléments les plus proches, c'est-à-dire que l'on cherche la distance minimale. Dans le cas d'une similarité, on cherche les éléments les plus similaires, c'est-à-dire l'indice de similarité maximal.

Cette mesure se fait en plusieurs étapes, qui seront développées ci-dessous. Tout d'abord, une matrice de similarité cumulée est construite de manière séquentielle. Ensuite, un chemin de similarité maximale est recherché.

Alors que la méthode originelle de Levenshtein [116] mesure une distance entre deux séries, celle développée par Needleman [132] permet de mesurer la similarité, et elle a été largement utilisée avec des brins d'ADN pour la comparaison de génomes ou l'alignement de séquences. Dans cette dernière méthode, le calcul de la matrice de similarité cumulée suit l'équation I.2.5, où $\star$ se réfère à la corrélation normalisée définie dans l'équation I.2.4. Des explications plus

FIGURE I.2.4.9 DTW classique.

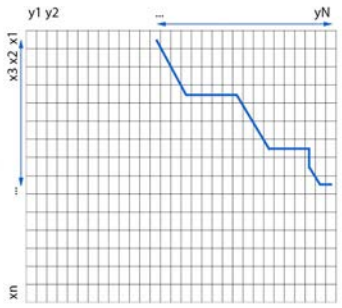


FIGURE I.2.4.10 DTW avec superposition.

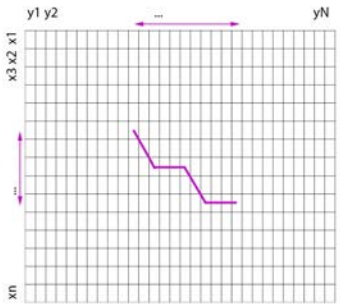


FIGURE I.2.4.11 DTW avec alignement local.

détaillées sur le DTW peuvent être trouvées dans le Chapitre 4 de [129].

$$D(i,j) = \max \begin{cases} D(i-1,j) - (1 - [BOW_A(i-1) \star BOW_A(i)]) \\ D(i,j-1) - (1 - [BOW_B(j-1) \star BOW_B(j)]) \\ D(i-1,j-1) + [BOW_A(i) \star BOW_B(j)] \end{cases} \quad \text{(I.2.5)}$$

Une fois que la matrice de similarité cumulée est calculée, le chemin de similarité la plus grande est recherché. Habituellement, comme l'on souhaite mettre en correspondance deux séries temporelles ensemble, le chemin est calculé de telle sorte que toute la séquence A soit alignée avec toute la séquence B, comme montré dans la figure I.2.4.9. Le chemin est obtenu en prenant la valeur de la case correspondant à la fin des deux séries temporelles, et en remontant les cases qui ont permis de donner le résultat.

Mais il peut être nécessaire d'ignorer le début d'une séquence et la fin de l'autre, car cela permet d'avoir des gestes de différente longueur, entourés de gestes qui ne font pas partie de l'imitation. Comme il est montré dans la figure I.2.4.10, la variante de détection de superposition permet ceci. Les seules modifications à l'algorithme originel consistent à mettre à zéro la première ligne et la première colonne de la matrice de similarité cumulée, et à rechercher le score dans la dernière colonne ou la dernière ligne.

Cependant, si un geste est précédé ou suivi d'éléments perturbateurs dans les deux séries, il peut être intéressant de vouloir isoler les morceaux d'interaction correspondant. La méthode d'alignement local permet ceci, et a été présentée par Smith-Waterman dans [170]. Elle est présentée dans la figure I.2.4.11. Elle permet de comparer des gestes courts dans des parties de videos. Pour cela, on modifie l'équation I.2.5 en ajoutant une quatrième option, un zéro. $D(i,j)$ est le maximum des trois options précédentes, ou est égal à 0. Ainsi, si les trois nombres devaient être négatifs (ce qui signifie que l'on n'a plus du tout de similarité), le 0 permet un "reset" du chemin, permettant d'avoir un nouveau point de départ, comme montré dans la figure I.2.4.11. De plus, le score maximal est maintenant recherché dans toute la matrice. Dans la suite, seule la variante de l'alignement local a été utilisée.

Enfin, on peut choisir de ne pas prendre une matrice carrée, afin d'autoriser les recherches pour un certain délai, comme illustré sur la figure I.2.4.12. Dans cet exemple, on pourra par exemple parler de DTW 75-75 si les deux fenêtres d'analyse ont une longueur de 75 échantillons,

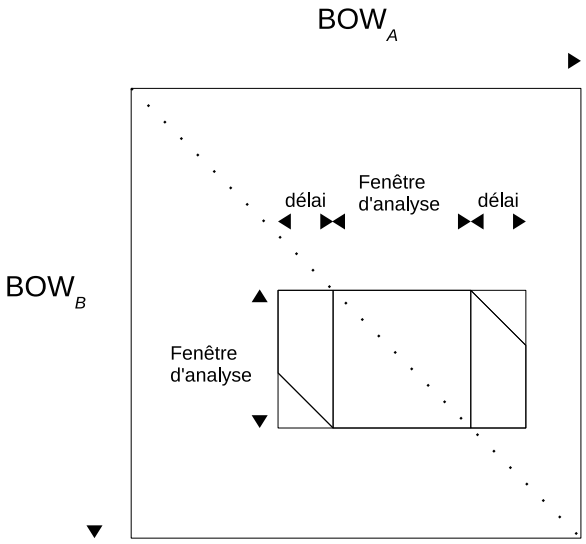


FIGURE I.2.4.12 Illustration des fenêtres considérées pour le DTW.

ou de DTW 75-125 si on autorise un délai de 25 échantillons (125 = 25+75+25).

I.2.4.3 Les paramètres importants

Suite aux travaux de Feldman et al. [72], un consensus a établi que le délai moyen entre un parent et son enfant pendant une interaction est considéré autour de 3 secondes. Cependant, dans le cadre d'une interaction naturelle, le délai n est inconnu et de plus varie dans le temps.

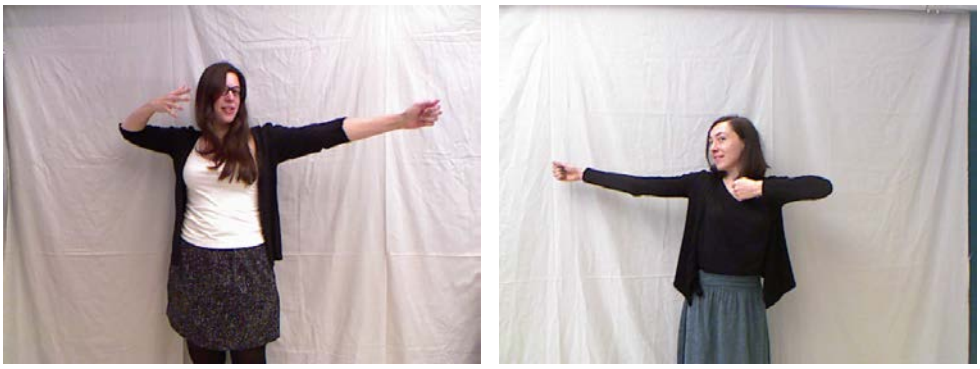
Les corrélations sont donc estimées pour plusieurs délais n et le degré de l'imitation est estimé en chaque temps t avec :

$$degré(t) = \max_{n\in[-n_{max},\ldots,+n_{max}]}([BOW_A \star BOW_B]_n(t)) \quad \text{(I.2.6)}$$

Le délai maximum n_{max} et la longueur de la fenêtre d'analyse vont alors jouer un rôle majeur dans l'évaluation de la méthode pour la reconnaissance de l'imitation, et sont illustrés sur la figure I.2.4.8. Leur influence sur la mesure sera étudiée dans la section I.2.6.7 .

I.2.5 Création d'une base de données

Comme il a été présenté dans la section I.1.3 , un certain nombre de bases de données existent. Malheureusement, peu de bases à part celle de Sun et al. [173] s'intéressent à l'étude de l'imitation et possèdent une annotation. Cependant cette base, principalement dédiée à l'analyse des expressions faciales, présente très peu de gestes et de mouvement, et ne peut donc pas être utilisée dans cette étude. La base de Delaherche et al. introduite dans [54] visait quand à elle à séparer le temporel du spatial. Ainsi, les gestes des participants sont effectués de manière simultanée, ce qui est une image pauvre de la réalité, dans laquelle les interactions naturelles exercent des gestes avec un certain délai entre eux. Cette absence de délai entre les partenaires justifie l'abandon de cette base pour l'évaluation de notre méthode.



(a) Participant 1 (b) Participant 2

FIGURE I.2.5.13 Participant 1, initiant un geste, et participant 2 l'imitant en miroir.

Le manque de base de données concernant l'imitation nous a mené à créer une nouvelle base, en suivant plusieurs conditions, qui semblaient indispensables à une évaluation objective de la méthode : les gestes effectués par les participants doivent être libres, que ce soit dans le temps, ou dans la forme (aucun geste n'est prédéfini). De plus, la base se déroulera dans une interaction dyadique. Cette liberté laissée aux participants, à la fois en forme et en temps, mène à une base de donnée plus naturelle qui sera utile aux chercheurs voulant développer de nouveaux algorithmes tout en s'épargnant la lourde tâche de la récolte de données.

Ainsi, le manque de base de données nous a amené à créer une nouvelle base respectant les conditions suivantes : l'imitation gestuelle est effectuée dans une interaction dyadique, sans gestes prédéfinis.

I.2.5.1 Protocole

Deux participants se tiennent face à face, à une distance d'environ 2 mètres. L'un d'entre eux, le meneur, effectue librement des gestes pendant que l'autre a pour instruction de l'imiter, en miroir. Les gestes ne sont pas appris ou choisis parmi un dictionnaire, et ne sont pas liés nécessairement à une sémantique. De plus, ils ne sont pas segmentés dans le temps, et sont effectués de manière naturelle, sans retour à une position neutre de repos. Cette liberté de l'exécution des actions, à la fois en temps et en forme, nous a conduit à implémenter une méthode non supervisée. En effet, la créativité des participants a permis de voir des gestes variés, comme des mouvements de la vie de tous les jours (manger, boire, se brosser les dents, se laver les cheveux, etc), des instruments de musique, des imitations d'animaux, ou même des gestions ou actions sans interprétation possible. La seule contrainte imposée aux participants dans le choix des gestes était de rester sur place afin de rester dans l'angle de vue des caméras. Un exemple d'imitation est visible sur la figure I.2.5.13.

I.2.5.2 Composition détaillée

La base de donnée est composée de 9 dyades. Chaque personne de la dyade est filmée par une caméra (640x480 pixels à 25 images par seconde) et les deux vidéos ont été enregistrées

(a) Image 1 (b) Image 2 (c) Image 3

FIGURE I.2.5.14 Trois exemples d'images extraites de la base externe d'apprentissage.

et synchronisées ensemble. Chaque dyade participe à deux sessions, pour lesquelles chaque participant a été d'abord meneur puis imitateur. Ainsi, un total de 18 interactions durant approximativement 2 minutes a été enregistré.

Les 18 participants étaient des volontaires. La base est composée de 8 femmes et de 10 hommes en dyades mixtes. La plupart d'entre eux (16) ont accepté la diffusion et l'utilisation de leur image. L'âge moyen est de 22,9 ans.

Il doit être noté que l'hypothèse d'imitation parfaite durant toutes les vidéos est contestable. En effet, elle est mise en défaut à certaines occasions, principalement lorsque des mouvements rapides et complexes sont effectués par le meneur. Ces occasions restent cependant assez rares pour que l'hypothèse d'imitation parfaite reste valable et ne remette pas en cause le protocole.

Pour finir, afin de ne pas favoriser les vidéos les plus longues et biaiser les résultats, seules les 2700 premières images de chaque interaction ont été gardées.

I.2.5.3 Base externe d'apprentissage

Pour tous les résultats qui seront présentés dans la section I.2.6 , les dictionnaires de mots visuels et de centres spatiaux ont été appris sur une base externe, qui a été présentée dans [54]. Cette base contient 10 minutes de vidéo, et présente une vue frontale de participants faisant des gestes, comme illustré dans la figure I.2.5.14.

La luminosité, les gestes, et la composition expérimentale sont différents de la base que nous avons réalisée (voir section I.2.5), à part pour le point de vue (frontal) qui est similaire. Cette base propose des dyades exécutant des gestes identiques, tout en étant synchronisés. Cette particularité ne sera pas utilisée ici, puisque les vidéos seront utilisées de manière indépendante, et ne serviront à aucun moment à l'évaluation.

I.2.6 Résultats

Une fois les dictionnaires spatiaux et visuels appris sur la base externe, la méthode va pouvoir être évaluée sur la base présentée à la section I.2.5 .

I.2.6.1 Méthodologie de mesure

Afin de pouvoir prendre une décision sur la présence ou non d'imitation, il est important d'avoir un score de référence auquel comparer les scores d'imitation. A cet effet, la création de vidéos de non-imitation est nécessaire, et sera donc présentée ci dessous. Pour chaque couple de vidéos (d'imitation et de non imitation), la méthode complète est appliquée, permettant d'obtenir à chaque instant un score mesurant la similarité entre les vidéos. Ce score est ensuite seuillé pour savoir si de l'imitation existe ou non. A partir de ces résultats, diverses métriques (présentées ci dessous) peuvent être appliquées afin d'évaluer la méthode.

I.2.6.1.a Imitation et non imitation

Dans la base que nous avons créée (section I.2.5), seules des vidéos d'imitation existent. En effet, il est difficile de demander à des participants de "ne pas s'imiter". Si cela avait été fait, nous aurions obtenu des vidéos où les participants font volontairement des gestes différents de leur partenaire. Or, ne pas s'imiter n'exclue pas la possibilité de faire le même geste dans un horizon temporel proche de son partenaire. Cette imitation due au hasard est dure à créer dans un dispositif expérimental avec deux participants en face à face. Afin d'éviter des conflits théoriques sur ce que n'est pas l'imitation, nous avons mis en place l'idée présentée par Bernieri dans [18], où des vidéos de pseudo-interactions sont créées en joignant des vidéos de différentes dyades. Ceci permet d'obtenir 153 vidéos de "non imitation" créées de toute pièce. Il est important de noter que ces vidéos peuvent contenir de l'imitation (sporadique), présente par hasard.

I.2.6.1.b Métrique

Une fois le score d'imitation seuillé, nous obtenons des décisions binaires sur la présence ou l'absence d'imitation dans les fenêtres temporelles concernées. Une matrice de confusion est alors calculée, en supposant que de l'imitation était présente à chaque instant dans les vidéos d'imitation, et absente à chaque instant dans les vidéos de non-imitation. Ainsi, on remplit les quatre cases de la matrice de confusion présentée dans la table I.2.1 :

Vrai Positif (VP) : une fenêtre temporelle d'un couple de vidéos d'imitation, dans laquelle de l'imitation a été détectée

Vrai Négatif (VN) : une fenêtre temporelle d'un couple de vidéos de non-imitation, dans laquelle de l'imitation n'a pas été détectée

Faux Positif (FP) : une fenêtre temporelle d'un couple de vidéos de non-imitation, dans laquelle de l'imitation a été détectée

Faux Négatif (FN) : une fenêtre temporelle d'un couple de vidéos d'imitation, dans laquelle de l'imitation n'a pas été détectée

A partir de la matrice de confusion, estimée pour tous les temps et tous les couples de vidéos, plusieurs indices peuvent être calculés, dont les plus utilisés sont :

La précision, égale à $\frac{VP}{VP+FP}$

Le rappel, aussi appelé Taux de Vrai Positif (TVP) ou sensibilité, égal à $\frac{VP}{VP+FN}$

Taux de Faux Positif (TFP), égal à 1− spécificité $= \frac{FP}{FP+VN}$

Le score F1, égal à $\frac{2*VP}{2*VP+FP+FN}$

TABLE I.2.1 Matrice de confusion

Décision / Vérité	Positif	Négatif
Positif	VP	FN
Négatif	FP	VN

Toutes ces métriques dépendant du seuil de décision évoqué en début de paragraphe. Une façon de s'en affranchir est d'utiliser les courbes ROC qui consistent à tracer le TVP en fonction du TFP en faisant varier le seuil de décision (chaque seuil donne un point).

Plus un système est efficace, plus il s'approche du point en haut à gauche de coordoonées (0 ;1). La diagonale représente un système prenant des décisions aléatoires. L'Aire Sous la Courbe (AUC pour Area Under the Curve) permet de quantifier l'efficacité du système, et de comparer deux systèmes ou deux jeux de paramètres.

Dans le cas de classes déséquilibrées entre exemples positifs et négatifs, il a été montré par [99] que l'aire sous la courbe ROC (AUC) est moins biaisée que le score F1. Ainsi, nos résultats seront présentés sous forme de courbes ROC et les scores données en AUC.

I.2.6.2 Position des clusters spatiaux

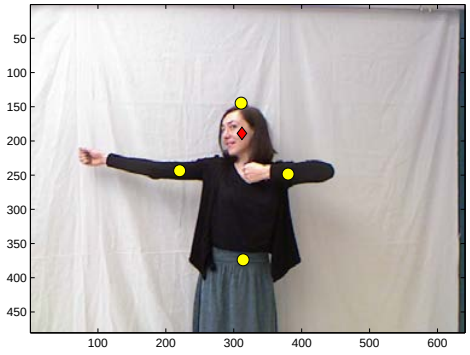
Nous avons vu dans la section I.2.3.4.c que les clusters spatiaux appris sur la base externe correspondent à des zones spatiales d'intérêt. Il est possible de faire varier le nombre de centres spatiaux appris, comme illustré dans la figure I.2.6.15.

Les centres spatiaux obtenus possèdent une symétrie droite/gauche forte, montrant que les gestes effectués dans la base d'apprentissage n'étaient pas uni-latéraux. De plus, les points possèdent une symétrie forte par rapport à la position des épaules. Ceci s'explique par le fait que l'espace atteignable par les bras des participants (qui concentrent la plupart des points d'intérêts) est lui aussi centré autour des épaules. Enfin, la figure I.2.6.15c possédant 32 clusters spatiaux permet de voir que la densité des points proches du visage et des épaules est plus grande que ceux en limite atteignable par les mains des participants.

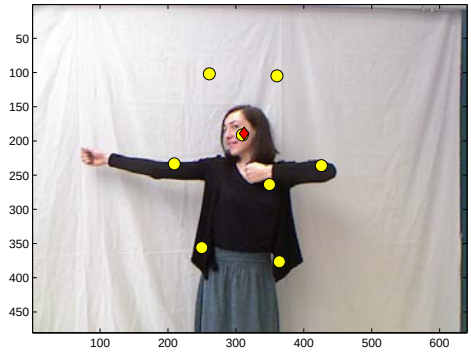
I.2.6.3 Validation du protocole

Afin d'être sûrs que la méthode est capable de séparer l'imitation de la non-imitation, un premier test a été effectué avec la corrélation avec une fenêtre d'interaction de longueur 75 images (3 secondes). A chaque instant, pour chaque couple de vidéos, le score de corrélation est calculé, tel que défini équation I.2.6. Les résultats sont affichés sur la figure I.2.6.16 pour un couple de vidéos d'imitation. Afin d'évaluer la séparabilité des deux classes, les distributions des scores pour les couples de vidéos d'imitation et de non-imitation ont été calculées et représentées sur la figure I.2.6.17. Un test statistique de Student (T-test) permet de vérifier, comme l'on peut le voir sur la figure, que les deux séries sont bien séparables (h=1 avec $p < 0,05$). Le processus est donc applicable pour la mesure non supervisée de l'imitation.

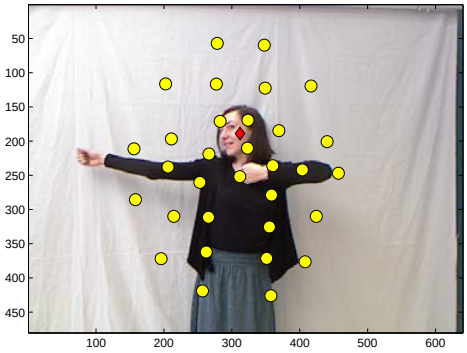
La figure I.2.6.17 permet aussi de voir que même lorsqu'aucune imitation volontaire n'est



(a) 4 clusters spatiaux.



(b) 8 clusters spatiaux.



(c) 32 clusters spatiaux.

FIGURE I.2.6.15 Centres des clusters spatiaux appris pour différents nombres de clusters. Le point rouge est le centre du visage détecté.

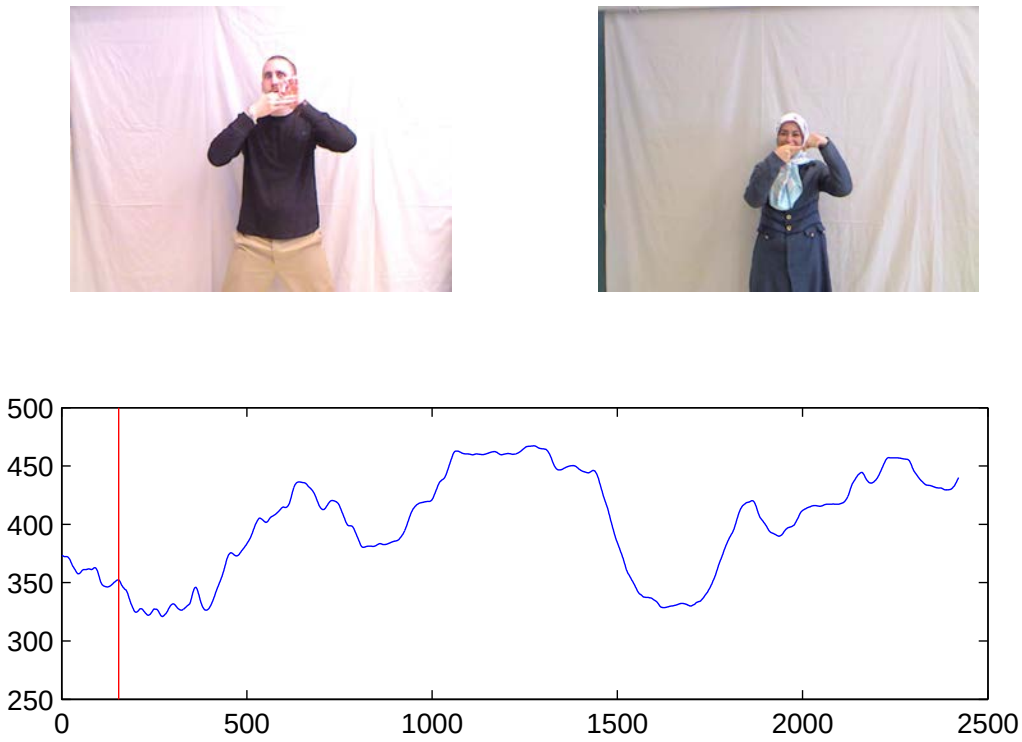


FIGURE I.2.6.16 Scores de la corrélation pour un couple de vidéos d'imitation. Le score est tracé en fonction du temps (courbe bleu). La barre verticale rouge montre le temps actuel correspondant aux deux images.

présente (cas des pseudos-interactions de non-imitation), il existe toujours une quantité non nulle d'imitation (la distribution des données de non-imitation possède une moyenne de $0,41$). Ceci peut s'expliquer grâce à plusieurs facteurs. Tout d'abord, il est difficile de réaliser deux gestes sans qu'aucune ressemblance n'existe. Ensuite, le codage des sacs-de-mots est fait de manière souple, ce qui signifie que chaque élément du descripteur contient des éléments non nuls, même si faibles, ce qui mène à une mesure de similarité non nulle.

I.2.6.4 Choix de la mesure de similarité

Le processus étant applicable, nous avons tout d'abord décidé d'évaluer les performances de trois méthodes de mesures : la corrélation avec une fenêtre d'analyse de longueur 75 frames (3 secondes), le DTW local avec la même fenêtre d'analyse de longueur 75 frames (DTW 75-75) et le DTW local autorisant en plus un délai de 25 frames (DTW 75-125). Pour cela, les points ont été décrits à l'aide du descripteur HOG-HOF, avec un dictionnaire de 64 mots, sans dictionnaire spatial. La corrélation est évaluée sur une fenêtre d'interaction de longueur 75 images, avec un délai maximum autorisé de 25 images.

Les courbes ROC pour les trois méthodes sont représentées sur la figure I.2.6.18. Ainsi, même si la corrélation et le DTW 75-125 ont un score AUC comparable, cette dernière possède

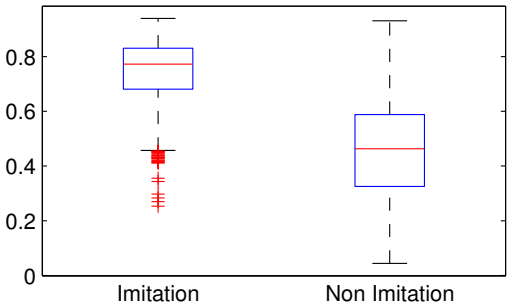


FIGURE I.2.6.17 Distributions des scores de corrélation (un T-test avec $p < 0,05$ pour h=1 prouve que ces deux distributions sont bien séparables).

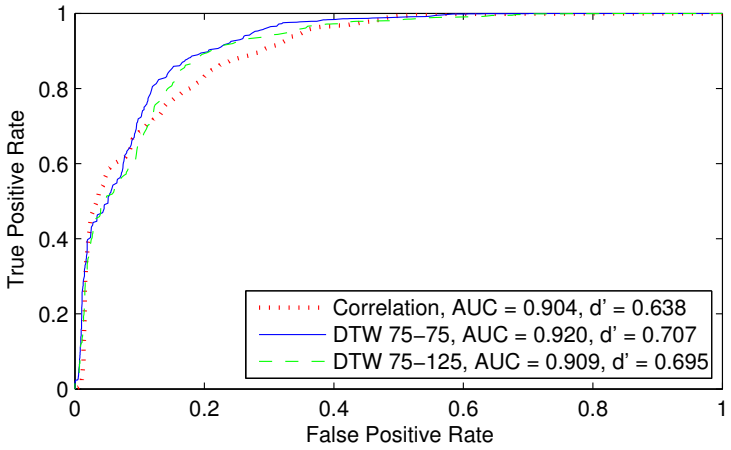


FIGURE I.2.6.18 Courbes ROC pour les trois méthodes.

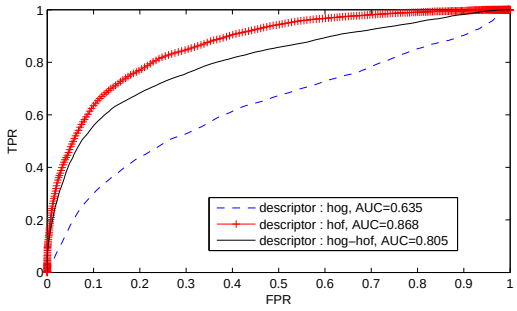


FIGURE I.2.6.19 Résultats pour les descripteurs HOG, HOF et HOG-HOF.

une courbe avec un point plus optimal (plus proche du point [0,1]). On peut aussi remarquer que les résultats sont légèrement meilleurs avec le DTW 75-75, ce qui peut s'expliquer par le fait que l'on limite la zone de recherche (voir la section I.2.6.7 pour une analyse plus détaillée sur ce paramètre). Cependant, il faudra noter deux points en défaveur du DTW. Le premier est que le temps de calcul nécessaire au DTW est quadratique à la longueur de la fenêtre d'analyse. Ensuite, l'analyse et l'interprétation des paramètres de l'interaction est complexifiée avec le DTW : le délai est variable (on peut cependant calculer un délai moyen), et la taille de la fenêtre d'interaction aussi. Ainsi, dans la suite, nous garderons la corrélation comme mesure de similarité.

I.2.6.5 Choix du descripteur

Nous évaluons ici les performances obtenues par les différents descripteurs présentés dans la section I.2.3.2 , à savoir HOG, HOF et la combinaison des deux HOG-HOF.

Pour cela, le délai maximum autorisable et la taille de la fenêtre d'interaction ont été choisis à des valeurs permettant un compromis entre le temps de calcul et une mesure globale de l'imitation. Ainsi le délai a été fixé à 5 secondes (125 images) et la taille de la fenêtre d'interaction à 10 secondes (250 images). De plus, le nombre de mots visuels a été fixé à 64, sans ajout d'information spatiale.

La figure I.2.6.19 présente les résultats obtenus pour les trois descripteurs. Le descripteur HOG est celui qui a les plus mauvaises performances. En effet, il prend seulement en compte l'apparence locale autours des points d'intérêts. La combinaison HOG-HOF a du coup de meilleures performances, car elle tient compte du mouvement. Cependant, HOF a de meilleurs résultats que HOG-HOF. Ceci peut s'expliquer par le fait que les histogrammes de gradients peuvent être perturbés par l'aspect local des vêtements. Le Flot Optique n'est quand à lui pas perturbé par ce phénomène, et conserve donc seulement de l'information utile, ce qui mène à de meilleurs résultats.

Ainsi, dans la suite, les HOF seront utilisés, et l'influence du nombre de mots visuels et de clusters spatiaux pourra être étudié.

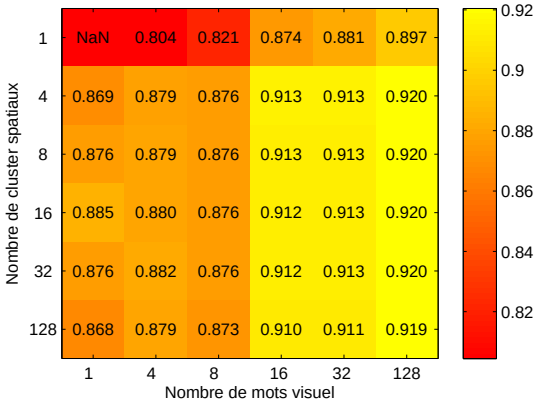


FIGURE I.2.6.20 AUC en fonction des nombres de mots visuels et spatiaux.

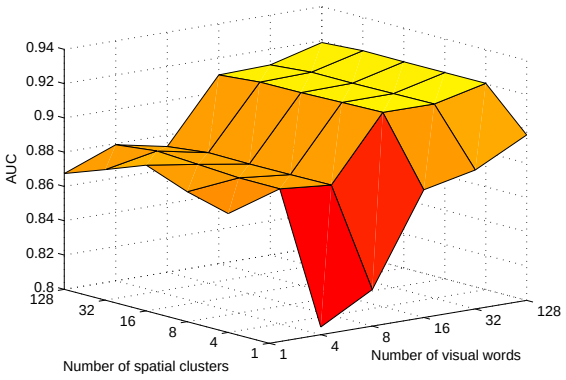


FIGURE I.2.6.21 AUC en fonction du nombre de mots visuels et spatiaux, représentation sous forme de surface.

I.2.6.6 Nombre de mots visuels et de clusters spatiaux

Nous allons ici étudier les résultats obtenus pour différentes combinaisons de nombre de mots visuels et de clusters spatiaux. Nous allons faire varier le nombre d'éléments dans les deux dictionnaires entre 1 (ce qui signifie que le dictionnaire n'est pas utilisé) et 128, comme montré dans la figure I.2.6.20. Des valeurs plus hautes n'ont pas été testées puisque l'amélioration obtenue était insignifiante comparée à la complexification du dictionnaire. Par exemple, un dictionnaire combiné de 32 clusters spatiaux et de 128 mots visuels contient 4096 possibilités.

La figure I.2.6.20 permet de voir plusieurs résultats intéressants. Tout d'abord, si l'information spatiale n'est pas inclue (1 seul cluster spatial), on peut voir sur la figure I.2.6.20 que les résultats restent faibles tant que le nombre de mots visuels n'augmente pas de manière drastique (128). Le fait d'inclure l'information spatiale, même de manière faible (4 clusters spatiaux) suffit à obtenir une amélioration significative quel que soit le nombre de mots visuels utilisé.

D'un autre côté, on pourrait penser que la seule localisation spatiale des points pourrait suffire à mesurer l'imitation entre les partenaires. Cependant, comme il peut être vu sur la figure I.2.6.20, pour 1 mot visuel, l'information spatiale seule a des performances assez limitées, ce qui prouve que l'information sur l'apparence est aussi nécessaire.

Enfin, on peut remarquer que la combinaison de 4 clusters spatiaux et de 16 mots visuels marque un tournant dans les résultats, l'AUC passant de 0,88 à 0,91. Après cette combinaison, bien que les dictionnaires augmentent en taille de manière significative (et la complexité de la combinaison de manière encore plus importante), les résultats quand à eux n'ont qu'une amélioration limitée ($< 0,01\%$).

Ainsi, dans la suite, 4 clusters spatiaux et 16 mots visuels seront utilisés dans les dictionnaires.

fenêtre d'analyse (secondes)

délai (secondes)	0.2	0.4	1	2	4	5	6	8	12	16	20	24	28	32	36	40	50	60	70	80	90	100
0	0.776	0.780	0.795	0.823	0.859	0.875	0.882	0.897	0.920	0.935	0.946	0.951	0.951	0.960	0.965	0.959	0.971	0.971	0.972	0.977	0.978	0.981
0.2	0.788	0.792	0.807	0.834	0.870	0.888	0.891	0.908	0.928	0.945	0.955	0.958	0.959	0.966	0.963	0.967	0.975	0.977	0.978	0.979	0.979	0.983
0.4	0.799	0.802	0.816	0.843	0.879	0.895	0.898	0.916	0.934	0.950	0.959	0.963	0.965	0.969	0.966	0.970	0.977	0.979	0.979	0.980	0.978	0.985
0.6	0.805	0.808	0.822	0.846	0.882	0.895	0.899	0.918	0.935	0.952	0.960	0.964	0.966	0.970	0.967	0.971	0.977	0.981	0.981	0.981	0.979	0.984
0.8	0.807	0.809	0.822	0.845	0.881	0.893	0.898	0.917	0.934	0.951	0.960	0.963	0.966	0.969	0.966	0.970	0.976	0.980	0.981	0.981	0.979	0.983
1	0.805	0.807	0.820	0.842	0.879	0.891	0.896	0.915	0.932	0.950	0.959	0.963	0.964	0.968	0.966	0.970	0.976	0.980	0.981	0.981	0.979	0.983
1.2	0.802	0.803	0.816	0.839	0.876	0.888	0.894	0.913	0.930	0.948	0.958	0.962	0.963	0.967	0.965	0.969	0.976	0.979	0.980	0.980	0.978	0.983
2	0.785	0.787	0.801	0.827	0.863	0.875	0.887	0.906	0.924	0.941	0.953	0.959	0.960	0.964	0.963	0.967	0.973	0.977	0.977	0.977	0.975	0.981
4	0.752	0.754	0.770	0.797	0.839	0.852	0.865	0.884	0.912	0.929	0.942	0.946	0.950	0.957	0.956	0.963	0.970	0.973	0.974	0.977	0.973	NaN
8	0.712	0.715	0.731	0.761	0.806	0.821	0.833	0.854	0.886	0.908	0.921	0.931	0.937	0.942	0.946	0.950	0.956	0.962	0.967	0.969	0.973	NaN
20	0.668	0.671	0.687	0.717	0.762	0.781	0.796	0.822	0.857	0.880	0.891	0.901	0.909	0.914	0.920	0.927	0.944	0.957	NaN	NaN	NaN	NaN
40	0.622	0.624	0.635	0.666	0.709	0.731	0.751	0.777	0.805	0.833	0.851	0.858	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

FIGURE I.2.6.22 AUC en fonction du délai maximum autorisé entre les partenaires et de la longueur de la fenêtre d'analyse.

I.2.6.7 Délai maximum et durée de la fenêtre d'analyse

Dans la section I.2.4.1 , nous avons présenté deux paramètres utilisés dans le calcul de la corrélation, à savoir le délai maximum admissible et la longueur de la fenêtre d'analyse (voir figure I.2.4.8). Comme nous allons le voir, ces deux paramètres ont une influence sur les résultats, mais ont aussi une interprétation dans le cadre de l'interaction.

Concernant la fenêtre d'interaction, la question est la suivante : pendant combien de temps est-il nécessaire d'observer une interaction afin de décider si il y a, ou non, de l'imitation entre les partenaires ? Il parait évident que plus l'observation sera longue, plus fiable sera la mesure. Cependant le but est d'obtenir une mesure locale et instantanée de l'imitation, ce qui suppose une fenêtre temporelle la plus courte possible. Ainsi, un compromis doit être réalisé. Le second paramètre est le délai temporel maximal entre les partenaires. Là aussi, si un délai trop court est rédhibitoire car les personnes mettent un certain temps à réagir dans des interactions naturelles, un délai trop long n'est pas non plus envisageable : si deux personnes font le même geste à 20 secondes d'écart, peut-on considérer que c'est de l'imitation et qu'il s'agit d'une réponse à un stimulus ? Là aussi, un compromis doit être trouvé. Bien souvent, les psychologues [72] utilisent un délai ad hoc de 3 secondes, mais est-ce optimal ? Afin d'optimiser ces paramètres, mais aussi d'étudier leur importance, nous faisons varier le délai maximum autorisé entre 0 et 40 secondes et la durée de la fenêtre d'analyse entre 5 et 100 secondes. Les résultats (AUC) sont présentés figure I.2.6.22.

Le premier résultat, auquel on pouvait s'attendre, est que plus la fenêtre d'analyse est longue, meilleur le score est. Le second résultat est l'existence d'un délai optimal, indépendamment de la durée de la fenêtre d'étude. Dans notre cas ce délai de 15 images, correspond à 0,6 seconde de décalage temporel entre les deux partenaires. Ce temps est court relativement aux 3 secondes habituellement utilisées par les psychologues dans l'étude des interactions naturelles [72]. Cependant ce résultat est cohérent avec la construction de notre base : il était demandé explicitement aux participants de s'imiter, ce qui se fait dans des temps très courts. Conformément à nos attentes, l'augmentation du délai maximum autorisé diminue les résultats. Cet effet prévisible est dû à la flexibilité introduite qui permet, pour les données artificielles, de trouver de l'imitation là où il n'y en a pas, alors que dans le cas des vidéos d'imitation les résultats ne changent pas, le délai optimal situé autour de 15 images étant déjà trouvé.

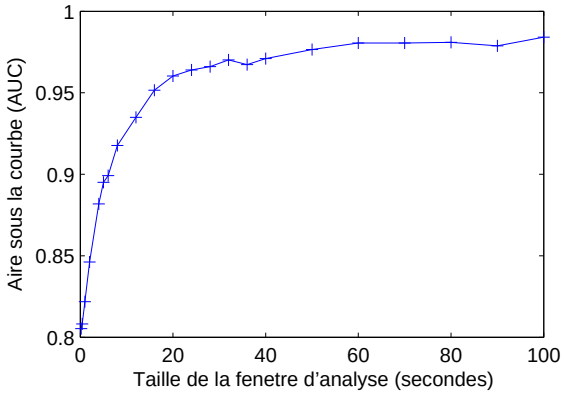


FIGURE I.2.6.23 AUC en fonction de la taille de la fenêtre d'analyse, pour un délai fixé à 15 images.

La figure I.2.6.23 montre l'évolution de l'AUC en fonction de la taille de la fenêtre d'analyse pour un délai maximum égal au délai optimal de 15 images.

Comme nous l'avions déjà constaté, l'AUC augmente avec la taille de la fenêtre. Cependant, au-delà de 750 images (30 secondes), cette augmentation est faible. Ainsi, il n'est pas nécessaire d'observer l'interaction sur une très longue durée pour détecter l'imitation de manière fiable. Le choix de la longueur de la fenêtre d'étude, qui est un compromis entre la qualité et la continuité de la mesure d'imitation, peut maintenant être fait en fonction de l'application, en connaissance de cause. Par exemple, une fenêtre de 10 secondes permet déjà d'avoir une bonne mesure de l'imitation (AUC $>$ 0.92) tandis qu'une une fenêtre de 20 secondes amène à de meilleurs résultats (AUC $>$ 0.96), proches du maximum (AUC $=$ 0.98), mais moins de réactivité.

I.2.6.8 Orientation dans la relation

Dans cette section, nous allons analyser les délais optimaux obtenus par la méthode. Les délais optimaux correspondent aux délais pour lesquels la corrélation était maximale, et qui sont donc compris entre moins le délai maximum autorisé et plus le délai maximum autorisé. En effet, cette analyse est importante afin de s'assurer que la méthode est capable de faire la différence entre le meneur et l'imitateur, même si les cas proposés ici sont simplifiés car les rôles du meneur et de l'imitateur sont fixes.

Cette analyse est faite seulement sur les vidéos d'imitation, avec un délai maximum autorisé de 100 images (4 secondes), et une fenêtre d'analyse de 500 images (20 secondes). À chaque instant, le délai optimal est calculé et est reproduit sur la figure I.2.6.24. Comme le meneur est toujours situé sur la caméra 1, le délai optimal devrait toujours être positif, si l'imitation était parfaite et constante.

La figure I.2.6.25 montre la probabilité d'avoir un certain délai pour les vidéos de cette base. Dans l'imitation, comme l'un des partenaires mène l'interaction, on peut s'attendre à avoir un maximum positif et à ce que le délai entre les deux partenaires soit assez court, ce qui est validé par l'étude de la figure I.2.6.25, qui montre un maximum autour de 15 images. La non-imitation quand à elle a une probabilité constante quel que soit le délai, ce qui était aussi attendu. On pourra noter que les deux pics pour les valeurs maximales du délai dans la non-imitation sont

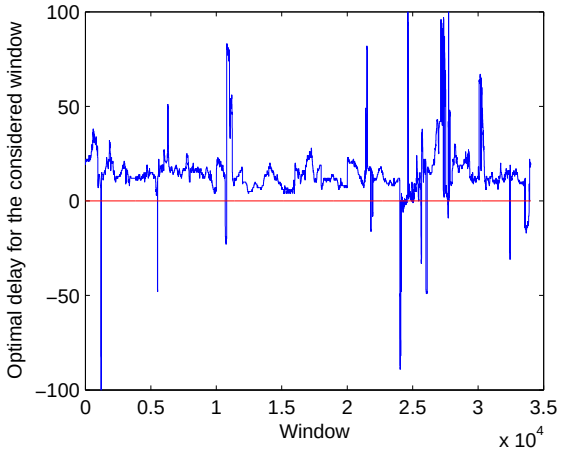


FIGURE I.2.6.24 Variations du délai optimal pour toutes les vidéos d'imitation.

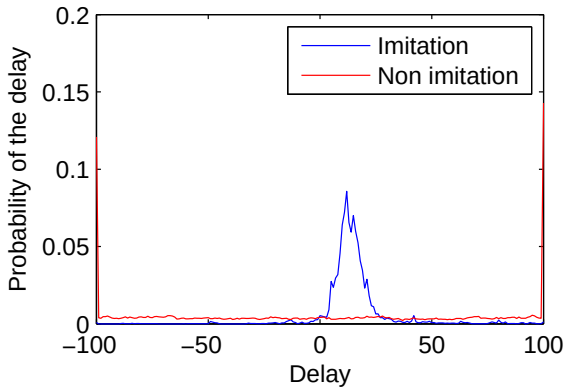


FIGURE I.2.6.25 Distribution de la probabilité de chaque délai pour l'imitation et la non imitation.

un effet de bord.

Le score moyen par délai a été calculé pour les vidéos d'imitation et de non-imitation. Ils sont affichés sur la figure I.2.6.26. Comme attendu, le score moyen pour chaque délai possible pour la non-imitation est plus faible que pour l'imitation, et presque constant pour tous les délais. Quand à l'imitation, deux phénomènes peuvent être notés. Le premier est que le score est maximal pour un délai autour de 15 images (0,6 seconde), ce qui confirme que la plupart de l'imitation de notre base se fait avec ce délai. Le second est que les scores moyens pour des délais positifs sont plus grands que ceux pour des délais négatifs. Ceci est aussi expliqué par la composition de la base, où le meneur était fixé, menant à des délais positifs. Cependant, un résultat intéressant est que les délais négatifs n'ont pas un score moyen égal à ceux de la non-imitation comme l'on aurait pu s'y attendre. Ceci est dû au fait que certains gestes de la base sont répétitifs (faire coucou par exemple), ce qui peut mener parfois à de bons scores d'imitation avec des délais négatifs.

Enfin, une nouvelle optimisation des paramètres (délai maximum admissible et longueur de la fenêtre d'analyse) peut être faite dans le but d'optimiser la reconnaissance de l'orientation, comme montré dans la figure I.2.6.27. Les résultats montrent que le délai maximum autorisé n'a

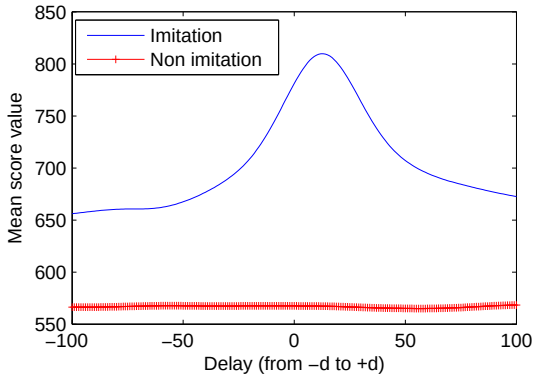


FIGURE I.2.6.26 Score moyen en fonction de chaque délai possible.

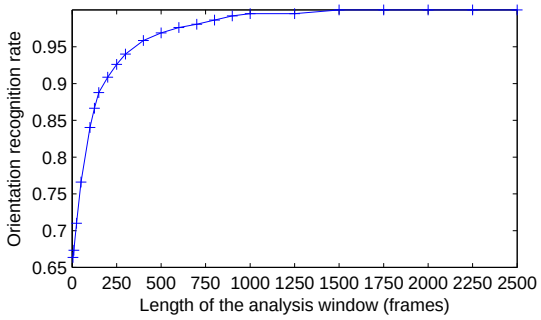


FIGURE I.2.6.27 Taux de reconnaissance de l'orientation en fonction de la longueur de la fenêtre d'analyse. Le délai maximum autorisé a été fixé à 20 images.

presque pas d'impact (moins de 2 % de variation) sur le taux de reconnaissance de l'orientation. Il a donc été fixé à 20 images (0,8 seconde) pour le tracé de la figure I.2.6.27.

Comme prévu, la figure I.2.6.27 montre que plus la fenêtre d'analyse est longue, plus le résultat est bon ($> 70\%$ pour 25 images (1 seconde), $> 90\%$ pour 250 images (10 secondes) et $> 96\%$ pour 500 images (20 secondes)). Cette progression est due à deux facteurs. Le premier est que plus la fenêtre d'analyse est longue, plus on est sûr du résultat, car l'on prend en compte plus de gestes. Le second facteur est un biais de notre mesure : nous supposons que pour chaque fenêtre, même les plus courtes, de l'imitation existe. Cependant, comme nous l'avons expliqué dans la section I.2.5.1, les gestes sont effectués de manière libre, avec la possibilité de faire des pauses entre deux gestes, ce qui peut mener à des fenêtres contenant peu de mouvements. Lorsque la taille des fenêtres augmente, ce phénomène est lissé grâce aux gestes environnants.

Comme nous l'avions montré dans la section I.2.6.7, un délai maximum admissible optimal existe ici aussi, et varie entre 10 et 15 images. Cependant, comme nous pouvons le voir dans la figure I.2.6.25, la plupart des délais se répartissent entre 0 et 30 images. Ce délai optimal est ici aussi un compromis entre autoriser un délai entre les partenaires et empêcher la recherche d'imitation là où elle n'existe pas.

Conclusion

Dans cette partie, plusieurs contributions majeures ont été apportées. Tout d'abord, face au manque de base de données d'imitation motrice dans la littérature, nous avons introduit une nouvelle base d'imitation dyadique gestuelle. Celle-ci est composée de 9 dyades réalisant une imitation gestuelle libre à tour de rôle (meneur/imitateur), ce qui conduit à 18 séquences d'imitation. Des pseudo-séquences de non imitation sont ensuite automatiquement générées en mixant les différentes dyades.

La seconde contribution est l'introduction d'une méthode de mesure d'imitation entre deux partenaires, issue des techniques de reconnaissance d'action non supervisée. Cette méthode, qui permet d'estimer le degré d'imitation à chaque instant de la séquence, utilise des modèles sac-de-mots dans lesquels l'information spatiale a été réintroduite. Une mesure de similarité entre les différents sac-de-mots conduit à l'estimation de l'imitation. Contrairement à la plupart des approches proposées dans la littérature, aussi bien la temporalité que la forme des gestes sont pris en compte. De plus, nous accédons à une mesure continue de l'imitation fort utile pour une modélisation fine des interactions.

Dans un troisième temps, nous nous sommes intéressés au délai temporel maximum admissible entre les partenaires : si la réponse à un stimulus ne peut pas être instantanée, nous avons montré que laisser trop de latitude dans le délai détériore les résultats en facilitant la détection d'imitation fortuite. Ainsi, il est possible, pour un scénario, ou un type d'interaction donné, de définir un délai maximum optimal.

La quatrième contribution répond à la question de temporalité : pendant combien de temps est-il nécessaire d'observer une interaction pour décider de la présence ou non d'imitation ? Nous avons ainsi confirmé notre première intuition : plus la fenêtre d'observation est longue, plus la mesure est fiable. Nous avons également montré que passé 20 secondes, les améliorations sont minimes et que dès 10 secondes, des mesures relativement fiables sont obtenues. Le choix final devra être réalisé en fonction de l'application et du compromis réactivité du système versus qualité de la mesure.

Pour finir, la dernière contribution concerne la détermination de l'orientation entre les partenaires. Dans le cas particulier de notre base, où les rôles sont fixés et l'orientation ne change pas de manière dynamique, notre méthode est capable de déterminer l'orientation dans la relation. Ici encore les résultats dépendent du délai maximum autorisé et de la taille de la fenêtre d'analyse. Nous avons montré que même une fenêtre de 1 seconde est suffisante afin de déterminer l'orientation de manière significative, mais les résultats deviennent vraiment très bons pour des fenêtres de plus de 10 secondes.

Les futurs challenges qui se présentent à nous sont d'étudier la méthode proposée sur une

base de données plus naturelle, dans laquelle l'imitation est plus diffuse, moins présente et où la dynamique de l'interaction intervient (orientation changeant dans le temps). Enfin, ces travaux sur l'imitation vont être étendus à des interactions naturelles et ainsi évoluer vers des mesures de synchronie.

Deuxième partie

Classification non supervisée des interactions

Introduction

Les interactions humaines sont riches, et prennent place dans des situations variées, allant du repas en famille à la réunion de travail, en passant par les rencontres de la vie de tous les jours. Elles se font à travers plusieurs canaux de communication en parallèle (verbal et non verbal, audio et visuel, ...) pour établir des relations, pour transmettre des idées, des pensées ou des émotions. Elles peuvent être définies comme des séquences dynamiques d'actions sociales entre des individus qui modifient et adaptent leur comportement en fonction de celui de leur partenaire.

Alors que l'analyse automatique des interactions est un domaine récent de la recherche en traitement du signal social, une littérature fournie existe sur le sujet dans le domaine de la psychologie sociale. Du point de vue de la recherche en traitement du signal social, le but est de comprendre le comportement social humain en observant les signaux échangés entre des personnes au moyen de capteurs multimodaux permettant de capturer les nombreuses dimensions des interactions sociales. Grâce aux évolutions faites dans le développement de capteurs à bas coûts, des salles multimodales ont pu être créées pour capturer les informations transmises sur les différents canaux de communication. Ces avancées permettent de revisiter à l'aide de techniques automatiques d'analyses de nombreux résultats obtenus en psychologie sociale.

Ainsi ces dernières années, l'intérêt dans l'analyse de schémas comportementaux a considérablement augmenté, en particulier l'interprétation des modalités de communication appelées signaux sociaux. La compréhension des principes fondamentaux qui gouvernent le statut d'une personne dans un groupe est d'une importance capitale pour les sciences sociales, et permettrait de mettre en place un grand nombre d'outils pour aider les recherches en psychologie sociale et organisationnelle.

Ces mêmes outils seront tout aussi utile pour le développement de la robotique sociale. La compréhension par la machine du contexte est également nécessaire pour que les robots interagissent avec les humains de manière socialement adaptée. Afin de comprendre le contexte, les systèmes informatiques doivent être capables de maintenir un modèle décrivant l'environnement, ses occupants et leurs activités. Les situations sont alors considérées comme des abstractions sémantiques déduites d'indices contextuels bas niveau.

La modélisation des individus permet d'étudier des éléments tels que la dominance ou la personnalité. La modélisation du groupe quant à elle, en utilisant les comportements non verbaux, permet de comprendre et modéliser les relations existant entre les membres et le groupe dans son ensemble, et ainsi mettre en avant l'intérêt, l'interactivité, les rôles ou la centralité dans les groupes.

Alors que des individus différents auront des styles de parole, d'expression gestuelle ou de

contact visuel différents, la dynamique d'un groupe évolue au dessus de ces styles individuels. Tandis que dans certains groupes les personnes parlent ou interrompent beaucoup, d'autres verront les personnes être plus silencieuses. Tandis que certains groupes amènent à une vraie discussion entre tous les intervenants, d'autres verront des différences de statut qui mènent à des différences dans le niveau de participation.

Ainsi, on peut distinguer des profils de groupe génériques permettant de décrire le groupe dans son ensemble, sans prendre en compte l'identité des participants ou le sens de la discussion.

L'analyse des groupes peut être traitée à travers deux approches différentes. La première tente de reconnaître les actions des individus de manière indépendante, et fusionne toutes les réponses à un niveau plus haut pour reconnaître l'interaction [200]. Cependant, dans une interaction, les comportements des uns sont liés aux comportements des autres, et ne sont pas complètement indépendants. La seconde approche [15] vise à reconnaître le type de groupe directement, en intégrant toutes les observations des individus dans un modèle unique. La modélisation est alors considérée plus fine et pertinente que la simple somme des individus.

Des applications potentielles d'une telle analyse incluent l'identification de comportements irresponsables, l'identification automatique et la quantification du leadership [161, 160], ou encore l'estimation de la cohésion d'une équipe [92]. Ainsi, d'un point de vue des ressources humaines, cela pourrait permettre de mettre en oeuvre à des moments clés des exercices de construction d'équipe ou des changements de leadership. Le suivi de groupe peut aussi permettre de voir si les personnes sont plutôt engagées dans des comportements coopératifs ou compétitifs.

Une autre application concerne la recherche de documents [201]. Les analyses de groupes permettent d'aider des personnes ayant manqué une réunion ou des personnes présentes voulant se rappeler certains détails à retrouver des moments clés dans des archives d'une réunion sans devoir écouter et revoir la totalité des enregistrements, les changements de dynamique d'un groupe et de sa configuration étant de bons indicateurs de changement d'activité du groupe.

Enfin, ces techniques peuvent servir de médiateur externe [104]. Le but principal d'un médiateur externe est d'obtenir une bonne vue d'ensemble d'une situation en analysant les dynamiques de l'interaction, et de fournir aux partenaires un retour durant des négociations, permettant ainsi d'améliorer ou faciliter l'interaction.

Deux grands problèmes font apparaître la difficulté à étudier le signal social : les caractéristiques à extraire et les modèles mathématiques à appliquer sur ces caractéristiques afin d'obtenir des motifs d'interaction ou de comportement. Ces deux problèmes seront traités dans la suite de cette thèse, juste après l'état de l'art concernant l'analyse des groupes dans le domaine de la psychologie sociale et dans le traitement du signal social. Une conclusion fermera le chapitre.

Chapitre II.1

État de l'art

Sommaire

Longtemps étudiée par les sociologues et les psychologues, l'étude des groupes et des interactions est devenue un point central du traitement du signal social [180]. En effet, le développement récent de capteurs performants et à bas prix a permis de développer des salles multimodales intelligentes [128, 186] ou d'utiliser des capteurs portables comme des badges [133].

Une fois les données acquises, des caractéristiques sont extraites. Bien que la communication verbale soit le premier mode de communication, le non verbal s'est révélé être porteur de nombreuses informations, difficiles à fausser, et permettant d'identifier de nombreuses caractéristiques des personnes [138, 107].

II.1.1 L'étude des groupes en psychologie sociale

Dans toute la suite, la notion de groupe est définie comme un ensemble de trois personnes minimum.

L'étude des petits groupes a une longue histoire dans la recherche en psychologie sociale [10, 124]. Les travaux de Bales en 1951 [10] puis en 1970 [11] sont centraux dans l'étude des groupes. Il développe une approche systématique pour l'observation et la description de ceux ci, en mettant en avant les processus cognitifs en place chez les individus. McGrath dans ses travaux de 1984 [124] met en avant les processus temporels qui interviennent dans les interactions de groupe.

Les travaux en psychologie sociale s'effectuent à plusieurs échelles. La première, celle de l'individu, étudie la construction individuelle. Goldberg propose dans [81] un modèle de personnalité se basant sur cinq catégories : Ouverture, Consciensosité, Extraversion, Agréabilité et Neuroticisme. Elle est aujourd'hui encore largement analysée et utilisée [100].

La seconde échelle s'intéresse à la construction sociale de l'individu, comme par exemple la dominance, le statut ou l'influence. L'article de Hall et al. [85] propose un état de l'art de la relation de cette construction aux comportements non verbaux.

La troisième échelle s'intéresse au mode de communication, et notamment à la communication non verbale. Ainsi, dans [62] Dovidio et al. étudient la relation forte qui existe entre le mode ou la nature des communications et la place ou l'interaction dans un groupe. Dans leur livre [119], Manusov et Patterson recensent les études majeures du domaine qui portent sur la communication non verbale, son histoire, l'évolution des différentes théories ou des méthodes, ainsi que ses facteurs d'influence et ses fonctions. Le livre de Knapp et al. [107], édité pour la première fois en 1978 et réédité pour la 8ème fois en 2013 complète le livre de Manusov et Patterson en offrant une décomposition différente. Knapp et al. décrivent d'abord l'environnement de la communication, puis les acteurs de la communication. Enfin, ils décrivent les différents comportements non verbaux : gestes et postures, toucher, visage, contact visuel et indices vocaux.

II.1.2 L'étude des groupes en traitement du signal social

Suivant les travaux effectués en psychologie sociale, les modélisations computationnelles s'intéressent aussi aux différentes échelles de l'étude de l'interaction.

Ainsi, on retrouve à la première échelle les études concernant les individus et leur construction individuelle. Divers éléments liés à la personnalité (extraversion, traits de personnalité) ont été prédits dans [141] par Pianesi et al. à l'aide de régression à vecteurs support sur des séquences de 1 minute de la base Mission Survival [140].

Jayagopi et al. [98] extraient de nombreuses caractéristiques (audio et vidéos) individuelles et de groupe, et les corrèlent à de nombreuses variables liées à la construction individuelle des personnes (extraversion, dominance, agréabilité, ...). Ces très nombreux tests permettent de voir l'ensemble des relations existant entre les caractéristiques et les variables individuelles.

La seconde échelle concernant la construction sociale de individu est largement étudiée. Dans [151], Rienks et al. prédisent les comportements dominants en calculant des caractéristiques liées aux tours de parole (temps total parlé, tours, interruptions). Le corpus de Aran et al. [134] a été par ailleurs spécialement créé pour l'étude de la dominance, du statut et du genre.

Dans [96], Jayagopi et al. souhaitent détecter la personne la plus dominante d'un meeting. Pour cela, des caractéristiques vocales et visuelles intra-personnelles et inter-personnelles sont extraites sur les données de la base AMI [123]. Chaque caractéristique est ensuite utilisée pour classifier la dominance des personnes, en utilisant la fonction "maximum" ou "minimum" suivant si la caractéristique est positivement corrélée à la mesure de dominance effectuée par des annotateurs sur des données d'exemple. Sans grande surprise, la personne la plus dominante est la mieux reconnue lors de l'utilisation de la caractéristique "temps total parlé" et le score de reconnaissance est alors de 70.8%.

Dans [161, 160] Sanchez-Cortes et al. souhaitent trouver le leader émergent. Pour cela, ils extraient des caractéristiques, et une simple règle de classement leur permet d'évaluer chacune des caractéristiques en corrélation avec une annotation manuelle de leadership. Le meilleur

résultat (63.5%) est obtenu pour la caractéristique "nombre total d'interruptions réussies pour les tours de parole de plus de 2 secondes". Une fusion des caractéristiques est ensuite effectuée au niveau de la décision (somme des classements pour plusieurs caractéristiques), ce qui permet d'améliorer la reconnaissance (72.5%) qui est obtenu pour un mélange de 5 caractéristiques comprenant la durée moyenne d'un tour de parole, le nombre total d'interruptions réussies, le nombre total d'interruptions réussies pour les tours de parole de plus de 2 secondes ainsi que la valeur médiane et la variance de l'énergie.

La dernière échelle, la plus haute et concernant le groupe dans son ensemble, a été centrale dans le domaine du traitement du signal social. C'est à ce niveau d'information que nous nous intéresserons dans le chapitre II.2, et nous présentons donc ci-dessous plus en détail les publications liées à cette échelle. Deux grandes familles d'études s'opposent, suivant la nature supervisée ou non des méthodes de reconnaissance.

II.1.2.1 Études supervisées de l'interaction dans les groupes

Dans [48], Cristani et al. essaient de catégoriser les interactions. Pour cela, ils disposent de trois configurations :

des échanges entre deux adultes ;
des échanges entre un adulte et un enfant ;
des disputes entre deux adultes.

La base est constituée de 38 conversations d'environ 9 minutes chacune. Seuls les enregistrements sonores sont disponibles. A partir de ces enregistrements, les Périodes de Conversation Stable (PCS [139]) sont extraites. Enfin, plusieurs modèles sont comparés : une gaussienne multidimensionnelle, le Modèle d'Influence de Tours de Paroles [42] et le modèle des auteurs basé sur une modélisation des PCS grâce à un modèle de mixture de gaussienne (GMM [65]) suivi par un Modèle d'Influence Observé [15]. Les auteurs effectuent tout d'abord une analyse qualitative des chaînes intra et inter partenaire lors d'une interaction adulte-enfant. Ils font ensuite des tâches de classification, où plusieurs combinaisons sont testées : échange (adulte-adulte) VS dispute (adulte-adulte), échanges (adulte-adulte et adulte-enfant) VS disputes (adulte-adulte), échanges (adulte-adulte) VS échanges (adulte-enfants), et enfin toutes les catégories les unes contre les autres. Les auteurs obtiennent un score de classification de 73% sur cette dernière tâche (la plus difficile des quatre). Ils décident ensuite d'appliquer un algorithme de clustering afin de trouver de manière non supervisée des catégories d'interactions. Pour cela, ils utilisent un dendrogramme, et choisissent un nombre de clusters égal à 3, correspondant au nombre de types d'interactions. La méthode permet d'atteindre une précision de 75.63%, les interactions étant bien rassemblées par type.

Dans [51], Dai et al. souhaitent analyser les interactions de groupes dans divers contextes. Pour cela, ils s'appuient sur une base de leur création composée de 6 meetings, quatre meetings comprenant 4 participants, un meeting avec 3 participants et le dernier meeting avec 5 participants. La base est constituée de données vidéo, audio et de métadonnées. Plusieurs niveaux d'informations sont utilisées. Tout d'abord les informations sur le type de scénario (présentation, discussion, pause) permettent d'avoir une échelle globale. Ensuite, les informations sur l'interaction ont un horizon temporel plus court et sont des sortes de "sous scénarios" ayant

lieu dans l'interaction (présentation en cours, questions et réponses, silence, pause, une personne étant en train de parler). Le plus bas niveau est défini par les informations d'entité, qui sont caractérisées par les mouvements individuels des personnes, leur activité vocale et leurs actions. Les rôles jouent un niveau intermédiaire entre les deux dernières échelles. Des caractéristiques de l'environnement sont extraites, comme par exemple le nombre de personnes, l'utilisation ou non du vidéo projecteur ou la direction vers laquelle le locuteur parle. D'autres caractéristiques, plus bas niveaux, sont elles aussi extraites, comme la position du corps, de la main droite, de la tête, ainsi que la pose de cette dernière. Les auteurs proposent un nouveau modèle, nommé Réseau Bayesien Dynamique - Motivé par des Évènements Multi-échelles basé sur les DBN [130]. Le premier test réalisé par les auteurs consiste à segmenter une interaction en ses trois types de scénario (présentation, discussion, pause), puis en ses sous scénarios, ainsi qu'en deux types de rôles "celui qui pose les questions", "celui qui n'est pas à son siège car il est en train de présenter" et "celui qui parle". Les résultats, calculés à partir de l'application successive de la méthode sur les 6 meetings, montrent une précision de 86% pour le type de scénario, 84.8% pour les sous-scénarios, et 87.4% 86.7% et 84.7 % pour les rôles définis précédemment.

Dans [31], Brdiczka et al. mettent en place un système de reconnaissance des configurations qui prennent place dans un groupe. Quatre configurations sont prédéfinies : (ABCD), (AB)(CD), (AC)(BD) et (AD)(BC), où les parenthèses dénotent les participants qui parlent ensemble. À l'aide de deux meetings enregistrés dans une salle multimodale, d'une durée de deux à trois minutes, ils extraient les tours de parole des quatre participants. Ces tours de paroles sont envoyés à un HMM [144] afin d'apprendre à reconnaître les configurations. Leur système atteint une précision de 84,8%, mais a été évalué seulement sur trois meetings d'une durée moyenne de 15 minutes.

McCowan et al. ont effectué plusieurs travaux liés à l'analyse de l'interaction. Parmi eux, le papier [122] s'intéresse à la reconnaissance d'actions de groupes telles que monologue, présentation, tableau, discussion ou prise de notes (il y a 8 actions pour un groupe de 4 personnes). Pour cela, une salle multimodale leur permet d'accéder à de nombreuses informations dont ils tirent des caractéristiques audio (tours de paroles), acoustiques (fréquence fondamentale, énergie et taux) ainsi que visuelles (position, orientation et mouvements de la tête et des mains). Ils testent ensuite plusieurs systèmes pour l'apprentissage, basés sur des variantes des HMM [144]. Les meilleurs résultats sont obtenus à l'aide du système basé sur les Modèles de Markov Cachés Asynchrones [16] et ils atteignent un pourcentage d'erreur de seulement 9.2%.

Cupillard et al. [50] souhaitent reconnaître des comportements à risque dans le métro. Pour cela, ils mettent tout d'abord en place un système permettant de fusionner les données de multiples caméras afin d'obtenir des données de haut niveau plus précises. Ils tentent alors de reconnaître deux scénarios prédéfinis : "un groupe est en train de se battre", ou "un groupe bloque le passage d'autres individus". Les interactions sont décrites par des caractéristiques de plusieurs niveaux. Tout d'abord, les états décrivent la vitesse des groupes, leur agitation, leur présence dans des zones de blocage (zone d'entrée des escalators par exemple). Pour chacun de ces états, deux événements sont définis : "commencent à être dans l'état", ou "finissent d'être dans l'état". Les scénarios sont alors reconnus grâce à des automates à états finis. Des scores de reconnaissance de 70% sont obtenus sur le scénario de bagarre et de 95% sur celui de blocage.

Burger et al. proposent dans [33] une base de données de meetings avec de multiples scénarios : planification de projet, planification de travail, exercice militaire, jeu, discussion informelle et discussion d'un sujet. Pour cela, une salle multimodale leur permet d'enregistrer la vidéo, l'audio pour chaque participant, ainsi que des métadonnées. Les données audio sont retranscrites manuellement par les auteurs. Un total de 104 meetings est disponible, pour environ 103 heures de réunion. Dans cet article, les auteurs mettent en avant l'effet du type de réunion sur les caractéristiques linguistes et acoustiques des participants. Par exemple, lors d'une discussion informelle, les participants ont tendance à parler plus vite et aucun leader n'apparaît. Dans les discussions sur un sujet, le rythme de parole est rapide, avec beaucoup d'échanges. Cependant, contrairement à la discussion informelle, un grand nombre d'assertion des participants est présent, pour montrer leur accord avec des opinions formulées par d'autres.

Enfin, dans [97] Jayagopi et al. apprennent à différencier les meetings coopératifs des meetings compétitifs. Pour cela, deux bases de données sont utilisées, la base AMI [123] pour les meetings coopératifs et la base "The Apprentice" [145] pour les meetings compétitifs. Des caractéristiques sont extraites à la fois au niveau individuel (activité vocale, énergie, interruptions) et au niveau du groupe (temps total parlé par le groupe par exemple). L'apprentissage et la prédiction sont ensuite effectués à l'aide de deux modèles, un basé sur le ratio de vraisemblance-logarithmique et le second à l'aide d'un SVM (avec un noyau linéaire et un quadratique). Un score de reconnaissance de 100% est obtenu lors de l'utilisation des caractéristiques "ratio du nombre d'interruption par le nombre de tours de paroles pour le groupe" et "mesure du nombre de tours de paroles à une distribution uniforme" avec le SVM à noyau quadratique.

II.1.2.2 Études non supervisées de l'interaction dans les groupes

Contrairement aux méthodes précédentes, les méthodes non supervisées n'ont pas besoin de données étiquetées.

Ainsi, dans [15, 42], Basu et al. cherchent à mesurer l'influence qu'ont les participants les uns sur les autres. Pour cela, ils se placent dans une salle multimodale expérimentale, qui leur permet d'avoir des données audio et vidéo. Ils extraient des enregistrements des caractéristiques aussi bien audio (taux de parole, fréquence fondamentale, énergie) que vidéo (quantité d'énergie de mouvement, suivi de blob). Ils revisitent alors le Modèle d'Influence [6] afin de construire un Modèle d'Influence simplifié. Une fois ce modèle appris sur les données, ils obtiennent alors des matrices d'influence entre les différents intervenants des meetings, qu'ils analysent qualitativement.

Raducanu et al. [145] essayent de déterminer le présentateur ainsi que la personne qui sera éliminée lors d'un jeu télévisuel "The Apprentice". Pour cela, des caractéristiques intra-personnelles (nombre de tours de parole, temps total de parole et nombre d'interruptions) et inter-personnelles (matrice d'interruption et de tours de paroles) sont extraites des enregistrements audio. Les caractéristiques intra-personnelles sont testées directement comme prédiction du statut (resp. de la personne éliminée) et le meilleur score de 85.7% est obtenu pour le nombre d'interruption (resp. 92.8% pour le nombre de tours de paroles). Les caractéristiques inter-personnelles sont elles associées à trois mesures de centralité [86] (centralité de degré entrants, centralité de degrés sortant, centralité de proximité). La meilleure reconnaissance,

atteignant 85.7% (resp. 78.5%) est obtenue pour la centralité de degré sortant (resp. entrant) pour la reconnaissance de la personne avec le plus haut statut (resp. la personne éliminée).

Dans [162], Sanchez-cortes et al. possèdent deux types de meetings : soit les 4 participants sont dans la même salle, soit une personne est à distance. Une des tâches proposées par l'article est d'essayer de reconnaître les deux types de meetings à partir de caractéristiques audio individuelles (temps de parole, nombre d'interruptions, nombre de tours de paroles...) ou de groupes (pourcentage du temps de silence, de paroles superposées, ...). Ils supposent que les meetings où les 4 participants sont dans la même salle auront des caractéristiques soient minimales ou maximales. La caractéristique "nombre de tours de paroles du groupe" est la plus pertinente, puisqu'elle permet d'avoir une précision de 81.1%.

Jayagopi et Gatica-Perez [95] extraient de la base AMI [123] des caractéristiques à partir des enregistrement audio, à la fois individuelles (temps de parole, nombre de tours de paroles, nombre d'interruptions ,...) ainsi que de groupe (distributions du temps de parole, du nombre de tours, des interruptions, quantité totale de temps parlé, nombre total de tours de paroles, ...). Ils quantifient ensuite ces caractéristiques. D'une part, les caractéristiques qui concernent le groupe dans sa totalité, sans différentiation des membres, sont quantifiés en 5 catégories : silence, un, deux, autre ou égal, en fonction de si une ou plusieurs personnes possèdent la plus grosse part des distributions de parole. D'autre part, les caractéristiques concernant les membres de manière individuelle sont quantifiées en fonction de si le maximum de parole est attribué au leader, à un autre membre, ou à personne (si personne ne parlait). Par exemple TempsParole-Leader (qui s'oppose à TempsParole-NonLeader et TempsParole-Silence) signifie que dans la tranche considérée, c'était le leader qui avait le maximum de temps de parole. Au final, un vocabulaire de 50 mots est obtenu, et chaque tranche temporelle contient exactement 12 éléments actifs. Cette représentation a l'avantage d'être robuste au nombre de participants et permet ainsi de comparer des groupes de différentes tailles. Un modèle Sac-de-mot permet ainsi de décrire chacun des meetings. A partir de ces sac-de-mots, une modélisation par Allocation de Dirichlet Latente (LDA [25]) est effectuée. Cette technique non supervisée permet de rassembler des tranches temporelles en un nombre fixé de catégories (ici 3). Les auteurs montrent que chacune de ces catégories est caractérisée par un ensemble de mots spécifiques, permettant d'analyser ces catégories et reconnaître trois classes de leadership définies par Lewin et al. [117] : autocratique, participative et "roue libre". Afin de confirmer la validité de cette catégorisation, les 8 meetings les plus typiques de chaque classe ont été annotés ; la classe autocratique est ainsi reconnue à 62.5%, la participative à 100% et la "roue libre" à 75%.

II.1.3 Discussions et orientations

Cet état de l'art a permis de mettre en évidence la présence majoritaire de méthodes supervisées. Un grand nombre de méthodes cherchent à relier des caractéristiques bas niveau à une ou des variables représentant la construction individuelle ou de groupe. Des données annotées permettent d'entraîner les systèmes de reconnaissance.

Quelques approches se démarquent en utilisant des méthodes non supervisées, utilisant un classement ou la corrélation entre lesdites caractéristiques et variables. Cependant, ces systèmes permettent seulement de mettre en évidence les liens existants entre elles.

De plus, peu de travaux tentent de structurer l'interaction sans avoir *a priori* une classification déjà pré-établie lors de la création des bases de données (meeting conflictuel VS coopératif, discussion adulte-adulte VS adulte-enfant,...). Nous proposons ainsi dans cette thèse de catégoriser les interactions, sans *a priori* sur les catégories qui seront obtenues.

D'autre part, les caractéristiques utilisées, qu'elles soient audio ou vidéo, sont très souvent les mêmes d'une étude à une autre, et sont toutes obtenues à partir d'une expertise humaine. Nous proposons donc dans cette thèse de répondre à la question suivante : est-il possible de trouver de manière non supervisée de nouvelles caractéristiques permettant une description riche de l'interaction ?

Chapitre II.2

Codage et classification des interactions

Sommaire

II.2.1 Introduction

L'approche que nous proposons dans ce chapitre part du constat suivant : dans une conversation, les séquences de tour de paroles constituent des indices observables de l'existence sous-jacente d'une organisation dans le groupe. Ainsi, notre approche s'appuiera sur trois grandes étapes : la première est la reconnaissance et séparation des locuteurs, permettant ainsi de séparer les enregistrements en tours de paroles. La seconde étape permet l'extraction d'indices intra et inter personnels, ainsi que des indices de groupe (voir la section II.2.3), qui permettront de former des vecteurs caractéristiques. La troisième et dernière étape consiste en la classification non supervisée des interactions, et elle est décrite dans la section II.2.4 .

II.2.2 Présentation de la base AMI

La base AMI [123] est une collection de réunions enregistrées dans des salles de réunion contenant un grand nombre de capteurs, permettant d'obtenir les signaux audio et vidéo pour chaque participant. Une vue d'ensemble présentée sur la figure II.2.2.2 montre l'emplacement

des caméras permettant de capturer à la fois l'ensemble de la pièce (voir figures II.2.2.2b et II.2.2.2c), ainsi qu'un gros plan de chaque participant (voir figure II.2.2.2a). De même, des micros capturent l'ensemble de la pièce, alors que d'autre de proximité permettent de capturer le son d'un seul participant à la fois (voir le schéma II.2.2.2).

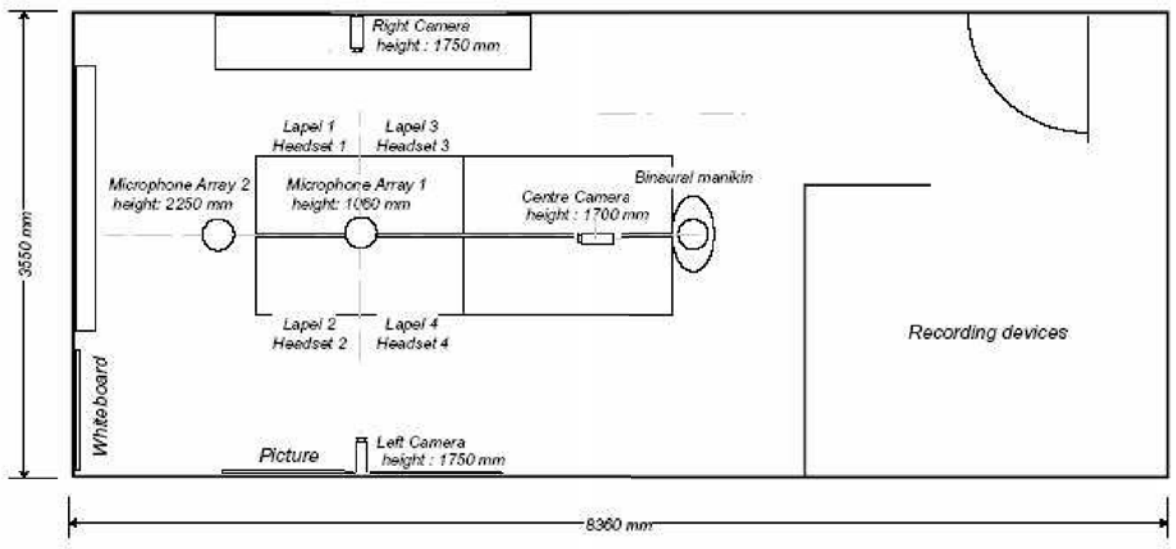


FIGURE II.2.2.1 Vue d'ensemble du dispositif (extrait de [123]).

Le corpus contient des réunions scénarisées ou des réunions sans scenarii. Seules les premières seront utilisées dans cette thèse. Dans celles-ci, quatre participants agissent comme des membres d'une équipe chargée de la création d'une nouvelle télécommande. Chaque participant joue durant une réunion entière un (et un seul) rôle parmi les quatre suivant :

Manager de Projet (PM)
Expert Marketing (ME)
Concepteur de l'Interface Utilisateur (UI)
Designer Industriel (ID)

Cependant, d'une réunion à une autre les participants peuvent être amenés à changer de rôle, mais tous les rôles sont présents dans toutes les réunions.

La base est constituée de 138 enregistrements scénarisés, pour un total de 45 heures et 38 minutes.

La base est fournie avec un grand nombre d'annotations, à plusieurs niveaux (rôles, transcription manuscrite des paroles, mouvements des participants,...).

II.2.3 Extraction des signaux

Afin de pouvoir étudier les interactions, il est important de représenter les différentes informations sous forme de signaux utilisables dans les algorithmes. Comme nous l'avons vu précédemment dans la section II.2.2 , la base AMI permet d'avoir accès à un grand nombre de signaux, comme la vidéo et le son pour chaque participant, des vidéos générales, des annota-

(a) Gros plan de chaque partici pant.

(b) Vue depuis un coin de la pièce

(c) Vue du dessus.

FIGURE II.2.2.2 Images extraites de la base AMI (extrait de [123]).

tions de gestes, etc. À partir de la vidéo, il est possible d'extraire un grand nombre de signaux ou d'événements comme par exemple les hochements de tête ou la quantité de mouvement du corps d'un participant. Ces informations ne sont cependant pas fiables car régulièrement, des personnes se lèvent pour aller au tableau, ou même, changent de place. Aussi, nous n'utilisons que le son dans cette thèse. D'autre part, afin de pouvoir généraliser la méthode à tous types d'interactions, nous avons également choisi de ne pas utiliser le contenu lexical disponible dans la base AMI.

A partir des signaux audio, nous extrayons les tours de parole de chaque participant, sous forme d'une suite binaire dont chaque bit donne l'information "parle" ou "ne parle pas".

Le processus complet est illustré sur la figure II.2.3.3. Tout d'abord, le signal vocal est extrait (signal mono-voix, figure II.2.3.3a). Ensuite la puissance du signal est calculée comme le carré de l'intensité audio (figure II.2.3.3b). Afin de pouvoir séparer les périodes où le participant parle de celles où il ne parle pas, il est important de supprimer les espaces entre les mots (qui durent de l'ordre de quelques centaines de millisecondes). Pour cela le signal de puissance est moyenné à l'aide d'une fenêtre de largeur 0,5 seconde (figure II.2.3.3c).

Pour séparer les périodes de parole de celles de silence, un simple seuillage pourrait être utilisé. Cependant, comme la puissance des quatre micros est très variable, ainsi que la hauteur de voix de chaque personne, il faudrait déterminer manuellement ce seuil pour chaque personne et chaque scénario. Nous avons donc préféré utiliser un modèle de mixtures de gaussiennes à 2 classes. Les paramètres des gaussiennes sont appris pour chaque signal, pour lequel l'apparte-

nance est ensuite calculée pour savoir si l'échantillon est plus probablement dû au bruit de fond ou à la parole du participant. On obtient alors l'information binaire "parle" ou "ne parle pas" désirée (figure II.2.3.3d).

Une fois les tours de paroles extraits pour tous les participants, on peut calculer des indices de plus haut niveaux. Ces indices peuvent être estimés sur des horizons temporels divers : soit sur la globalité de l'interaction (pourcentage de temps parlé pour chaque participant), soit sur des fenêtres temporelles finies. De plus, comme nous allons le voir dans la suite, ces indices peuvent être liés à une personne (temps de parole total), à un couple de personnes (interruptions d'une personne par une autre), voire être liés au groupe dans sa globalité (fraction du temps avec silence total).

II.2.3.1 Indices intra-personnels

Un certain nombre d'indices peuvent être extraits, pour chaque personne, et sont indépendants des autres participants. Nous n'aborderons ici en détail que les indices liés aux tours de paroles, mais il existe des indices pour la prosodie (fréquence fondamentale, rythme, énergie,...) ou concernant la gestuelle (mouvements de la tête, du corps, quantité de mouvement,...). Bien entendu, ces indices peuvent être calculés sur des fenêtres temporelles finies ou sur la globalité de l'interaction. Pour les tours de paroles, des indices habituels de la littérature ont été extrait [95, 160] :

TTP-A, Temps Total Parlé : on compte le nombre de secondes où le participant A parle ;

NbTP-A, nombre de tours de paroles : on compte le nombre d'interventions du participant, chaque intervention étant définie comme une période durant laquelle le participant ne s'interrompt pas ;

DuréeTP-A, temps moyen d'un tour de parole de la personne A : moyenne de tous les tours de parole du participant ;

NbInter-A, nombre de fois où la personne A réussi à interrompre un autre participant ;

NbNonInter-A, nombre de fois où la personne A échoue à interrompre un autre participant ;

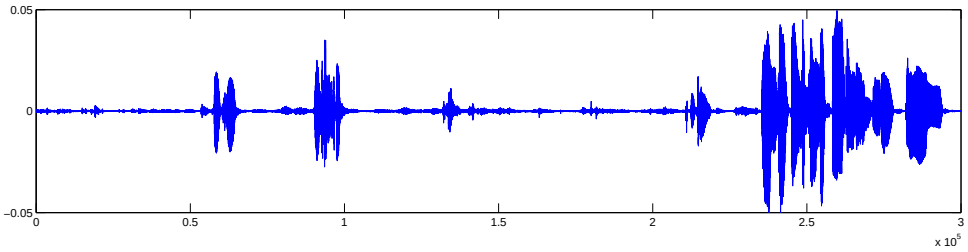
NbRep-A, nombre de fois où la personne A répond à une autre personne.

Afin de garantir la normalisation des indices quelle que soit la taille de la fenêtre, les indices TTP, NbTP, NbInter, NbNonInter et NbRep-A (tous sauf DuréeTP) sont normalisés par la longueur de la fenêtre d'observation considérée. Étant donné qu'il y a 4 personnes dans les vidéos de la base AMI, nous avons donc un total de 24 indices personnels (6 indices x 4 personnes).

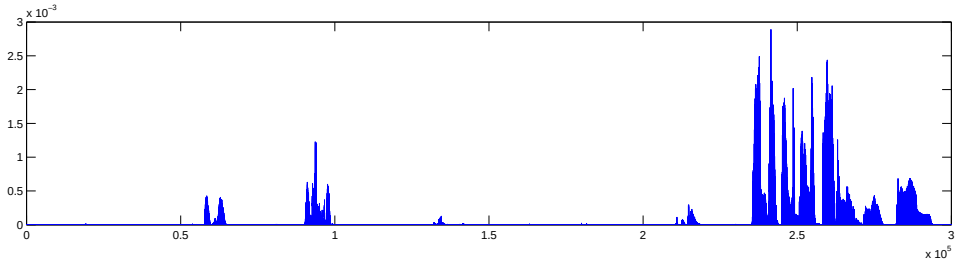
II.2.3.2 Indices inter-personnels

Comme les tours de paroles ont lieu dans le cadre de meetings mettant en jeu 4 participants, il est possible d'extraire un certain nombre d'indices qui donnent des informations sur les relations entre les différentes personnes. Les indices suivants ont été extraits, en se basant sur les indices de la littérature [95, 146] :

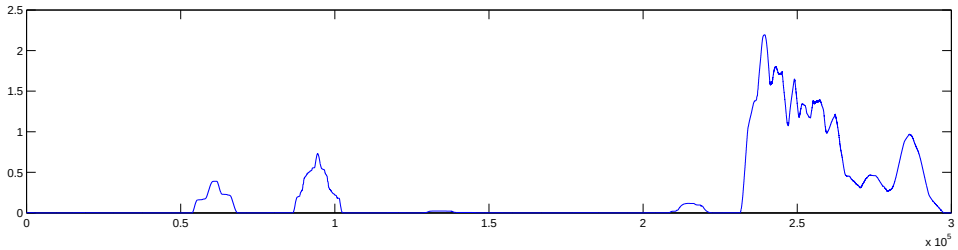
NbInter-A-B, nombre d'interruptions réussies de la personne A par la personne B : une interruption réussie de la personne A par la personne B est définie comme le fait que la personne A était en train de parler, quand la personne B lui a coupé la parole. La



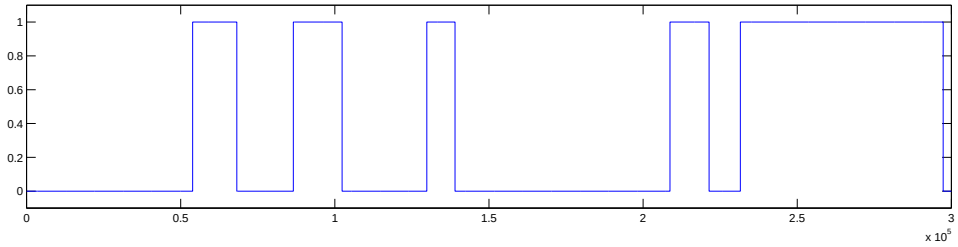
(a) Signal audio brut.



(b) Puissance du signal audio.

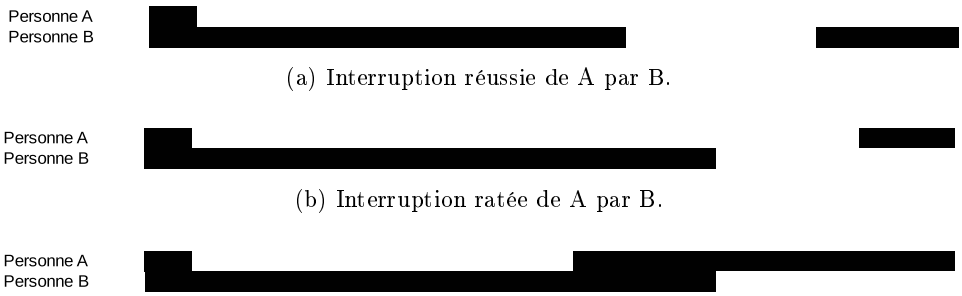


(c) Puissance moyennée sur une fenêtre de 0,5 seconde.



(d) Information binaire : parle ou ne parle pas.

FIGURE II.2.3.3 Étapes de l'extraction des tours de paroles.



(a) Interruption réussie de A par B.

(b) Interruption ratée de A par B.

(c) Réponse de B à A.

FIGURE II.2.3.4 Illustration des indices interactifs (une case blanche signifie que la personne parle).

personne A s'arrête donc de parler pendant que la personne B continue (voir figure II.2.3.4a) ;

NbNonInter-A-B, nombre d'interruptions ratées de la personne A par la personne B : une interruption ratée de la personne A par la personne B est définie comme le fait que la personne A était en train de parler, quand la personne B lui a coupé la parole. La personne A continue de parler et la personne B doit donc arrêter (voir figure II.2.3.4b) ;

NbRep-A-B, nombre de réponse de la personne A à la personne B : la personne B était en train de parler, elle s'arrête, et après une durée inférieure à 1 seconde la personne A se met à parler (voir figure II.2.3.4c).

Afin de garantir la normalisation des indices quelle que soit la taille de la fenêtre, tous ces indices sont normalisés par la longueur de la fenêtre d'observation considérée. Étant donné qu'il y a 4 personnes dans les meetings de la base AMI, les indices NbInter et NbNonInter ici possèdent 12 composantes (16 possibilités en tout, moins les 4 composantes liées à la personne avec elle même). L'indice NbRep possède 16 composantes (car on conserve le fait qu'une personne puisse reprendre la parole alors qu'elle était la dernière personne à avoir parlé). Ce qui mène à un total de 40 indices.

II.2.3.3 Indices de groupe

Il existe une dernière catégorie d'indices, liés cette fois à la dynamique globale du groupe. Ils donnent des informations importantes sur le type d'interaction qui a lieu [95] :

DuréePP, Temps pendant lequel plusieurs personnes (au moins 2) parlent en même temps ;

Durée1P, Temps pendant lequel une seule personne parle ;

Durée0P, Temps total de silence (où personne ne parle) ;

TTP, Temps total parlé : on somme les temps de parole des 4 participants ;

NbTP, nombre de tours de parole ;

NbInter, nombre d'interruptions réussies ;

NbNonInter, nombre d'interruptions échouées ;

RatioInter, ratio du nombre d'interruption réussie par tour de paroles ;

RatioNonInter, ratio du nombre d'interruption échouées par tour de paroles.

Il y a 9 indices de groupes. Afin de garantir la normalisation des indices quelle que soit la taille de la fenêtre, tous les indices sauf les deux derniers sont normalisés par la longueur de la fenêtre d'observation considérée.

Les trois types d'indices permettent donc d'obtenir un vecteur de taille 73 (24 personnels + 40 interactifs + 9 de groupes) représentant l'interaction. Une fois tous ces indices obtenus, nous souhaitons faire émerger des spécificités liées à chaque interaction. Notre but sera donc d'essayer, à l'aide d'une méthode non supervisée, de découvrir différents types d'interaction, et de voir ce à quoi ils correspondent.

II.2.4 Classification non supervisée des interactions

Une fois les indices extraits, chaque interaction peut être représentée par un vecteur (si on considère une seule fenêtre pour toute l'interaction), ou une matrice (si l'on considère plusieurs fenêtres temporelles pour toute l'interaction). Étant donné que l'on souhaite catégoriser les interactions en plusieurs classes, et que ces classes sont inconnues, il est nécessaire d'appliquer une méthode non supervisée. Pour cela, plusieurs méthodes existent, en allant des techniques de clustering (k-moyennes [76], regroupement hiérarchique [101],...) à des modèles à variables latentes (Analyse en Composantes Principales [102], Analyse en Composantes Indépendantes [44], Décomposition en Valeurs Singulières [83], Analyse Sémantique Latente Probabiliste [89] (PLSA), Factorisation en Matrices Non-négatives [114] (NMF pour Non-negative Matrix Factorisation)).

Les travaux de Lee et Seung [114] ont comparé les NMF, l'ACP et la Quantification Vectorielle (VQ) (cf figure II.2.4.5).

Les trois méthodes cherchent à mettre les entrées sous une forme factorisée (V=WH) avec différentes contraintes sur W et H. Une image d'entrée (un visage dans ce cas nommé "original" sur la figure 3.3) est approximée par une combinaison linéaire des images du dictionnaire.

On peut tout d'abord constater que la quantification vectorielle amène à une factorisation très creuse qui vise à choisir un mot du dictionnaire. La reconstruction est donc parfois très mauvaise.

L'ACP (ou PCA en anglais) amène à un dictionnaire (image de gauche figure II.2.4.5) qui n'est pas facilement interprétable. L'entrée, qui est bien reconstruite, est une somme pondérée des images du dictionnaire utilisant des coefficients qui peuvent être positifs ou négatifs. Ainsi, un visage correspondra à la somme de quantités positives de certains visages et de quantité négatives d'autres visages. Mais comment interpréter une telle somme ?

La factorisation en matrices non négatives présente plusieurs avantages. Le premier est que chaque image du dictionnaire correspond à une partie du visage. Le dictionnaire est donc facilement interprétable. Le second avantage est que les éléments de H sont tous positifs. Ainsi, l'entrée est décomposée par addition des éléments du dictionnaire : un visage est créé par l'addition d'un nez, d'une oreille gauche, d'un oeil droit,... Cette décomposition, sous forme de quantité de présence des éléments du dictionnaire est beaucoup plus facilement interprétable que les résultats de l'ACP.

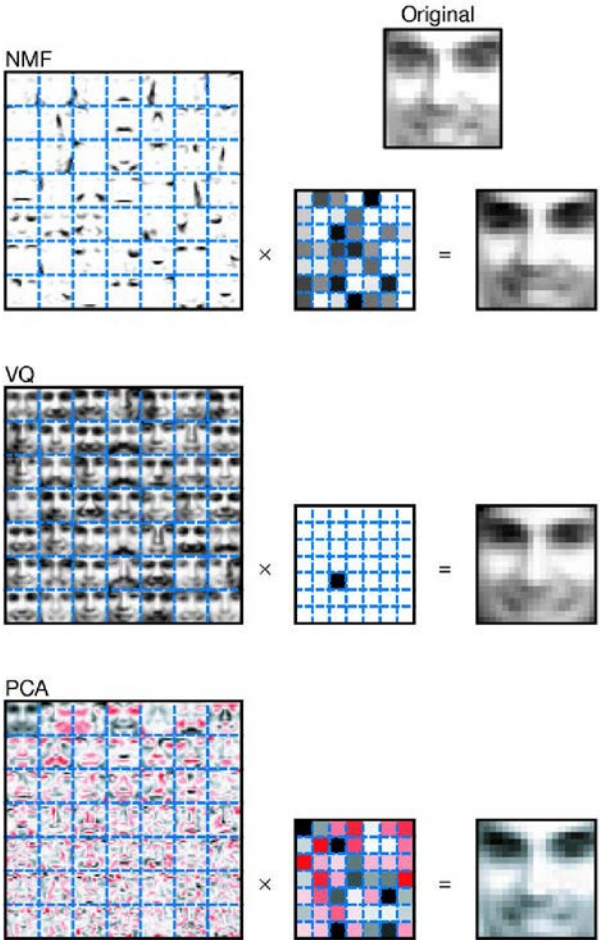


FIGURE II.2.4.5 Comparaison de la décomposition effectuée par NMF, ACP et VQ (tiré de Lee et al. [114]). L'image originale est factorisée sous la forme V=WH. Le dictionnaire W est représenté à gauche pour les trois méthodes, l'excitation H est représenté au centre (les valeurs positives sont illustrées en nuances de noir tandis que les valeurs négatives sont illustrées en nuance de rouge) et l'image reconstruite V est représentée à droite.

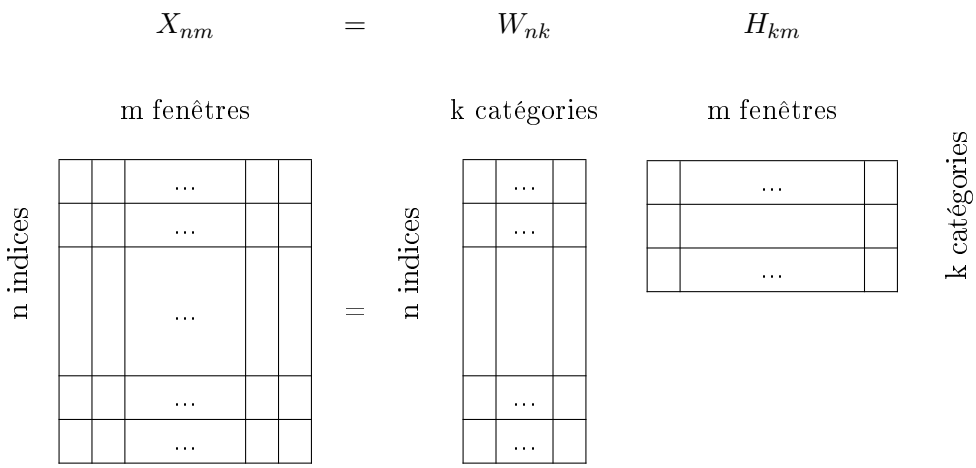


FIGURE II.2.4.6 Factorisation en matrices non-négatives de X = W.H

II.2.4.1 Factorisation en matrices non-négatives

La factorisation en matrices non-négatives (NMF pour Non-negative Matrix Factorisation) permet la factorisation d'une matrice en deux matrices non-négatives, comme illustré sur la figure II.2.4.6. Bien qu'il soit nécessaire de choisir la dimension dans laquelle on souhaite se projeter, il s'agit d'un algorithme non supervisé ne nécessitant pas d'exemple d'entraînement avec une classification pré-établie. Ainsi, tout comme l'Analyse en Composante Principale (ACP), elle permet une réduction de dimension grâce à une factorisation.

La NMF décompose une matrice X (n-par-m) en deux matrices W (n-par-k) et H (k-par-m), comme illustré dans la figure II.2.4.6. Dans la NMF, la contrainte principale consiste à garder des matrices non négatives, ce qui signifie que tous les coefficients de W et H doivent être positifs ou nuls.

La matrice X, dans notre cas, consiste en l'activation de certains indices (ou mots) en fonction du temps (dans certaines fenêtres temporelles). La matrice W représente le regroupement de ces n indices en k catégories, et la matrice H l'excitation de ces k catégories dans le temps (m fenêtres).

Dans l'algorithme NMF, plusieurs paramètres doivent être fixés : l'initialisation, le choix de la fonction coût et le choix des critères de convergence.

L'initialisation des matrices peut être faite de plusieurs manières, de la plus simple comme l'initialisation aléatoire (positive) à des méthodes plus complexes comme la NNSVD (Non-Negative double Singular Value Decomposition). Toutes ces initialisations et leurs implications ont été traitées par Boutsidis et al. dans [30]. Dans le cas de l'implémentation de Berry et al. [20] utilisée dans la suite, plusieurs initialisations aléatoires sont utilisées, afin d'éviter des maximas locaux dans la convergence de l'algorithme (seule la meilleure solution est conservée).

Le choix de la fonction coût dépend de plusieurs paramètres : la nature des données, les contraintes que l'on veut imposer entre les différents vecteurs du dictionnaire (orthogonalité, sparsité, ...). Tous ces problèmes ont été traités dans de nombreux documents, comme par exemple le livre de Cichocki et al. [43], l'article original de Lee et al. [115] ou le tutoriel de Hopke [90]. L'implémentation de Berry et al. [20] utilisée cherche à minimiser la fonction coût

définie dans l'équation II.2.1, où 'fro' représente la norme de Frobenius, n la première dimension de X et m sa seconde dimension.

$$D = \frac{\sqrt{||X - W * H||_{fro}}}{n * m} \tag{II.2.1}$$

Du choix de la fonction coût découle des règles de mise à jour des matrices. Les règles les plus communes sont les règles Alternating Least Square (ALS), les règles additives et les règles multiplicatives. Elles sont traitées dans le chapitre 2 de Cichocki [43]. Dans notre cas, la méthode Alternating Least Square (ALS) est utilisée, définie dans les équations II.2.2 et II.2.3, où le symbole † représente le pseudo-inverse de Moore-Penrose, et $[x]_+ = \max(\epsilon, x)$ permet de ne garder que les éléments positifs.

$$W \leftarrow [XH^T(HH^T)^{-1}]_+ = [XH^\dagger]_+ \tag{II.2.2}$$

$$H \leftarrow [(W^TW)^{-1}W^TX]_+ = [W^\dagger X]_+ \tag{II.2.3}$$

Quant aux critères de convergence, on peut citer le nombre d'itérations, la fonction coût qui atteint un certain seuil, les variations de la fonction coût qui sont sous un certain seuil, ... Ces éléments sont aussi traités dans les références précédentes. Dans cette thèse, tous ces critères sont combinés pour limiter les temps de calcul tout en assurant une bonne approximation de la factorisation.

Ces différentes étapes vont donc maintenant pouvoir être appliquées aux indices extraits de la base AMI.

II.2.4.2 Application des NMF à la catégorisation des interactions

Pour cette première modélisation, les indices sont calculés sur une fenêtre unique, contenant la totalité de chaque vidéo. Ainsi chaque vidéo est représentée par un vecteur unique constitué de tous les indices extraits pour l'ensemble des 4 participants de cette vidéo.

Ces vecteurs sont ensuite concaténés afin de constituer une matrice d'observation X de 73 indices par 138 vidéos, comme illustré sur la figure II.2.4.7. Cette matrice est ensuite décomposée grâce à une NMF, pour donner une matrice dictionnaire W (73 indices par k classes) et une matrice H (k classes par 138 vidéos), comme illustré sur la figure II.2.4.7. La matrice W permet de représenter l'importance qu'ont les indices dans chaque classe. La matrice H quant à elle montre l'appartenance de chacune des vidéos aux classes trouvées grâce à la décomposition.

Comme il a été expliqué dans la section II.1.2.2, Lewin et al. [117] considèrent trois grands types de leadership : autocratique, participatif et "free rein" (roue libre). Suivant cette idée, le nombre de classes est fixé à 3 pour la suite. Le résultat du NMF est présenté dans la figure II.2.4.8.

Afin de mieux comprendre la factorisation obtenue, nous allons tout d'abord essayer de visualiser quelques vidéos représentatives de chaque classe (section II.2.4.2.a). Puis nous analyserons la matrice dictionnaire W afin de voir quels sont les indices qui caractérisent chaque classe (sections II.2.4.2.b et II.2.4.2.c).

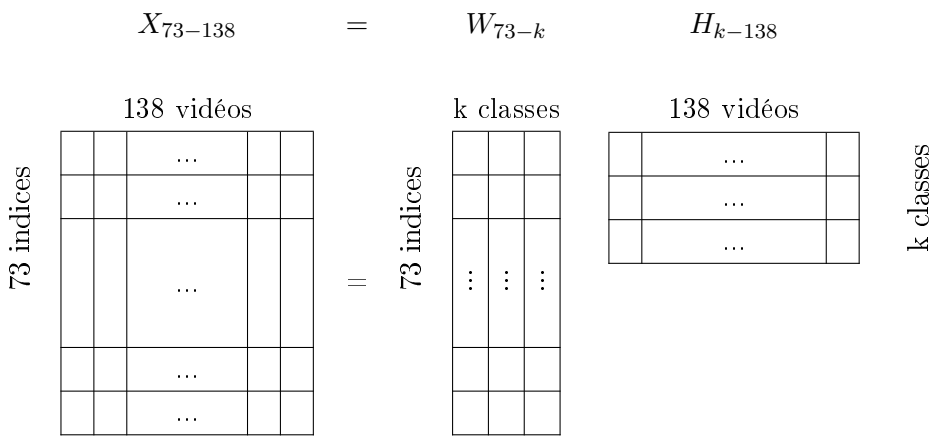
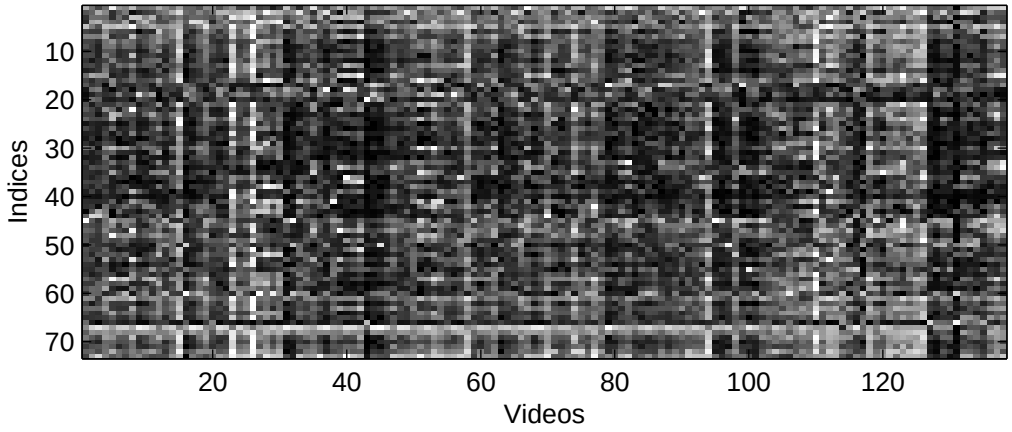
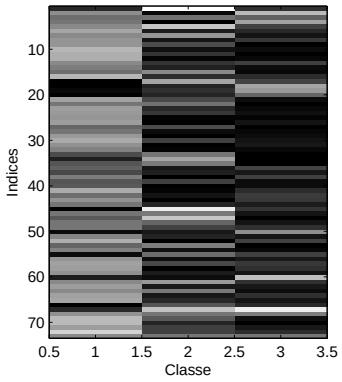


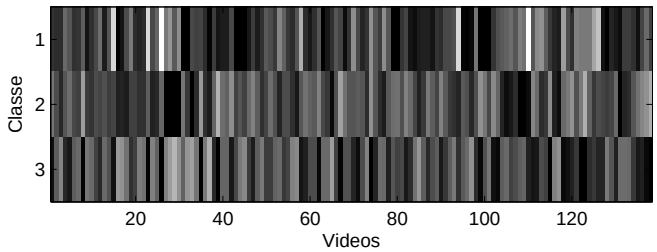
FIGURE II.2.4.7 Factorisation en matrices non-négatives des 73 indices de la base AMI.



(a) Matrice d'observation X, entrée des NMF.

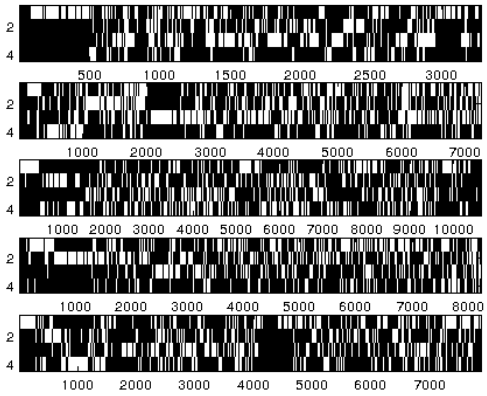


(b) Matrice dictionnaire W.

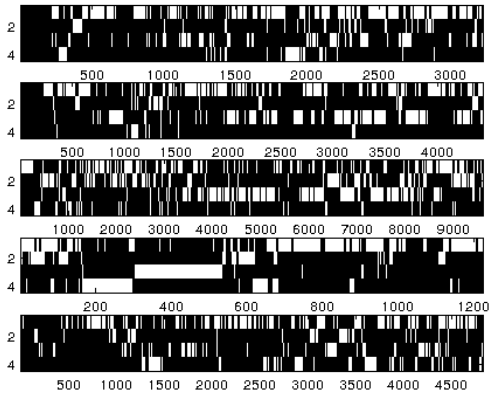


(c) Matrice d'excitation H.

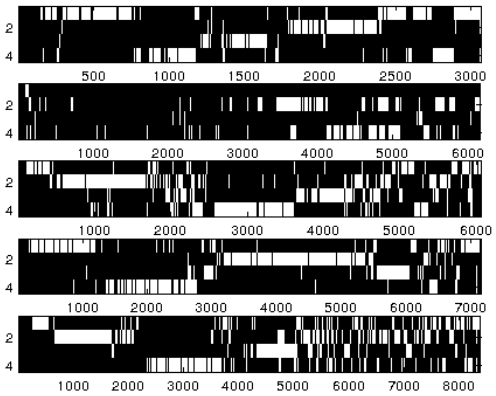
FIGURE II.2.4.8 Résultat de la décomposition des indices des vidéos en 3 classes par NMF.



(a) Classe 1.



(b) Classe 2.



(c) Classe 3.

FIGURE II.2.4.9 Présentation de 5 meetings représentatifs par classe.

II.2.4.2.a Visualisation de vidéos types pour chaque classe

Les figures II.2.4.9a, II.2.4.9b et II.2.4.9c présentent des exemples de tours de paroles de meetings appartenant à la même classe.

La visualisation des 3 classes de vidéo semble cohérente. En effet, la classe 1 (figure II.2.4.9a) semble être caractérisée par un grand nombre d'échanges, d'interruptions, ce qui fait penser à une sorte de débat et correspond donc à une interaction de type participatif.

Dans la classe 2 (figure II.2.4.9b), le Manager de Projet (première ligne) a l'air très présent, soit par de longs tours de parole, soit par de très nombreuses interventions. La classe 2 correspond donc à une interaction de type autocratique du Manager de Projet (PM)

La classe 3 (figure II.2.4.9c) quand à elle a l'air de ne pas être lié à un rôle en particulier, mais plutôt de se caractériser par le fait qu'un rôle autre que le Manager de Projet domine, avec de très longs tours de paroles. Elle correspond aux interactions de type "roue libre" ou "frein rein" tel que défini par [117].

Les analyses suivantes permettront de confirmer (ou d'infirmer) ces premières impressions.

TABLE II.2.1 Indices ayant le plus de poids pour chaque classe. Les indices en orange sont les indices personnels, les indices bleus des indices interactifs et les indices verts les indices de groupe.

Classe 1		Classe2		Classe 3	
Indice	Poids	Indice	Poids	Indice	Poids
RatioInter	4.18	TTP-PM	5.12	Durée1P	4.73
NbInter	3.71	NbRep-PM-PM	4.80	NbRep-UI-UI	3.80
NbTP	3.70	NbRep-PM-ME	3.99	TTP-ID	3.75
NbNonInter-UI	3.62	NbTP-PM	3.96	DuréeTP-ID	3.31
NbInter-ID	3.60	Durée1P	3.86	TTP-UI	3.24
NbRep-ID-UI	3.51	NbNonInter-PM-ME	3.35	DuréeTP-ME	3.04
NbInter-UI	3.48	DuréeTP-PM	3.32	NbRep-ID-ID	2.85
NbInter-ME	3.45	NbRep-PM	3.12	TTP	2.51
NbNonInter-ME-UI	3.40	NbTP-ME	3.03	RatioNonInter	2.40
NbInter-UI-PM	3.39	NbNonInter-ME	2.79	DuréeTP-UI	2.01
NbTP-ID	3.33	NbRep-ME-PM	2.49	TTP-ME	1.93
DuréePP	3.29	NbRep-ME	2.48	NbNonInter-ID-PM	1.40
NbInter-PM-ID	3.23	NbNonInter-PM-UI	2.46	DuréeTP-PM	1.37
NbRep-UI-PM	3.21	NbRep-PM-ID	2.42	NbNonInter-ID-UI	1.34
NbRep-UI	3.20	RatioNonInter	2.35	NbRep-ID-UI	1.32

II.2.4.2.b Analyse des coefficients de W

La matrice W contient la description des indices contenus dans chaque classe. Ainsi, pour chaque classe, il est possible de regarder les indices avec les poids les plus grands. Le tableau II.2.1 répertorie les 15 indices ayant les poids les plus importants dans chaque classe. Les poids de tous les indices, ordonnés par importance, sont tracés sur la figure II.2.4.10.

Comme nous pouvons le voir dans le tableau II.2.1, la classe 1 (interaction participative) est dominée par des indices liés à la dynamique du groupe : ratio global d'interruption par tour élevé (RatioInter), nombre d'interruptions réussies (NbInter) et nombre de tours de paroles (NbTP). Tout ceci indique que la conversation est très dynamique, avec beaucoup d'échanges et de changements de locuteurs. Cela confirme le premier *a priori* que l'on avait en observant la figure II.2.4.9a : c'est une classe de style "participatif". Neufs des 15 premiers indices sont liés aux interruptions (NbInter, NbNonInter, NbInter, NbNonInter).

L'analyse de la figure II.2.4.10a permet de voir que cette classe n'a pas de prédominance dans les poids, et que peu d'indices ont un poids nul : DuréeTP-PM, DuréeTP-ID, DuréeTP-UI, responses-PM-PM, et Durée0P. Ces poids nuls montrent que dans ce type d'interaction les tours de paroles sont plutôt courts (DuréeTP), et qu'il y a rarement de silence (Durée0P).

La classe 2 (interaction autocratique du project manager) est dominée par des indices liés au Manager de Projet (10 indices sur 15). Les indices les plus importants montrent que le PM parle longtemps (TTP-PM et DuréeTP-PM) ou souvent (NbTP-PM) et répond souvent (NbRep-PM, NbRep-PM).Tout ceci montre bien que la caractéristique de cette classe est qu'elle est centrée

sur l'omniprésence du PM.

L'analyse de la figure II.2.4.10b permet de voir des paliers. Ainsi les deux premiers indices (TTP-PM et responses-PM-PM) sont loin devant le seconde groupe d'indices constitués de NbRep-PM-ME, NbTP-PM et Durée1P. Plusieurs autres paliers existent ensuite, de moins en moins marqués. Les indices avec les poids nuls permettent de voir une absence d'interaction entre les rôles ID et UI. De plus, aucun de ces indices à poids nuls ne concerne le PM.

La dernière classe ("roue libre") quant à elle est caractérisée par de longs monologues (DuréeTP pour les 4 rôles, appuyé par de nombreuses interruptions ratées marquées par RatioNonInter et deux NbNonInter), avec peu de personnes parlant en même temps (Durée1P), et les trois rôles autre que le PM parlant beaucoup (TTP pour les trois). Cela confirme les observations faites lors de l'examen de la figure II.2.4.9c.

L'analyse de la figure II.2.4.10c permet là aussi de voir plusieurs groupes d'indices se détacher. Tout d'abord le premier indice, Durée1P, est loin devant. Puis vient un ensemble d'indices dont le poids décroit de manière linéaire jusqu'à environ 2, ce sont les indices identifiés précédemment. Les indices de poids nul sont ici aussi caractéristiques : NbInter-ME, NbInter-PM-ME, NbInter-ME-PM, NbNonInter-PM-ME, NbNonInter-PM-UI, responses-PM-ME, responses-ME-PM, et NbInter. Ils montrent que ce type d'interaction est caractérisé par peu d'interruptions (NbInter), et peu d'interaction avec le PM (les autres indices). Le palier d'indices les plus faibles confirme ceci avec des indices complémentaires (DSI et NbNonInter).

Comme nous l'espérions, ces trois types d'interaction sont liés aux trois types de leadership décrits par Lewin et al. [117] : autocratique, participatif et roue-libre. En effet, dans AMI, il existe un Project Manager. Bien que ce rôle ne soit pas forcément lié à la personne qui sera le vrai leader de l'interaction (il ne l'est que dans 65% des cas [96]), un leader finit toujours par émerger [160]. Le style du leader mène alors à un style d'interaction particulier.

Dans le tableau II.2.1, certains indices ont des poids forts pour plusieurs classes (par exemple Durée1P et RatioNonInter pour les classes 2 et 3). Ils ne permettent donc pas de discriminer les classes. Nous allons donc voir les indices qui sont les plus différents selon les classes.

II.2.4.2.c Indices les plus différents selon les classes

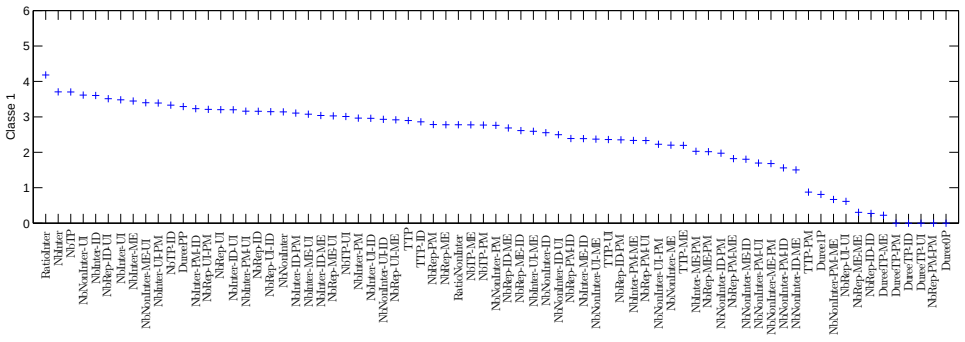
Afin de mieux identifier les classes, nous allons chercher les indices qui sont les plus discriminants, c'est à dire ceux qui ont des valeurs très différentes selon les classes.

Pour cela, le poids de chaque indice de chaque classe est divisée par la somme de cet indice sur toutes les classes (voir équation II.2.4). On appellera poids discriminant (noté $w_{discriminant}$) le résultat. Ainsi, le poids discriminant de l'indice i, pour la classe c, est défini par :

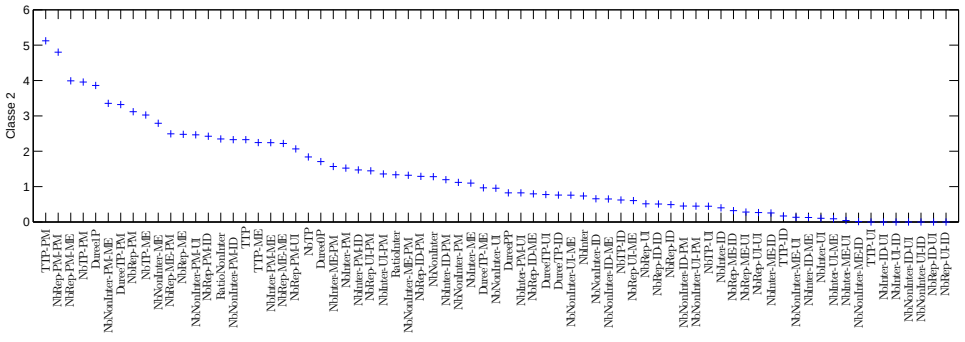
$$w_{discriminant}(i,c) = \frac{w(i,c)}{\sum_{s=1}^{k} w(i,s)}, \text{ avec ici i}\in[1\,;73],\ c\in[1,2,3] \text{ et k=3} \qquad \text{(II.2.4)}$$

Ainsi, un score proche de 1 pour $w_{discriminant}(i,c)$ signifie que l'indice numéro i est très présent pour la classe c, et absent dans les autres classes. Un score égal à $\frac{1}{3}$ signifie que cet indice est aussi présent dans une classe que dans l'autre, il n'est donc pas discriminant.

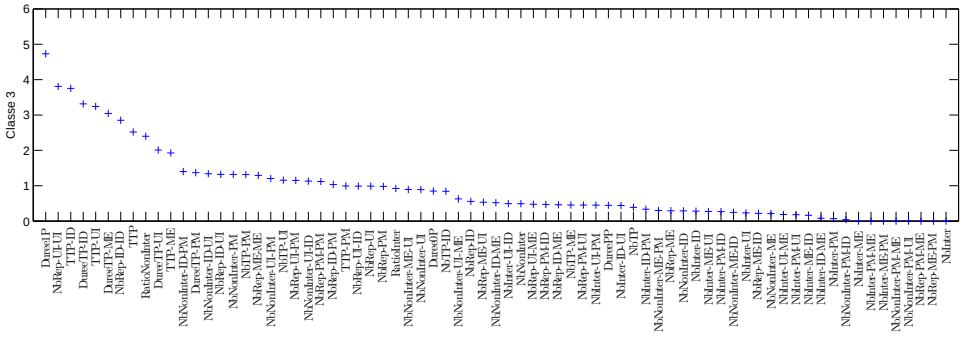
Nous pouvons donc visualiser la matrice W remaniée selon ces poids, sous forme du tableau II.2.2.



(a) Poids des indices de la classe 1.



(b) Poids des indices de la classe 2.



(c) Poids des indices de la classe 3.

FIGURE II.2.4.10 Poids des indices dans la matrice W, pour chaque classe, ordonnés par ordre de grandeur.

TABLE II.2.2 Indices triés par poids discriminant pour chaque classe. Les indices en orange sont les indices personnels, les indices bleus des indices interactifs et les indices verts les indices de groupe.

Classe 1		Classe2		Classe 3	
Indice	Poids	Indice	Poids	Indice	Poids
NbInter-ID-ME	0.94	NbNonInter-PM-ME	0.84	DuréeTP-ID	0.81
NbInter-UI	0.91	NbRep-PM-PM	0.81	NbRep-UI-UI	0.81
NbInter-ME-UI	0.91	TTP-PM	0.73	NbRep-ID-ID	0.78
NbInter-UI-ME	0.90	DuréeTP-PM	0.71	DuréeTP-UI	0.72
NbInter-ID-UI	0.88	NbRep-PM-ME	0.69	DuréeTP-ME	0.72
NbNonInter-ME-ID	0.88	Durée0P	0.67	TTP-UI	0.58
NbInter-UI-ID	0.86	NbNonInter-PM-ID	0.59	TTP-ID	0.55
NbInter-ME-ID	0.85	NbNonInter-PM-UI	0.59	Durée1P	0.50
NbInter-ID	0.84	NbRep-ME-ME	0.58	NbNonInter-ID-PM	0.37
NbInter	0.83	NbRep-ME-PM	0.55	NbNonInter-ID-UI	0.35
NbRep-ME-ID	0.83	NbNonInter-ME	0.54	NbRep-ME-ME	0.34
NbRep-ME-UI	0.79	NbTP-PM	0.49	Durée0P	0.33
NbNonInter-ME-UI	0.77	NbInter-PM-ME	0.49	TTP	0.33
NbRep-UI-ID	0.76	NbTP-ME	0.48	RatioNonInter	0.32
NbInter-PM-UI	0.76	NbRep-PM-ID	0.46	NbNonInter-UI-PM	0.31

Cette visualisation des données, complémentaire de la visualisation par poids bruts, permet de mettre en avant les indices qui possèdent des valeurs très différentes suivant les classes, et caractérisent donc les-dites classes. Ainsi, on peut voir que la classe 1 (interaction participative) est caractérisée par les indices liés aux interruptions (12 sur 15). La classe 2 (interaction autocratique du project manager) quand à elle possède 11 indices liés au Manager de Projet. Enfin la classe 3 ("roue libre") confirme ses indices avec peu de changements, gardant des indices liés aux longs tours de paroles des rôles autre que le Manager de Projet.

Ces deux analyses ont permis de montrer que chaque classe possède des indices la caractérisant, et que certains indices sont caractéristiques de chaque classe. Cependant, la question de la cohérence de ces indices à l'intérieur de chaque classe se pose : est-ce qu'un indice avec un fort poids dans la matrice W est omniprésent dans toutes les vidéos appartenant à cette classe ? Existe t'il des indices, peut être avec un coefficient moins fort, mais dont la cohérence à l'intérieur de chaque classe est plus grande ?

II.2.4.2.d Indices les plus cohérentes à l'intérieur de chaque classe

Afin de vérifier si les indices avec les poids les plus forts sont bien des invariants de la classe, nous avons tracé des diagrammes boite à moustache sur la figure II.2.4.11 représentant la distribution des valeurs prises par chaque indice pour toutes les vidéos de la classe. Ces diagrammes sont assez difficiles à interpréter, aussi, nous en extrayons quatre exemples spécifiques à certains indices (II.2.4.12). Ces quatre exemples permettent de voir que les indices que l'on avait

trouvé dans la section II.2.4.2.c comme étant des indices typiques à chaque classe le sont bien aussi dans la réalité. Par exemple, l'indice TTP-PM qui figurait parmi les indices permettant de discriminer la classe 2, est bien statistiquement supérieur dans cette classe que dans les deux autres classes (voir la figure II.2.4.12a).

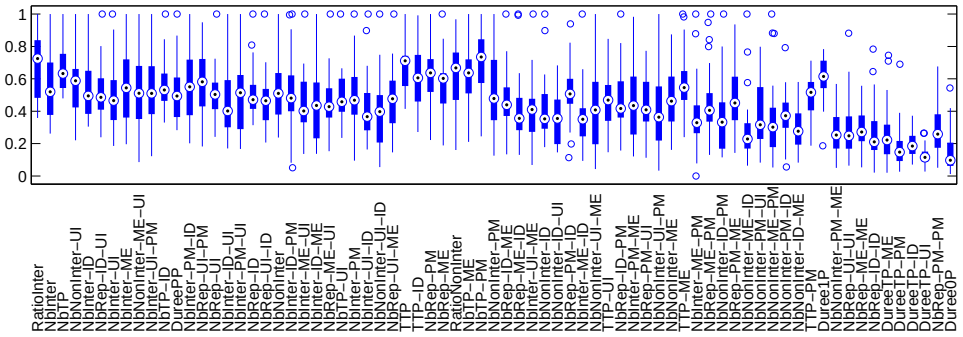
II.2.4.3 Évolution du type d'interaction au cours du temps

Jusqu'ici, les indices étaient calculés sur une unique fenêtre comprenant les interactions complètes. Cela menait à un vecteur représentant chaque interaction. Cependant les interactions changent au cours du temps, et catégoriser une interaction entière dans une seule classe est réducteur et ne permet pas de voir la dynamique changeante des interactions. Afin d'augmenter la finesse d'analyse, les indices sont alors calculés sur des fenêtres temporelles beaucoup plus courtes que l'interaction elle même. Cependant, étant donné qu'un bon nombre d'indices sont des statistiques (durée moyenne, nombre de tours de paroles,...), il est nécessaire d'avoir des fenêtres temporelles qui ne soient pas trop courtes. Ainsi, tout comme Jayagopi et al. font dans [95], des fenêtres de longueur 5 minutes ont été utilisées, décalées entre elles de 1 minute (la première fenêtre concerne les temps de 0 à 5 minutes, la seconde de 1 minute à 6 minutes, etc). Une fois les indices calculés sur ces fenêtres, comme tous les indices sont normalisés par la longueur de la fenêtre, il est donc possible de réutiliser le dictionnaire appris précédemment afin de reconnaître les probabilités d'appartenance de chaque fenêtre temporelle à chacune des classes en calculant la matrice H, comme illustré sur la figure II.2.4.13.

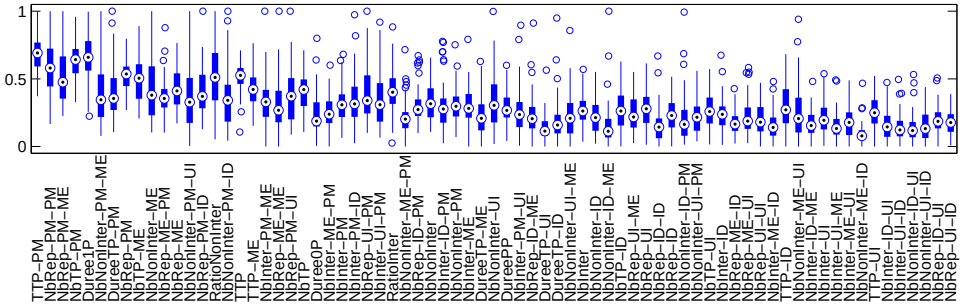
Un exemple du résultat est présenté sur la figure II.2.4.14. Sur cette figure, on peut voir qu'entre 1500 et 2500 échantillons (375 secondes à 625) le participant 2 parle quasiment tout seul, ce qui mène donc à un indice DuréeTP-ME très important. Or nous avions vu que cet indice est caractéristique de la classe 3 (tableau II.2.2). Ceci est bien reconnu par la NMF, dont les coefficients de la classe 3 dans la matrice d'excitation H (bas de la figure II.2.4.14) sont très importants pour les fenêtres 5 à 11. Si l'on regarde plus loin dans la matrice X (haut de la figure II.2.4.14), autour de 6000 échantillons ou 9800 échantillons, on voit de très nombreux échanges et interruptions entre les participants, ce qui est caractéristique de la classe 1 (cf tableau II.2.2). Ainsi, notre méthode permet de visualiser dans le temps les variations du type d'interaction.

Conclusion

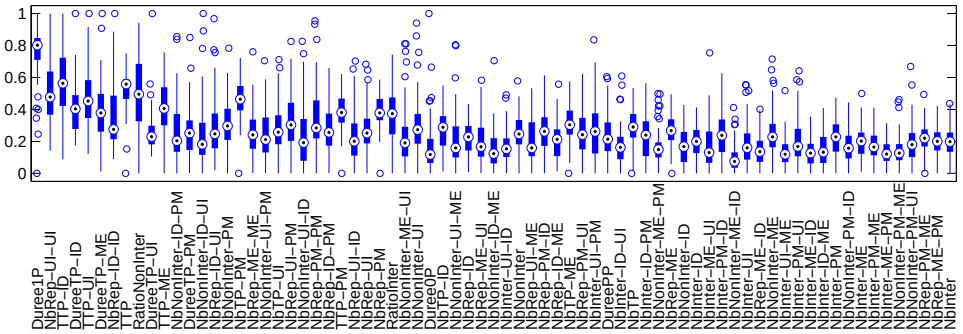
Dans cette partie, nous avons tout d'abord extrait différents indices ad-hoc permettant de décrire l'interaction. Ces indices se répartissent en trois catégories : des indices intra-personnels, des indices inter-personnels et des indices de groupe. A partir de ces indices, nous avons souhaité caractériser de manière automatique le type d'interactions et faire ressortir des interactions type. Après avoir répertorié toutes les méthodes qui pouvaient être envisagées, nous avons montré que



(a) Variabilité des indices de la classe 1.

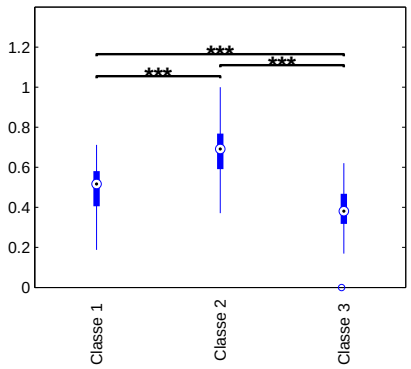


(b) Variabilité des indices de la classe 2.

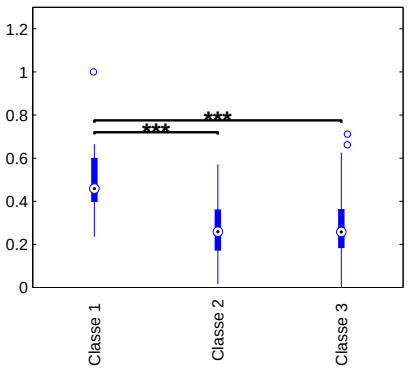


(c) Variabilité des indices de la classe 3.

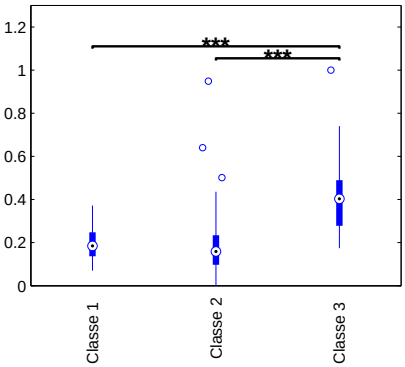
FIGURE II.2.4.11 Variabilité des indices dans la matrice W, pour chaque classe, ordonnés par ordre de grandeur.



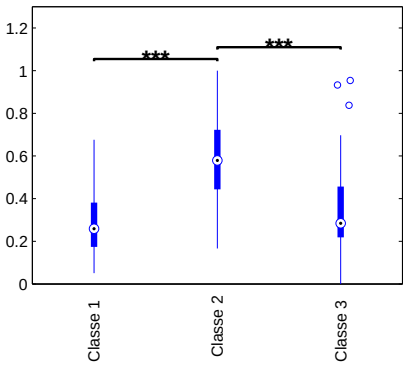
(a) Variabilité de l'indice TTP PM pour les trois classes. T test p<0.05 entre toutes les classes.



(b) Variabilité de l'indice NbInter UI pour les trois classes. T test p<0.05 entre les classes 1 et 2, puis 1 et 3. Cet indice est bien spécifique à la classe 1.



(c) Variabilité de l'indice DuréeTP ID pour les trois classes. T test p<0.05 entre les classes 1 et 3, puis 2 et 3. Cet indice est bien spécifique à la classe 3.



(d) Variabilité de l'indice NbRep PM PM pour les trois classes. T test p<0.05 entre les classes 1 et 2, puis 2 et 3. Cet indice est bien spécifique à la classe 2.

FIGURE II.2.4.12 Variabilité de quelques indices comparés pour les différentes classes. Des tests statistiques permettent de vérifier la séparabilité des indices pour les classes.

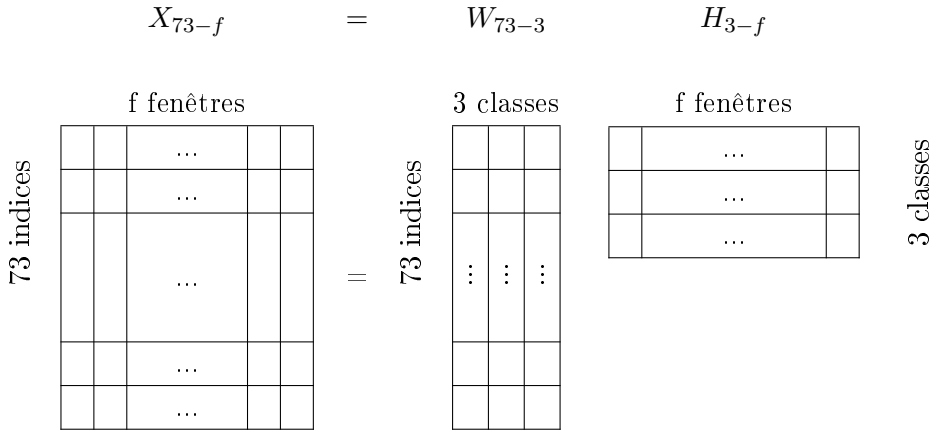
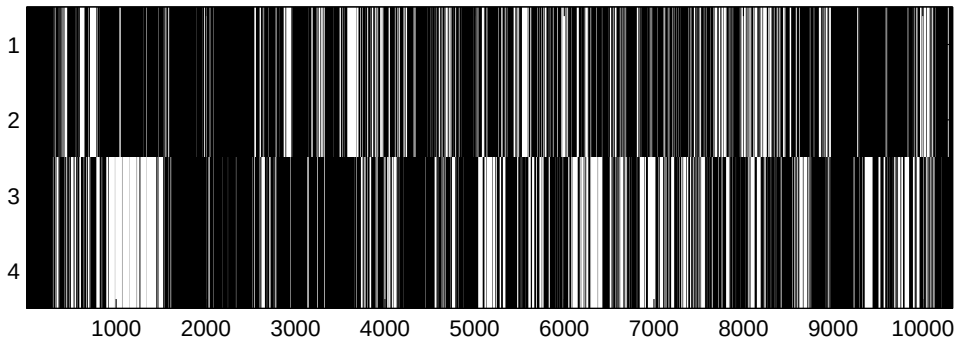
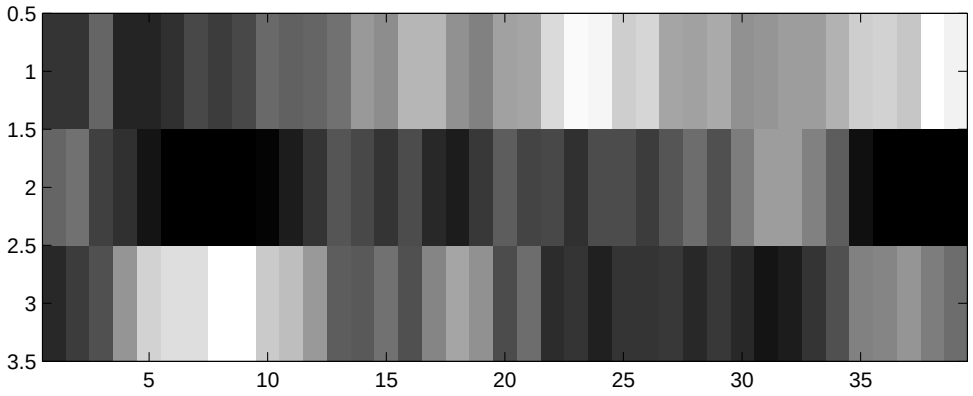


FIGURE II.2.4.13 Décomposition de la matrice de données X contenant les 73 indices calculés sur f fenêtres d'une interaction. Le but est de calculer la matrice H donnant le degré d'appartenance de chaque fenêtre à chacune des classes, en réutilisant la matrice dictionnaire W calculée précedemment.



(a) Matrice X.



(b) Matrice H.

FIGURE II.2.4.14 Probabilités d'appartenance aux classes qui varient au cours du temps, pour le meeting 113.

l'application d'une factorisation en matrices non négatives permettait de découvrir de manière non supervisée les différents types d'interactions, qui rejoignent les grands types de leadership définis par Lewin et al. [117]. On retrouve ainsi des interactions où le premier participant est autocratique, des interactions participatives, une sorte de débat où tous les participants interagissent et des interactions de type "roue libre" qui ne répondent pas aux deux premiers scénarios.

Chacun de ces trois type d'interaction possède des caractéristiques propres qui permettent de les différencier, comme par exemple un grand nombre d'interruptions et des temps moyens de tour de parole courts pour les interactions participatives, une omniprésence du Project Manager dans les interactions autocratiques, et de longs discours d'un autre rôle que le PM dans le cas des interactions de type "roue libre". Cependant, une interaction réelle est rarement d'un seul type et varie au cours du temps. Nous avons donc analysé sur des fenêtres temporelles de 5 minutes la variation d'appartenance aux trois classes. Ceci nous a permis de visualiser les changements de type d'interactions qui s'opèrent au cours du temps.

Il est assez difficile de valider les méthodes non supervisées dans le sens où on aurait pu obtenir d'autres interactions types qui auraient aussi une cohérence. De plus, même si les résultats obtenus sont logiques et correspondent à nos attentes, il est très difficile de quantifier leur qualité. La suite de ces travaux portent donc sur la reconnaissance de rôle, problème bien connu dans la littérature où une vérité de terrain est présente et permet donc de valider les résultats.

Troisième partie

Reconnaissance de rôles

Introduction

De nos jours, les réunions occupent une place de plus en plus importante dans la vie des entreprises. Il a été démontré [63] que les cadres passent plus de 50% de leur temps de travail en réunion. Cependant, les professionnels s'accordent à dire que jusqu'à 50% du temps passé est improductif et 25% passé à discuter de sujets non pertinents. La compréhension des interactions sociales devient un enjeu sociétal et économique majeur. Ainsi, dans un grand nombre d'entreprises, des coachs sont mis en place afin d'améliorer et faciliter les interactions.

La communauté du traitement du signal social s'intéresse donc à l'étude des interactions, et se pose les questions du Où ? Avec qui ? Comment ? Sur quel sujet ? Dans quel but ?. Deux grandes approches existent. La première provient de la linguistique et tente de résoudre le problème des interactions sociales de la perspective de la compréhension des dialogues. L'analyse des conversations permet d'examiner la manière dont les partenaires gèrent les tours de paroles, les difficultés de communication, le choix des mots... Tous ces éléments permettent de donner une image de la relation entre les partenaires, leurs croyances ou leur affiliation à des groupes. La seconde approche considère l'interprétation de la communication non verbale. A l'intérieur de cette dernière, les expressions faciales, la prosodie, les gestes du corps ou les tours de paroles sont utilisés afin d'obtenir des indices sur les comportements individuels et de groupe comme l'engagement, la dominance, la persuasion...

Les interactions sociales peuvent être définies comme une séquence dynamique d'échange de signaux sociaux entre des individus qui modifient et adaptent leur comportement en fonction de leurs partenaires. Les interactions sont définies et influencées par un grand nombre de paramètres, parmi lesquels il est possible de citer le type de meeting, les positions dans un groupe, la structure du groupe (existence d'une hiérarchie par exemple), le degré de familiarité entre les personnes, la charge émotionnelle (qui se reflète dans l'humeur)... La psychologie et la sociologie ont déterminé que les rôles sont une variable très importante pour caractériser les interactions sociales. En effet, Tischler disait dans son *Introduction à la Sociologie* [177] : "les gens n'interagissent pas entre eux comme des être anonymes. Ils viennent ensemble dans le contexte d'un environnement spécifique, avec des buts spécifiques. Leurs interactions impliquent des comportements associés à des statuts définis et des rôles particuliers. Ces statuts et ces rôles aident à façonner nos interactions sociales et fournissent une prévisibilité".

Les rôles sont en effet un aspect majeur dans la compréhension des interactions sociales. Tout d'abord, les rôles sont associés à des attentes partagées que les personnes ont à propos de leur propre comportement ainsi que des comportements des autres. Ainsi, ils contribuent à la prévisibilité générale dans les interactions, ce qui constitue une condition majeure lorsque l'on souhaite faire des suppositions convenables à propos des autres et ainsi participer aux échanges

sociaux. D'autre part les rôles résultent en des motifs caractéristiques du comportement qui peuvent être identifiés et reconnus par des partenaires dans une interaction. La caractérisation et la reconnaissance des rôles sont donc des étapes clé dans l'étude des interactions et leurs applications sont multiples.

Parmi les applications possibles de la reconnaissance des rôles, la structuration de documents audio [192, 193] est primordiale. En effet, de nos jours de très grandes quantités de données audio sont archivées, et les recherches à l'intérieur de ces documents peuvent s'avérer fastidieuses. Afin de les simplifier, la reconnaissance des rôles peut permettre de résumer efficacement les dits documents (les rôles étant souvent représentatifs d'un type de contenu particulier). Ainsi dans des navigateurs de données, le rôle d'une personne peut permettre d'identifier rapidement les segments d'intérêt (le présentateur d'un journal permet d'obtenir un résumé d'un sujet, alors qu'un journaliste permet d'obtenir une version détaillée du sujet). Les rôles peuvent aussi n'être utilisés que comme une information supplémentaire permettant d'enrichir la description d'un document (avec son titre, ses sujets, ...).

Enfin, comme évoqué au début de ce paragraphe, la compréhension des rôles dans un groupe peut permettre de mettre en place des outils automatiques d'amélioration des relations dans un groupe, en améliorant la conscience de soi tout comme le font les coachs. En effet, le but n'est pas de décrire des événements précis, mais de donner un retour sur des périodes de temps plus globales, sur le comportement ou la position des personnes dans le groupe, comme par exemple : "au début de la réunion, tu ne laissais personne parler" ou "dans la deuxième partie de la réunion, tu as été trop passif".

Dans la suite de cette partie, nous ferons tout d'abord un état de l'art des travaux concernant l'analyse et la reconnaissance des rôles, à la fois en psychologie sociale et en traitement du signal social. Des bases de données liées aux rôles seront alors présentées, avant de détailler des méthodes computationnelles de reconnaissance. Nous nous concentrerons ensuite sur la découverte non supervisée de motifs d'interactions, puis nous présenterons et évaluerons un système de reconnaissance des rôles utilisant lesdits motifs, avant de conclure.

Chapitre III.1

État de l'art

Sommaire

Ce chapitre a pour objectif de présenter une revue des travaux liés aux rôles des individus dans une interaction. Nous commencerons par les travaux effectués en psychologie sociale, et présenterons trois grandes catégories de rôles de ce domaine. Nous verrons ensuite les bases de données existantes pour chacune des trois grandes catégories, puis nous finirons par les travaux du domaine du traitement du signal social qui traitent de la reconnaissance de rôles.

III.1.1 Définition des rôles

III.1.1.1 Approche en psychologie sociale

La psychologie sociale a longtemps étudié les groupes [10, 124]. L'étude des rôles que jouent les individus au sein d'un groupe est une composante centrale dans la compréhension du fonctionnement de ce groupe.

Le terme de rôle a été traité dans la littérature de psychologie sociale de nombreuses manières. Par exemple, Goffman et al. [80] ou Tishler et al. [177] voyaient les rôles d'un point de

vue dramatique, décrivant les gens comme des acteurs dans le grand théâtre de la vie. Cependant, trois grandes tentatives de définition des rôles ont émergé dans les interactions en petits groupes, et sont détaillées dans les études de Salazar [159] ou de Hare [87] :

1. les rôles **formels** sont liés à l'attente qu'ont les autres d'un comportement spécifique qu'une personne est censée effectuer ;
2. les rôles **informels** sont liés à des caractéristiques associées à la position d'une personne dans un groupe (son statut) ;
3. les rôles **fonctionnels** sont liés au comportement adopté par une personne dans une situation particulière.

La première définition est par exemple soutenue par Bormann [28], qui dit que "quand les autres membres savent ce qu'une personne va faire, et que cette personne sait ce que l'on attend d'elle, alors cette personne assume un rôle". Cependant, selon Salazar [159], cette approche souffre de plusieurs limitations [159]. Tout d'abord, il se peut que la personne ne joue aucun rôle dans le groupe. Ensuite, comme les rôles formels sont liés aux attentes des autres, le comportement de la personne joue une part mineure dans la détermination du rôle de la personne. Enfin, les attentes liées à réalisation du rôle ne sont pas forcément les mêmes pour l'acteur que pour les autres membres.

La seconde définition est partagée par de nombreux auteurs tels que Katz et Kahn [103] ou McGrath [124] et considère que les rôles sont équivalents à la position (statut) dans un groupe ou une organisation. Ainsi ces rôles informels ne sont pas liés aux caractéristiques d'une personne mais aux caractéristiques de comportement incombant à la personne dans une position particulière [124]. Pour Katz et al. [103], un rôle est défini par une séquence de quatre étapes : attentes, envoyer des attentes, recevoir des attentes, et se comporter. Mais cette vue a aussi ses limites [159]. La première vient de l'hypothèse que les rôles sont attribués. C'est à dire qu'un rôle correspond à une position à laquelle une personne est associée et dont le comportement doit correspondre. La seconde limitation est liée au fait que le rôle d'une personne doit être le résultat du processus d'interaction dans un groupe, et qu'il n'est pas juste associé à sa position.

La troisième définition a été soutenue par Biddle [21] qui considère les rôles comme "des caractéristiques comportementales d'une ou plusieurs personnes dans un contexte". C'est à dire que les rôles fonctionnels sont liés aux individus, qu'ils apparaissent dans un certain contexte et sont donc limités par des spécificités contextuelles. Ces observations ont mené à la conclusion que les rôles consistent en des "comportement modaux", spécifiques à une catégorie de rôle pour une personne particulière [159]. Ils peuvent d'autre part changer en réponse à des changements de contexte spatiaux ou temporels.

Cette troisième définition s'est développée et les rôles correspondant sont nommés "rôles fonctionnels" [159]. Ainsi Benne et Sheats [17] ont séparé ces rôles fonctionnels en trois classes : tâche-orientés, maintenance-orientés et individus-orientés. Les deux premiers types sont tournés vers les besoins d'un groupe : les rôles tâche-orientés fournissent une coordination et facilitation en vue de l'accomplissement d'une tâche, alors que les rôles maintenance-orientés contribuent à la structure sociale et aux relations interpersonnelles afin de réduire les tensions et de maintenir un fonctionnement de groupe harmonieux. Le troisième type de rôles, les individus-orientés, sont plus tournés vers les besoins et objectifs des individus.

Bales [11] a étendu l'approche de Benne et Sheats [17] et proposé l'Analyse de Processus d'Interaction (IPA), une méthode pour étudier de petits groupes en classifiant les comportements individuels dans un espace de rôles à deux dimensions, la première correspond aux Tâches et la seconde à l'aspect Socio-Émotionnel. Cette dernière dimension provient d'activités qui supportent ou affaiblissent les relations interpersonnelles. Par exemple, complimenter une personne est un comportement socio-émotionnel positif alors qu'insulter une personne est à l'opposé et détruit la cohésion du groupe. La dimension correspondant aux Tâches est plutôt liée à du management et à donner des solutions pour les problèmes auxquels le groupe fait face, comme par exemple poser et répondre à des questions.

III.1.1.2 Les rôles en science computationnelle

La reconnaissance des rôles dans le domaine du Traitement du Signal Social se place principalement dans deux cadres : des émissions (radio ou télévisuelles) [14, 13, 118, 181, 69, 145, 156, 198, 146, 157] ou des réunions réunissant un nombre modéré de participants [199, 60, 77, 69, 140, 113, 156, 22, 185, 194, 164, 61, 163]. Ces différents cadres mènent à des analyses de différents types de rôles, qui suivent la classification établie dans le domaine de la psychologie sociale.

La première catégorie de rôle est liée à la position (ou statut) qu'occupent les personnes dans un groupe (cf deuxième approche en psychologie sociale présentée précédemment). Cette catégorie peut être séparée en deux sous catégories : les rôles formels et informels. Les rôles **formels** sont des rôles qui correspondent à des fonctions spécifiques, par exemple les différents intervenants dans des émissions comme le présentateur, le second présentateur, le journaliste, l'invité, l'interviewé ou le présentateur météo. Les rôles **informels** quant à eux sont liés à une position dans un système social, sans que cela ne corresponde à des fonctions spécifiques. On trouve ainsi des rôles plus "diffus", car bien qu'ils correspondent à une position, cette position n'a pas forcément de particularité, comme les rôles d'Expert Marketing, de Designer Industriel et de concepteur de l'Interface Utilisateur présents dans la base AMI [36]. Ces rôles sont souvent fixes au cours de l'interaction.

La seconde catégorie est liée aux comportements modaux qu'expriment les participants (cf troisième approche en psychologie sociale). Dans cette catégorie, une classification se basant sur les travaux de Bales [11] a été proposée par Pianesi [140] et est utilisée tout ou partie dans de nombreuses études [199, 60, 140]. Elle comporte deux sous-catégories, les **rôles liés à la tâche** et les **rôles socio-émotionnels**. Cette catégorisation s'est démontrée être fiable en terme d'évaluation, puisque la consistance inter-juge est bonne (coefficient de Cohen κ=0.7 pour la tâche et κ=0.6 pour le socio-émotionnel [140]) .

Les rôles liés à la tâche sont décomposés en 5 éléments :

l'orienteur : oriente le groupe en donnant le planning, définit des buts et procédures, garde le groupe concentré et résume les arguments les plus importants et les décisions du groupe ;

le donneur : fournit des informations factuelles, répond aux questions et donne ses croyances et attitudes à propos d'une idée en exprimant des valeurs personnelles et des informations factuelles ;

le demandeur : demande des suggestions, des informations, ainsi que des clarifications afin d'améliorer l'efficacité de prise de décision du groupe ;
le greffier, ou technicien procédural : utilise les ressources disponibles pour le groupe, les gérant pour le groupe ;
le suiveur : écoute juste, sans participer activement à l'interaction.

Les rôles socio-émotionnels sont eux aussi composés de 5 éléments :

l'attaquant : minimise le statut des autres, exprime son désaccord, attaque le groupe ou les problèmes ;
le portier : est le modérateur du groupe, arbitre les relations, encourage ou facilite la participation et régule le flux de communication ;
le protagoniste : prend la parole, conduit la conversation en assumant une perspective personnelle et en affirmant son autorité ;
le supporter : a une attitude coopérative démontrant la compréhension, l'attention et l'acceptation tout en fournissant un support technique et relationnel ;
la personne neutre : accepte de manière passive les idées des autres, servant d'audience à la discussion du groupe.

Chaque participant va jouer plusieurs rôles au cours de l'interaction, mais chacun ne joue à chaque instant qu'un seul rôle lié à la tâche et un seul rôle socio-emotionnel.

III.1.2 Bases de données

Un grand nombre de bases existent dans la littérature pour la détection de rôles (voir la table III.1.1). Ici encore, les trois catégories de rôles (formels, informels et fonctionnels) apparaissent, et bien que certaines bases aient été créées avec un type de rôle précis, des annotations supplémentaires peuvent être ajoutées afin de détecter un autre type de rôle, comme par exemple la base AMI [123] dont les quatre rôles sont informels (liés à une position dans une entreprise virtuelle) mais sur laquelle un certain nombre d'études ont préféré détecter des rôles fonctionnels [60, 164, 163].

III.1.2.1 Rôles formels

Le Linguistic Data Consortium [45] est de loin le plus gros fournisseur de bases de données audios trouvables à l'adresse `https://catalog.ldc.upenn.edu/`. Parmis toutes ces bases, la base TREC-7 SDR contient 35 demi-heures de diffusion du programme radio "All Things Considered". C'est une sous base de la base DARPA HUB-4 [75], elle aussi étant un sous ensemble de la base DARPA Broadcast News. Les rôles de cette base sont les rôles classiques d'émissions radios ou télévisuelles, à savoir : présentateur, second présentateur, journaliste, invité, interviewé et météo. Une autre base est la Broadcast Conversation, enregistrée dans le programme DARPA GALE. Elle contient des interactions spontannées dans des programmes radio tels que des débats, des interviews, des directs,... Les sujets de la base sont variés, allant de la politique, l'économie jusqu'aux problèmes sociaux. Un total de 36 bulletins ont été annotés manuellement, chacun d'entre eux comportant un nombre d'intervenants compris entre 3 et 18.

Dans [12], Banerjee et al. créèent la base CMU (Carnejie Mellon University) de meetings

à l'aide d'une salle multimodale, grâce à laquelle l'audio (un micro-cravate pour chaque personne), la vidéo (une caméra globale), et plusieurs informations supplémentaires telles que les présentations ou le tableau blanc sont enregistrés de manière synchronisée. Les rôles de cette base sont liés au meeting, à savoir présentateur, participant cible, participant passif.

La base TDT4 [109] contient environ 300 heures de diffusion audio provenant de multiples radios (CNN, ABC, PRI, ...). Elle a été enrichie d'annotations sur les sujets et des transcriptions par Strassel et al. [172]. Les rôles contenus sont là aussi liés aux émissions radios.

Dans [181], Vinciarelli présente une base de diffusions radio, composée de 96 bulletins de la Radio Suisse Romande. La durée totale de cette base s'élève à 18 heures 56 minutes et 23 secondes. Chaque bulletin est composé d'environ 12 participants et dure en moyenne 11 minutes 50. Une seconde base de cette radio est composée de 27 débats durant chacun une heure, et implique en moyenne 25 participants.

Le corpus EPAC [68] est composé de 100 heures de données audio, sélectionnées parmi 1700 heures de contenu audio pour leur nature conversationnelle. Ces données sont tirées de trois radios Françaises (France Inter, RFI, France Culture) et ont été manuellement transcrites et annotées en terme de segmentation de locuteur et de rôles (animateur, journaliste, invité...).

III.1.2.2 Rôles informels

La base AMI [123], et notamment sa sous base contenant seulement les meetings scénarisés [36] est une collection de 138 meetings pour une durée totale de 45 heures et 38 minutes de matériel audio (micros cravate pour chaque participant), vidéos (globale et une centrée sur chaque participant) et métadonnées (tableau, présentations, ...). Elle met en jeu 4 personnes jouant chacune un rôle particulier : le Manager de Projet, l'Expert Marketing, l'Expert d'Interface Utilisateur et le Designer Industriel. La même personne, apparaissant dans plusieurs meetings, peut jouer des rôles différents selon les meetings, mais les rôles sont fixés à l'intérieur de chaque meeting.

L'émission de télévision "The Apprentice" de la chaîne NBC est décrite dans [145, 146]. Elle met en jeu six à dix-huit personnes participant à une compétition avec des éliminations, le vainqueur gagnant un salaire de 250 000 dollars et dirigeant une des compagnies du magnat Donald Trump. Étant une émission de télé-réalité, les réactions des participants sont naturelles. Environ 90 minutes de données audios ont été traitées par Raducanu et al [145, 146].

Le corpus de meetings ICSI décrit dans [94] par Janin et al. consiste en des enregistrements audio de réunions naturelles ayant pris place au sein de l'International Computer Science Institute à Berkeley. Ces réunions étant naturelles, le nombre de participants varie grandement d'un enregistrement à l'autre (de 3 à 9), avec en moyenne 6 participants. Un total de 75 réunions sont publiées, pour un total d'environ 72 heures. Des métadatas sur les participants sont fournies, comme l'âge, le genre, la langue native et le niveau d'éducation (lycéen, étudiant, doctorant, professeur ou autre).

III.1.2.3 Rôles fonctionnels

La base Mission Survival Corpus de Pianesi et al. [140] met en scène onze groupes de personnes devant réaliser une tâche de survie. À l'origine conçue par la NASA (Administration

Aéronautique et Spatiale Nationale) pour entraîner les astronautes avant les premiers atterrissages sur la lune, cette tâche de survie se démontra être un bon indicateur des processus de prises de décision des groupes [84]. Les meetings sont filmés par des caméras centrées sur chacun des quatre participants, et l'activité gestuelle de chaque participant est calculée. Plusieurs microphones permettent d'extraire l'activité vocale des participants. Une annotation des rôles fonctionnels (liés à la tâche et socio-émotionnels) est aussi fournie.

TABLE III.1.1 Bases liées aux rôles.

Nom de la base	Auteurs	Types de rôles	Type de données	Modalités	Durée
Mission survival corpus	Pianesi et al. [140]	fonctionnels	Meeting	Audio + Vidéos + Annotations	≈ 4h30
AMI	McCowan et al. [123]	informels	meeting	Audio + Vidéo + Texte	45h 38min
The Apprentice	Raducanu et al. [145, 146]	informels	Émission TV	Audio + Vidéo	1h30
ICSI	Janin et al. [94]	informels	Meeting	Audio + Texte	72h
DARPA HUB-4	Fiscus et al. [75]	formels	Radio	Audio + Texte	97h
Canergie Mellon University	Banerjee et al. [12]	formels	Meeting	Audio + Vidéo + Texte	inconnue
TDT4	Kong et al. [109, 172]	formels	Radio	Audio + Texte	300h
Radio Suisse Romande	Vinciarelli [181]	formels	Radio	Audio	18h 56min
Radio Suisse Romande	Salamin et al. [156]	formels	Radio	Audio	27h
EPAC	Estève et al. [68]	formels	Radio	Audio + Texte	100h
DARPA GALE	Linguistic Data Consortium [45]	formels	Radio	Audio + Texte	inconnue

III.1.3 Systèmes de reconnaissance de rôles

Le traitement du signal social s'intéresse à l'analyse automatique des signaux échangés durant une interaction. Une application en est la reconnaissance de rôle qui va demander d'une part à extraire les signaux pertinents et d'autre part à les modéliser.

III.1.3.1 Rôles formels

Barzilay et al. [14] utilisent la reconnaissance de rôles afin de structurer des diffusions audio. Pour cela, ils s'appuient sur une base composée de 35h de programmes radio, sous ensemble tiré de la base TREC-7 SDR [75], et considèrent la reconnaissance de trois rôles : présentateur, journaliste et invité. Des caractéristiques lexicales sont extraites, et sont ensuite résumées en n-grammes afin d'apprendre des phrases signatures caractérisant chaque rôle. Deux classifeurs sont ensuite utilisés : BoosTexter [165] et un classifieur modélisant l'entropie maximale (MaxEnt). MaxEnt a des résultats légèrement supérieurs au BoosTexter. D'autre part, les résultats sur les transcriptions audio automatiques sont très légèrement inférieurs (précision de 77%) à ceux obtenus avec les transcriptions manuelles.

Banerjee et Rudnicky [13] essayent quant à eux de modéliser à la fois les états d'un meeting (discussion / présentation / briefing) et les rôles (présentateur / participants, définis différemment selon le type de meeting). Ils s'appuient sur une base créée à l'université de Carnegie Mellon [12], sur laquelle ils extraient des caractéristiques liées aux tours de paroles dans des fenêtres temporelles limitées : nombre de changement de locuteurs, nombre de participants ayant parlé, nombre de discours superposés et temps moyen de superposition, durée totale de parole pour chaque participant et durée totale de parole pour chaque participant superposé avec un autre participant. Une fois ces informations extraites, un algorithme d'arbre de décision C4.5 [195] est utilisé sur la base annotée. Le score maximum de reconnaissance de rôle est atteint pour une fenêtre temporelle de 15 secondes avec une précision de 53%.

Liu [118] identifie lui aussi les rôles (présentateur, reporter, autre) liés à des émissions radios tirées de la base TDT4 [172], mais en madarin. Pour cela, il utilise la transcription manuelle fournie sur laquelle des caractéristiques lexicales sont extraites. Deux classifeurs, HMM et MaxEnt, sont ensuite comparés pour la reconnaissance de rôles et MaxEnt permet d'atteindre une précision de 77.4%. La combinaison des deux classifieurs permet d'augmenter la précision jusqu'à 82.97%.

Dans [181], Vinciarelli utilise deux classifieurs originaux pour le domaine. Le premier se base sur la théorie des graphes, et un de ses modèle appelé l'Analyse de Réseau Social (ARS ou SNA pour Social Network Analysis [189]) utilise une mesure de centralité. Le second se base sur l'analyse des distributions des durées des interventions, et emploie des règles bayésiennes. Ces méthodes sont appliquées sur une base de bulletins radio suisses, dans lesquels il apprend à reconnaître les rôles formels : présentateur, second présentateur, invité, interviewé, abstrait, météo. Les deux méthodes sont alors combinées, permettant ainsi d'obtenir une précision de 85.1%.

Favre et al. [69] reconnaissent à la fois des rôles formels et informels. Pour cela, ils s'appuient sur trois bases, dont les deux premières sont des émissions, et la dernière est la base AMI [123].

Ils extraient tout d'abord un Réseau d'Affiliation Sociale (RAS ou SAN pour Social Affiliation Network [189]) ainsi que des n-grammes à partir desquels un modèle de language statistique et un HMM permettent d'apprendre les rôles. Sur la première base, l'approche Bayésienne donne les meilleurs résultats (82.5%) alors que c'est la combinaison HMM+3-gramm qui est optimale pour la seconde base (86.1%). Les rôles informels de la troisième base sont quand à eux les plus difficiles à reconnaître (48.0%).

De nombreuses autres études ont été faites [156, 22, 198, 158, 157], mais ne seront pas détaillées ici. Elles apparaissent cependant dans la table récapitulative III.1.2.

TABLE III.1.2 Études liées aux rôles formels

Auteurs	Indices comportementaux (TP = Tours de Parole)	Classifieur	Score
Barzilay et al. [14]	Lexical	BoosTexter / Entropie maximum	77%
Banerjee et al. [13]	TP	Arbre de décision C4.5	53%
Liu et al. [118]	Lexical	HMM / Entropie maximum	82,97%
Vinciarelli [181]	TP	Analyse de Réseau Social (ARS) / Speaker Duration Analysis	85,1%
Favre et al. [69]	RAS + Lexical	SLM + HMM	82,5% (base 1) 86,1% (base 2) 48% (base 3)
Salamin et al. [156]	Réseau d'Affiliation Sociale (RAS) / Speaker Duration Analysis	Estimation bayesienne	97% (base 1) 98,1% (base 2)
Bigot et al. [22]	TP + Prosodie	K-means / SVM / Hierarchique	92%
Yaman et al. [198]	Lexique	Dynamic Bayesian Network	89,5%
Salamin et al. [158, 157]	TP + Prosodie	Conditionnal Random Fields	99,1% (base 1) 96,9% (base 2)

III.1.3.2 Rôles informels

Dans [192], Weng et al. analysent des films afin de reconnaître les deux personnages principaux ainsi que les communautés de personnes les entourant. Pour cela, ils s'appuient sur l'Analyse de Réseau Social (ARS) afin d'identifier les liens entre les personnages, ce qui permet ensuite de détecter les scénarios.

Garg et al. [77] s'appuient sur la base AMI [123] et tentent de reconnaître les rôles de Manager de Projet, Expert Marketing, Designer Industriel et conception de l'Interface Utilisateur.

Pour cela, ils utilisent à la fois les informations lexicales de la base apprises avec BoosTexter ainsi que l'Analyse de Réseau Social, et la combinaison des deux leur permet d'atteindre une précision de 67.9% (78% si l'extraction lexicale est manuelle).

Laskowski et al. [113] veulent reconnaître les rôles de la base AMI [123], ainsi que le statut universitaire (étudiant, doctorant ou professeur) dans la base ICSI [94]. Pour cela, les tours de paroles sont extraits et les rôles appris à partir d'un modèle d'estimation de maximum de vraisemblance. Peu de données étant utilisées en test, la précision de 53% sur les rôles d'AMI et de 67% sur le statut universitaire ne sont pas forcément significatives et comparables avec les autres études.

D'autres études ont été faites [145, 146, 164], mais ne seront pas détaillées ici. Elles apparaissent cependant dans la table III.1.3.

TABLE III.1.3 Études liées aux rôles informels.

Auteur	Indices comportementaux (TP = Tours de Parole)	Classifieur	Score
Weng et al. [192]	Social Network	Analyse de Réseau Social (ARS)	97,6% (communauté) 100% (scénario)
Garg et al. [77]	Lexical + Social Network	BoosTexter + MLE	78%
Laskowski et al. [113]	TP	MLE	53% (rôles) 67% (statut)
Salamin et al. [156]	Réseau d'Affiliation Sociale (RAS) / Speaker Duration Analysis	Estimation bayesienne	56% (base 3)
Raducanu et al. [145, 146]	TP	Mesure de centralité / Modèle d'influence	85,7% (présentateur) 92,8% (viré)
Sapru et al. [164]	TP + Prosodie + Lexical	BoosTexter	74%

III.1.3.3 Rôles fonctionnels

Zancanaro et al. [199] utilisent le Schéma de Codage des Rôles Fonctionnels (FRCS) et tentent de reconnaître à la fois les rôles liés à la tâche et les rôles socio-émotionnels. Tout d'abord, ils présentent leur base "Survival Corpus", qui est annotée pour les rôles. Ensuite, ils extraient sur des fenêtres temporelles (centrée ou alignées à gauche) les tours de parole ainsi que les tremblements des participants. Enfin un SVM est utilisé sur les fenêtres, dont on fait varier la longueur. Ainsi, pour les rôles de tâche, une fenêtre de 7 secondes suffit à battre un classement par classe majoritaire (qui lui n'apporte aucune information), et plus la fenêtre est longue, meilleur est le résultat (la précision atteint 65% pour une fenêtre de 14 secondes). Pour les rôles socio-émotionnels, la précision pour une fenêtre maximale de 14 secondes atteint 70%.

Dong et al. démontrent dans [60] que le Modèle d'Influence (MI) possède plusieurs avantages sur les SVM et les HMM en terme de reconnaissance de rôles fonctionnels, à savoir qu'il évite

le sur-apprentissage et le fléau de la dimension, tout en permettant d'assurer une meilleure robustesse et généralisation. Pour cela, ils s'appuient sur le Mission Survival Corpus [140], pour lequel les tours de parole sont calculés ainsi que le fidgeting. Les trois classifieurs SVM, HMM et MI sont ensuite entraînés, soit sur les caractéristiques d'un seul participant, soit sur les caractéristiques des quatre. Le MI se révèle alors efficace avec une précision de 75%.

Pianesi et al. présentent leur propre base Mission Survival Corpus dans [140], dans laquelle les tours de parole et l'activité du corps des participants sont extraits. Un SVM est ensuite entraîné afin de reconnaître les rôles basés sur le Schéma de Codage des Rôles Fonctionnels (FRCS). Les scores sont évalués sur plusieurs tailles de fenêtres, et atteignent 90% pour les rôles liés à la tâche et 92% pour les rôles socio-émotionnels (fenêtres de 9 secondes).

De nombreuses autres études ont été faites [185, 194, 164, 61, 163], mais ne seront pas détaillées ici. Elles apparaissent cependant dans la table III.1.4.

TABLE III.1.4 Études liées aux rôles fonctionnels.

Auteurs	Indices comportementaux (TP = Tours de Parole)	Classifieur	Score
Zancanaro et al. [199]	TP + Visuel	SVM	65% (tâche) 70% (socio-emotionnel)
Dong et al. [60]	TP + Visuel	SVM / HMM / Modèle d'Influence	75% (socio-emotionnel)
Pianesi et al. [140]	TP + Visuel	SVM	90% (tâche) 92% (socio-emotionnel)
Vinciarelli et al. [185]	TP + Prosodie	Dynamic Bayesian Network	68% (socio-emotionnel)
Wilson et al. [194]	TP + Prosodie	CRF	33,3% (socio-emotionnel)
Sapru et al. [164]	TP + Prosodie + Lexical	BoosTexter	66% (socio-emotionnel)
Dong et al. [61]	TP + Visuel	SVM / HMM / Modèle d'Influence	58% (tâche) 71% (socio-emotionnel)
Sapru et al. [163]	TP + Prosodie + Lexical	HRCF	70% (socio-emotionnel)

III.1.4 Résultats de reconnaissance de rôles informels sur la base AMI

Dans le chapitre suivant III.2, une méthode de reconnaissance de rôles informels s'appuyant sur la base AMI sera décrite. Afin de pouvoir évaluer ses performances vis à vis de l'état de

l'art, nous allons présenter ici les travaux s'appuyant sur cette même base et considérant les quatre rôles informels, à savoir :

le Manager de Projet
l'Expert Marketing
l'Expert d'Interface Utilisateur
le Designer Industriel

La même personne, apparaissant dans plusieurs réunions, peut jouer des rôles différents selon les réunions, mais les rôles sont fixés à l'intérieur de chaque réunion.

Cette base a été choisie pour le challenge qu'elle représente. Tout d'abord, les rôles étant informels, ils correspondent à une position dans un système social donné et ne sont pas nécessairement associés à des modèles de comportement stables comme dans le cas de rôles formels (comme par exemple les rôles liés à des émissions télévisées). De plus, les participants jouent un rôle qui ne correspond pas à leur métier réel, dans une réunion scénarisée, et interprètent donc un rôle ce qui complexifie encore la tâche de reconnaissance. Enfin, bien que le rôle de manager de projet se rapproche très fortement d'un rôle formel d'un vrai manager (une personne dominante qui articule le débat), Jayagopi et al. ont démontré dans [96] que le manager de projet n'est la personne dominant le meeting que dans 65% des cas, ce qui complexifie la reconnaissance de ce rôle.

Ainsi, comme nous allons le voir, deux grandes catégories d'études existent, suivant si elles utilisent le lexique comme caractéristique d'entrée ou non. Bien que le modèle que nous présenterons dans le chapitre suivant III.2 n'utilise pas le lexique, les deux types d'approches sont présentées ici.

III.1.4.1 Modèles n'utilisant pas l'information lexicale

Dans [113], Laskowski et al. extraient les tours de paroles pour ensuite en déduire un nombre varié d'indices afin de caractériser les tours de parole :

la probabilité qu'une personne parle à un instant t alors que personne ne parlait à l'instant t-1
la probabilité qu'une personne continue à parler à un instant t alors que personne d'autre que lui ne parlait à l'instant t-1
la probabilité qu'une personne i commence à parler à l'instant t alors que la personne j parlait à l'instant t-1
la probabilité qu'une personne i continue de parler à l'instant t alors que la personne j parlait à l'instant t-1
la probabilité qu'une personne parle à un instant t
les moyennes des indices précédents pour chaque participant

Ainsi, pour la base AMI, ils obtiennent 28 indices caractérisant un seul participant et 24 indices caractérisant les participants deux à deux. Chaque indice est ensuite modélisé par un modèle gaussien. Ils utilisent enfin un modèle d'estimation de maximum de vraisemblance afin de reconnaître les rôles. Peu de données étant utilisées en test (20 réunions sur les 138 disponibles), la précision de 52.5% sur les rôles d'AMI n'est pas forcément significative et comparable avec les autres études présentées ici.

Laskowski et al. essayent ensuite de ne reconnaître que le manager de projet, et obtiennent un résultat de 75% pour cette tâche (toujours évaluée sur 20 réunions).

Dans [70], Favre et al. segmentent tout d'abord les réunions en tours de paroles. Une Analyse de Réseau Social [189] permet de capturer les caractéristiques des interactions entre les rôles. Deux types de distributions sont ensuite utilisées dans la modélisation : Bernoulli ou multinomiales. Un simple classifieur bayésien permet ensuite de reconnaître les rôles. Le score maximal obtenu est de 43,6% pour les distributions de Bernoulli.

Dans [69], Favre et al. extraient tout d'abord les tours de parole. Une Analyse de Réseau Social [189] est suivie d'une Analyse en Composante Principale [24] afin de réduire la dimension du vecteur de description. Un Modèle de Markov Caché permet d'estimer la séquence des rôles ayant parlé, la probabilité *a priori* des rôles étant elle fournie par un modèle n-gram [154]. Trois longueurs de n-gram sont testées (1, 2 et 3). Un apprentissage par un simple modèle de Bayes permet de comparer les résultats des systèmes proposés. Le meilleur résultat de 48% est obtenu pour le 2-gram.

Dans [156], Salamin et al. extraient tout d'abord les tours de parole des participants. Un premier système, basé sur l'Analyse de Réseau Social [189], permet de comprendre les interactions entre les différents rôles au cours des réunions. Le second système permet de modéliser les durées de parole de chaque rôle par des distributions gaussiennes. Enfin les deux systèmes sont combinés afin de reconnaître les rôles, par une modélisation des rôles soit de manière indépendante, soit de manière dépendante. La modélisation s'appuie sur deux types de distribution possible : Bernoulli ou multinomiale [24]. Toutes les combinaisons de distribution et dépendance des rôles sont testées, avec des tours de parole extraits automatiquement ou manuellement. Le meilleur résultat 56% est obtenu pour les annotations manuelles, avec les rôles considérés comme dépendants, et une modélisation par des distributions de Bernoulli. Les résultats sont détaillés dans la table III.1.5.

Dans [49], Cristani et al. construisent à partir des tours de paroles des périodes de conversation stable (PCS). Ces dernières sont construites à partir des périodes continues, que ce soit en terme de silence ou de parole. Ceci permet de mettre en avant la dynamique de la conversation. A partir de ces PCS, un modèle de mixtures de gaussiennes [65] est construit, suivi d'un modèle d'influence observé [15]. Afin d'appliquer cette méthodologie à la reconnaissance de rôles, un modèle est appris pour chaque correspondance possible entre les 4 personnes et les 4 rôles (soit 24 modèles possibles). Lors de la reconnaissance sur un exemple inconnu, le modèle le plus probable (au sens du maximum de vraisemblance) est choisi. Un taux de reconnaissance de 58,75% est obtenu sur 20 exemples de test.

III.1.4.2 Modèles utilisant l'information lexicale

Dans [77], Garg et al. utilisent tout d'abord un système de reconnaissance du locuteur (97% de classification correcte) ainsi qu'une reconnaissance automatique de parole (taux d'erreur sur les mots d'environ 40%) afin d'obtenir la transcription des réunions de la base. A partir de là, deux systèmes sont mis en place. Le premier système, une Analyse de Réseau Social [189] se base sur les tours de parole et une modélisation par des distributions de Bernouilli discrètes permet ensuite de reconnaître les rôles grâce à une estimation par maximum de vraisemblance.

Le second système utilise le lexique et le célèbre classifieur BoosTexter [165]. Les deux systèmes sont testés séparément puis conjointement sur l'extraction automatique des transcriptions, ainsi que sur les données annotées fournies avec la base. Les différentes combinaisons sont détaillées dans le tableau final III.1.5, mais sans lexique le meilleur score global annoncé est de 49.5%, contre 78% pour la combinaison des deux systèmes.

Dans [96], Jayagopi et al. extraient un grand nombre d'indices variés afin de caractériser l'interaction :

vocaux : liés à une seule personne (énergie, temps de parole, nombre de tours de parole), ou liés à plusieurs personnes (nombre d'interruptions réussies ou non, nombre de prise de parole après une autre personne) ;

visuels : activité (longueur, tours, interruptions) et attention (reçue, donnée).

Chacun des indices est utilisé pour prédire le manager de projet. Les deux meilleurs résultats sont obtenus pour deux indices liés aux tours de paroles : le nombre de fois où un participant parle après un autre participant (66,7%) et le nombre de tours de paroles (63,2%). Une mesure de centralité [27] basée sur le nombre de fois où un participant i parle après un des autres permet d'améliorer les résultats jusqu'à 68,4%.

III.1.5 Discussions et orientations

Cet état de l'art a soulevé plusieurs grandes tendances du domaine du traitement du signal social.

Tout d'abord, les travaux du domaine suivent les grandes approches décrites par la socio-psychologie, à savoir la catégorisation des rôles en trois grandes approches : les rôles formels, informels et fonctionnels. Les rôles formels sont les rôles les plus simples à reconnaître, et un très grand nombre d'études ont été faite sur le choix du classifieur afin d'améliorer les résultats de reconnaissance. Plusieurs modalités sont utilisées en combinaison pour améliorer les résultats. Les rôles fonctionnels sont les seconds plus étudiés, là aussi grâce à une reconnaissance facilité (le lien entre les caractéristiques du rôle et le rôle étant très fort dans cette approche). Ici encore, les efforts de la littérature portent principalement sur le choix du classifieur, avec plusieurs combinaisons des modalités. Les rôles informels sont eux les plus difficiles à reconnaître, car ils ne correspondent pas forcément à un lien direct avec des caractéristiques.

Il est notable que peu de travaux portent leurs efforts sur l'amélioration des caractéristiques extraites, alors même que ce sont ces caractéristiques qui permettront au classifieur de relier plus ou moins facilement les participants aux rôles.

Ainsi, nos travaux porteront sur les rôles informels de la base AMI, en se focalisant sur l'amélioration de l'extraction de caractéristiques, sans utilisation du lexique.

TABLE III.1.5 Comparaison des résultats pour la base AMI.

Comparaison des résultats pour la base AMI, avec et sans utilisation du lexique. "B" signifie Bernouilli, "M" Multinomiale, "I" indépendants, "D" dépendants et "ARS" est utilisé pour Analyse de Réseau Social.

	Global	PM	ME	UI	ID	Moyenne des 4 rôles
Laskowski et al. [113] (20 exemples)	52.5%	60%	40%	70%	40%	52.5%
Favre et al. [70] Bernoulli	43,6%	79,4%	19,5%	33%	13%	36,2%
Multinomiale	42,8%	76,4 %	14,4%	30,2%	22,5%	35,9%
Favre et al. [69] HMM	44,4%	63,4%	28,4%	40,1%	22,6%	38,6%
HMM + 1-gram	40,7%	70,6 %	16,1%	32%	12,4 %	32,8%
HMM + 2-gram	48%	63,3%	35,8%	34,6%	28,3 %	40,5%
HMM + 3-gram	46,9%	61,8 %	39,4%	33,8%	25%	40%
Bayes	43,5%	75,3%	15,1 %	40%	15,1 %	36,4%
Salamin et al. [156] B,I (auto.)	46%	79,6 %	13,1 %	41,4%	20,3 %	38,6%
B,D (auto.)	46,4 %	68,7%	26%	32,9%	25,7 %	38,3%
M,I (auto.)	39,3%	67,4 %	18%	19,3%	25,6 %	32,6%
M,D (auto.)	47,3%	67,4 %	28,7%	22%	24,3%	35,6%
B,I (man.)	51,2 %	83,3%	15,9 %	42%	29%	42,6%
B,D (man.)	56%	76,1 %	37,7%	40,6 %	41,3%	48,9%
M,I (man.)	43,7 %	67,4%	17,4%	39,1 %	21,7%	36,4%
M,D (man.)	52,6 %	76,8%	29%	34,1 %	33,3%	43,3%
Garg et al. [77] ARS (auto)	43,1 %	75,7%	16,4%	41,2 %	13,4%	36,7%
ARS (man.)	49,5 %	79%	20,3%	44,9%	24,6 %	42,2%
Cristani et al. [49] (20 exemples)	58,75%	85 %	60%	50%	40%	58,75%
Garg et al. [77] lex. (auto.)	67,1 %	78,3%	71,9%	38,1%	53%	60,3%
ARS + lex. (auto.)	67,9 %	84%	69,8%	38,1 %	50,1%	60,5%
lex (man.)	76,7 %	92%	70,3 %	60,1%	60,9 %	70,825%
ARS + lex. (man.)	78%	95,7%	68,8%	60,1 %	61,6%	71,6%

Chapitre III.2

Reconnaissance de rôles

Sommaire

III.2.1 Introduction

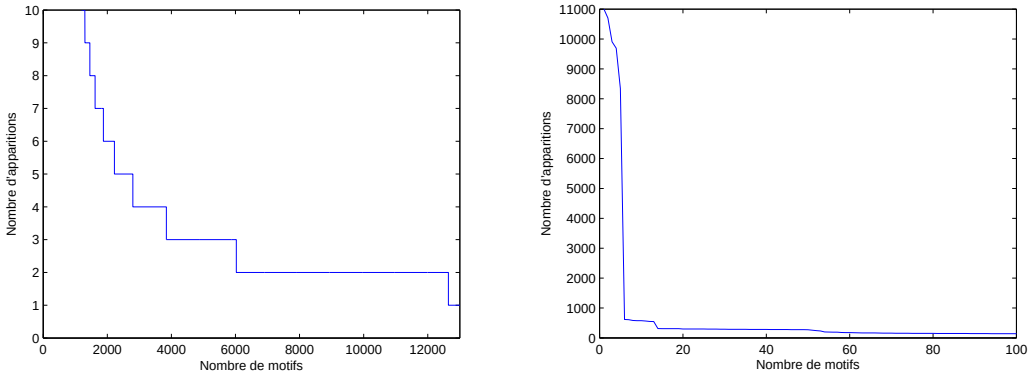
Dans ce chapitre, nous allons nous intéresser à la reconnaissance de rôles qui est fondamentale pour la compréhension des interactions qui prennent place lors d'un meeting. Pour cela, nous allons nous appuyer sur la base AMI présentée dans le chapitre précédent, qui se compose de quatre rôles prenant place dans 138 meetings : le Manager de Pojet (PM), l'Expert Marketing (ME), le concepteur de l'Interface Utilisateur (UI) et le Designer Industriel (ID).

Notre méthode se démarque des approches usuelles par l'aspect non supervisé que nous introduirons dans la partie III.2.2 à travers les patterns interactifs et personnels, qui serviront à coder l'interaction. Un SVM avec une architecture bien choisie permettra ensuite de reconnaître les rôles de la base dans la section III.2.3.

III.2.2 Découvrir des motifs d'interaction de manière non supervisée

Nous avons vu dans le chapitre II.2 qu'une factorisation non supervisée de différents indices pouvait permettre de faire apparaître des types d'interactions qui ont un sens interprétable. Cependant, les indices utilisés au départ et présentés section II.2.3 étaient le fruit d'une expertise humaine et étaient en nombre limité. Afin d'augmenter le nombre d'éléments pour représenter

FIGURE III.2.2.1 Définition des motifs élémentaires.



(a) Tracé pour les 13000 premiers motifs.

(b) Tracé pour les 100 motifs les plus présents.

FIGURE III.2.2.2 Nombre d'apparition des 125 280 motifs uniques. Triés par motifs les plus fréquents.

les interactions et d'arriver à des indices réellement pertinents et discriminants, une méthode non supervisée s'appuyant sur les données brutes est mise en place.

III.2.2.1 Extraction naïve des motifs

Dans la suite, nous appellerons motif un patch extrait directement des tours de parole entre les partenaires interactifs. Ce motif est caractérisé par deux éléments : le nombre de participants, et la durée, comme illustré sur la figure III.2.2.1.

Si l'on considère des motifs de 2,5 secondes (10 échantillons) avec 4 participants, il existe $2^{4*10} = 2^{40} \approx 10^{12}$ motifs possibles.

Or, même dans une base aussi vaste que la base AMI, certains motifs n'apparaissent jamais. Les différents motifs existant ont donc été répertoriés : ils se limitent à 125 280 motifs. Cependant, la fréquence d'apparition de ces motifs est disparate. Ainsi, comme on peut le voir sur la figure III.2.2.2a qui trace le nombre d'apparition des motifs, alors qu'environ 12 643 motifs apparaissent au moins 2 fois dans la base, seuls environ 1041 apparaissent plus de 13 fois (soit dans presque 10% des vidéos).

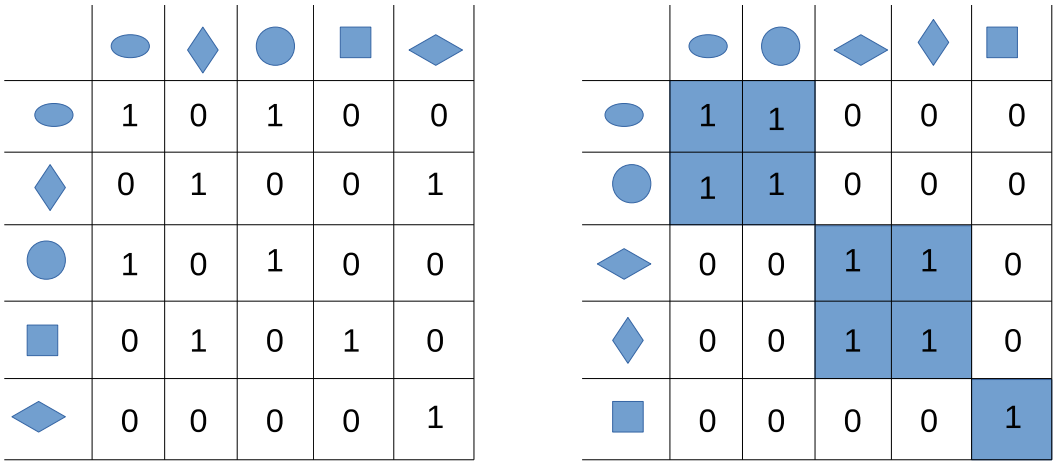
La figure III.2.2.3 permet de visualiser les 100 motifs les plus fréquents dans la base. Plusieurs remarques sautent aux yeux. Tout d'abord, dans ces motifs, il y a peu de motifs d'interaction comme par exemple un élément tel que "la personne A interrompt la personne B". D'autre

part, un grand nombre de motifs se ressemblent, étant soit différent d'un seul élément (deux premiers motifs de la seconde ligne), soit étant simplement décalés dans le temps (motifs 2 et 3 de la premières ligne).

Il n'est évidemment pas possible de garder les 125 280 motifs. Cependant, garder seulement les 100 ou 200 premiers motifs les plus fréquents n'est pas non plus souhaitable car ils sont répétitifs et n'apportent pas assez de diversité d'information pour décrire les interactions (voir la figure III.2.2.3).

Il est donc nécessaire de mettre en place une méthode permettant, à partir des motifs présents dans la base, d'apprendre des motifs d'intérêt se démarquant des autres, et permettant au mieux de représenter la diversité des échanges dans les interactions. Ces représentants pourront ensuite être utilisés pour décrire les interactions dans une application de reconnaissance de rôles (voir la section III.2.3).

FIGURE III.2.2.3 Aperçu des 100 motifs les plus présents.



(a) Calcul de la ressemblance entre les motifs.

(b) Rassemblement des motifs en catégories.

FIGURE III.2.2.4 Principe du rassemblement des motifs en catégories.

III.2.2.2 Extraction de motifs pertinents

Étant donné qu'un grand nombre de motifs se ressemblent, nous les rassemblons en catégories contenant des motifs semblables (différence minime entre les motifs ou décalage temporel léger entre les motifs).

III.2.2.2.a Principe

Afin de trouver des motifs discriminants, nous allons mesurer la similarité entre les motifs existants, comme le montre l'exemple de la figure III.2.2.4a. Une fois la matrice de similarité entre tous les motifs calculés, on va rassembler ceux ci dans des catégories comme illustré dans la figure III.2.2.4b.

La réalisation de cette méthode suppose de calculer une matrice de similarité de n-par-n, où n est le nombre total de motifs. Or il existe plus 125 000 motifs ! Le calcul d'une telle matrice nécessite $\frac{k^2-k}{2}$ calculs (seul le triangle supérieur de la matrice doit être calculé, car elle est symétrique et tous ses éléments diagonaux sont égaux à 1), et est impossible pour 125 000 motifs. La question du choix des motifs se pose donc de nouveau. Comme nous l'avons vu précédemment, ne garder que les motifs les plus fréquents ne permettrait pas une bonne représentation des données. Afin de garder un ensemble représentatif tout en réduisant le nombre de motif, seuls les motifs apparaissant au moins 3 fois ont été gardés pour le calcul de la matrice de similarité, soit 5000 motifs. Ce choix sera évalué plus tard dans la section III.2.2.2.e.

Une fois ces motifs sélectionnés, il faut alors calculer la matrice de similarité, puis rassembler les motifs en groupes d'équivalence. Pour cela, un éventail d'algorithmes non supervisés existe, et sera présenté dans la section III.2.2.2.c, juste après l'étude du codage utilisé.

(a) Motif A. (b) Motif B. (c) Motif C.

FIGURE III.2.2.5 Comparaison de deux types de codages. Codage [0 ;1] : Corr(A,B)=6 et Corr(B,C)=5. Alors qu'en codage [-1 ;1] : Corr(A,B)=10 et Corr(B,C)=22

III.2.2.2.b Choix du codage des motifs

Les motifs sont des matrices de taille nombres de personnes par longueur temporelle du motif, et chaque case (i,j) représente si la personne i parle au temps j.

Afin de calculer la matrice de similarité entre les motifs, le calcul de corrélation-croisée (équation I.2.4) de la partie I.2.4.1 a été utilisé, rappelé ci-dessous dans l'équation III.2.1, où M_A et M_B représentent deux motifs, L la largeur des motifs et n le délai entre les motifs.

$$[M_A \star M_B]_n = \sum_{k=0}^{L} \frac{M_A(k)^T . M_B(k+n)}{||M_A(k)||_2 * ||M_B(k+n)||_2} \tag{III.2.1}$$

Ici le délai est limité à la moitié de la taille des motifs (à la fois en positif et en négatif), afin de garder un minimum de ressemblance entre les motifs.

Cependant, afin de ne pas favoriser le score de corrélation entre deux motifs qui se ressemblent peu mais où les personnes parlent beaucoup par rapport à deux motifs qui se ressemblent mais où les personnes parlent peu, il est nécessaire de changer de codage. En effet, si l'on considère les exemples de la figure III.2.2.5, on voit que la corrélation entre les motifs A et B codés par "0 = noir = ne parle pas" et "1 = blanc = parle" est de 6 . Or si on prend les motifs B et C codés par le même codage, on obtient une corrélation égale à 5, plus faible alors que les motifs B et C sont beaucoup plus ressemblants que A et B. Cependant, si le codage choisi est "-1 = ne parle pas" et "1 = parle", alors la nouvelle corrélation pour A et B est de 10, et de 22 pour B et C. Cette nouvelle mesure correspond mieux à ce qui est souhaité. En effet, le codage [-1 ;1] pénalise les différences entre les motifs, alors que le codage [0,1] les ignore. Le codage [-1 ;1] sera donc utilisé dans la suite.

III.2.2.2.c Rassemblement des motifs en groupe d'équivalence.

Une fois la matrice de similarité calculée, plusieurs méthodes existent afin de rassembler les motifs proches. Parmi elles, les dendrogrammes, les k-moyennes ou les médoïdes peuvent être cités. Ces méthodes ont été écartées pour plusieurs raisons. Tout d'abord les données sont de taille non négligeable (matrice de 5000 par 5000), ce qui exclue les dendrogrammes. Ensuite, il est souhaitable que les groupes obtenus ne soient pas équiprobables (certains groupes peuvent ne contenir que quelques motifs tandis que d'autres peuvent en contenir plusieurs centaines),

excluant ainsi les k-moyennes qui ont tendance à créer des groupes de même taille. Enfin, le représentant obtenu doit être non binaire, ce qui exclue les médoïdes dont le représentant est l'élément le plus central de chaque groupe.

La méthode qui va être présentée ci-dessous est une méthode spectrale qui s'appuie sur le papier présenté par Weiss [191], qui combine les techniques de Scott et Malik [168] et Shi et Longuet-Higgins [169]. Elle a l'avantage de pouvoir gérer un très grand nombre de données, et de créer des groupes d'équivalence dont la seule contrainte est la similarité entre les éléments, permettant ainsi d'avoir des groupes de taille très hétérogène.

Tout d'abord, la matrice de corrélation (ou matrice d'affinité) A est calculée entre les différents motifs (figure III.2.2.6a). On obtient ainsi une matrice de taille 5000 par 5000. De là est extraite la matrice D de degré de A (elle aussi de taille 5000 par 5000, étape 2). La matrice d'affinité normalisée N (taille 5000 par 5000) est ensuite calculée en suivant l'étape 3. Puis les valeurs et vecteurs propres de celle ci sont extraits et la matrice V est construite à partir des k premiers vecteurs propres (matrice de taille 5000 par k, étape 4), et elle est ensuite normalisée (étape 5). Enfin, la matrice Q (taille 5000 par 5000) est calculée suivant l'étape 6 (figure III.2.2.6b), puis rendue binaire par seuillage pour donner l'appartenance de chaque motif à un groupe d'équivalence (étape 7, figure III.2.2.6c).

La méthode est résumée par les étapes suivantes :

1. Calcul de la matrice d'affinité A
2. $D = diag(sum(A, 2))$
3. $N = D^{-1/2} * A * D^{-1/2}$
4. Calcul des k premiers vecteurs propres de N, placés dans V
5. Normalisation de V
6. $Q = V * V^T$
7. $Q_{binaire} = Q(Q > Q_{seuil})$

Une fois les groupes d'équivalence calculés, il est possible de réorganiser les matrices A, Q et $Q_{binaire}$ selon les groupes d'équivalence afin de voir les blocs d'équivalence effectués (figures III.2.2.7a,III.2.2.7b et III.2.2.7c).

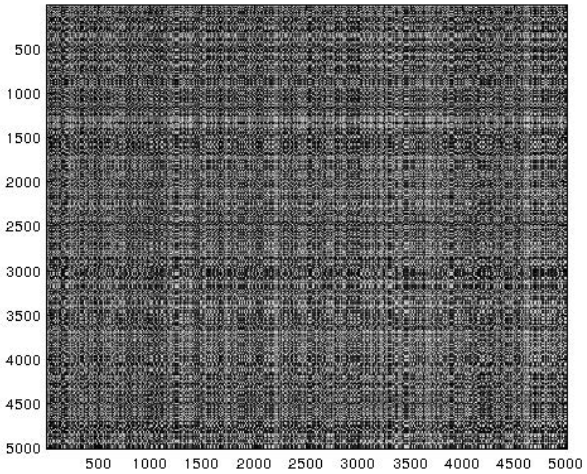
L'observation des groupes d'équivalence montre que la méthode a été capable de rassembler dans des groupes d'équivalence des motifs se ressemblant (figure III.2.2.8). Ainsi, avec les paramètres Q_{seuil}=0.6 et k=50 choisis de manière ad hoc, 465 groupes d'équivalences sont obtenus. Des prototypes de chaque groupe peuvent alors être calculés en recalant les motifs dans le temps et en en calculant la moyenne, ce que nous verrons un peu plus tard (voir figure III.2.2.14).

Les nombres k et Q_{seuil} ont été choisis de manière ad-hoc. Nous allons donc maintenant étudier leur influence sur le nombre de groupes d'équivalences obtenus, et sur leur qualité.

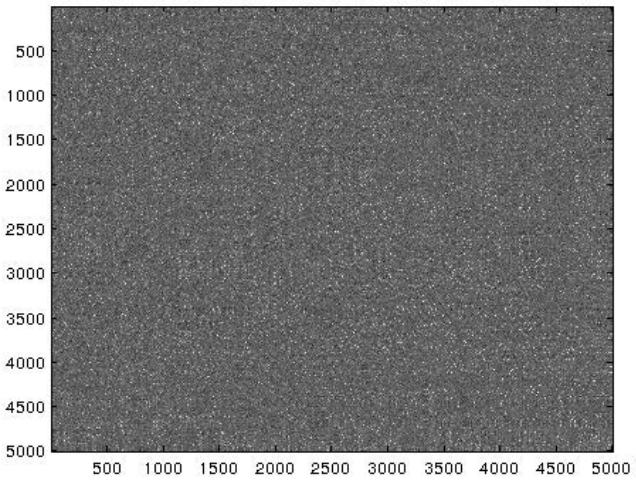
III.2.2.2.d Influence des paramètres de la méthode

Le nombre de groupes d'équivalence obtenus dépend de deux paramètres : le nombre de valeurs propres à conserver k et le seuil Q_{seuil}.

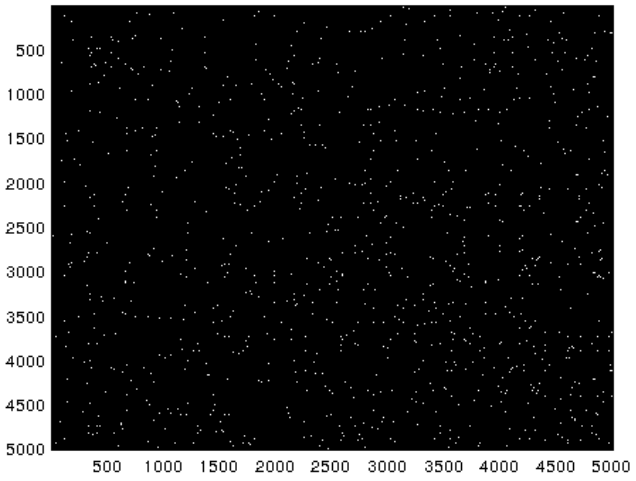
Nous avons donc tracé les valeurs absolues des valeurs propres (figure III.2.2.9a) ainsi que l'inertie cumulée des valeurs propres (figure III.2.2.9b).



(a) Matrice d'affinité A (corrélation entre les motifs).

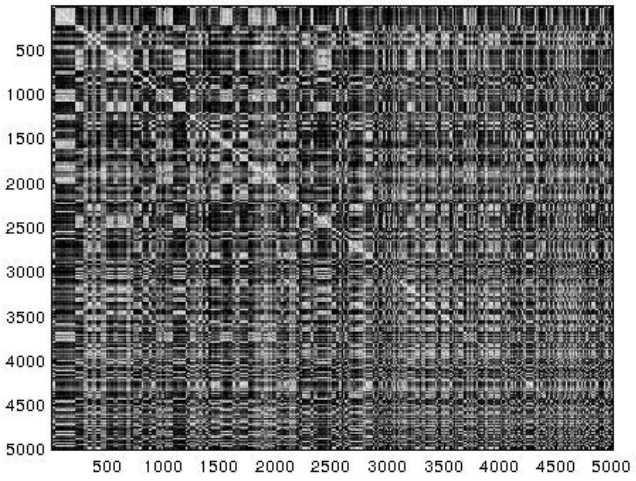


(b) Matrice Q d'appartenance

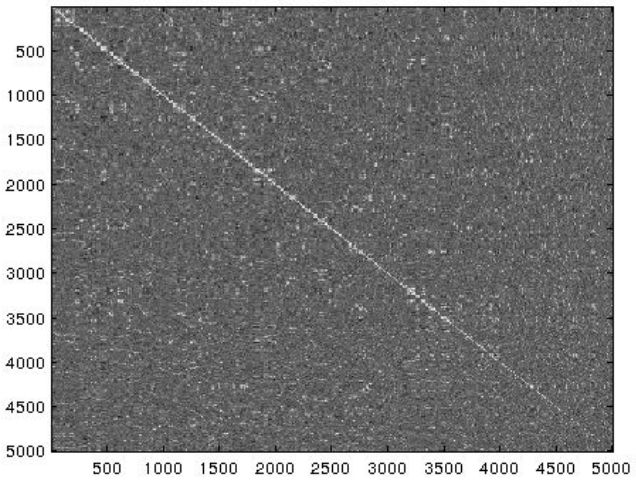


(c) Matrice $Q_{binaire}$.

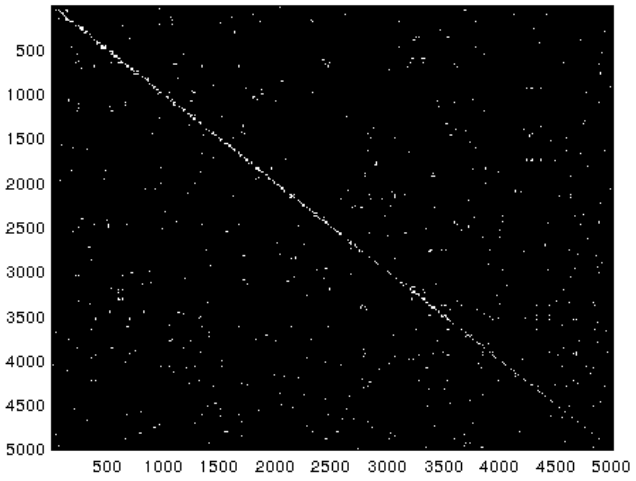
FIGURE III.2.2.6 Matrices A, Q et $Q_{binaire}$ obtenues lors de l'application de la méthode.



(a) Matrice d'affinité A réorganisée par groupe d'équivalence.

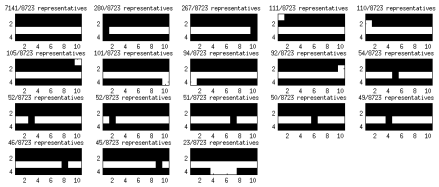


(b) Matrice Q d'appartenance réorganisée par groupe d'équivalence.

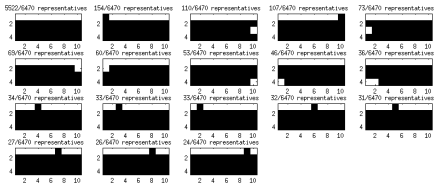


(c) Matrice $Q_{binaire}$ d'appartenance réorganisée par groupe d'équiva lence.

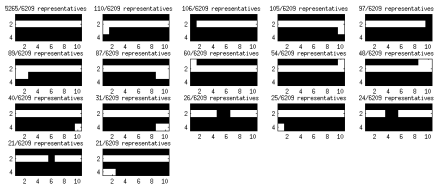
FIGURE III.2.2.7 Matrices A, Q et $Q_{binaire}$ réorganisées par groupe d'équivalence.



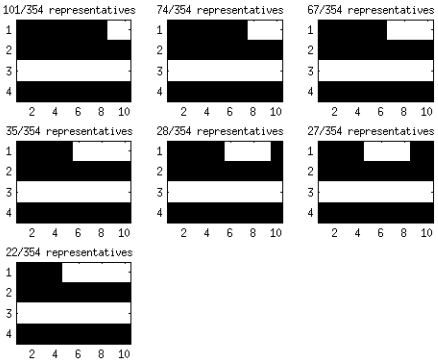
(a) Groupe d'équivalence 11.



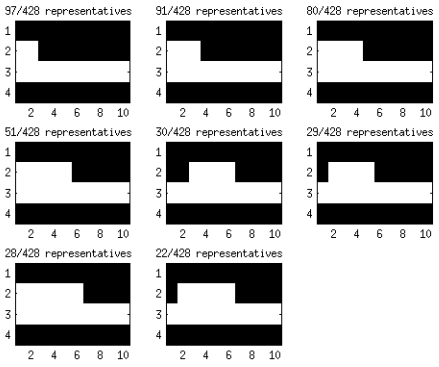
(b) Groupe d'équivalence 12.



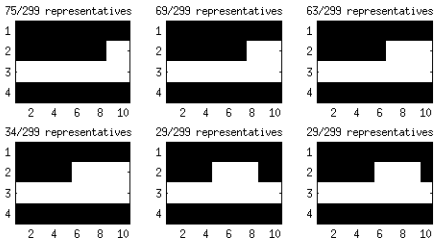
(c) Groupe d'équivalence 13.



(d) Groupe d'équivalence 24.

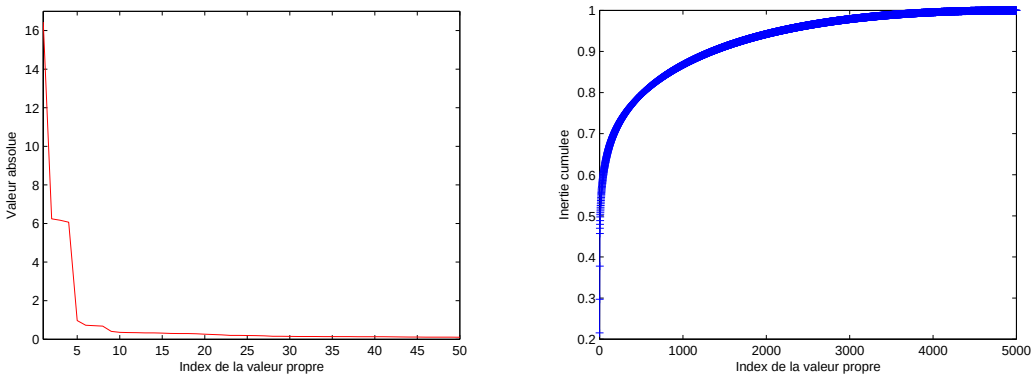


(e) Groupe d'équivalence 25.



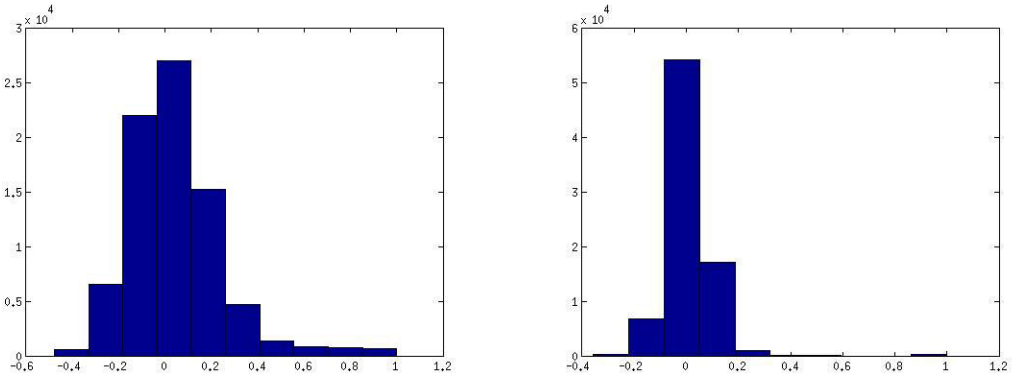
(f) Groupe d'équivalence 36.

FIGURE III.2.2.8 Six exemples de groupes d'équivalence obtenus par la méthode. Chaque figure présente les motifs composant le groupe.



(a) Valeurs absolue des 50 premières valeurs propres. (b) Inertie cumulée des valeurs propres.

FIGURE III.2.2.9 Valeurs propres, et inertie cumulée.



(a) Cas k=24. (b) Cas k=117.

FIGURE III.2.2.10 Comparaison de la répartition des valeurs de Q pour k=24 et k=117.

Ces deux premières figures permettent de se rendre compte que beaucoup de valeurs propres ont de très faibles valeurs. Par exemple, la 300^{eme} valeur propre vaut 0.0228. L'étude de la figure III.2.2.9b permet de se rendre compte que si l'on devait sélectionner k en appliquant le critère de Jolliffe [102] (c'est à dire prendre k tel que 95% de l'inertie cumulée totale soit conservée), cela nous mènerait à devoir prendre k=2161 ! Un autre critère possible est celui de Cattell [37], qui considère que l'on ne doit prendre en compte que les valeurs propres avant l'apparition du "coude". Ce critère nous mènerait à k=28, ce qui est une valeur faible. En effet, si l'on regarde l'influence qu'a le paramètre k sur les valeurs contenues dans la matrice Q, on peut voir sur la figure III.2.2.10 que plus k est petit, moins les groupes d'équivalences seront bien définis, un grand nombre de valeurs se situant entre 0,6 et 0,8 ce qui signifie que le seuil Q_{seuil} pourra potentiellement séparer des éléments qui auraient du être ensemble. Au contraire, plus k sera grand plus il y aura de groupes d'équivalences au final, et plus des motifs qui auraient du se retrouver ensemble se retrouvent séparés.

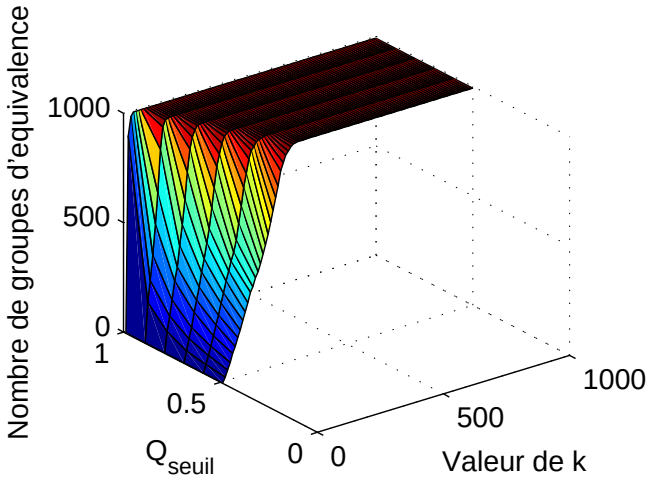


FIGURE III.2.2.11 Tracé de la variation du nombre de groupes d'équivalence en fonction de k et Q_{seuil}.

Ainsi il est nécessaire de choisir k et Q_{seuil} afin d'avoir un nombre raisonnable de groupes d'équivalences. La figure III.2.2.11 montre l'effet des valeurs des paramètres k et Q_{seuil} sur le nombre de groupes d'équivalences qui sera créé.

Une méthode de sélection de k et Q_{seuil} est de s'appuyer sur l'indice silhouette [155]. La valeur de la silhouette pour chacun des 5000 motifs initiaux est une mesure de la similarité de ce motif aux motifs de son propre groupe comparé à la similarité du motif avec les motifs des autres groupes. Cette valeur se situe entre -1 et +1, où -1 signifie que le motif est très mal classé alors que +1 signifie que le motif est très cohérent dans son groupe, et très différent des autres groupes. La valeur est calculée selon la formule III.2.2, où a_i est la distance moyenne du motif i aux autres motifs de son cluster, et b_i et la distance moyenne du motif i aux motifs d'un autre cluster, minimisé sur tous les clusters.

$$S_i = \frac{b_i - a_i}{max(a_i, b_i)} \tag{III.2.2}$$

Plusieurs valeurs de Q_{seuil} (de 0.3 à 0.95) sont alors testées pour k fixé à 5. On obtient alors le résultat présenté dans la figure III.2.2.12 : pour chaque valeur de Q_{seuil}, une boite à moustache montre la répartition des 5000 valeurs de silhouettes obtenues. La valeur optimale obtenue est alors $Q_{seuil} = 0.55$.

Avec les paramètres $k = 5$ et $Q_{seuil} = 0.55$, on obtient alors seulement 8 groupes d'équivalences ! Les représentants de ces groupes sont tracés sur la figure III.2.2.13. On voit que ces représentants ne vont pas permettre d'avoir une représentation riche de l'interaction.

Le paramètre k est alors fixé de manière ad hoc à $k = 50$ (avec $Q_{seuil} = 0.55$), après analyse du nombre de groupes, de leur cohérence, et de leur richesse de représentation. On obtient alors 465 groupes d'équivalences.

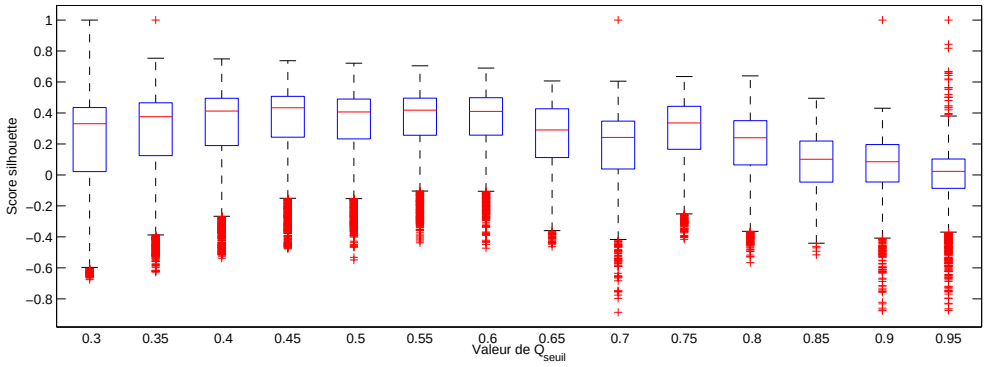


FIGURE III.2.2.12 Boite à moustache des scores silhouette pour Q_{seuil} avec $k = 5$.

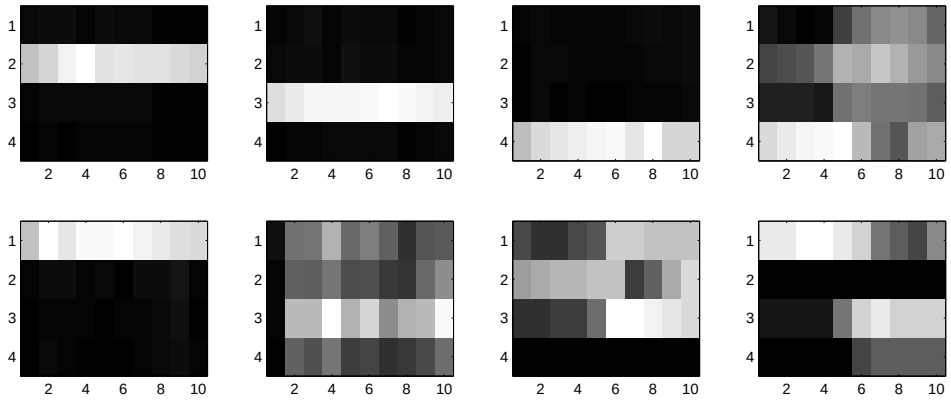


FIGURE III.2.2.13 Représentants des 8 groupes d'équivalence obtenus avec les paramètres $k = 5$ et $Q_{seuil} = 0.55$.

TABLE III.2.1 Comparaison de la répartition des motifs et des groupes d'équivalence dans les catégories.

	Catégorie 0	Catégorie 1	Catégorie 2	Catégorie 3	Catégorie 4
Nombre de motifs	1	2575	39657	55154	27893
Pourcentage des motifs	0%	2%	31.6%	44%	22.2%
Nombre de groupe d'équivalence	1	30	225	126	84
Pourcentage des groupes d'équivalence	0.2%	6.4%	47.9%	26.8%	17.9%

III.2.2.2.e Analyse des groupes d'équivalence obtenus

La figure III.2.2.14 présente les représentants des 100 premiers groupes d'équivalence. Beaucoup de ces groupes concernent les mêmes participants, et la plupart de ces groupes concernent une à deux personnes maximum.

Sachant que l'on souhaite caractériser l'interaction entre personnes, il serait préférable de conserver des groupes relatifs aux interactions, même s'ils correspondent à des motifs très peu fréquents. La table III.2.1 étudie quelque statistiques en fonction du nombre de personnes impliquées dans l'interaction. Ainsi, on appellera Catégorie 1 les motifs où une seule personne intervient, Catégorie 2 lorsqu'il y a deux personnes, Catégorie 3 pour trois personnes et Catégorie 4 lorsque les 4 participants ont parlé dans le motif. La Catégorie 0 concerne le motif unique où personne ne parle. Le tableau III.2.1 résume les pourcentages de motifs dans chaque catégorie, ainsi que le pourcentage de groupe d'équivalences. On voit à travers ce tableau que certaines Catégories sont sous représentées (catégorie 3) alors que d'autres sont sur représentées (catégories 1 et 2), ce que nous laissait présager les motifs de la figure III.2.2.14.

Cette première étude statistique permet de voir que les motifs impliquant 3 ou 4 personnes sont sous-représentés. Mais au delà de cette sous-représentation, il est intéressant de voir si les groupes représentant ces catégories les représentent bien. Pour cela, pour chaque motif la distribution du taux de parole est calculée. Par exemple, le motif où seul P1 parle sera caractérisé par le vecteur [1 ;0 ;0 ;0]. Si P2 et P3 parlent 1/4 et 3/4 du temps respectivement dans un autre motif, celui-ci sera caractérisé par la distribution [0 ;0.25 ;0.75 ;0]. La figure III.2.2.15 représente un graphe tri-dimensionnel avec pour premier axe le pourcentage de la personne qui parle le plus, deuxième (resp. troisième) axe celui qui a le deuxième (resp. troisième) taux le plus important.

Ainsi, la Catégorie 1, qui contient seulement une personne qui parle, se retrouve forcément comme un point situé en [1,0,0] sur le graphe (point rouge sur la figure III.2.2.15). La Catégorie 2 quand à elle possède deux personnes qui parlent. Ils peuvent aller de parler à part égale (point [0.5,0.5,0]), jusqu'à avoir un quasi monologue d'une personne (point [1-ϵ,ϵ,0]). Il s'agit donc d'une droite continue (points bleus sur la figure III.2.2.15). La catégorie 3 possède trois personnes qui parlent. Ici encore un grand nombre de cas est possible, mais la complexité est plus grande car les trois peuvent parler. Ainsi si ils parlent à parts égales on se trouve au point [0.33,0.33,0.33], si un seul parle en quasi monologue on est au point [1-ϵ,$\frac{\epsilon}{2}$,$\frac{\epsilon}{2}$] et si les deux

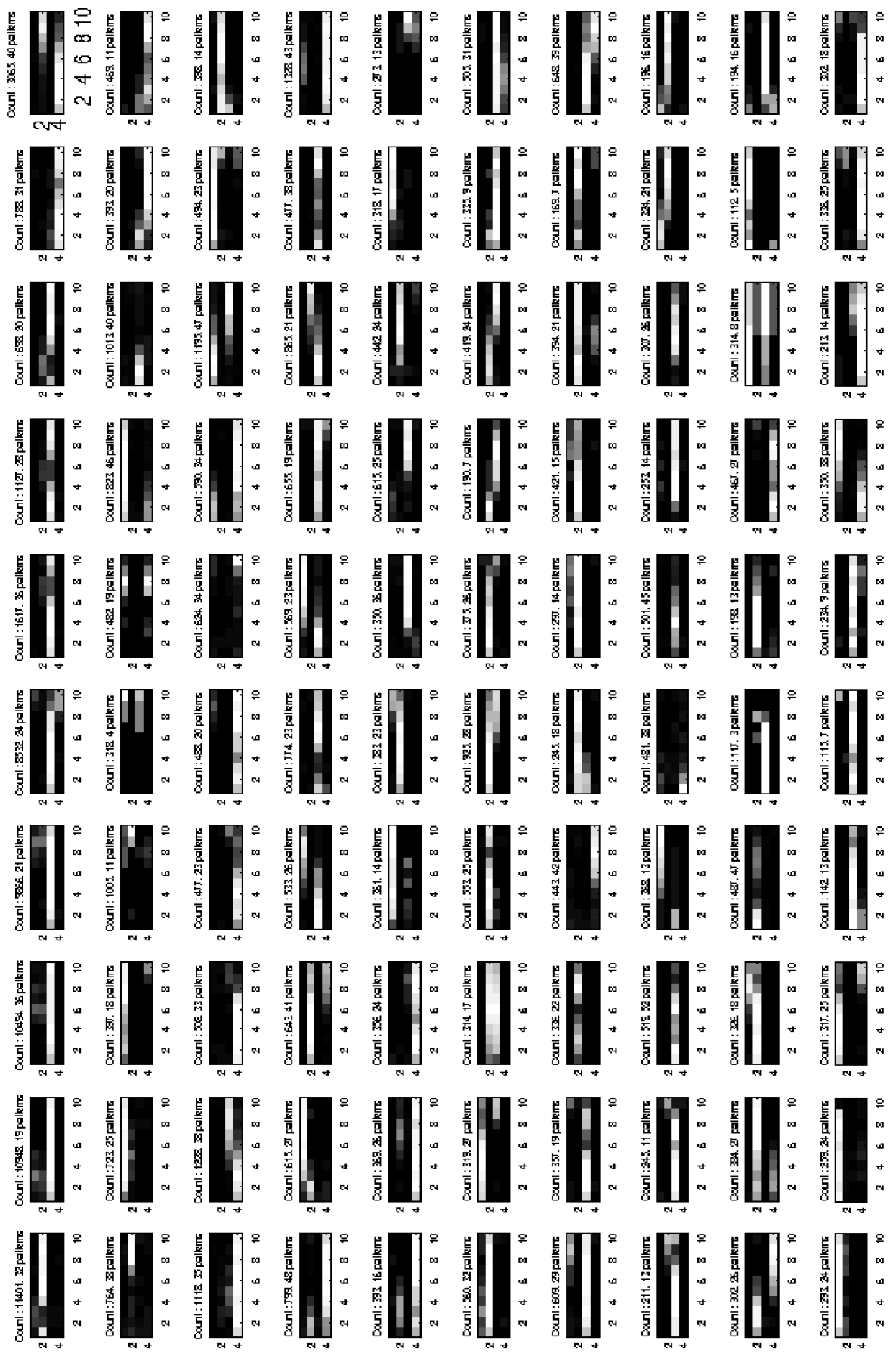


FIGURE III.2.2.14 Représentants des 100 premiers groupes d'équivalence.

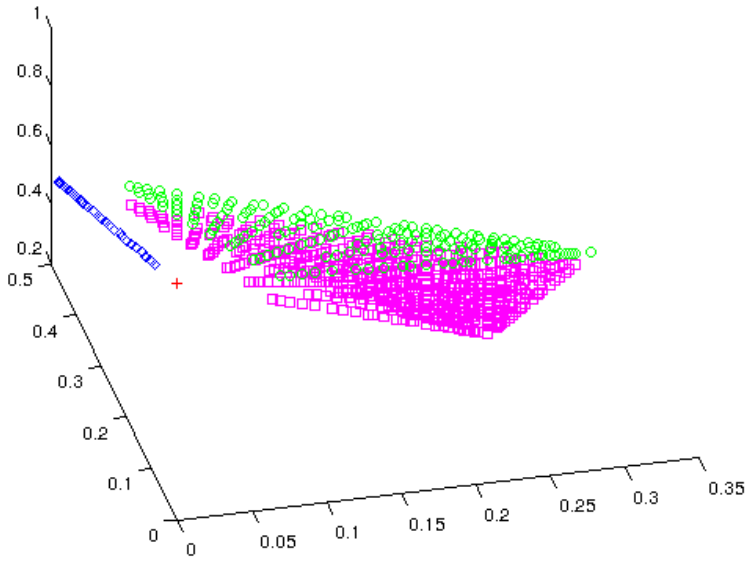


FIGURE III.2.2.15 Description de l'espace par les différentes catégories : catégorie 1 (en rouge), catégorie 2 (en bleu), catégorie 3 (en vert), catégorie 4 (en rose).

premiers gardent toute la parole [0.5-$\frac{\epsilon}{2}$,0.5-$\frac{\epsilon}{2}$,ϵ]. La surface est donc un triangle (points verts sur la figure III.2.2.15). Enfin, la Catégorie 4 rajoute à cela le fait que le 4ème participant peut parler. Ceci rajoute un point [0.25,0.25,0.25] lorsque les 4 parlent de manière équitable. L'espace décrit par cette catégorie est donc une pyramide (points roses sur la figure III.2.2.15).

Le même graphe tri-dimensionnel est tracé sur la figure III.2.2.16 avec les groupes d'équivalence. On peut voir que les catégories 1 et 2 sont bien représentées, mais que les catégories 3 et 4 sont très faiblement représentées.

Ce problème de perte de description des scénarios interactifs est très pénalisant pour la suite où l'on souhaite étudier l'interaction. Il est dû en grande partie à la toute première étape de la méthode où l'on ne conserve que les 5000 motifs les plus fréquents. Or, les motifs interactifs sont par définition peu fréquents puisqu'ils correspondent aux instants de transitions qui sont bien moins nombreux que les régimes établis (une personne qui parle, parle en général plusieurs secondes alors qu'elle ne cesse de parler que durant un échantillon).

III.2.2.2.f Apprentissage de patterns personnels et d'interaction

Afin de contrecarrer le problème évoqué précédemment, nous avons donc séparé les motifs originels selon chaque catégorie. Puis, dans chaque catégorie, la méthode précédente est appliquée afin de tirer des groupes d'équivalence. Nous avons ainsi des groupes d'équivalence représentant chaque catégorie (53 pour la catégorie 1, 136 pour la catégorie 2, 157 pour la 3 et 241 pour la 4). Une analyse de la représentation des distributions des taux de paroles a été effectuée et a permis de confirmer la richesse de représentation des groupes d'équivalences appris par groupe (figure III.2.2.17), ce que l'on peut visualiser sur les 100 premiers représentants de

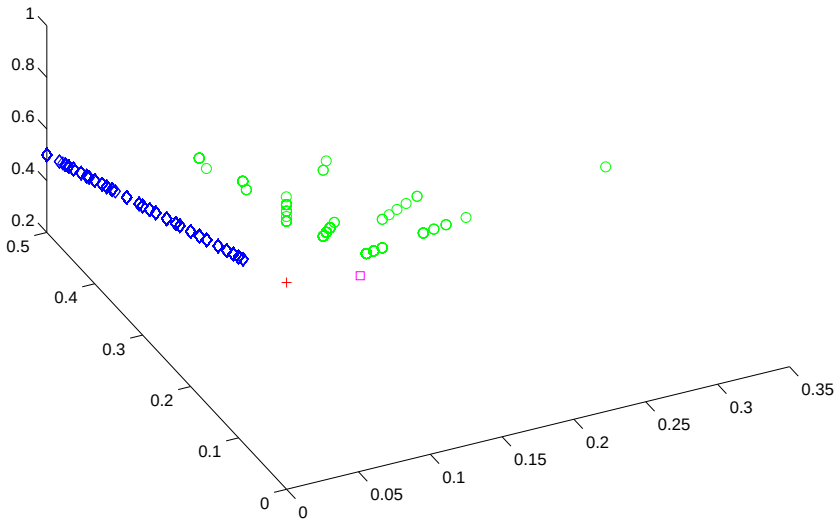


FIGURE III.2.2.16 Description de l'espace occupé par les groupes d'équivalence : catégorie 1 (en rouge), catégorie 2 (en bleu), catégorie 3 (en vert), catégorie 4 (en rose).

chaque Catégorie représentés dans les figures III.2.2.18a (Catégorie 1), III.2.2.18b (Catégorie 2), III.2.2.18c (Catégorie 3), et III.2.2.18d (Catégorie 4).

III.2.3 Système de reconnaissance des rôles dans la base AMI

Notre étude se base sur la base AMI, présentée dans la partie II.2.2. Rappelons que dans cette base, quatre rôles existent :

le Manager de Projet, PM
l'Expert Marketing, ME
le concepteur de l'Interface Utilisateur, UI
le Designer Industriel, ID

Dans chaque meeting, chacun des quatre participants joue un des rôles présentés ci dessus. Le but de notre méthode est de reconnaître le rôle de chacun des participants, en se basant seulement sur les tours de paroles de ceux-ci.

Pour cela, les tours de paroles sont tout d'abord extraits selon la méthode présentée dans la section II.2.3. A partir des tours de paroles sont ensuite extraits :

des indices intra-personnels, 24 indices (section II.2.3.1) ;
des indices inter-personnels, 40 indices (section II.2.3.2) ;
des indices de groupe, 9 indices (section II.2.3.3) ;
des patterns non supervisés appris globalement, 465 patterns (section III.2.2.2) ;
des patterns intra-personnels et inter-personnels, 587 patterns (section III.2.2.2.f).

Les rôles ne changeant pas à l'intérieur de chaque interaction, chaque vidéo peut être encodée de manière globale, sur une seule fenêtre temporelle. Ainsi, chaque interaction est représentée

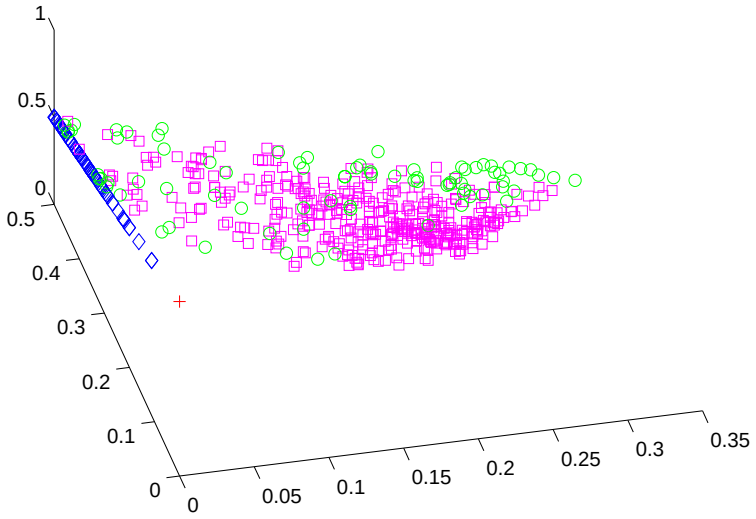


FIGURE III.2.2.17 Description de l'espace occupé par les groupes d'équivalence appris par catégorie : catégorie 1 (en rouge), catégorie 2 (en bleu), catégorie 3 (en vert), catégorie 4 (en rose).

par un vecteur contenant le nombre d'apparition des indices ou patterns.

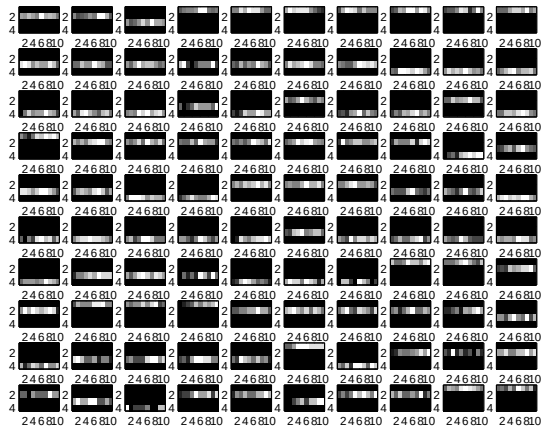
Chaque vecteur est ensuite utilisé en entrée de l'algorithme Machines à Vecteurs de Support (SVM) afin d'apprendre puis reconnaître les rôles de chaque participant. Après avoir fait quelques rappels sur les SVM dans la section III.2.3.1, nous proposons dans la section III.2.3.2 deux architectures pour l'apprentissage des rôles : SVM-4 et SVM-24. Nous appliquerons ensuite ces méthodes dans la section résultats III.2.3.3.

III.2.3.1 Rappels sur les Machines à Vecteurs de Support

Les Machines à Vecteurs de Support (SVM) [46] sont un modèle d'apprentissage supervisé. Nous présentons ici rapidement le principe des SVM linéaires, leur extension en non-linéaire, ainsi que le cas des SVM multi classes. Plus d'informations peuvent être trouvées dans le tutoriel de Burges [34].

III.2.3.1.a SVM linéaire

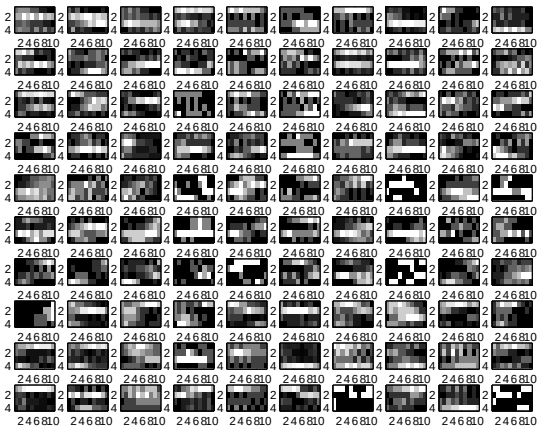
Dans un problème d'apprentissage à deux classes, on cherche à séparer ces deux classes à l'aide d'un hyperplan. Si les données sont linéairement séparables, alors il existe plusieurs hyperplans de séparation. Par exemple, sur la figure III.2.3.19, les exemples de la classe 1 (les ronds noirs), sont linéairement séparables des exemples de la classe 2 (les ronds blancs). Le premier hyperplan H_1 ne sépare pas les classes, puisque des exemples noirs sont présents des deux côtés de l'hyperplan. Le second hyperplan H_2, sépare bien les deux classes, mais on voit qu'il passe très près d'exemples des deux classes. Le dernier hyperplan H_3, sépare lui aussi



(a) Représentants des groupes d'équivalence appris pour la catégorie 1.

(b) Représentants des groupes d'équivalence appris pour la catégorie 2.

(c) Représentants des groupes d'équivalence appris pour la catégorie 3.



(d) Représentants des groupes d'équivalence appris pour la catégorie 4.

FIGURE III.2.2.18 Représentants des groupes d'équivalence appris pour chacune des catégories.

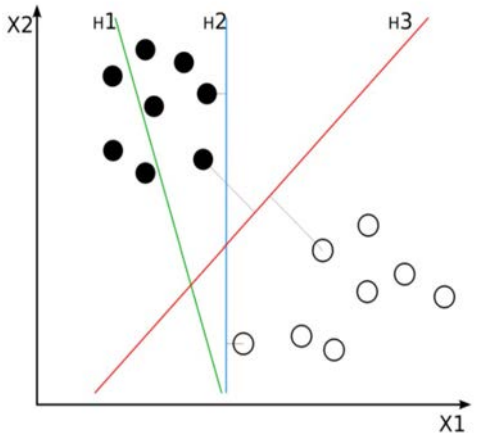


FIGURE III.2.3.19 Plusieurs hyperplans possibles.

les deux classes, et les exemples sont éloignés de l'hyperplan. C'est l'hyperplan avec la marge maximale.

Le problème consiste à optimiser l'orientation de l'hyperplan afin de maximiser la marge. La solution de ce problème d'optimisation est grandement détaillé dans la littérature [34] et ne sera donc pas traité ici.

III.2.3.1.b Marge Souple

De plus, il se peut que quelques exemples rendent le système non linéairement séparable. Afin de contrer ce problème, Cortes et Vapnik mirent en place en 1995 dans [46] un terme supplémentaire de pénalité, autorisant des exemples à être mal classés et permettant ainsi d'avoir une marge souple. Un paramètre C permet de contrôler le compromis entre le nombre d'erreurs de classement et la largeur de la marge.

III.2.3.1.c SVM non linéaire

Il se peut cependant que les données ne soient pas séparables linéairement. Afin de résoudre ce problème, Boser et al. proposèrent dans [29] d'appliquer le stratagème du noyau d'Aizerman et al. [1]. Pour cela, on applique aux données X une transformation non linéaire ϕ. On transpose alors le problème dans un nouvel espace $\phi(X)$, appelé espace de redescription. La fonction noyau définie par

$$K(X_i, X_j) = \phi(X_i)^T.\phi(X_j) \tag{III.2.3}$$

permet alors d'éviter de calculer des produits scalaires dans le nouvel espace $\phi(X)$ de haute dimension. De nombreuses fonctions noyaux existent dans la littérature, mais les plus courantes sont :

le noyau linéaire $K(X_i, X_j) = X_i^T.X_j$
le noyau polynomial $K(X_i, X_j) = (X_i^T.X_j + 1)^d$, où d est le degré

le noyau gaussien $K(X_i, X_j) = exp\left(-\frac{||X_i - X_j||^2}{2\sigma^2}\right) = exp\left(-\gamma||X_i - X_j||^2\right)$,
où σ (ou son équivalent $\gamma = \frac{1}{2\sigma}$) est un paramètre à régler

Dans la suite, le noyau gaussien sera utilisé.

III.2.3.1.d SVM multi-classes

Jusqu'ici, seul le cas binaire avait été traité. Cependant, la littérature a étendu les SVM à des problèmes multi classes. Un grand nombre d'approches existent, et sont traitées en détail dans la littérature [91, 64]. Deux méthodes simples mettant en jeux des SVM binaires peuvent être utilisées, le un-contre-tous et le un-contre-un.

Dans le premier cas, si l'on souhaite avoir N-classes, N-SVM binaires vont être construits, chacun d'entre deux devant reconnaître les données d'une classe, opposées aux données de toutes les autres classes (donc un-contre-tous). Le classifieur ayant la marge la plus grande remporte le vote, et la classe associée à ce classifieur est donc attribuée à la donnée inconnue.

Dans le second cas, le un-contre-un, $N(N-1)/2$ SVM binaires sont construits, chacun opposant une des classes à une des autres classes. En phase de test, la donnée à classer est analysée par chaque classifieur et un vote majoritaire permet de déterminer sa classe.

Dans la suite, l'implémentation des SVM multi-classe utilisée est la célèbre bibliothèque libSVM [38]. Cette implémentation utilise la méthode un-contre-un suite aux résultats publiés par Hsu et Lin [91] qui ont montré que le un-contre-un et le un-contre-tous ont des résultats similaires, mais où le un-contre-un possède un temps d'apprentissage plus court, et des classes non déséquilibrées (dans le un-contre-tous les exemples négatifs représentent $\frac{N-1}{N}\%$, où N est le nombre de classes).

III.2.3.2 Application à la reconnaissance de rôles (SVM 4 classes contre SVM 24 classes)

Comme il a été dit précédemment, chaque interaction de la base AMI est composée de quatre rôles. Notre tâche consiste à apprendre à reconnaître le rôle de chaque participant. Plusieurs stratégies sont possibles.

La première méthode, SVM-4, va apprendre les rôles de chaque participant de manière indépendante. Elle est composée de quatre SVM, un pour chaque position. Chacun d'entre eux prend en entrée le vecteur total, et donne en sortie le rôle de la position auquel il est associé (voir figure III.2.3.20). Chacun des 4 SVM est un SVM à 4 classes (une classe pour chaque rôle). Ainsi, cette méthode est capable de trouver, pour chaque position le rôle plus probable. Avec cette méthode, il est donc possible d'avoir plusieurs fois le même rôle, à des positions différentes, et ne pas voir apparaître certains rôles.

La seconde méthode, SVM-24, va considérer le groupe comme un agencement ordonné des rôles, comme par exemple les deux permutations [ME,UI,ID,PM] ou [PM,ID,ME,UI]. On apprendra alors un SVM à 24 classes, chacune correspondant à une permutation particulière des quatre rôles (voir la figure III.2.3.20).

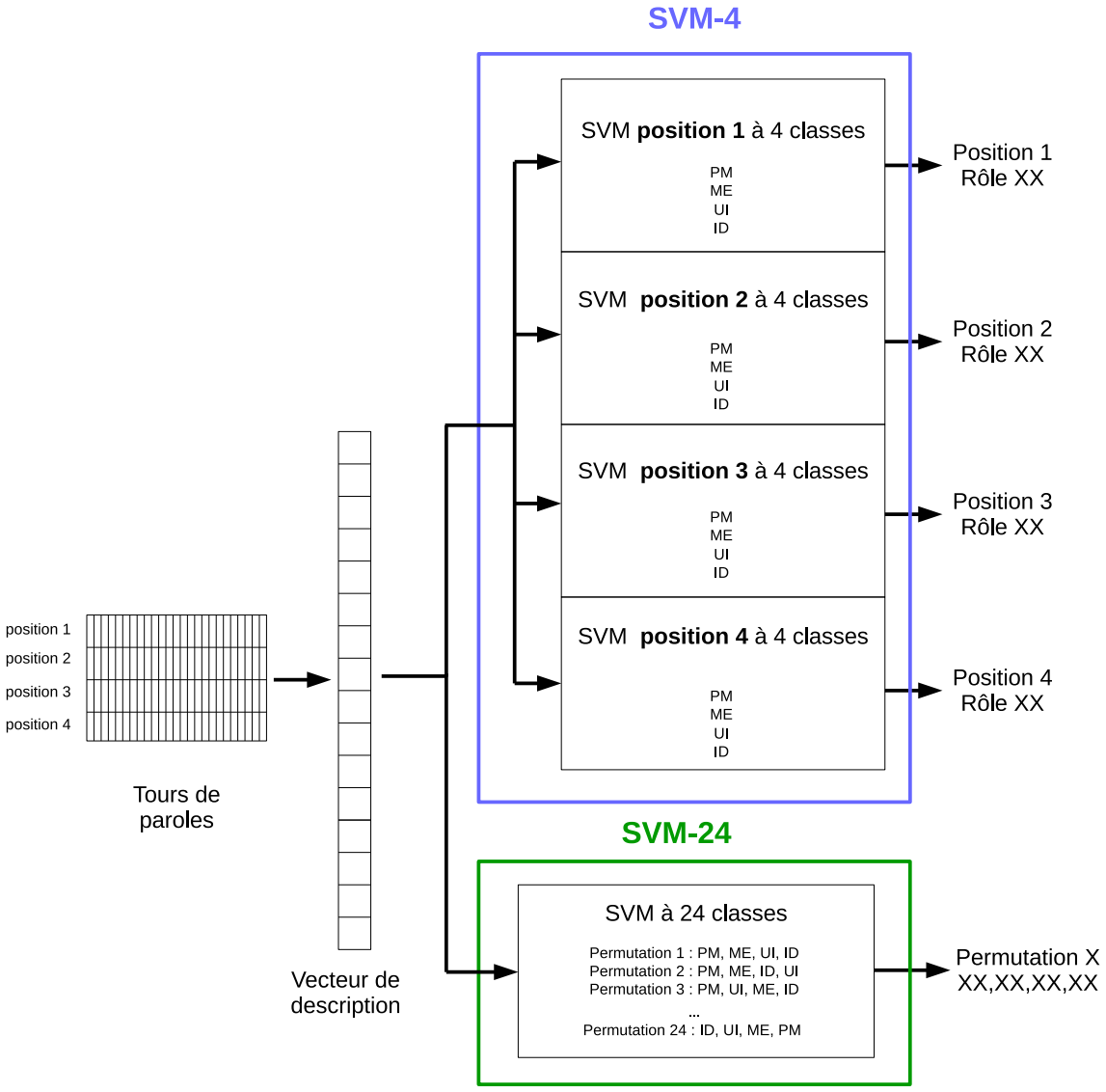


FIGURE III.2.3.20 Illustration des SVM-4 et SVM-24.

III.2.3.3 Résultats

Les méthodes présentées ci dessus ont été appliquées sur les 138 interactions de la base AMI. Or ce nombre de données n'est pas suffisant pour un bon apprentissage d'un SVM. Deux solutions sont donc mises en places afin de régler ce problème. Tout d'abord, nous avons vu que dans la base AMI le Manager de Projet est toujours en position 1, l'Expert Marketing en position 2, etc... Afin que les SVM puissent apprendre différentes situations, il est nécessaire de mélanger ces données en permutant aléatoirement les rôles. Si l'on garde les 24 permutations possibles pour les 138 interactions, on obtient alors un ensemble de 3312 interactions. Cependant, dans un problème d'apprentissage, les données sont souvent séparées en deux parties, une pour l'apprentissage et une autre pour le test. Afin de ne pas perdre une grande partie des données, une approche par K-folder est appliquée. Dans cette approche, les données sont séparées en K groupes, dont K-1 sont utilisés pour l'apprentissage, et le dernier est utilisé pour le test. On change ensuite de folders pour faire un nouveau passage d'apprentissage et de test. Ainsi, K tests sont effectués, chacun avec un nouveau système appris sur les K-1 folders restants, et c'est l'ensemble des résultats moyennés qui permet d'évaluer le système. Ici, chaque folder correspond à l'ensemble des permutations d'une interaction, et contient 24 éléments. De plus étant donné que la base AMI possède 138 interactions, il y a donc 138 folders.

Tout d'abord, les méthodes SVM-4 et SVM-24 (présentées dans la section III.2.3.2) vont être appliquées sur les indices intra-personnels, inter-personnels et de groupe (définis dans la section II.2.3.1). Ensuite, les motifs appris dans les sections III.2.2.2 et III.2.3.3.b remplaceront les indices personnels et interactifs. Enfin, un tableau récapitulatif (section III.2.3.3.c) permettra de comparer les différents résultats entre eux. Ces résultats seront comparé au meilleur résultat de l'état de l'art effectué dans des conditions similaires (pas d'utilisation de lexique et test effectué sur les 138 interactions en K-folder) : Salamin et al. [156]. Une analyse sur les rôles et la base sera alors faite.

III.2.3.3.a SVM 4 classes contre SVM 24 classes

Dans un premier temps, nous souhaitons comparer les méthodes SVM 24 classes et SVM 4 classes. Dans les deux cas, les paramètres C (paramètre de réglage de la marge souple) et γ (paramètre du noyau gaussien) des SVM sont fixés par validation-croisée à l'intérieur de l'ensemble des données d'apprentissage par recherche en grille : un ensemble de valeurs possibles pour chacun des paramètres est défini, et toutes les combinaisons de ces deux ensembles de valeurs sont alors évaluées dans la validation-croisée. Cette étude comparative est réalisée en utilisant les indices intra-personnel, interpersonnels et de groupe de la section II.2.3.1 , soit un vecteur de dimension 73 représentant chaque interaction.

Une matrice de confusion est alors créée avec les rôles réels et les rôles trouvés pour chaque personne (table III.2.2). Étant donné que le SVM-4 attribue un rôle à chacune des positions (voir la figure III.2.3.20), il faut noter que cette méthode peut attribuer plusieurs fois le même rôle dans une interaction, d'où la non symétrie dans les sommes verticales et horizontales. Remarquons que les rôles PM et ME sont assez fortement liés, ainsi que les rôles ID et UI. Les scores de reconnaissance des différents rôles sont alors calculés et reportés dans la table III.2.4. Le PM est sans surprise le rôle le mieux reconnu. Le SVM-4 a des performances moyennes

TABLE III.2.2 Matrice de confusion pour la méthode SVM-4, obtenue pour $C = 10^{-3}$ et $\gamma = 10^{-2}$.

	PM	ME	UI	ID	Somme
PM	2346	536	248	182	3312
ME	812	1187	627	677	3312
UI	538	622	1159	993	3312
ID	549	493	1052	1218	3312
Somme	4254	2838	3086	3070	13248

TABLE III.2.3 Matrice de confusion pour la méthode SVM-24, pour $C = 10^{-2}$ et $\gamma = 10^{-2}$.

	PM	ME	UI	ID	Somme
PM	2525	368	151	268	3312
ME	424	1266	727	895	3312
UI	233	751	1483	845	3312
ID	130	927	951	1304	3312
Somme	3312	3312	3312	3312	13248

sur trois des quatre rôles. En effet, bien que les indices donnent des informations sur toutes les personnes, le rôle de chaque personne est évalué de manière indépendante, sans utiliser les possibles informations apportées par l'évaluation des rôles des autres personnes. Cependant, cette information est utilisée dans le modèle SVM-24, que nous allons donc maintenant évaluer.

La même méthodologie est donc appliquée avec la méthode SVM-24. La matrice de confusion est présentée dans la table III.2.3. La méthode donnant en sortie une permutation (donnant la correspondance entre les personnes et les rôles), chaque rôle ne se voit attribué qu'une seule fois par interaction, ce qui donne des sommes horizontales et verticales égales. Les scores de reconnaissances calculés sont alors reportés dans la table III.2.4. Contrairement à la méthode SVM-4, la méthode SVM-24 utilise le fait que les rôles soient uniques dans chaque interaction, et tous distribués. Ainsi, il est possible de voir que les scores de reconnaissance de tous les rôles sont améliorés par rapport à SVM-4.

Cependant, les indices calculés sont peu nombreux et ont été définis de manière ad-hoc. Ils peuvent donc ne pas rendre complètement compte de la réalité des interactions. Dans la suite, nous allons garder la méthode SVM-24 et l'appliquer aux groupes d'équivalence calculés de manière non supervisée à partir des 5000 motifs les plus présents dans la base (voir la section III.2.2.2).

III.2.3.3.b Utilisation des patterns

Nous avions vu dans la section III.2.2.2 que les groupes d'équivalence permettaient une généralisation non supervisée des indices et qu'ils permettaient de décrire l'interaction. Le SVM-24 est donc appliqué à ces groupes, obtenus dans un premiers temps avec les 5000 motifs les plus fréquents. Comme il y a 465 groupes, le vecteur de caractéristiques d'une interaction, qui est

TABLE III.2.4 Comparaison des résultats pour les méthodes SVM-4 et SVM-24.

	Global	PM	ME	UI	ID
Indices, SVM-4	44,61%	70,8%	35,8%	35,0%	36,8%
Indices, SVM-24	49,66%	76,2%	38,2%	44,8%	39,4%
Salamin et al. [156]	48.93%	76.1%	37.7%	40.6%	41.3%

TABLE III.2.5 Matrice de confusion pour les patterns avec SVM-24, pour $C = 10^{-8}$ et $\gamma = 10^{-4}$.

	PM	ME	UI	ID	Somme
PM	2756	343	45	168	3312
ME	304	1335	778	895	3312
UI	133	447	1638	1094	3312
ID	119	1187	851	1155	3312
Somme	3312	3312	3312	3312	13248

l'histogramme des groupes d'équivalence sur toute la durée de l'interaction, est maintenant de dimension 465. La matrice de confusion de la table III.2.5 est alors obtenue. Les taux de reconnaissance de rôles sont reportés dans la table III.2.7. On peut voir que les scores obtenus sont meilleurs que ceux obtenus avec les premiers indices, ce qui signifie que les groupes d'équivalence comportent bien plus d'informations que les indices. Cependant, nous avions aussi vu que ces groupes d'équivalence ne décrivent pas précisément la complète réalité des données. Afin de palier ce problème, nous avions alors appris des groupes d'équivalence par catégorie (section III.2.2.2.f). Ceux-ci amènent à la matrice de confusion III.2.6. Le score de reconnaissance des rôles global (présenté dans la table III.2.7) avec ces nouveaux motifs appris par catégorie est encore amélioré de 3%.

III.2.3.3.c Tableau récapitulatif

L'ensemble des résultats de ce chapitre ont été réunis dans la table III.2.8. Les résultats de l'état de l'art sans lexique ont été ajoutés à cette table, ainsi que quelques uns avec lexique.

TABLE III.2.6 Autre matrice de confusion pour les patterns avec SVM-24, obtenue pour $C = 10^{-6}$ et $\gamma = 10^{-3}$.

	PM	ME	UI	ID	Somme
PM	2809	229	68	206	3312
ME	138	1510	696	968	3312
UI	191	622	1663	836	3312
ID	174	951	885	1302	3312
Somme	3312	3312	3312	3312	13248

TABLE III.2.7 Comparaison des résultats pour les patterns globaux et les patterns appris par classe.

	Global	PM	ME	UI	ID
Groupes d'équivalence initiaux	51,98%	83,2%	40,3%	49,5%	34,9%
Groupes d'équivalence par catégorie	54,98%	84,8%	45,6%	50,2%	39,3%
Salamin et al. [156]	48.93%	76.1%	37.7%	40.6%	41.3%

Nous pouvons voir que la méthode proposée permet de battre l'état de l'art actuel sur cette base, dans le cas où l'on n'utilise pas le lexique.

Le rôle de PM est le mieux reconnu, car ce rôle correspond à une réalité fonctionnelle dans le groupe. Ceci n'est pas vrai par exemple pour le ID qui lui n'a qu'une expertise technique, et dont la fonctionnalité dans le groupe n'a pas de particularité, le rendant ainsi plus compliqué à reconnaître.

La méthode SVM-24 permet de s'enrichir de la reconnaissance plus facile de certains rôles afin d'appréhender les autres rôles en utilisant cette information. La méthode non supervisée d'obtention de motifs, appris par catégorie, permet d'obtenir un vecteur descripteur assez riche permettant de décrire précisément les interactions et ainsi de reconnaître les différents rôles plus facilement.

Conclusion

Dans ce chapitre, nous avons abordé le problème de la reconnaissance de rôles dans des interactions à 4 participants. La première étape consiste à représenter l'interaction de manière pertinente. Plusieurs solutions ont été testées. Tout d'abord, nous avons extraits des indices "ad-hoc" de l'état de l'art réputés particulièrement discriminants [156]. Ils sont cependant peu nombreux et on peut d'autre part se demander si ce sont vraiment les indices qui caractérisent le mieux les interactions. Ainsi, nous avons proposé une nouvelle méthode, non supervisée, permettant d'extraire des "motifs" aptes à caractériser les interactions. Les deux types d'indices ont été testés selon deux architectures de reconnaissance que nous avons mises en place : les SVM 4 classes et les SVM 24 classes. Les résultats obtenus, au niveau de l'état de l'art, permettent de valider à la fois l'architecture de reconnaissance proposée et la sélection de caractéristiques pertinente pour décrire une interaction.

TABLE III.2.8 Comparaison des résultats pour les différentes méthodes de reconnaissance de rôles.

	Moyenne	PM	ME	UI	ID
Indices, SVM-4	44,61%	70,8%	35,8%	35,0%	36,8%
Indices, SVM-24	49,66%	76,2%	38,2%	44,8%	39,4%
Patterns globaux	51,98%	83,2%	40,3%	49,5%	34,9%
Patterns par classe	54,98%	84,8%	45,6%	50,2%	39,3%
Salamin et al. [156] B,D (man.)	48,93%	76,1%	37,7%	40,6%	41,3%
Favre et al, [69] HHM + 2-gram	40,5%	63,3%	35,8%	34,6%	28,3%
Garg et al. [77] ARS (man.)	42,2 %	79%	20,3%	44,9%	24,6 %
Laskowski et al, [113] (20 exemples)	52,5%	60%	40%	70%	40%
Cristani et al. [49] (20 exemples)	58,75%	85 %	60%	50%	40%
Garg et al. avec lexique (auto) [77]	60,5%	84%	69,8%	38,1%	50,1%
Garg et al. avec lexique (manuel) [77]	71,55%	95,7%	68,8%	60,1%	61,6%

Conclusion et perspectives

Conclusions

Le fil conducteur de cette thèse est l'étude des interactions sociales qui est un thème en plein essor ces dernières années. En effet, le développement des robots et de la puissance des ordinateurs font que ceux-ci (robots et avatars) sont prêts à intégrer notre univers de tous les jours, si ce n'est, qu'ils manquent cruellement d'intelligence en ce qui concerne la sociabilité. Ainsi, par exemple, un robot trop présent et qui nous suit partout devient vite intrusif. Un avatar qui parle d'un ton trop monotone, ou au contraire trop enjoué est vite lassant. Il faut, de la même façon, adapter les émotions faciales, la prosodie, les gestes, ... à la relation en cours qui dépend de nombreux facteurs tels que la personnalité de l'interlocuteur, son état émotionnel ou encore l'environnement. Tant de paramètres et de signaux sociaux entrent en jeu que le traitement du signal social et son interprétation n'en sont encore qu'à leur début, même si de nombreuses avancées ont été réalisées. Ainsi, certains chercheurs [32, 105, 152] se focalisent sur les signaux sociaux pertinents tels que les états émotionnels. D'autres équipes [150, 3, 149, 167, 53] étudient la temporalité aussi appelée synchronie entre les signaux aussi bien intra-personnels qu'inter-personnels. L'étude des groupes de personnes a elle aussi connue des avancées récentes [134, 160, 60] avec, par exemple, la reconnaissance de rôles [163, 61, 157].

Cette thèse se focalise sur deux aspects principaux. Le premier est l'étude d'un signal social particulier : l'imitation. Celle-ci, qui peut être audio (le débit et la hauteur de la parole ont tendance à s'harmoniser au cours d'une interaction) ou gestuelle (même ampleur des gestes, mêmes gestes, mêmes postures) joue un rôle important dans le maintien de l'interaction. Le second aspect traité dans cette thèse est lié à l'étude des groupes. Ainsi, nous avons recherché à répertorier des interactions et à déterminer des motifs spécifiques permettant de la décrire. Nous reprenons ci-dessous les principales contributions de cette thèse.

Synthèse des principales contributions

Dans la première partie de cette thèse nous avons mis en place un système d'évaluation automatique de l'imitation. Pour cela, une base d'imitation dyadique a été créée faisant intervenir 9 dyades dans un jeu d'imitation gestuelle. Un système de reconnaissance automatique des phases d'imitation et de non imitation a ensuite été créé en utilisant les développements récents des méthodes de reconnaissance non supervisée de gestes. Ce système permet d'arriver, avec une latence plus ou moins grande, aux trois grandeurs caractéristiques de l'imitation telles que définies par les psychologues, à savoir : le degré d'imitation entre les partenaires, le délai

entre les partenaires et l'orientation de l'imitation. De plus, le compromis entre la durée de la fenêtre d'observation (et donc la latence) et la qualité des résultats a été analysé.

Dans une seconde partie, nous nous sommes intéressés aux interactions de groupe et plus particulièrement à leur catégorisation. Ainsi, à partir de plusieurs interactions de groupe, est-on capable de dégager des tendances principales en étudiant seulement des indices non verbaux afin de ne pas être influencé par ce qui est dit ? Nous avons pour cela mis au point une méthode utilisant en entrée des signaux sociaux ad-hoc de la littérature, aussi bien intra-personnels qu'inter-personnels. Elle utilise une factorisation en matrices non négatives qui fait ressortir de manière non supervisée 3 types principaux de groupes qui rejoint les types de leadership définis par Lewin et al. [117]. Cette factorisation en matrice non négative permet également de déterminer, parmi les signaux sociaux d'entrée, quels sont les signaux déterminants pour effectuer la classification, ce qui amène à des interprétations intéressantes. Enfin la méthode a été appliquée sur des fenêtres temporelles afin de visualiser l'évolution du type d'interaction au cours du temps.

La dernière partie de cette thèse vise à trouver de manière automatique les signaux sociaux (appelés motifs) pertinents à extraire à partir d'un groupe de personnes. En effet, lors de l'étape précédente de catégorisation des interactions, les signaux en entrée ont été trouvé de manière ad-hoc, sur la base du bon sens, mais sont-ils vraiment les plus pertinents, en existe-t-il d'autre ? Dans un premier temps, nous avons montré qu'en considérant uniquement les tours de parole de 4 participants, pendant 2,5 secondes, le nombre de motifs d'interactions possible est voisin de 10^{12}, ce qui n'est pas raisonnable. Nous avons donc proposé une méthode spectrale pour sélectionner, de manière non supervisée, les motifs les plus importants pour décrire les interactions. Ces motifs, et ceux déterminés de manière ad-hoc dans la partie précédente, ont été comparés dans une tâche de reconnaissance de rôle. Ceci a permis d'une part, de valider les nouveaux motifs proposé et d'autre part, la méthode de reconnaissance de rôle mise en place.

Perspectives

Perspectives à court terme

En ce qui concerne la mesure automatique de l'imitation, la base créée possède une très grande quantité d'imitation, ce qui n'est pas représentatif de la réalité, et l'orientation ne change pas au cours du temps. La méthode devra donc être testée pour des interactions naturelles, où l'imitation est plus diffuse, moins présente, et où l'orientation change de manière dynamique au cours du temps. D'autre part, une adaptation devra être trouvée pour des imitations statiques où les deux participants adoptent la même pose mais ne bougent plus.

Pour la catégorisation non supervisée des groupes, nous nous sommes confrontés au problème de l'évaluation : les résultats sont seulement interprétables et nous ne disposons d'aucune vérité de terrain *a priori*. Il serait donc intéressant de limiter le cadre des interactions et de travailler sur des scénarios type, définis en collaboration avec des psychologues afin de valider à la fois la méthode proposée et les indices extraits.

L'extraction automatique d'indices de groupe pertinents a été effectuée seulement sur les tours de parole et une extension naturelle consiste à prendre en compte plusieurs signaux et

modalités simultanément. La complexité sera d'autant plus étendue et on pourra alors tester les limites de la méthode proposée. Nous avons d'autre part validé la méthode en mettant en place une reconnaissance de rôle mais pour être complet, ces mêmes indices pourraient être testés dans d'autres tâches comme par exemple, la reconnaissance de la personne dominante.

Perspectives à long terme

Concernant l'imitation, nous nous sommes seulement intéressés aux gestes dans le cadre de cette thèse mais il serait pertinent d'étendre ces travaux aux signaux audio en observant par exemple le débit de la voix ou son volume qui ont aussi tendance à s'adapter lors des interactions. Peut-être également que la mesure d'imitation deviendrait alors une mesure d'adaptation. Dans un second temps, cette mesure pourra être utilisée comme une caractéristique de l'interaction, et des travaux mesurant sa corrélation avec divers indices de l'interaction comme la qualité, la dominance, ou des traits de personnalités pourront être entrepris.

Le travail proposé sur les groupes de personnes, visant à extraire des motifs pertinents à la description des groupes, n'est qu'une étape préliminaire et beaucoup de perspectives restent à explorer. Ainsi, l'étude de la temporalité dans les groupes de personnes, de l'influence des personnes l'une sur l'autre ou encore de l'importance de la personnalité des personnes dans le comportement du groupe restent des problèmes non résolus à explorer.

Publications

[127] Stéphane Michelet, Koby Karp, Emilie Delaherche, Catherine Achard et Mohamed Chetouani. Automatic imitation assessment in interaction. In *Human Behavior Understanding*, volume 7559 de *Lectures Notes in Computer Science*, pages 161-173. Springer Berlin Heidelberg, 2012.

[54] Emilie Delaherche, Sofiane Boucenna, Koby Karp, Stéphane Michelet, Catherine Achard, et Mohamed Chetouani. Social coordination assessment : Distinguishing between shape and timing. In *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction*, volume 7742, pages 9-18. Springer Berlin Heidelberg, 2013.

[126] Stéphane Michelet, Catherine Achard et Mohamed Chetouani. Evaluation automatique de l'imitation dans l'interaction. Dans *Actes de la conférence RFIA 2014*, France, 2014.

[9] Marie Avril, Chloë Leclère, Sylvie Viaux, Stéphane Michelet, Catherine Achard, Sylvain Missonnier, Miri Keren, David Cohen, et Mohamed Chetouani. Social Signal processing for studying parents-infant interaction. *Frontiers in psychology*, 5, 2014.

Bibliographie

[1] A Aizerman, Emmanuel M Braverman, and LI Rozoner. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and remote control*, 25 :821 837, 1964.

[2] U Altmann. Studying movement synchrony using time series and regression models. In *Program and Abstracts of the COST 2102 Final Conference held in conjunction with the 4th COST 2102 International Training School on Cognitive Behavioural Systems*, 2011.

[3] Uwe Altmann. Investigation of Movement Synchrony Using Windowed Cross-Lagged Regression. In Anna Esposito, Alessandro Vinciarelli, Klï£¡ra Vicsi, Catherine Pelachaud, and Anton Nijholt, editors, *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues*, number 6800 in Lecture Notes in Computer Science, pages 335 345. Springer Berlin Heidelberg, January 2011.

[4] Nalini Ambady and Robert Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences : A meta-analysis. *Psychological bulletin*, 111(2) :256 274, 1992.

[5] Michael Argyle. *The psychology of interpersonal behaviour*. Penguin UK, 1994.

[6] C Asavathiratham. *The Influence Model : A Tractable Representation for the Dynamics of Interacting Markov Chains*. PhD thesis, Massachusetts Institute of Technology, 2000.

[7] Kathleen T Ashenfelter, Steven M Boker, Jennifer R Waddell, and Nikolay Vitanov. Spatiotemporal symmetry and multifractal structure of head movements during dyadic conversation. *Journal of Experimental Psychology : Human Perception and Performance*, 35(4) :1072 1091, 2009.

[8] Véronique Aubergé. La glu socio-affective : enjeux et risques du robot compagnon. In *JA-SFTAG 2014*, pages 13 14, 2014.

[9] Marie Avril, Chloé Leclère, Sylvie Viaux, Stéphane Michelet, Catherine Achard, Sylvain Missonnier, Miri Keren, David Cohen, and Mohamed Chetouani. Social signal processing for studying parent-infant interaction. *Frontiers in psychology*, 5(1437) :1 14, 2014.

[10] Robert F Bales. *Interaction process analysis ; a method for the study of small groups*. Addison-Wesley, 1950.

[11] Robert Freed Bales. *Personality and interpersonal behavior*. Holt, Rinehart & Winston, 1970.

[12] Satanjeev Banerjee, Jason Cohen, Thomas Quisel, Arthur Chan, Yash Patodia, Ziad Al Bawab, Rong Zhang, Alan Black, Richard M Stern, Alexander I Rudnicky, and others.

Creating multi-modal, user-centric records of meetings with the carnegie mellon meeting recorder architecture. In *Proceedings of the ICASSP 2004 MEeting Recognition Workshop*, 2004.

[13] Satanjeev Banerjee and Alexander I Rudnicky. Using simple speech-based features to detect the state of a meeting and the roles of the meeting participants. In *Proceedings of the 8th Internationnal Conference on Spoken Language Processing*, 2004.

[14] Regina Barzilay, Michael Collins, Julia Hirschberg, and Steve Whittaker. The rules behind roles : Identifying speaker role in radio broadcasts. In *AAAI/IAAI*, pages 679 684, 2000.

[15] Sumit Basu, Tanzeem Choudhury, Brian Clarkson, and Alex (Sandy) Pentland. Learning Human Interactions with the Influence Model. Technical report, Mit Media Laboratory Technical Note, 2001.

[16] Samy Bengio. An Asynchronous Hidden Markov Model for Audio-Visual Speech Recognition. In *Advances in Neural Information Processing Systems, NIPS 15*. MIT Press, 2003.

[17] Kenneth D Benne and Paul Sheats. Functional roles of group members. *Journal of Social Issues*, 4 :41 49, 1948.

[18] F. J. Bernieri, J. S. Reznick, and R. Rosenthal. Synchrony, pseudo synchrony, and dissynchrony : Measuring the entrainment process in mother-infant interactions. *Journal of Personality and Social Psychology*, 54(2) :243 253, 1988.

[19] Frank J Bernieri and Robert Rosenthal. *Interpersonal coordination : Behavior matching and interactional synchrony*, pages 401 432. Cambridge University Press, 1991.

[20] Michael W Berry, Murray Browne, Amy N Langville, V Paul Pauca, and Robert J Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1) :155 173, 2007.

[21] Bruce Jesse Biddle. *Role theory : Concepts and research*. Krieger Pub Co, 1979.

[22] Benjamin Bigot, Isabelle Ferrané, Julien Pinquier, and Régine André-Obrecht. Speaker role recognition to help spontaneous conversational speech detection. In *Proceedings of the 2010 international workshop on Searching spontaneous conversational speech*, pages 5 10. ACM, 2010.

[23] S. Bilakhia, S. Petridis, and M. Pantic. Audiovisual detection of behavioural mimicry. In *ACII*, pages 123 128, Geneva, Switzerland, September 2013.

[24] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[25] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3 :993 1022, 2003.

[26] Steven M Boker, Jennifer L Rotondo, Minquan Xu, and Kadijah King. Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological Methods*, 7(3) :338 355, 2002.

[27] Phillip Bonacich and Paulette Lloyd. Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, 23(3) :191 201, 2001.

[28] E Bormann. *Communicating in small groups : Theory and practice*. New York : Harper and Row, 1990.

[29] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144 152. ACM, 1992.

[30] C. Boutsidis and E. Gallopoulos. SVD based initialization : A head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4) :1350 1362, 2007.

[31] Oliver Brdiczka, Jérôme Maisonnasse, and Patrick Reignier. Automatic detection of interaction groups. In *Proceedings of the 7th international conference on Multimodal interfaces*, pages 32 36. ACM, 2005.

[32] Tom Brøndsted, Thomas Dorf Nielsen, and Sergio Ortega. Affective MultiModal interaction with a 3d agent. In *The Eighth International Workshop On the Cognitive Science Of Natural Language Processing*, pages 102 109, 2000.

[33] Susanne Burger, Victoria MacLaren, and Hua Yu. The ISL meeting corpus : the impact of meeting type on speech style. In *Interspeech*, 2002.

[34] Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2) :121 167, 1998.

[35] N. Campbell. Multimodal processing of discourse information ; the effect of synchrony. In *Second International Symposium on Universal Communication, 2008.*, pages 12 15, 2008.

[36] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. The AMI Meeting Corpus : A Pre-announcement. In Steve Renals and Samy Bengio, editors, *Machine Learning for Multimodal Interaction*, number 3869 in Lecture Notes in Computer Science, pages 28 39. Springer Berlin Heidelberg, January 2006.

[37] Raymond B Cattell. The scree test for the number of factors. *Multivariate behavioral research*, 1(2) :245 276, 1966.

[38] Chih-Chung Chang and Chih-Jen Lin. LIBSVM : A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3) :1 27, 2011. Software available at http ://www.csie.ntu.edu.tw/ cjlin/libsvm.

[39] Tanya L Chartrand and John A Bargh. The chameleon effect : The perception-behavior link and social interaction. *Journal of personality and social psychology*, 76(6) :893 910, 1999.

[40] Tanya L Chartrand and Amy N Dalton. *Mimicry : Its ubiquity, importance, and functionality*, chapter 22, pages 458 483. Oxford, 2009.

[41] Clara Michelle Cheng and Tanya L Chartrand. Self-monitoring without awareness : using mimicry as a nonconscious affiliation strategy. *Journal of personality and social psychology*, 85(6) :1170 1179, 2003.

[42] Tanzeem Choudhury and Sumit Basu. Modeling Conversational Dynamics as a Mixed-Memory Markov Process. In *NIPS*, pages 281 288, 2004.

[43] Andrzej Cichocki, Shun-ichi Amari, Rafal Zdunek, and Anh Huy Phan. *Nonnegative Matrix and Tensor Factorizations : Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation.* Wiley-Blackwell (an imprint of John Wiley & Sons Ltd), September 2009.

[44] Pierre Comon. Independent component analysis, a new concept ? *Signal processing*, 36(3) :287 314, 1994.

[45] Linguistic Data Consortium.

[46] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3) :273 297, 1995.

[47] John B Cortes and Florence M Gatti. Physique and self-description of temperament. *Journal of Consulting Psychology*, 29(5) :432 439, 1965.

[48] M. Cristani, A. Pesarin, C. Drioli, A. Tavano, A. Perina, and V. Murino. Auditory dialog analysis and understanding by generative modelling of interactional dynamics. In *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, pages 103 109, June 2009.

[49] Marco Cristani, Anna Pesarin, Carlo Drioli, Alessandro Tavano, Alessandro Perina, and Vittorio Murino. Generative modeling and classification of dialogs by a low-level turn-taking feature. *Pattern Recognition*, 44(8) :1785 1800, 2011.

[50] Frédéric Cupillard, FranÃğois Brémond, and Monique Thonnat. Group behavior recognition with multiple cameras. In *Applications of Computer Vision, 2002.(WACV 2002). Proceedings. Sixth IEEE Workshop on*, pages 177 183. IEEE, 2002.

[51] Peng Dai, Huijun Di, Ligeng Dong, Linmi Tao, and Guangyou Xu. Group interaction analysis in dynamic context. *Systems, Man, and Cybernetics, Part B : Cybernetics, IEEE Transactions on*, 39(1) :34 42, 2009.

[52] Charles Darwin, Paul Ekman, and Phillip Prodger. *The expression of the emotions in man and animals.* Oxford University Press, USA, 1998.

[53] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen. Interpersonal synchrony : A survey of evaluation methods across disciplines. *Affective Computing, IEEE Transactions on*, 3(3) :349 365, 2012.

[54] Emilie Delaherche, Sofiane Boucenna, Koby Karp, Stéphane Michelet, Catherine Achard, and Mohamed Chetouani. Social coordination assessment : Distinguishing between shape and timing. In *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction*, volume 7742, pages 9 18. Springer Berlin Heidelberg, 2013.

[55] Emilie Delaherche and Mohamed Chetouani. Multimodal coordination : exploring relevant features and measures. In *Proceedings of the 2nd international workshop on Social signal processing*, pages 47 52. ACM, 2010.

[56] Emilie Delaherche, Guillaume Dumas, Jacqueline Nadel, and Mohamed Chetouani. Automatic measure of imitation during social interaction : a behavioral and hyperscanning-EEG benchmark. *Pattern Recognition Letters*, 66(0) :118 126, 2014.

[57] Karen Dion, Ellen Berscheid, and Elaine Walster. What is beautiful is good. *Journal of personality and social psychology*, 24(3) :285 290, 1972.

[58] Trinh Minh Tri Do and Daniel Gatica-Perez. Human interaction discovery in smartphone proximity networks. *Personal and Ubiquitous Computing*, 17(3) :413 431, 2011.

[59] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005*, pages 65 72. IEEE, 2005.

[60] Wen Dong, Bruno Lepri, Alessandro Cappelletti, Alex Sandy Pentland, Fabio Pianesi, and Massimo Zancanaro. Using the influence model to recognize functional roles in meetings. In *Proceedings of the 9th international conference on Multimodal interfaces*, ICMI '07, pages 271 278, New York, NY, USA, 2007. ACM.

[61] Wen Dong, Bruno Lepri, Fabio Pianesi, and Alex Pentland. Modeling Functional Roles Dynamics in Small Group Interactions. *IEEE Transactions on Multimedia*, 15(1) :83 95, 2013.

[62] John F Dovidio, Michelle Hebl, Jennifer A Richeson, and J Nicole Shelton. *Nonverbal communication, race, and intergroup interaction*, chapter 25, pages 481 500. Sage Thousand Oaks, CA, 2006.

[63] Michael Doyle and David Straus. *How to Make Meetings Work !* Berkley, New York, reprint edition edition, September 1993.

[64] Kai-Bo Duan and S Sathiya Keerthi. Which is the best multiclass SVM method ? An empirical study. In *Multiple Classifier Systems*, pages 278 285. Springer, 2005.

[65] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2012.

[66] Guillaume Dumas, Jacqueline Nadel, Robert Soussignan, Jacques Martinerie, and Line Garnero. Inter-brain synchronization during social interaction. *PLoS ONE*, 5(8) :e12166, 2010.

[67] Paul Ekman and Wallace V Friesen. The repertoire of nonverbal behavior : Categories, origins, usage, and coding. *Semiotica*, 1(1) :49 98, 1969.

[68] Yannick Esteve, Thierry Bazillon, Jean-Yves Antoine, Frédéric Béchet, and Jérôme Farinas. The EPAC Corpus : Manual and Automatic Annotations of Conversational Speech in French Broadcast News. In *LREC*, pages 1686 1689, 2010.

[69] Sarah Favre, Alfred Dielmann, and Alessandro Vinciarelli. Automatic role recognition in multiparty recordings using social networks and probabilistic sequential models. In *Proceedings of the 17th ACM international conference on Multimedia*, MM '09, pages 585 588, New York, NY, USA, 2009. ACM.

[70] Sarah Favre, Hugues Salamin, John Dines, and Alessandro Vinciarelli. Role recognition in multiparty recordings using social affiliation networks and discrete distributions. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 29 36. ACM, 2008.

[71] Sebastian Feese, Bert Arnrich, Gerhard Tröster, Bertolt Meyer, and Klaus Jonas. Quantifying behavioral mimicry by automatic detection of nonverbal cues from body motion. In *PASSAT 2012 and SocialCom 2012*, pages 520 525, 2012.

[72] Ruth Feldman. Infant mother and infant father synchrony : The coregulation of positive arousal. *Infant Mental Health Journal*, 24(1) :1 23, 2003.

[73] Ruth Feldman. Parent infant synchrony and the construction of shared timing ; physiological precursors, developmental outcomes, and risk conditions. *Journal of Child Psychology and Psychiatry*, 48(3-4) :329 354, 2007.

[74] Tiffany Field, Paul Greenwald, Connie Morrow, Brian Healy, Tamar Foster, Moshe Guthertz, and Patricia Frost. Behavior state matching during interactions of preadolescent friends versus acquaintances. *Developmental Psychology*, 28(2) :242 250, 1992.

[75] Jonathan Fiscus, John Garofolo, Mark Przybocki, William Fisher, and David Pallett. English broadcast news speech (HUB4), 1997.

[76] Edward W Forgy. Cluster analysis of multivariate data : efficiency versus interpretability of classifications. *Biometrics*, 21 :768 769, 1965.

[77] N. P. Garg, Sarah Favre, Hugues Salamin, Alessandro Vinciarelli, and Dilek Hakkani Tür. Role Recognition for Meeting Participants : an Approach Based on Lexical Information and Social Network Analysis. In *ACM International Conference on Multimedia*, pages 693 696, Vancouver, Canada, 2008.

[78] Daniel Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups : A review. *Image and Vision Computing*, 27(12) :1775 1787, 2009.

[79] Steffen Gauglitz, Tobias Höllerer, and Matthew Turk. Evaluation of Interest Point Detectors and Feature Descriptors for Visual Tracking. *Int. J. Comput. Vision*, 94(3) :335 360, 2011.

[80] Erving Goffman and others. *The presentation of self in everyday life*. Anchor Books, 1959.

[81] Lewis R Goldberg. An alternative" description of personality" : the big-five factor structure. *Journal of personality and social psychology*, 59(6) :1216 1229, 1990.

[82] Daniel Goleman. *Social intelligence*. Random house, 2007.

[83] Gene H Golub and Christian Reinsch. Singular value decomposition and least squares solutions. *Numerische mathematik*, 14(5) :403 420, 1970.

[84] Jay Hall and Wilfred Harvey Watson. The effects of a normative intervention on group decision-making performance. *Human relations*, 23(4) :299 317, 1970.

[85] Judith A Hall, Erik J Coats, and Lavonia Smith LeBeau. Nonverbal behavior and the vertical dimension of social relations : a meta-analysis. *Psychological bulletin*, 131(6) :898 924, 2005.

[86] Robert A Hanneman and Mark Riddle. *Introduction to social network methods*. University of California Riverside, 2005.

[87] A Paul Hare. Types of Roles in Small Groups A Bit of History and a Current Perspective. *Small Group Research*, 25(3) :433 448, 1994.

[88] C. Harris and M. Stephens. A Combined Corner and Edge Detector. In *Proceedings of the 4th Alvey Vision Conference*, pages 147 151, 1988.

[89] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289 296. Morgan Kaufmann Publishers Inc., 1999.

[90] Philip K Hopke. A guide to Positive Matrix Factorization. *Journal of Neuroscience*, 2(10) :1 16, 1991.

[91] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, 13(2) :415 425, 2002.

[92] Hayley Hung and Daniel Gatica-Perez. Estimating cohesion in small groups using audio-visual nonverbal behavior. *Multimedia, IEEE Transactions on*, 12(6) :563 575, 2010.

[93] William James. *The Principles of Psychology*. H. Holt and Company, 1890.

[94] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and others. The ICSI meeting corpus. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 1, pages 1 364. IEEE, 2003.

[95] D. B. Jayagopi and D. Gatica-Perez. Mining Group Nonverbal Conversational Patterns Using Probabilistic Topic Models. *Trans. Multi.*, 12(8) :790 802, 2010.

[96] Dinesh Babu Jayagopi, Sileye Ba, Jean-Marc Odobez, and Daniel Gatica-Perez. Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 45 52. ACM, 2008.

[97] Dinesh Babu Jayagopi, Bogdan Raducanu, and Daniel Gatica-Perez. Characterizing conversational group dynamics using nonverbal behaviour. In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pages 370 373. IEEE, 2009.

[98] Dineshbabu Jayagopi, Dairazalia Sanchez-Cortes, Kazuhiro Otsuka, Junji Yamato, and Daniel Gatica-Perez. Linking speaking and looking behavior patterns with group composition, perception, and performance. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, ICMI '12, pages 433 440, New York, NY, USA, 2012. ACM.

[99] Laszlo A. Jeni, Jeffrey F. Cohn, and Fernando De La Torre. Facing Imbalanced Data Recommendations for the Use of Performance Metrics. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 245 251, 2013.

[100] Oliver P John and Sanjay Srivastava. The Big Five trait taxonomy : History, measurement, and theoretical perspectives. *Handbook of personality : Theory and research*, 2(1999) :102 138, 1999.

[101] Stephen C Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3) :241 254, 1967.

[102] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[103] Daniel Katz and Robert Louis Kahn. *The social psychology of organizations*. John Wiley & Sons, 1978.

[104] Taemie Kim, Agnes Chang, Lindsey Holland, and Alex Sandy Pentland. Meeting mediator : enhancing group collaboration with sociometric feedback. In *CHI'08 Extended Abstracts on Human Factors in Computing Systems*, pages 3183 3188. ACM, 2008.

[105] Yelin Kim and Emily Mower Provost. Emotion classification via utterance-level dynamics : A pattern-based approach to characterizing affective expressions. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3677 3681, 2013.

[106] Alexander Kläser. *Learning human actions in video*. PhD thesis, Université de Grenoble, jul 2010.

[107] Mark Knapp, Judith Hall, and Terrence Horgan. *Nonverbal communication in human interaction*. Cengage Learning, 2013.

[108] Günther Knoblich and Natalie Sebanz. The social nature of perception and action. *Current Directions in Psychological Science*, 15(3) :99 104, 2006.

[109] Junbo Kong and David Graff. TDT4 multilingual broadcast news speech corpus, 2005.

[110] JessicaL. Lakin, ValerieE. Jefferis, ClaraMichelle Cheng, and TanyaL. Chartrand. The Chameleon Effect as Social Glue : Evidence for the Evolutionary Significance of Nonconscious Mimicry. *Journal of Nonverbal Behavior*, 27(3) :145 162, 2003.

[111] Ivan Laptev and Tony Lindeberg. Space-time interest points. In *IN ICCV*, pages 432 439, 2003.

[112] Ivan Laptev and Tony Lindeberg. Local descriptors for spatio-temporal recognition. In *Spatial Coherence for Visual Motion Analysis*, volume 3667, pages 91 103. Springer Berlin Heidelberg, 2006.

[113] Kornel Laskowski, Mari Ostendorf, and Tanja Schultz. Modeling Vocal Interaction for Text-independent Participant Characterization in Multi-party Conversation. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, SIGdial '08, pages 148 155, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[114] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755) :788 791, October 1999.

[115] Daniel D Lee and H. Sebastian Seung. Algorithms for Non-negative Matrix Factorization. *IN NIPS*, 13 :556 562, 2000.

[116] VI Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8) :707 710, 1966.

[117] Kurt Lewin, Ronald Lippitt, and Ralph K White. Patterns of aggressive behavior in experimentally created "social climates". *The Journal of Social Psychology*, 10(2) :269 299, 1939.

[118] Yang Liu. Initial study on automatic identification of speaker role in broadcast news speech. In *Proceedings of the Human Language Technology Conference of the NAACL,*

Companion Volume : Short Papers, pages 81 84. Association for Computational Linguistics, 2006.

[119] Valerie Manusov and Miles L Patterson. *The Sage handbook of nonverbal communication*. Sage, 2006.

[120] B Mathon. Les neurones miroirs : de l'anatomie aux implications physiopathologiques et thérapeutiques. *Revue Neurologique*, 169(4) :285 290, 2013.

[121] Leslie Z McArthur and Reuben M Baron. Toward an ecological theory of social perception. *Psychological review*, 90(3) :215 238, 1983.

[122] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang. Automatic analysis of multimodal group actions in meetings. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(3) :305 317, March 2005.

[123] Iain McCowan, Jean Carletta, W Kraaij, S Ashby, S Bourban, M Flynn, M Guillemot, T Hain, J Kadlec, V Karaiskos, and others. The AMI meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, volume 88, pages 137 140, 2005.

[124] Joseph Edward McGrath. *Groups : Interaction and performance*, volume 14. Prentice-Hall Englewood Cliffs, NJ, 1984.

[125] Maurice Merleau-Ponty and Colin Smith. *Phenomenology of perception*. Motilal Banarsidass Publishe, 1996.

[126] Stéphane Michelet, Catherine Achard, and Mohamed Chetouani. Evaluation automatique de l'imitation dans l'interaction. In *Actes de la conférence RFIA 2014*, France, June 2014.

[127] Stéphane Michelet, Koby Karp, Emilie Delaherche, Catherine Achard, and Mohamed Chetouani. Automatic imitation assessment in interaction. In *Human Behavior Understanding*, volume 7559 of *Lecture Notes in Computer Science*, pages 161 173. Springer Berlin Heidelberg, 2012.

[128] Darren Moore. The IDIAP smart meeting room. Technical report, IDIAP, 2002.

[129] Meinard Müller. *Information Retrieval for Music and Motion*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

[130] Kevin Patrick Murphy. *Dynamic bayesian networks : representation, inference and learning*. PhD thesis, University of California, Berkeley, 2002.

[131] Clifford Ivar Nass and Scott Brave. *Wired for speech : How voice activates and advances the human-computer relationship*. MIT press Cambridge, 2005.

[132] Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3) :443 453, 1970.

[133] Daniel Olguín, Benjamin N Waber, Taemie Kim, Akshay Mohan, Koji Ara, and Alex Pentland. Sensible organizations : Technology and methodology for automatically measuring organizational behavior. *Systems, Man, and Cybernetics, Part B : Cybernetics, IEEE Transactions on*, 39(1) :43 55, 2009.

[134] Daniel Gatica-perez Oya Aran, Hayley Hung. A multimodal corpus for studying dominance in small group conversations. In *LREC MMC'10*, pages 22 26, 2010.

[135] Wei Pan, Wen Dong, M. Cebrian, Taemie Kim, J.H. Fowler, and A.S. Pentland. Modeling dynamical influence in human interaction : Using data to make better inferences about influence within social systems. *Signal Processing Magazine, IEEE*, 29(2) :77 86, 2012.

[136] Maja Pantic and Alessandro Vinciarelli. *Social Signal Processing*, chapter 7, pages 84 93. Oxford library of Psychology, 2014.

[137] A. Pentland. Social Signal Processing [Exploratory DSP]. *IEEE Signal Processing Magazine*, 24(4) :108 111, 2007.

[138] Alex Pentland and Trac Heibeck. *Honest Signals, How They Shape Our World*. The MIT Press, 2008.

[139] Anna Pesarin, Marco Cristani, Vittorio Murino, Carlo Drioli, Alessandro Perina, and Alessandro Tavano. A statistical signature for automatic dialogue classification. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1 4. IEEE, 2008.

[140] Fabio Pianesi, Massimo Zancanaro, Bruno Lepri, and Alessandro Cappelletti. A multimodal annotated corpus of consensus decision making meetings. *Language Resources and Evaluation*, 41(3-4) :409 429, December 2007.

[141] Fabio Pianesi, Massimo Zancanaro, Elena Not, Chiara Leonardi, Vera Falcon, and Bruno Lepri. Multimodal support to group dynamics. *Personal and Ubiquitous Computing*, 12(3) :181 195, 2008.

[142] Steven M Platek, Feroze B Mohamed, and Gordon G Gallup. Contagious yawning and the brain. *Cognitive brain research*, 23(2) :448 452, 2005.

[143] R. Q Quiroga, T. Kreuz, and P. Grassberger. Event synchronization : a simple and fast method to measure synchronicity and time delay patterns. *Physical Review E*, 66(4) :041904.1 041904.9, 2002.

[144] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257 286, 1989.

[145] B. Raducanu, J. Vitria, and D. Gatica-Perez. You are fired ! Nonverbal role analysis in competitive meetings. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009*, pages 1949 1952, April 2009.

[146] Bogdan Raducanu and Daniel Gatica-Perez. Inferring competitive role patterns in reality TV show through nonverbal analysis. *Multimedia Tools and Applications*, 56(1) :207 226, 2012.

[147] Mika Raento, Antti Oulasvirta, and Nathan Eagle. Smartphones an emerging tool for social scientists. *Sociological methods & research*, 37(3) :426 454, 2009.

[148] Fabian Ramseyer and Wolfgang Tschacher. Synchrony : A core concept for a constructivist approach to psychotherapy. *Constructivism in the human sciences*, 11(1) :150 171, 2006.

[149] Fabian Ramseyer and Wolfgang Tschacher. Nonverbal synchrony in psychotherapy : coordinated body movement reflects relationship quality and outcome. *Journal of consulting and clinical psychology*, 79(3) :284 295, 2011.

[150] Dennis Reidsma, Anton Nijholt, Wolfgang Tschacher, and Fabian Ramseyer. Measuring Multimodal Synchrony for Human-Computer Interaction. In *Proceedings of the 2010 International Conference on Cyberworlds*, CW '10, pages 67 71, Washington, DC, USA, 2010. IEEE Computer Society.

[151] RJ Rienks and D Heylen. Automatic dominance detection in meetings using easily detectable features. In *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, 2005.

[152] F. Ringeval, A. Sonderegger, B. Noris, A. Billard, J. Sauer, and D. Lalanne. On the Influence of Emotional Feedback on Emotion Awareness and Gaze Behavior. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 448 453, September 2013.

[153] Giacomo Rizzolatti. The mirror neuron system and its function in humans. *Anatomy and embryology*, 210(5-6) :419 421, 2005.

[154] Roni Rosenfield. Two decades of statistical language modeling : Where do we go from here ? *Proceedings of the IEEE*, 88(8) :193 207, 2000.

[155] Peter J Rousseeuw. Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20 :53 65, 1987.

[156] H. Salamin, S. Favre, and A. Vinciarelli. Automatic Role Recognition in Multiparty Recordings : Using Social Affiliation Networks for Feature Extraction. *IEEE Transactions on Multimedia*, 11(7) :1373 1380, November 2009.

[157] H. Salamin and A. Vinciarelli. Automatic Role Recognition in Multiparty Conversations : An Approach Based on Turn Organization, Prosody, and Conditional Random Fields. *IEEE Transactions on Multimedia*, 14(2) :338 345, April 2012.

[158] Hugues Salamin, Alessandro Vinciarelli, Khiet Truong, and Gelareh Mohammadi. Automatic role recognition based on conversational and prosodic behaviour. In *Proceedings of the international conference on Multimedia*, pages 847 850. ACM, 2010.

[159] Abran J Salazar. An analysis of the development and evolution of roles in the small group. *Small Group Research*, 27(4) :475 503, 1996.

[160] D. Sanchez-Cortes, O. Aran, M.S. Mast, and D. Gatica-Perez. A Nonverbal Behavior Approach to Identify Emergent Leaders in Small Groups. *IEEE Transactions on Multimedia*, 14(3) :816 832, 2012.

[161] Dairazalia Sanchez-Cortes, Oya Aran, Dinesh Babu Jayagopi, Marianne Schmid Mast, and Daniel Gatica-Perez. Emergent leaders through looking and speaking : from audio-visual data to multimodal recognition. *Journal on Multimodal User Interfaces*, 7 :39 53, 2012.

[162] Dairazalia Sanchez-Cortes, Dinesh Babu Jayagopi, and Daniel Gatica-Perez. Predicting remote versus collocated group interactions using nonverbal cues. In *Proceedings of the ICMI-MLMI'09 Workshop on Multimodal Sensor-Based Systems and Mobile Phones for Social Computing*, pages 3.1 3.4. ACM, 2009.

[163] Ashtosh Sapru and Hervé Bourlard. Automatic social role recognition in professional meetings using conditional random fields. In *Proceedings of Interspeech*, 2013.

[164] Ashtosh Sapru and Fabio Valente. Automatic speaker role labeling in AMI meetings : recognition of formal and social roles. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 5057 5060. IEEE, 2012.

[165] Robert E Schapire and Yoram Singer. BoosTexter : A boosting-based system for text categorization. *Machine learning*, 39(2) :135 168, 2000.

[166] Cordelia Schmid, Roger Mohr, and Christian Bauckhage. Evaluation of interest point detectors. *Int. J. Comput. Vision*, 37(2) :151 172, 2000.

[167] R.C. Schmidt, Samantha Morr, Paula Fitzpatrick, and MichaelJ. Richardson. Measuring the Dynamics of Interactional Synchrony. *Journal of Nonverbal Behavior*, 36(4) :263 279, 2012.

[168] Guy L Scott and Hugh Christopher Longuet-Higgins. Feature grouping by'relocalisation'of eigenvectors of the proximity matrix. In *BMVC*, pages 1 6. Citeseer, 1990.

[169] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8) :888 905, 2000.

[170] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1) :195 197, March 1981.

[171] Yale Song, Louis-Philippe Morency, and Randall Davis. Multimodal human behavior analysis : learning correlation and interaction across modalities. In *ICMI*, pages 27 30, 2012.

[172] Stephanie Strassel, Junbo Kong, and David Graff. TDT4 Multilingual Text and Annotations, 2005.

[173] Xiaofan Sun, Jeroen Lichtenauer, Michel Valstar, Anton Nijholt, and Maja Pantic. A multimodal database for mimicry analysis. In *Affective Computing and Intelligent Interaction*, volume 6974, pages 367 376. Springer Berlin / Heidelberg, 2011.

[174] Xiaofan Sun, Anton Nijholt, and Maja Pantic. Towards Mimicry Recognition during Human Interactions : Automatic Feature Selection and Representation. In Antonio Camurri and Cristina Costa, editors, *Intelligent Technologies for Interactive Entertainment*, volume 78 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 160 169. Springer Berlin Heidelberg, 2012.

[175] Xiaofan Sun, Anton Nijholt, Khiet P. Truong, and Maja Pantic. Automatic visual mimicry expression analysis in interpersonal interaction. In *CVPRW 2011*, pages 40 46. IEEE Computer Society, 2011.

[176] Xiaofan Sun, Khiet P. Truong, Maja Pantic, and Anton Nijholt. Towards visual and vocal mimicry recognition in human-human interactions. In *SMC 2011 : Special Session on Social Signal Processing*, pages 367 373. IEEE Computer Society, 2011.

[177] Henry L. Tischler. *Introduction to Sociology*. Harcourt School, Fort Worth, 3 sub edition edition, April 1990.

[178] Lyn M Van Swol. The effects of nonverbal mirroring on perceived persuasiveness, agreement with an imitator, and reciprocity in a group discussion. *Communication Research*, 30(4) :461 480, 2003.

[179] G. Varni, G. Volpe, and A. Camurri. A system for real-time multimodal analysis of nonverbal affective social interaction in user-centric media. *IEEE Transactions on Multimedia*, 12(6) :576 590, 2010.

[180] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder. Bridging the gap between social animal and unsocial machine : A survey of social signal processing. *IEEE Transactions on Affective Computing*, 3(1) :69 87, 2012.

[181] Alessandro Vinciarelli. Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. *Multimedia, IEEE Transactions on*, 9(6) :1215 1226, 2007.

[182] Alessandro Vinciarelli, Anna Esposito, Elisabeth André, Francesca Bonin, Mohamed Chetouani, Jeffrey F Cohn, Marco Cristani, Ferdinand Fuhrmann, Elmer Gilmartin, Zakia Hammal, et al. Open challenges in modelling, analysis and synthesis of human behaviour in human human and human machine interactions. *Cognitive Computation*, 7 :397 413, 2015.

[183] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing : Survey of an emerging domain. *Image and Vision Computing*, 27(12) :1743 1759, 2009. Visual and multimodal analysis of human spontaneous behaviour.

[184] Alessandro Vinciarelli, Maja Pantic, Hervé Bourlard, and Alex Pentland. Social signal processing : state-of-the-art and future perspectives of an emerging domain. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 1061 1070. ACM, 2008.

[185] Alessandro Vinciarelli, Fabio Valente, Sree Harsha Yella, and Ashtosh Sapru. Understanding social signals in multi-party conversations : Automatic recognition of socio-emotional roles in the ami meeting corpus. In *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*, pages 374 379. IEEE, 2011.

[186] Alex Waibel, Tanja Schultz, Michael Bett, Matthias Denecke, Robert Malkin, Ivica Rogina, Rainer Stiefelhagen, and Jie Yang. SMaRT : The smart meeting room task at ISL. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 4, pages IV.752 IV.755. IEEE, 2003.

[187] Heng Wang, Muhammad Muneeb Ullah, Alexander Klaser, Ivan Laptev, and Cordelia Schmid. Evaluation of local spatio-temporal features for action recognition. In *British Machine Vision Conference 2009*, pages 124.1 124.11, London, Royaume-Uni, 2009. BMVA Press.

[188] Shangfei Wang, Zhilei Liu, Zhaoyu Wang, Guobing Wu, Peijia Shen, Shan He, and Xufa Wang. Analyses of a Multimodal Spontaneous Facial Expression Database. *Affective Computing, IEEE Transactions on*, 4(1) :34 46, 2013.

[189] Stanley Wasserman and Katherine Faust. *Social network analysis : Methods and applications*, volume 8. Cambridge university press, 1994.

[190] Tomio Watanabe, Masamichi Ogikubo, and Yutaka Ishii. Visualization of respiration in the embodied virtual communication system and its evaluation. *International Journal of Human-Computer Interaction*, 17(1) :89 102, 2004.

[191] Y. Weiss. Segmentation using eigenvectors : a unifying view. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 975 982, 1999.

[192] Chung-Yi Weng, Wei-Ta Chu, and Ja-Ling Wu. Movie analysis based on roles' social network. In *Multimedia and Expo, 2007 IEEE International Conference on*, pages 1403 1406. IEEE, 2007.

[193] Chung-Yi Weng, Wei-Ta Chu, and Ja-Ling Wu. Rolenet : Movie analysis from the perspective of social networks. *Multimedia, IEEE Transactions on*, 11(2) :256 271, 2009.

[194] Theresa Wilson and Gregor Hofer. Using linguistic and vocal expressiveness in social role recognition. In *Proceedings of the 16th international conference on Intelligent user interfaces*, pages 419 422. ACM, 2011.

[195] Ian H Witten and Eibe Frank. *Data Mining : Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[196] Bo Xiao, Panayiotis G. Georgiou, Chi-Chun Lee, Brian Baucom, and Shrikanth S. Narayanan. Head motion synchrony and its correlation to affectivity in dyadic interactions. In *ICME*, pages 1 6, 2013.

[197] Xuehan Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, pages 532 539, 2013.

[198] Sibel Yaman, Dilek Hakkani-Tür, and Gökhan Tür. Social role discovery from spoken language using dynamic Bayesian networks. In *INTERSPEECH*, pages 2870 2873, 2010.

[199] Massimo Zancanaro, Bruno Lepri, and Fabio Pianesi. Automatic Detection of Group Functional Roles in Face to Face Interactions. In *Proceedings of the 8th International Conference on Multimodal Interfaces*, ICMI '06, pages 28 34, New York, NY, USA, 2006. ACM.

[200] Dong Zhang, Daniel Gatica-Perez, Samy Bengio, and Iain McCowan. Modeling individual and group actions in meetings with layered HMMs. *Multimedia, IEEE Transactions on*, 8(3) :509 520, 2006.

[201] Xingquan Zhu, Xindong Wu, Ahmed K Elmagarmid, Zhe Feng, and Lide Wu. Video data mining : semantic indexing and event detection from the association perspective. *Knowledge and Data Engineering, IEEE Transactions on*, 17(5) :665 677, 2005.

Résumé

Dans une interaction sociale, nous adaptons notre comportement à celui de nos interlocuteurs. L'étude et la compréhension des mécanismes sous-jacents à cette adaptation constituent le coeur du Traitement du Signal Social. Le but de cette thèse est de proposer des méthodes d'étude et des modèles pour l'analyse des signaux sociaux dans un contexte d'interaction en exploitant à la fois des techniques issues du traitement du signal et de la reconnaissance des formes.
Tout d'abord, une méthode non supervisée permettant de mesurer l'imitation entre deux partenaires en termes de délai et de degré est proposée en étudiant uniquement des données gestuelles. Dans un premier temps, des points d'intérêts spatio-temporels sont détectés afin de sélectionner les régions les plus importantes des vidéos. Ils sont ensuite décrits à l'aide d'histogrammes pour permettre la construction de modèles sac-de-mots dans lesquels l'information spatiale est réintroduite. Le degré d'imitation et le délai entre les partenaires sont alors estimés de manière continue grâce à une corrélation-croisée entre les deux modèles sac-de-mots.
La deuxième partie de cette thèse porte sur l'extraction automatique d'indices permettant de caractériser des interactions de groupe. Après avoir regroupé tous les indices couramment employés dans la littérature, nous avons proposé l'utilisation d'une factorisation en matrice non négative. En plus d'extraire les indices les plus pertinents, celle-ci a permis de regrouper automatiquement et de manière non supervisée des meetings en 3 classes correspondant aux trois types de leadership tels que définis par les psychologues.
Enfin, la dernière partie se focalise sur l'extraction non supervisée d'indices permettant de caractériser des groupes. La pertinence de ces indices, par rapport à des indices ad-hoc provenant de l'état de l'art, est ensuite validée dans une tâche de reconnaissance des rôles.

Mots clés Apprentissage non supervisé, Interaction, Traitement du Signal Social, Imitation, Reconnaissance de rôles, Extraction d'indices, Sac-de-Mots, Factorisation en Matrices Non-négatives

Abstract

In an social interaction, we adapt our behavior to our interlocutors. Studying and understanding the underlying mecanisms of this adaptation is the center of Social Signal Processing. The goal of this thesis is to propose methods of study and models for the analysis of social signals in the context of interaction, by exploiting both social processing and pattern recognition techniques
First, an unsupervised method allowing the measurement of imitation between two partners in terms of delay and degree is proposed, only using gestual data. Spatio-temporal interest point are first detected in order to select the most important regions of videos. Then they are described by histograms in order to construct bag-of-words models in which spatial information is reintroduced. Imitation degree and delay between partners are estimated in a continuous way thanks to cross-correlation between the two bag-of-words models.
The second part of this thesis focus on the automatic extraction of features permitting to characterizing group interactions. After regrouping all features commonly used in literature, we proposed the utilization of non-negative factorization. More than only extracting the most pertinent features, it also allowed to automatically regroup, and in an unsupervised manner, meetings in three classes corresponding to three types of leadership defined by psychologists.
Finally, the last part focus on unsupervised extraction of features permitting to characterize groups. The relevance of these features, compared to ad-hoc features from state of the art, is then validated in a role recognition task.

Keywords Unsupervised learning, Interaction, Social Signal Processing, Imitation, Role recognition, Feature extraction, Bag-of-words, Non-negative Matrix Factorization