Analyse de Signaux Sociaux pour la Modélisation de l'interaction face à face

Ammar Mahdhaoui

THESE DE DOCTORAT

de l'Université Pierre et Marie Curie – Paris 6

présentée par

Ammar MAHDHAOUI

pour l'obtention du grade de

Docteur de l'Université Pierre et Marie Curie – Paris 6

Spécialité

Informatique

Analyse de signaux sociaux pour la modélisation de l'interaction face à face

A soutenir le 13 Décembre 2010 devant le jury composé de

Rapporteurs	Alessandro VINCIARELLI	Professeur	IDIAP – University of Glasgow
	Laurent BESACIER	Professeur	LIG-lab – UJF Grenoble
Examinateurs	Jean-Claude MARTIN	Professeur	LIMSI – Paris Sud 11
	Maurice MILGRAM	Professeur Emerite	ISIR – UPMC Paris
Directeurs	Jean-Luc ZARADER	Professeur	ISIR – UPMC Paris
	Mohamed CHETOUANI	Maître de Conférences	ISIR – UPMC Paris
Invité	David COHEN	Professeur	ISIR – UPMC Paris

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Contexte et motivations

L'interaction se caractérise par une réciprocité des actions. Une interaction dite sociale est une relation dynamique de communication et d'échange d'information entre deux individus ou entre plusieurs individus à l'intérieur d'un groupe. Cette relation réciproque n'est pas toujours contrôlée, elle est parfois involontaire et peut se situer à un niveau qui échappe à la conscience du sujet. L'interaction sociale fait intervenir plusieurs canaux dont la communication, verbale et non verbale. La contribution au message des informations verbales (parole) et non-verbales (geste, expression faciale, posture) dépend énormément du contexte, des personnes et des cultures. De plus, l'ensemble de ces messages se complètent entre eux et permettent d'offrir une information plus riche.

L'apprentissage ainsi que le développement de l'être humain, et plus particulièrement celui de l'enfant, dépendent principalement de son interaction avec son environnement. Les interactions humaines constituent la base du développement intellectuel, cognitif et affectif. Les carences affectives, totales ou partielles, ont pour conséquence une altération du développement global. Ces carences affectent de façon plus ou moins sévères le développement de l'enfant.

L'analyse des interactions sociales passe généralement par l'extraction d'informations dans les différents flux audio et vidéo. Selon les domaines de recherche (informatique ou psychologie par exemple), les méthodologies mises en oeuvre font plus ou moins appel à des approches automatiques. Au-delà de ces approches particulières, il s'agit de structurer et d'interpréter ces interactions. Cette analyse est fondamentale notamment pour les applications liées à la santé où les comportements sont très souvent plus difficiles à comprendre. Par exemple l'autisme qui est un trouble envahissant du développement caractérisé par des variabilités importantes (spectre autistique), la compréhension de ces variabilités requiert la mise en oeuvre de méthodologies d'analyse spécifiques.

Dans le cadre de cette thèse, nous nous intéressons à la modalité audiovisuelle dans l'objectif de structurer des interactions sociales particulières : interactions parents-enfants avec une perspective développementale (longitudinale). Nos travaux consistent dans un premier temps à développer des méthodes automatiques pour l'extraction de signaux sociaux. Dans un deuxième temps, nous avons cherché à les analyser en proposant un modèle contextuel de l'interaction. Nos travaux de thèse se situent ainsi dans le domaine du

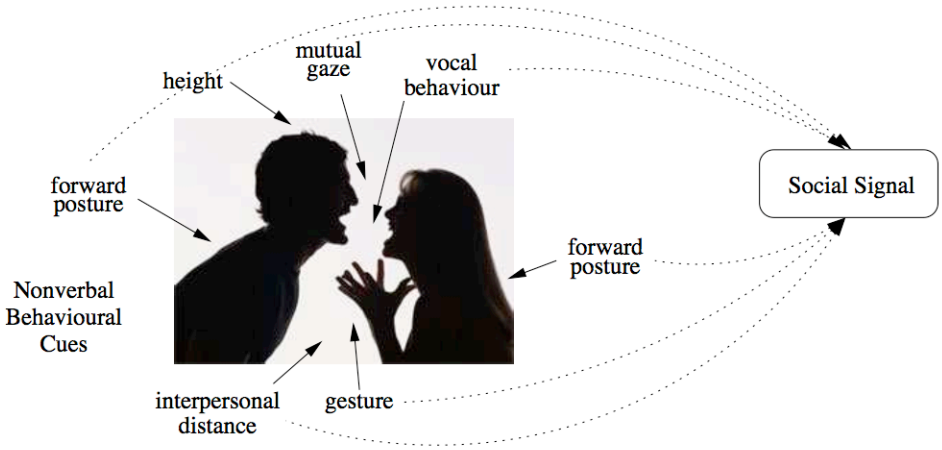


FIG. 1 – Signaux sociaux [Vinciarelli *et al.*, 2009c]

traitement du signal social.

Analyse des signaux sociaux

L'analyse des signaux sociaux[1] (*ASS*) est un domaine de recherche émergent [Pentland, 2007] [Vinciarelli *et al.*, 2009c] [Vinciarelli *et al.*, 2009b]. Il consiste à comprendre, interpréter, et parfois prédire les comportements humains. Dans le cadre d'une interaction face à face, les signaux sociaux se définissent par leurs rôles : informatifs ou communicatifs. Les signaux informatifs consistent à transmettre une information au destinataire, tandis que les signaux communicatifs permettent d'organiser la conversation et de partager les tours de parole entre les différents interlocuteurs. Ils donnent des informations sur la qualité de l'interaction, les émotions sociales, les comportements sociaux, les relations sociales, etc [Pantic *et al.*, 2009]. Les signaux sociaux sont transmis à travers des comportements non verbaux : gestes, posture, expressions faciales, etc (cf. figure 1).

L'interprétation de messages en exploitant uniquement les signaux non verbaux est une tâche complexe du fait de la dynamique et de l'interdépendance des signaux. Pour investiguer ces dépendances, Argyle [1967] a mené des expériences sur la perception de messages sociaux par des humains en exploitant uniquement des comportements non verbaux.

L'analyse des signaux sociaux consiste à étendre ce concept à l'interaction homme-machine [Pentland, 2007] [Vinciarelli *et al.*, 2009c]. Il s'agit de doter

[1] http ://sspnet.eu/

FIG. 2 – Interaction non verbale homme-robot

une machine de la capacité à percevoir des informations non verbales. Plusieurs travaux ont été proposés dans la littérature. Ces travaux consistent à analyser, comprendre et prédire les comportements humains dans le cadre des interactions sociales. Récemment, Aran et Gatica-Perez [2010] ont développé une méthode automatique qui permet d'identifier la personne dominante dans une conversation en se basant uniquement sur des signaux non verbaux : tour de parole et gestes. Dans le cadre d'interaction homme-robot (human-robot interaction *HRI*), AL Moubayed *et al.* [2009] ont développé une plateforme d'interaction en se basant uniquement sur des expressions faciales et des intonations de la voix (sourire, mouvement de la tête, proéminence acoustique). Le système consiste à produire automatiquement des "backchannels multimodaux" (liés aux fonctions phatiques et régulatrices des signaux de communication) sans avoir accès aux informations verbales (cf. figure 2).

Signaux émotionnels et interactions sociales

Les signaux émotionnels sont des outils très importants de la communication face à face. Ils jouent un rôle fondamental dans la vie quotidienne et exercent une influence non négligeable sur les comportements humains. Ils oeuvrent pour transmettre plus de sens aux échanges et renseignent sur les intentions et motivations. Les émotions se traduisent par des modifications du signal de parole et plus particulièrement les éléments non verbaux (para

et extra-lingustiques). Parmi ces éléments, on retrouve la qualité vocale ainsi que d'autres caractéristiques prosodiques comme le rythme de la parole ou l'intonation.

En revanche, la définition du phénomène émotionnel est un sujet controversé sur lequel se sont penchés de nombreux chercheurs. Les recherches menées montrent qu'il est difficile d'aboutir à un consensus sur la définition du phénomène émotionnel. Kleinginna et Kleinginna [1981] donnent même l'impression de l'absence d'un dénominateur commun entre les définitions classiques et les définitions modernes [De Bonis, 1996]. Même si les débats philosophiques et scientifiques concernant la nature, les propriétés et les fonctions des émotions se continuent, deux aspects sont généralement acceptés pour la définition du phénomène émotionnel :

- Les émotions sont des réactions à des situations considérées appropriées aux besoins, objectifs et attitudes d'un individu.
- Les émotions comportent des composantes physiologiques, affectives, comportementales et cognitives.

Dans les interactions sociales, les émotions sont des facteurs de régulation des comportements humains car elles permettent d'avoir un feedback des actions produites (en percevant comment un comportement est perçu). Elles permettent également de coordonner les échanges sociaux en renseignant sur les états émotionnels de chaque personne (parent-enfant dans notre cas). Cet échange d'informations se réalise à travers des signaux verbaux et non verbaux. Dans la grande majorité des cas, l'échange de messages verbaux (communication effectuée au moyen des mots) s'accompagne de changements continuels de l'expression du visage, de la direction du regard (mimique faciale) et d'un très grand nombre de mouvements de l'ensemble du corps (gestes et postures). Les émotions affectent également la dimension cognitive des individus en modifiant les mécanismes d'interprétation et en influençant les processus de prise de décision.

Dans cette thèse, nous avons cherché à caractériser une émotion dite sociale car elle est produite dans un contexte interactif afin d'induire une réponse de l'interlocuteur. Le support de cette émotion sociale est le *motherese* ou parole adressée au bébé (*infant-directed speech*). Ce registre de parole particulier est connu pour avoir un impact sur le comportement de l'enfant. Un des objectifs de cette thèse est une meilleure compréhension du rôle du *motherese* dans l'interaction sociale. Cette étude exploite une méthodologie souvent utilisée dans les recherches sur le développement de l'enfant. Il s'agit d'exploiter des interactions longitudinales, vécues dans des contextes réels et variés, regroupées ici dans des films familiaux : enregistrements filmés par les parents (non professionnels) d'interactions naturelles et spontanées. Les films familiaux de notre étude ont été réalisés sans l'objectif d'être analysés car collectés a posteriori,

quelques années après l'établissement des diagnostics.

Problématique et méthodologie

Notre étude vise plus particulièrement à analyser les comportements complexes d'enfants à devenir autistique. L'autisme étant défini comme un trouble précoce, global et sévère du développement de l'enfant caractérisé par des perturbations dans les domaines des interactions sociales réciproques, de la communication (verbale et non verbale) et des comportements, un champ d'intérêts et activités au caractère restreint, répétitif. Ces signes, lorsqu'ils sont associés, définissent le syndrôme autistique [Muratori et Maestro, 2007]. Ainsi, le développement de la communication et celui des fonctions cognitives sont très perturbés ce qui se traduit cliniquement par un isolement, une diminution des activités spontanées et des troubles du langage. Cependant, l'intensité de ces signes peut être très différente d'un enfant à l'autre et peut varier avec l'âge.

Les enfants autistes se caractérisent par une interaction sociale faible et un manque d'échange avec leurs parents. Ainsi, le langage peut être totalement absent ou se développer de façon déviante. Cela peut affecter l'engagement des parents dans l'interaction. Par conséquent, cette altération va renforcer le retard du développement social et retarder également le processus d'acquisition du langage. De plus, leur pauvre réponse aux sollicitations de leurs parents peut appauvrir l'incitation parentale.

Récemment, des chercheurs étudiant les interactions sociales et l'acquisition du langage ont étudié le rôle du *motherese* dans le développement de l'enfant autiste. Les résultats préliminaires de Laznik *et al.* [2005] ont montré que la plupart des séquences positives (réponses multimodales de l'enfant : vocalisation, expressions faciales, regard, etc.) sont initiées par le *motherese*. Cette analyse manuelle est impossible à généraliser à un grand nombre de films. Afin de mieux comprendre le rôle de ce registre de parole, il est important de proposer des modèles d'interactions estimés sur un grand nombre de données. L'étude des films familiaux demeure une des rares méthodes d'exploration des premières années.

Questions suscitées par les films familiaux ? : La littérature sur les liens entre interaction et développement de l'enfant ainsi que les études manuelles préliminaires suggèrent que le *motherese* joue le rôle de modérateur dans l'interaction parent-enfant. Les travaux présentés dans cette thèse se situent dans le cadre d'une collaboration entre le service de psychiatrie de l'enfant et de l'adolescent de l'hôpital de la Pitié-Salpêtrière dirigé par David

Cohen, le département psychiatrie et de neurologie de l'enfant de l'université de Pise dirigé par Filippo Muratori. Nous avons eu accès à une base de films familiaux décrivant des interactions parents-enfants naturelles et longitudinales. Certains comportements des enfants et des parents ont été annotés manuellement par des psychologues (cf. chapitre 4).

Pour étudier le rôle modérateur du *motherese*, nous avons proposé des modèles computationnels permettant de définir expérimentalement le *motherese* et de le détecter automatiquement dans un grand nombre de scènes audiovisuelles. Son rôle est ensuite étudié dans un modèle plus global de l'interaction multimodale et longitudinale. Nous déclinons par la suite les questions spécifiques à ces modèles (cf. figure 3).

Détection automatique de parole émotionnelle : La détection du *motherese* est une étape fondamentale dans notre travail. Comme nous le verrons dans le chapitre 2, le *motherese* possède certaines caractéristiques perceptuelles particulières (modulation exagérée de la prosodie, modification de la durée de phonèmes) permettant de préciser un premier modèle de reconnaissance. Cependant, comme pour la majorité des signaux de parole émotionnelle, il n'y a pas de consensus sur les caractéristiques acoustiques et prosodiques à utiliser mais également sur les techniques de classification. De plus, la qualité sonore très variable des enregistrements nécessite de développer des méthodes robustes. La littérature portant sur la détection de parole émotionnelle montre qu'il est nécessaire et important de passer par une phase d'annotation manuelle des données permettant de définir expérimentalement la catégorie recherchée. Ensuite, il s'agit de mettre en oeuvre des méthodes d'extraction de caractéristiques et de classification (le plus souvent supervisées) de signaux. Dans cette thèse, la problématique liée à la question de la détection du *motherese* réside dans le caractère interactif de la production car s'agissant d'une émotion sociale et non prototypique produite dans des contextes très variés (situation d'interaction : bain, jeux, repas, etc.).

Traitement de données partiellement étiquetées : L'objectif de la détection automatique est de prédire la classe d'appartenance des signaux de parole (*motherese* vs *non motherese*) en se basant sur des modèles le plus souvent optimisés par apprentissage sur des données étiquetées. L'annotation des données joue alors un rôle fondamental tout en posant le problème du traitement des données non étiquetées. Même avec une annotation de qualité (fidélité inter-juge importante) et des techniques robustes de classification, la généralisation à des situations hétérogènes reste un verrou scientifique important. La classification semi-supervisée offre une solution pertinente en permettant le renforcement des modèles. L'exploitation à la fois de données

étiquetées et non étiquetées dans les techniques de co-apprentissage consiste à combiner plusieurs ensembles du même problème de classification [Blum et Mitchell, 1998]. Ces ensembles sont supposés indépendants et forment des “vues" (extracteur de caractéristique + classifieur). Chacune des “vues" est alors optimisée sur les données étiquetées et utilisée en prédiction sur les données non étiquetées. Il est ainsi possible d'ordonner les données non étiquetées avec plusieurs points de vue différents. Le renforcement des classifieurs s'opère alors par le ré-apprentissage des modèles en exploitant des données étiquetées par les vues complémentaires. Dans le cadre du traitement de la parole émotionnelle et du fait de la multitude de caractéristiques potentiellement exploitables, nous avons proposé un algorithme de co-apprentissage permettant de fusionner de manière dynamique un grand nombre de classifieurs tout en renforçant la règle de catégorisation.

Modélisation de l'interaction longitudinale : En se basant sur des observations directes d'interaction, nous avons proposé un modèle computationnel de l'interaction sociale parent-enfant. Il consiste à modéliser les réponses des enfants par rapport aux stimulations des parents (ex : verbales, objet, toucher). La nature stochastique et multi-modale de l'interaction imposent de développer des approches robustes mais également lisibles pour une exploitation par des cliniciens. L'interdépendance et l'impact des signaux d'interaction ont été étudiés par un modèle statistique traditionnellement utilisé en reconnaissance de la parole (*n-gram*). Cette modélisation permet à la fois d'explorer les mécanismes d'émergence de comportements dyadiques mais également de quantifier l'influence des stimulations des parents sur les réponses des enfants.

La modélisation de l'interaction sociale représente un défi constant notamment à cause du très peu de connaissances actuelles sur les facteurs qui guident et régulent l'interaction. Ces travaux de recherche tout en étant ancrés dans le domaine du traitement du signal et de la reconnaissance des formes requiert une collaboration étroite avec des chercheurs en psychologie-psychiatrie spécialistes du développement de l'enfant. Et par la nature même de cette recherche, nos contributions se situent dans le domaine du traitement du signal social appliqué aux troubles du développement.

Organisation du document

Ce document est organisé autour de quatre chapitres (cf. figure 3) :

– Chapitre 1 : la première partie introduit le contexte de notre étude portant sur l'interaction sociale et l'analyse de signaux sociaux. Elle est associée à un état de l'art nous permettant de situer les approches mé-

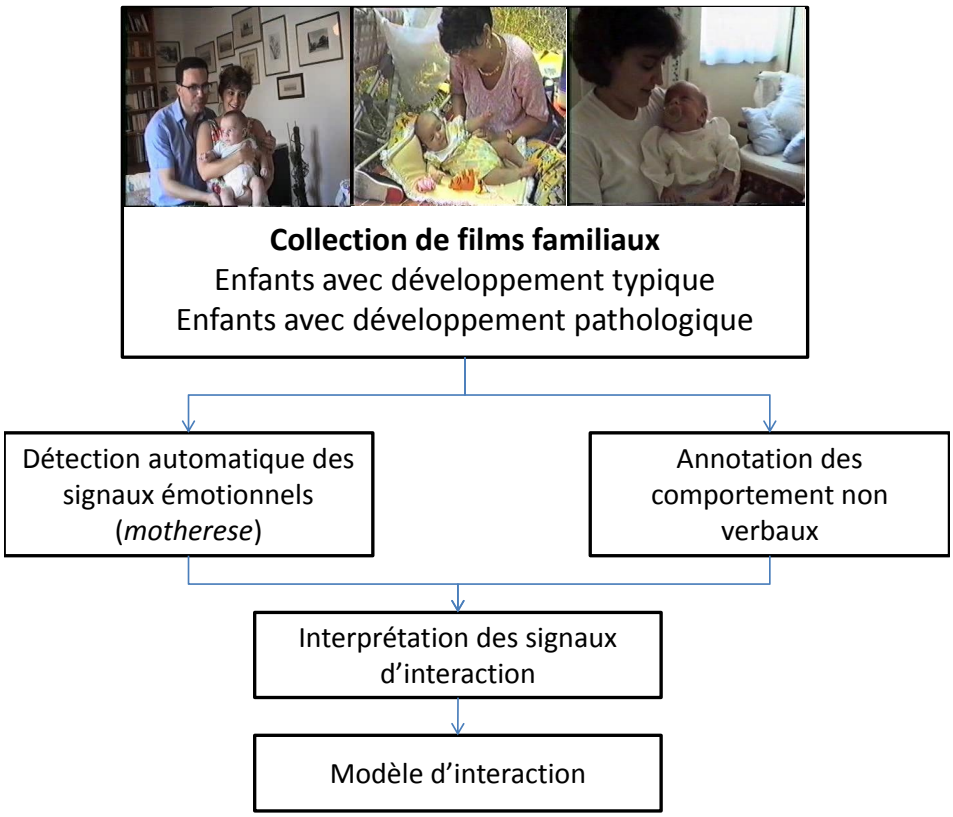


FIG. 3 – Méthodologie du travail

thodologiques employées. À partir d'une description des stratégies de communication chez l'adulte, nous précisons les spécificités de l'interaction avec les enfants typiques et pathologiques en développement.

– Chapitre 2 : la deuxième partie définit le concept de l'émotion sociale analysée : le *motherese*. Aussi nous présentons les étapes préalables nécessaires au développement d'un système automatique de reconnaissance de signaux émotionnels : l'acquisition et l'annotation d'un matériel d'étude des évènements émotionnels, l'extraction de caractéristiques et les différentes approches de classification. Un système de détection d'émotions est généralement composé de deux grandes parties : extraction des caractéristiques et classification. Nous avons utilisé différents descripteurs acoustiques afin de caractériser les manifestations émotionnelles dans le signal de parole : informations spectrales et prosodiques. La classification exploite des approches supervisées ainsi que de la fusion d'informations extraites à des échelles temporelles différentes (segmentales et supra-segmentales) permettant d'améliorer les performances

globales du système de détection.

- Chapitre 3 : la troisième partie propose une nouvelle approche pour la classification de signaux émotionnels. Dans cette partie, nous proposons un algorithme de co-apprentissage automatique en se basant sur la fusion des plusieurs classifieurs. La méthode proposée permet d'apprendre une règle de catégorisation à partir d'un petit nombre de données étiquetées. Cependant du fait de la variabilité des situations étudiées, la définition expérimentale du *motherese* n'est pas suffisante. Nous avons proposé une méthodologie consistant à faire collaborer différents classifieurs dans l'objectif de renforcer la règle de catégorisation.
- Chapitre 4 : la quatrième partie présente une modélisation computationnelle de l'interaction sociale (parent-enfant) exploitant différents signaux de communication. Nous situerons notre contribution en modélisation dans le domaine de l'analyse des signaux sociaux. La méthodologie employée est également complétée par des analyses quantitatives et statistiques afin d'identifier le rôle de ces signaux. En complément à l'étude de l'impact des signaux, nous avons proposé une structuration de l'interaction basée sur des méthodes de regroupement (clustering). Cette structuration requiert de définir une nouvelle représentation des dyades interactives inspirée du traitement automatique de documents.
- Conclusions et perspectives (partie 4.8) : Dans la dernière partie, nous dressons les principales conclusions de ce travail et nous présentons les perspectives de recherches qu'ouvrent les méthodologies proposées.

Chapitre 1

Interaction Sociale

Sommaire

1.1 Introduction

L'interaction sociale est un processus d'échanges réciproques d'informations et d'actions entre des humains. Chaque information ou action a un effet

sur l'interlocuteur produisant le plus souvent une réaction. Plusieurs stratégies de communication sont employées afin de transmettre ces informations (communication verbale et non-verbale) exploitant des signaux dits sociaux. Les signaux sociaux sont définis comme étant des signaux communicatifs ou informatifs qui, directement ou indirectement, fournissent des informations sur des "faits sociaux" comme les émotions sociales, les relations sociales [Brunet *et al.*, 2009].

L'analyse automatique de ces signaux est identifiée comme étant un verrou majeur pour la compréhension et la modélisation des interactions sociales. Un domaine émergent, appelé traitement du signal social (Social Signal Processing [1]) [Pentland, 2007] [Vinciarelli *et al.*, 2009b], se propose d'étudier avec une méthdologie inter-disciplinaire les comportements humains.

L'objectif de notre travail est de développer un modèle d'analyse des interactions longitudinales entre des parents et leurs enfants. Dans cet objectif, nous proposons dans ce chapitre de présenter dans un premier temps les grandes lignes du traitement du signal social en se focalisant sur les modes et signaux de communication (verbale et non verbale), l'analyse et la modélisation automatique de ces signaux. Dans un second temps, nous passerons en revue la spécificité du développement des enfants typiques et autistes afin d'identifier les signaux sociaux pertinents pour la modélisation de l'interaction longitudinale parent-enfant.

1.2 Communication

1.2.1 Définition

La communication est un processus de transfert d'information d'une source à l'autre. Elle peut être perçue comme un processus à double sens, où il y a échange et progression des pensées, des sentiments, des connaissances et des idées. Ainsi, elle est à la base de toute relation humaine car elle permet d'établir des relations avec autrui. Du fait de sa nature, la communication humaine est étudiée par des communautés différentes (anthropologie, psychologie, sciences du langage, sciences de l'ingénieur, etc). Comme tout processus de communication, la communication humaine requiert une source (émetteur), un message, un support ainsi qu'un destinataire. Le langage est reconnu pour être un vecteur d'information privilégié. Cependant, d'autres supports sont utilisés résultant dans une multitude de stratégies regroupées ici sous les termes de communication verbale et non-verbale.

[1] http ://sspnet.eu/

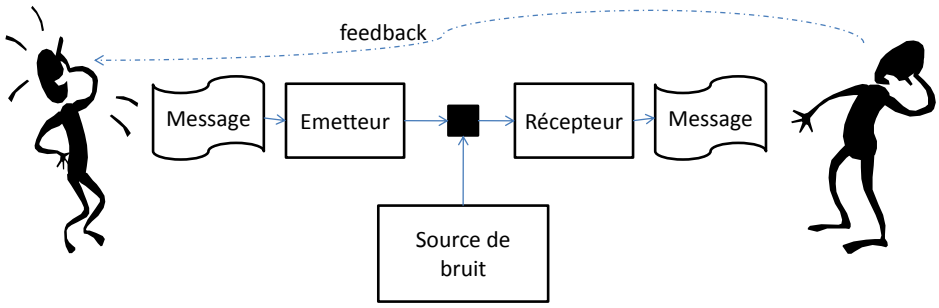


FIG. 1.1 – Architecture générale d'un processus de communication (schéma de Shannon)

1.2.2 Modèles de communication

Pour les premiers théoriciens, la communication se limite au transfert d'une information entre une source et une cible qui la reçoit. Elle est présentée comme un système linéaire et mécanique sans ancrage social. Il s'agit d'une conception télégraphique. Dans ce contexte, Shannon et Warren [1975] ont proposé un modèle de communication dans le but de transmettre un message à un destinataire avec un mimimum de distorsion. Ce modèle comprend six éléments (cf. figure 1.1) :

1. Une source d'information qui produit un message.
2. Un émetteur qui encode le message en signaux.
3. Un canal où circule les signaux codés.
4. Un récepteur qui décode le message à partir des signaux.
5. Une destination ou un destinataire qui reçoit le message.
6. Le bruit qui interfère dans la transmission du message.

Cependant, deux critiques majeures peuvent être adressées à ce modèle. Premièrement, il ignore totalement le fait que la communication est effectuée par des individus (ou des groupes) c'est-à-dire par des opérateurs sur lesquels vont intervenir de manière massive des facteurs psychologiques, des contraintes sociales, des systèmes de normes, des valeurs [Abric, 1996]. Deuxièmement, il pose la communication comme un processus linéaire et séquentiel, même si la réaction ou rétroaction du destinataire permet d'entretenir une boucle interactive [Abric, 1996].

Plus tard, pour tenir compte du fait qu'une communication est bidirectionnelle : le récepteur réagit et retourne de l'information à l'émetteur (réaction), Wiener [1964] a amélioré le schéma de communication proposé par Shannon

et Warren [1975] par l'ajout d'une boucle représentant le feedback (positif ou négatif) du destinataire. Dans ce cas l'émetteur et le récepteur sont considérés en interaction.

Ensuite, de nombreux modèles de communication ont été proposés dans la littérature, nous citerons le modèle de Lasswell [1948] qui s'est intéressé à la communication de masse (la communication de masse est l'ensemble des techniques qui permettent de mettre à la disposition d'un vaste public toutes sortes de messages). Selon lui, on peut décrire convenablement une action de communication en répondant aux questions suivantes : 'qui dit quoi ?, par quel canal ?, à qui ? et avec quel effet ?' L'intérêt essentiel de ce modèle est de dépasser la simple problématique de la transmission d'un message et d'envisager la communication comme un processus dynamique avec une suite d'étapes ayant chacune leur importance, leur spécificité et leur problématique. Il met aussi l'accent sur la finalité et les effets de la communication. Cependant, le processus de communication selon ce modèle est limité à la dimension persuasive. La communication est perçue comme une relation autoritaire. Il y a absence de toute forme de rétroaction, et le contexte sociologique et psychologique n'est pas pris en compte.

Enfin, en introduisant la notion de contexte et de rétroaction, certains chercheurs ont tenté de corriger les défauts de ces premiers modèles. Nous citons le modèle de Riley et Riley [1973]. Les auteurs ont apporté de nouvelles notions liées directement à la sociologie, le contexte et l'appartenance à un groupe. De plus ce modèle prend en compte la notion d'une boucle de rétroaction entre l'émetteur et le récepteur. Cela montre qu'il y a réciprocité et inter-influence entre les individus. Ce modèle est à l'origine des travaux sur la communication de groupe.

1.3 Stratégies de communication

La communication est un moyen très important pour partager les idées, elle permet d'exprimer à l'autre son attention, son amitié, l'amour, les sentiments, les connaissances, les opinions , l'intérêt et de transmettre des informations. La communication peut être divisée en deux catégories : communication verbale et communication non verbale. La communication verbale considère les informations en relation avec le langage, tandis que la communication non verbale utilisent des signaux dits non verbaux (gestes, posture, etc). La communication face à face ou multimodale exploitent de manière continue et multi-modale le verbal et le non verbal.

1.3.1 Communication verbale

La communication verbale (ou la parole) représente l'instrument le plus important que nous ayons à notre disposition pour rendre notre vie intéressante et bien transmettre nos messages. Sans la parole nous aurions beaucoup moins de possibilités de montrer aux autres ce que nous sommes, ce que nous pensons, ce que nous ressentons. Nous saurions aussi moins rapidement ce que les autres pensent, ce qu'ils ressentent vis-à-vis de nous et vis-à-vis des autres réalités, ce qu'ils désirent, ce qu'ils attendent de nous et de la vie en général. Par conséquent, nous développerions nos connaissances beaucoup plus lentement. Aussi, nos sentiments évolueraient plus lentement à cause du manque d'échanges avec les autres.

D'un point de vue ethnologique, la communication verbale est un ensemble de sons émis dans le but d'établir une communication avec autrui. Elle passe généralement par la parole. De plus, pour assurer une communication verbale, trois éléments sont particulièrement importants :

- L'opération de codage du message (mise en mots d'une pensée) : la bonne utilisation du vocabulaire et de la syntaxe disponibles sont importants. Ainsi, l'émetteur exprimera ses pensées avec précision.
- L'opération de décodage du message (opération inverse) : il faut que le récepteur maîtrise suffisamment le niveau de langage utilisé par l'émetteur (le code de communication doit être connu par l'émetteur et le récepteur).
- Le feedback (l'opération de rétroaction) qui permet les réajustements indispensables au passage effectif du message.

Paradoxalement, un résultat d'une analyse d'Albert Mehrabian's [Mehrabian's, 1972] montre que la communication verbale ne représente que 7% de ce qui est perçu par un individu, loin derrière la communication non verbale (mouvements, expression, façon de parler, regard), (voir figure 1.2).

1.3.2 Communication non verbale

La communication non verbale est simplement tout ce qui n'est pas en relation avec le langage. Il concerne les expressions faciales, gestes, postures, intonations de la voix ou émotions, apparences physiques, etc. Le but de la communication non verbale est de compléter le message verbal, ainsi elle permet de transmettre des messages complexes : un locuteur peut modifier et ajouter des informations au message verbal en utilisant des signaux non-verbaux. Par conséquent, la communication non verbale permet de renforcer et crédibiliser le message verbal lorsqu'elle est adaptée mais peut décrédibiliser ce même message si elle est inadaptée. Il est ainsi possible d'exploiter la concordance

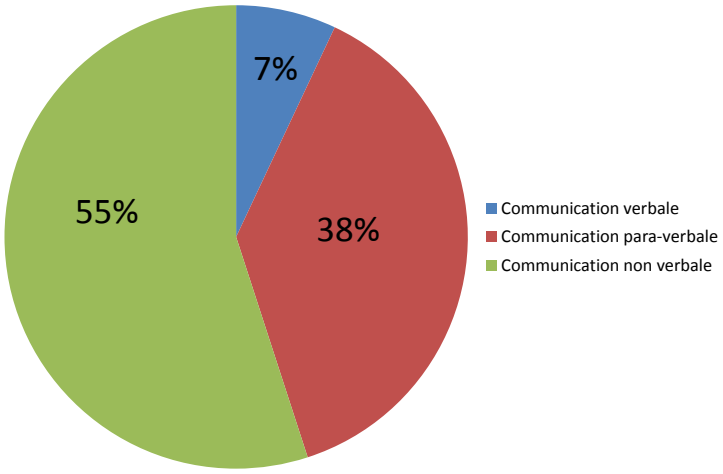


FIG. 1.2 – Distribution en pourcentage de communication verbale, para-verbale et non verbale

ou non du verbal et non verbal pour transmettre un message tout autre.

La communication non verbale comprend un ensemble vaste et hétérogène de processus ayant des propriétés communicatives, en commençant par les comportements plus manifestes et macroscopes comme l'aspect extérieur, les comportements de relation spatiale avec les autres (rapprochements, prises de distance) et les mouvements du corps (du tronc, des membres ou de la tête), jusqu'aux activités moins évidentes ou plus fugaces, comme les expressions faciales, les regards et les contacts visuels, les intonations vocales, etc. Les composantes non verbales de la parole vont au delà des mots prononcés, elles sont : le ton, la hauteur (aiguë ou grave) l'amplitude de la voix, le timbre (rauque ou clair), l'intensité (forte ou faible), les intonations, le rythme et le débit de la parole. Ces signaux ont un effet positif sur la communication [Mehrabian's, 1972].

Le message non verbal permet également d'établir une communication entre des personnes de langues différentes. Par contre, les signaux non verbaux ne sont pas tous universels et ils doivent être interprétés en fonction du contexte [Ekman, 1999b] : le rire et l'expression de la douleur sont les expressions non verbales les plus universelles. De plus, la signification d'un geste dépend de la situation, de l'émetteur, du récepteur, de la culture, de la religion, etc. Donc, un comportement non verbal peut avoir plusieurs significations d'une personne à l'autre. Selon Cormier [2000], il faut avoir une connaissance approfondie de la personne avec qui nous interagissons afin de

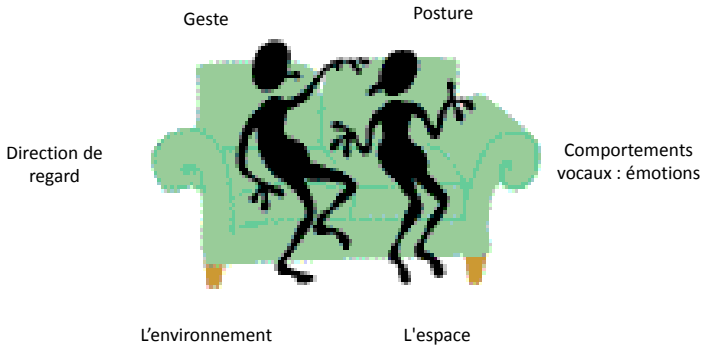


FIG. 1.3 – Communication face à face

permettre de décoder plus justement ces comportements non verbaux.

1.3.3 Communication face à face

La communication face à face demeure la première forme de communication, elle regroupe les deux formes de communication : la communication verbale et la communication non verbale. La communication face à face est une procédure d'échange, verbal et non verbal, synchrone et continu dans le temps préétablie entre deux interlocuteurs se faisant face (cf. la figure 1.3). Ces précisions quelque peu fastidieuses en apparence sont nécessaires pour distinguer la conversation proprement dite, dont les participants sont en contact à la fois visuel, auditif, mais aussi olfactif et parfois tactile (contrairement à la conversation téléphonique, par exemple, ou à l'échange de messages sur un babillard électronique), et qui donc met en oeuvre, outre le langage, plusieurs systèmes signifiants. Dans le processus de la communication face à face, une grande partie de l'information acquise par le destinataire est dérivée de signaux non verbaux tels que la posture et les expressions faciales. Par conséquent, l'information arrive complète et moins ambiguë au destinataire grâce à la complémentarité des messages verbaux et non verbaux. Les signaux non verbaux permettent d'exprimer des idées qui sont souvent mieux décrites par les gestes. Cependant, en dehors de cette forme de communication, le message est incomplet, par conséquent le destinataire sera obligé d'interpréter les informations manquantes.

L'évidence et l'intensité des relations entre parole et mouvements corporels sont étayées par les données et les analyses en provenance des sciences cognitives et des neurosciences. Dans le cadre de l'interaction face à face, l'interlocuteur a toujours tendance à parler et en même temps à bouger sa tête et/ou

son corps pour mieux attirer l'attention du destinataire [Rimé, 1991]. Cette corrélation entre la parole et les mouvements corporels est très importante, par exemple dans des cas de handicap comme l'aphasie, trouble du langage parlé, la parole et les mouvements labiaux s'accompagnent généralement par des gestes et des mouvements corporels pour éclaircir les informations.

L'avantage de la communication face à face est que le locuteur reçoit d'une façon continue un feedback multimodale du destinataire. Le feedback correspond à toute information qui, partant du destinataire, revient à l'émetteur ; donc, toute réaction verbale ou non verbale au comportement de l'autre constitue une forme de feedback. Vu sous cet angle, d'une manière implicite, le feedback est continuellement présent dans la communication face à face. Ces comportements encouragent l'émetteur à continuer son discours, par exemple quand le destinataire manifeste un intérêt pour la communication, il fait un hochement de tête ou garde un contact visuel.

1.4 Signaux de communication

Différents signaux : la parole, les gestes, les mouvements, les émotions (vocales et faciales) sont connues pour être les supports de la communication verbale et non-verbale.

1.4.1 Signaux verbaux de communication

La communication verbale est caractérisée par la mise en commun de signes linguistiques avec des significations précises. Le langage est le mode de communication verbal privilégié de l'humain. En partageant un protocole commun basé sur les phonèmes, les mots ou bien encore des phrases, la communication verbale a un statut social spécifique permettant de caractériser un individu par sa langue, sa culture ou bien encore son histoire. Du fait de son codage de l'information, la langue des signes appartient à la catégorie de la communication verbale.

1.4.2 Signaux non-verbaux de communication

La communication non-verbale est basée sur l'implicite et elle joue un rôle très important dans le processus d'interaction [Pantic *et al.*, 2006] [Pantic *et al.*, 2008]. Plusieurs supports sont utilisés pour l'expression d'un message non-verbal comme les expressions faciales, le contact visuel, des gestes, la posture et le langage corporel. Les vêtements et la coiffure jouent également un rôle important dans cette catégorie de communication. Les éléments para-

FIG. 1.4 – Expressions faciales

linguistiques incluant la qualité vocale, l'intonation et le style de la parole, peuvent influer sur la communication et compléter le message verbal.

1.4.2.1 Expressions faciales

Le visage contient la plupart des récepteurs sensoriels et des effecteurs sensori-moteurs (les yeux, les oreilles, la bouche et le nez). Il est impliqué dans plusieurs activités comme la reconnaissance de l'identité, la production de la parole, la communication des états affectifs par les expressions faciales, etc., Le visage est considéré comme un miroir authentique de la personnalité, l'âge, le sexe et d'autres caractéristiques peuvent êtres extraits à partir du visage [Ambady et Rosenthal, 1992]. Ainsi, c'est une interface multimodale (émetteur/récepteur) des échanges émotionnels, par l'intermédiaire surtout des expressions faciales, qui jouent un rôle majeur dans la communication non verbale [Ambady et Rosenthal, 1992] [Grahe et Bernieri, 1999] [Mehrabian et Ferris, 1967]. Le visage est donc la source de signaux essentiels et nécessaires pour assurer une communication interpersonnelle dans la vie sociale [Keltner et Ekman, 2000].

Selon Paul Ekman, il existe six expressions faciales distinctes (colère, dégoût, peur, joie, tristesse et surprise) [Cohn, 2006], voir la figure 1.4. Ekman a décrit précisément chacune de ces expressions. [Ekman et Friesen, 1978]. La

taxonomie FACS, Facial Action Coding System, est une méthode de description des mouvements du visage décomposés en unités d'action (UA). Il s'agit d'une norme commune qui permet de catégoriser systématiquement l'expression physique des émotions, et elle s'est avérée très utile pour les psychologues et les thérapeutes dans la prise en charge de patients n'ayant pas accès au langage (niveau de douleur, dépression). Les unités d'action (AU - Action Unit) correspondent à l'action visible d'un muscle ou groupe musculaire. La taxonomie FACS constitue une description objective des signaux faciaux décrits par 46 mouvements élémentaires indépendants ou unités actions.

1.4.2.2 Le comportement de la voix

Le langage parlé est un moyen complexe de communication. Il y a en effet ce que dit le sujet (le sens des mots) et la façon dont il le dit, laquelle peut être, ou ne pas être, en accord avec le sens des mots. Ces paramètres non verbaux sont appelés paralinguistiques ou encore paraverbaux. Généralement, on considère que les caractéristiques non verbales du langage parlé sont les suivantes :

- Qualité vocale
- Vocalisations non-linguistiques
- Silences
- Pauses
- Tour de parole

Qualité vocale La qualité vocale est une notion complexe à définir du fait de sa nature subjective. Cependant, elle est généralement reliée à des paramètres acoustiques et prosodiques tels que les formants, le pitch, le tempo ou bien encore l'intensité avec des corrélats physiques plus ou moins fidèles : spectre, fréquence fondamentale $F0$, durée et énergie. La dynamique de ces paramètres permet au locuteur de transmettre des informations différentes (par exemple : l'ironie). L'interlocuteur analyse continuellement ces paramètres pour extraire également des informations (par exemple l'état émotif). De plus, pour souligner un mot ou une idée l'interlocuteur aura tendance à modifier sa voix (le rythme et/ou l'intensité) [Hirschberg, 1993] pour structurer son discours [Hirschberg et Grosz, 1992].

Vocalisations non-linguistiques Les vocalisations non-linguistiques appelées aussi ségrégations jouent un rôle important pour compléter le message verbal. Il s'agit de mots comme “ehm", “ah-ah", “uhm"'. Les fonctions de ces vocalisations non-linguistiques sont multiples. Elles sont utilisées pour remplacer un mot que l'interlocuteur n'a pas pu trouver, par exemple, quand

l'interlocuteur ne sait pas comment répondre à une question, il prononce un "ehm" prolongée. Elles peuvent servir de rétroaction (feedback) afin d'entretenir l'interaction : montrer son accord avec le locuteur ou son intérêt à la communication [Shrout et Fiske, 1981].

Silences et pauses Le silence représente un instrument de communication qui joue un role très important dans l'interaction [Keltner et Haidt, 1999]. Cependant, il est difficile de l'interpréter, car il interfère avec les autres signes, avec le type de relation interpersonnelle, la situation communicationnelle et la culture de référence. Plusieurs études mettent en évidence l'importance du contexte dans l'interprétation du silence. Par exemple, si le silence est accompagné du détournement du regard ou de la tête, cela peut indiquer que l'on désire finir ou interrompre la communication. Par contre, le silence est indispensable pour la communication face à face car il joue souvent le rôle d'un instrument permettant d'attirer l'attention quand la personne qui est supposée commencer à parler, oblige le ou les partenaires à l'écouter plus attentivement. On distingue deux types de silence [Bruneau et Francine, 1973] [Sullivan *et al.*, 1990] :

- Silences psycholinguistes : cette forme est liée aux hésitations syntaxiques et grammaticales de courte durée ou aux ralentissements qui accompagne le décodage du discours. Le silence psycholinguiste est produit quand l'interlocuteur a besoin de temps pour réfléchir à ses idées et formuler ses phrases. Cette forme survient surtout au début de l'intervention parce que l'interlocuteur a besoin de temps pour reformuler ses mots. Dans ce sens, c'est un signe de difficulté de prise de parole.
- Silences interactifs : les silences interactifs sont des pauses dans une conversation. Ils peuvent être liés à des rapports affectifs ou personnels aussi bien qu'à l'échange d'informations et/ou à la résolution d'un problème. Ils sont constamment utilisés dans les échanges et communications au sein de petits groupes d'individus. De plus, les silences interactifs sont généralement plus longs que les silences psycholinguistiques. La différence entre le silence interactif et le silence psychologique réside principalement dans le fait que chaque participant a conscience du degré et de la façon dont on attend de lui qu'il participe à la communication. Cette forme de silence peut jouer un rôle de respect des autres pour les écouter, ou pour ignorer quelqu'un qu'on n'a pas envie de répondre, mais aussi pour attirer l'attention vers d'autres comportements non verbaux comme le regard ou les expressions faciales.

Tour de parole Un autre comportement vocal aussi important est le tour de la parole [sathas, 1995].Un tour de parole est défini par la possibilité dont

bénéficie un interlocuteur de prendre la parole dans le cadre d'une conversation. Le tour de parole a une double fonction : la régulation de la conversation et la coordination entre les différents interlocuteurs. Le nombre total de tours de parole est une indication globale de la participation de l'interlocuteur à la conversation. Ce nombre donne des indices sur la personnalité de l'interlocuteur et son degré de présence et de domination dans une conversation.

1.4.2.3 Gestes, postures et langages corporels

Le langage corporel se manifeste par des postures qui peuvent concerner : la tête, le buste, le bassin, les jambes et les bras. Par les gestes, nous nous exprimons et nous pouvons avoir un comportement de défense ou d'agression. Les gestes sont souvent utilisés pour synchroniser l'interaction ou transmettre un message (la parole n'est pas toujours nécessaire), par exemple le doigt pointé vers la porte pour demander de sortir, le signe de la main pour dire au revoir, le hochement de tête pour dire oui ou le battement de mains (applaudissement) pour montrer notre satisfaction devant une manifestation.

De nombreuses études ont été menées sur les postures et les gestes de l'homme pour communiquer ses intérêts et ses émotions [Coulson, 2004][Van den Stock *et al.*, 2007]. Toutes ces études affirment que les comportements corporels de l'homme changent avec son état émotionnel. Quelques recherches [Costa *et al.*, 2001] [Ekman et Rosenberg, 2005] ont montré également que des gestes comme l'inclinaison de la tête ou toucher le visage accompagnent souvent les états affectifs comme la honte et l'embarras. Cependant, le langage corporel peut prendre des orientations plus diverses, ce qui le rend moins formel et plus difficile à analyser que le langage verbal, mais le rend dans le même temps plus riche et complet.

La majorité des gestes sont produits en accompagnement du message verbal (parole) [McNeill, 1996]. Certains gestes répètent l'information verbale pour la rendre davantage compréhensible, ils sont bien souvent instructifs, et ponctuent les propos (par exemple, quand une personne indique un chemin à suivre, ses gestes miment le trajet à parcourir). Par contre, les gestes comme de nombreux comportements non verbaux ont parfois un effet inverse, d'autres peuvent être sans rapport avec le discours (manipuler un objet, ses cheveux, se gratter).

1.4.2.4 Le contact visuel

Le regard est souvent le premier sens utilisé pour entrer en contact face à face. Dans la communication publique, la fréquence des contacts visuels affecte le récepteur du message, ainsi l'audience préfère les interlocuteurs qui main-

tiennent un bon contact visuel. L'influence du regard pendant l'interaction est multiple comme par exemple :

- Un clin d'oeil : il indique que ce qui est dit ne doit pas être pris au sérieux,
- Un regard soutenu : intention hostile,
- Un regard panoramique : comme son nom l'indique, il parcourt lentement tous les interlocuteurs - (implication de tous dans un message apparemment individuel).

Ainsi, il est important de tenir compte de ces aspects lors d'une communication face à face afin d'éviter des malentendus. Dans une interaction face à face l'orientation du regard peut refléter facilement la pensée de l'interlocuteur.

Une autre capacité incluse dans le contact visuel est l'attention conjointe. Elle permet d'orienter l'attention d'autrui vers un objet, une action et constitue un moyen et un premier marqueur explicite de la communication intentionnelle.

1.4.2.5 Exploitation de l'espace et de l'environnement

La situation des interlocuteurs dans l'espace constitue aussi une part des communications non verbales. C'est un comportement non produit directement par l'interlocuteur mais qui est très important pour la communication face à face. Les distances interpersonnelles varient selon l'âge, le sexe, le statut social et la culture des interlocuteurs. Par ailleurs, la distance entre deux interlocuteurs reflète la qualité de leur relation. Hall [1959] a étudié les distances entre les interlocuteurs et a montré qu'elles ont une dimension sociale. Il a ainsi dénombré 4 zones dans lesquelles les humains opèrent en général (voir la figure 1.5) :

- Distance intime : de 0 à 40 cm, favorisant le contact immédiat.
- Distance personnelle : de 40 cm à 1 m, relations individuelles privées.
- Distance sociale : de 1 m à 2 m, qui se manifeste au sein des petits groupes et qui est souvent déterminée par le statut des personnes.
- Distance publique : à partir de 3 m, qui est celle des acteurs ou orateurs.

Les distances interpersonnelles varient beaucoup d'une culture à une autre.

1.4.2.6 Le toucher

Le toucher est un mode de communication qui est plus ou moins utilisé selon les cultures et les civilisations. Par exemple dans les sociétés occidentales, il est réservé aux intimes. Le toucher peut-être utilisé pour signifier une relation professionnelle, une relation sociale, une amitié, de l'intimité, de l'excitation sexuelle. Ainsi, dans la culture occidentale, une poignée de main molle

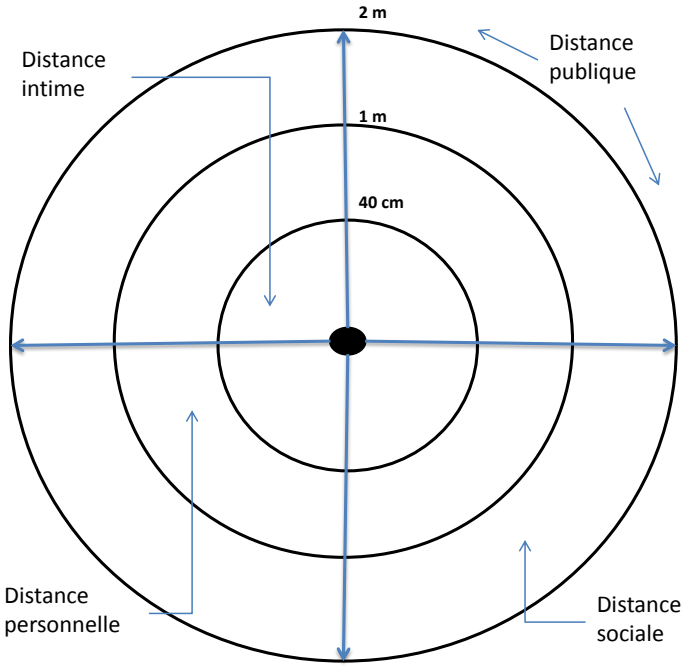


FIG. 1.5 – Situation des individus dans l'espace

évoque des sentiments négatifs, c'est une interprétation d'un manque d'intérêt ou de vitalité. Tandis qu'une main moite est souvent considérée comme un signe d'anxiété. Le toucher est essentiel pour le développement physique et psychologique des enfants et pour le bien-être émotionnel des adultes. Enfin, on utilise le toucher pour influencer les autres et pour améliorer la qualité de l'interaction.

1.4.2.7 Apparence physique

L'apparence physique est l'allure générale d'un individu que l'on voit en premier lieu. Il s'agit de la taille, la forme du corps, la physionomie, la couleur de la peau et des cheveux. D'autres caractéristiques artificielles telles que les vêtements, le maquillage peuvent être utilisées pour modifier le visage et le corps. La caractéristique la plus importante dans l'apparence physique est l'attractivité. Ainsi, des personnes sont jugées plus intelligentes que d'autres uniquement sur la base de leur attrait physique [Schneider *et al.*, 2005], ce phénomène est appelé l'effet de 'halo' [Thorndike, 1920]. Dans le milieu professionnel, la personne dont l'apparence physique est attirante remporte plus de succès professionnels et affectifs et est pleine de qualités. Par conséquent, les personnes attractives sont souvent considérées comme ayant un statut élevé

et une bonne personnalité, même s'il n'y a aucune base objective pour valider de tels jugements [Greenwald et Banaji, 1995], [Warner et D.B., 1986]. Par ailleurs, il existe d'autres caractéristiques nécessaires pour assurer l'attractivité, par exemple les choix des vêtements qui se font généralement en fonction de l'âge, de la morphologie et de la situation professionnelle, sachant que par le choix de notre tenue, nous donnons déjà une certaine image de nous-mêmes.

1.5 Analyse des signaux sociaux

1.5.1 Etat de l'art

Les signaux sociaux correspondent aux comportements qui sont produits et échangés par les humains dans le cadre des interactions sociales pour accompagner et compléter les messages verbaux. Par ailleurs, l'analyse et l'interprétation automatiques de ces signaux sont identifiées comme étant des verrous majeurs de la compréhension des interactions sociales. L'analyse des signaux sociaux regroupe divers domaines de recherche qui sont en relation avec le traitement automatique des comportements humains.

Les méthodologies mises en oeuvre pour l'analyse automatique de signaux sociaux exploitent, généralement, celles du traitement du signal et de la reconnaissance des formes. Les étapes communes à la majorité des systèmes sont (voir la figure 1.6) :

- l'acquisition des données,
- la détection des personnes dans les scènes,
- l'extraction des signaux sociaux,
- l'interprétation des signaux sociaux par rapport au contexte de l'interaction.

1.5.1.1 Acquisition de données

Le domaine de l'analyse des interactions sociales requiert des données, des enregistrements de situations nécessaires à la mise en place de modèles computationnels. L'acquisition de données interactionnelles en adéquation avec les objectifs de recherche constitue un sujet d'étude en soi. En revanche, pour avoir un bon modèle d'interaction, il est indispensable d'avoir des données réelles (real-life). Ainsi, la base de données doit idéalement illustrer des interactions vécues dans des contextes réels. Or, les interactions vécues en contextes réels sont imprévisibles et il est difficile de contrôler ce qui est collecté. Les microphones et les caméras sont les capteurs privilégiés pour l'acquisition des signaux sociaux avec des structuration diverses (capteur fixe, en mouvement, en réseau) [Vinciarelli et Odobez, 2006] [McCowan *et al.*, 2003] [Waibel *et al.*,

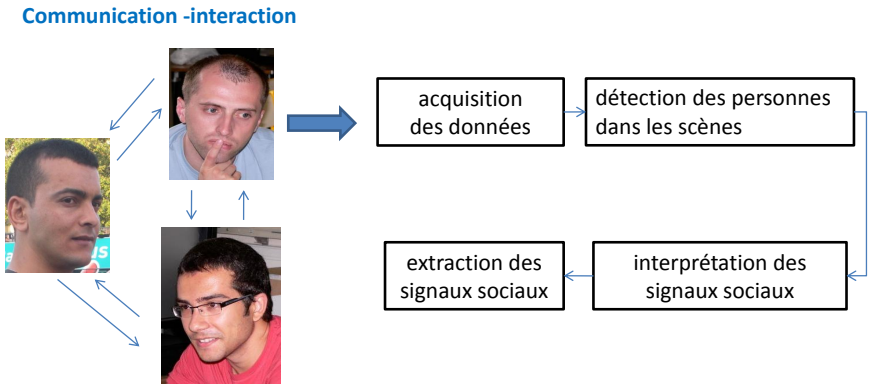


FIG. 1.6 – Différentes étapes pour l'analyse des signaux sociaux

2003]. La littérature montre que d'autres capteurs peuvent se révéler pertinents pour l'analyse des comportements : activités physiologiques [Gunes *et al.*, 2008a], téléphones cellulaires [Eagle et Pentland, 2006].

Après la phase d'acquisition des données d'interactions, la deuxième étape consiste à identifier les personnes et annoter leurs comportements (tour de parole, émotion, expressions faciales, gestes, etc). L'annotation se fait principalement de manière manuelle résultant dans une augmentation des ressources humaines. Des approches semi-automatiques sont récemment apparues, elles consistent à annoter manuellement une partie de données puis à se servir de ces annotations pour apprendre un système d'annotation automatique. L'annotation manuelle étant très coûteuse, nous avons proposé dans ce travail de thèse une méthodologie basée sur l'apprentissage semi-supervisé (chapitre 3), exploitant que très peu de données étiquetées.

1.5.1.2 Détection de personnes dans les scènes

La détection de personnes s'effectue à travers deux canaux : audio et vidéo. Concernant l'audio, il s'agit de segmenter le flux audio de telle sorte que chaque segment ne contienne qu'un seul locuteur. Tandis que pour le canal vidéo, la tâche consiste à détecter les personnes, leurs visages.

Détection audio Le phénomène de détection de personnes s'appelle segmentation en locuteurs ou encore indentification de locuteurs si on dispose déjà des informations à priori sur les différents interlocuteurs. La segmentation en locuteurs consiste à proposer pour un flux audio un descriptif définissant le

nombre de locuteurs et l'intervention de chacun d'entre eux dans l'interaction. Elle permet de segmenter le flux de parole en des fenêtres homogènes, c'est-à-dire ne contenant les productions verbales que d'un seul et même locuteur [Tranter et Reynolds, 2006].

La plupart des travaux en segmentation de locuteur utilisent des méthodes statistiques et probabilistes pour modéliser les différentes caractéristiques à rechercher. Les travaux fondamentaux définissant la segmentation en locuteurs ont été réalisés par Siu *et al.* [1991]. Les auteurs ont proposé une architecture reprise et complétée par la majorité des autres systèmes de segmentation. Le processus se décompose en trois phases distinctes : la paramétrisation, la détection de ruptures et la classification. De nombreux chercheurs ont complété cette architecture avec une phase finale de resegmentation du signal [Reynolds *et al.*, 2000] [Moraru, 2002]. Les éléments de cette architecture jouent les rôles suivants :

- La paramétrisation : convertir le document sonore (l'enregistrement) en paramètres acoustiques (le plus souvent en utilisant le codage *MFCC*).
- La détection de ruptures : délimiter le flux de parole en segments ne contenant les productions que d'un seul locuteur.
- La classification : regrouper les segments locuteur par locuteur jusqu'à découvrir le nombre de classes correspondant au nombre de locuteurs intervenant dans le flux de parole. A la fin de la classification, chaque classe contient l'ensemble des segments prononcés par un locuteur. La classe définit alors l'intervention du locuteur dans le signal audio.
- Enfin, la re-segmentation affine les frontières des segments. L'enregistrement est de nouveau découpé en segments en fonction des données contenues dans les classes.

Les phases de regroupement et de classification s'effectuent à l'aide des techniques décrites dans [Johnson et Woodland, 1998] [Reynolds *et al.*, 1998].

Détection vidéo La détection de personnes consiste à détecter leurs visages ou complètement leurs corps. La détection de visage est généralement la première étape d'analyse des expressions faciales [Pantic et Rothkrantz, 2000] ou de l'estimation de l'orientation de la tête [Bailly et Milgram, 2009]. La détection du corps est très étudiée dans le domaine de la vidéo surveillance automatique (station de métro, entrée d'une entreprise, etc.) [Gavrila, 1999] [Moeslund et Granum, 2001]. Dans le cadre de l'étude de l'interaction sociale, la détection de personnes permet d'identifier les interlocuteurs (localisation, caractérisation).

Dans la littérature, on distingue plusieurs travaux portant sur la détection de visage. Les approches classiques de détection du visage sont décrites dans [Rowley *et al.*, 1998] [Sung et Poggio, 1998]. L'approche la plus répandue

consiste à catégoriser les pixels de l'image en tant que visage ou non-visage. L'inconvénient de cette approche réside dans le temps de calcul qui ne permet pas souvent de faire des traitements en temps réel. Et de manière plus générale, la définition de la classe non visage est un problème connu pour être complexe. D'autres approches ont été développées pour la détection de visage basées sur la couleur de peau [Hsu *et al.*, 2002] [Garcia et Tziritas, 1999], ces approches cherchent à détecter des éléments caractéristiques du visage (yeux, bouche, nez, contour de la tête) puis à combiner les résultats de ces détections à l'aide de modèles géométriques et radiométriques, notamment via des modèles déformables. D'autres approches consistent à déterminer des points d'intérêts (maxima locaux d'un filtre gaussien aux dérivées secondes). Enfin, les méthodes dites par apprentissage, comme l'algorithme de Viola et Jones [2004] sont les plus exploitées. Cet algorithme est basé sur un classifieur de type Adaboost [Freund et Schapire, 1997] et est composé de deux étapes : l'apprentissage du classifieur à partir d'un grand nombre d'exemples positifs (les objets d'intérêts, i.e. des visages) et négatifs, et l'application de ce classifieur lors d'une phase de détection.

La détection automatique de personnes est un problème non complètement résolu. Cependant de nombreux systèmes requièrent des informations sur la présence ou l'absence de personnes dans leur environnement. Les applications potentielles d'un système de détection automatique sont nombreuses. Nous pouvons par exemple citer les systèmes de transport intelligent, la robotique, la surveillance, l'indexation d'images (ou de vidéos), l'analyse d'interactions. D'une manière générale, le principe de la détection de personnes dans une image ou une vidéo est le même. Pour une image, le détecteur ne dispose d'aucune information a priori et doit donc parcourir l'image entière avec une fenêtre de détection. Cependant, pour une vidéo, il est possible de simplifier le problème en se focalisant par exemple sur les zones de l'image où il y a eu un mouvement. Il est également possible de raisonner en fonction de la position des personnes présentes dans les images précédentes. À l'instar de la détection de visage, les méthodologies employées pour la détection de personnes exploitent soit des modèles explicites (2D ou 3D) du corps [Zhao *et al.*, 2006], soit des techniques d'apprentissage supervisé. À partir d'une base d'images, des caractéristiques de la forme du corps humain sont extraites et un modèle discriminant est construit. Nous pouvons citer par exemple les travaux de Papageorgiou *et al.* [1999] qui ont proposé un détecteur basé sur les ondelettes de *Haar* et les machines à vecteurs de support (*SVM*). Viola et Jones [2004] ont également proposé un détecteur basé sur les filtres de *Haar* et l'algorithme de boosting. Ensuite, Dalal et Triggs [2005] ont utilisé avec succès les histogrammes de gradients orientés dans le cas de la détection de piétons. D'autres méthodes détectent séparément différentes parties du corps

humain et fusionnent ensuite ces résultats [Wu et Nevatia, 2006].

1.5.1.3 Détection automatique de signaux sociaux

Après la détection de personnes dans les scènes d'interactions, la seconde étape consiste à extraire les informations audiovisuelles et les différents comportements des interlocuteurs (gestes, mouvements, émotions, etc.).

Détection de signaux sociaux à partir de l'apparence physique Très peu d'études s'intéressent à l'analyse de l'apparence physique de l'humain. Parmi les objectifs de ces études est la mesure automatique de beauté faciale [Aarabi *et al.*, 2001] [Eisenthal *et al.*, 2005] [Sepheri *et al.*, 2006]. Pour cela, Aarabi *et al.* [2001] ont développé un système qui détecte séparément les différents éléments du visage (nez, lèvres, yeux, etc.). Ensuite ils ont développé un modèle de visage basé sur la distance entre les différents éléments faciaux et leurs dimensions en utilisant des méthodes d'apprentissage automatique. D'autres approches ont été proposées dans la littérature, en se basant sur la forme du corps, la couleur de peau et les cheveux pour la tâche d'identification et de suivi de personnes [Ben Abdelkader et Yacoob, 2009] [Darrell *et al.*, 2000]. Par contre, tous ces travaux ne s'adressent pas au problème d'étude de l'interaction sociale.

Détection de signaux sociaux à partir des gestes et des postures La reconnaissance de gestes est un domaine très actif et représente une application très importante de la vision par ordinateur et de la reconnaissance des formes. L'interprétation automatique de l'implication du geste sur l'interaction sociale est un domaine très étudié. Par ailleurs, la plupart des applications de la reconnaissance des gestes sont dédiées à la commande automatique et à l'interaction homme-machine [Sepheri *et al.*, 2006] [Wu et T.S., 1999]. Dans le même contexte, la reconnaissance de la langue des signes est une catégorie d'application proche de l'analyse de l'interaction [Kong et Ranganath, 2008].

De nombreuses méthodes de reconnaissance de gestes existent dans la littérature [Pavlovic *et al.*, 1997] [Bretzner *et al.*, 2002]. Elles font le plus souvent apparaître une étape d'extraction et une étape de classification s'appuyant généralement sur un apprentissage effectué autour d'un alphabet réduit de gestes. L'analyse des gestes suppose que l'on fasse l'analyse spatiale et temporelle des mouvements. Nous avons retenu deux approches : une première basée sur l'histogramme des orientations du gradient [Freeman et Roth, 1995] et une seconde, utilisant la superposition des squelettes de la main et de l'avant-bras, calculés sur l'ensemble de la séquence. L'histogramme des orientations du gradient est utilisé comme une signature statique calculée sur la première et la

dernière image de la séquence. Ces signatures statiques permettent de délimiter le geste dynamique. Tandis que, le squelette est utilisé comme une signature dynamique calculée sur chaque image de la séquence. C'est cette signature qui assure la reconnaissance des gestes. Ensuite, l'interprétation de la dynamique d'une main effectuant un geste permet de définir une trajectoire en considérant un ensemble de paramètre dans un espace donné, ce qui permet d'inférer la signification de cette trajectoire.

Du point de vue reconnaissance des postures humaines, les travaux existants s'intéressent principalement à la vidéo surveillance [Gandhi et Trivedi, 2007] ou à la reconnaissance d'activité [Wang et Hu, 2003]. La reconnaissance des postures humaines a été également appliquée à l'étude de l'interaction sociale [Argyle, 1975] [Coulson, 2004] [Gunes *et al.*, 2008b] [Vinciarelli *et al.*, 2009c]. Coulson [2004] a défini une correspondance entre certaines postures statiques et les six émotions (peur, tristesse, joie, etc). Il suppose que la reconnaissance des émotions à partir de postures humaines est comparable à reconnaissance des émotions faciales ou vocales [Castellano *et al.*, 2007]. Certains travaux montrent que la détection de l'émotion 'triste' à partir des postures humaines serait plus performantes que celles des émotions 'fâché' ou 'peur' [Van den Stock *et al.*, 2007].

Détection des signaux sociaux à partir du regard et des expressions faciales Les travaux de recherche en psychologie ont démontré que les expressions faciales et les mouvements du regard jouent un rôle important dans la coordination de la conversation humaine [Boyle *et al.*, 1994], et ont un impact plus important sur l'auditeur que le contenu verbal du message exprimé. L'analyse automatique des expressions faciales et de l'orientation du regard constitue une étape importante pour l'étude des comportements. Les travaux de recherche sur l'analyse des expressions faciales et du mouvement du regard se situaient principalement dans le cadre de la psychologie. Les progrès effectués dans des domaines connexes tels que la détection, le suivi et la reconnaissance de visages [Zhao *et al.*, 2000] ainsi que dans le traitement d'images et la reconnaissance des formes, ont apporté une contribution significative permettant la proposition de solutions automatiques.

Avant de procéder à l'analyse d'une expression faciale sur une image ou sur une séquence vidéo, il faut d'abord détecter et suivre le visage observé [Yang *et al.*, 2002] [Viola et Jones, 2004] puis extraire les informations pertinentes pour l'analyse de l'expression faciale. Il existe différentes méthodes pour l'extraction de ces informations. Les méthodes globales modélisent le visage dans son intégralité alors que les méthodes reposant sur des traits caractéristiques permettent une modélisation locale du visage dans les régions susceptibles de se modifier selon les expressions faciales affichées. Après avoir détecté, le

visage et extrait les informations pertinentes, l'étape suivante consiste à identifier l'expression faciale affichée.

Dans le but de représenter le mouvement intérieur au visage, lorsque celui-ci est détecté dans l'image, Black et Yacoob [1997] utilisent des modèles locaux paramétriques. Ils estiment le mouvement relatif des traits faciaux dans le repère du visage. Les paramètres de ce mouvement servent par la suite à représenter l'expression faciale [Ekman, 1999a]. De manière similaire, Cohn *et al.* [1998] utilisent un algorithme hiérarchique pour effectuer le suivi des traits caractéristiques par estimation du flot optique. Les vecteurs de déplacement représentent l'information sur les changements d'expression faciale. Padgett *et al.* [1996] utilisent des gabarits d'oeil et de bouche, calculés par analyse en composantes principales d'un ensemble d'apprentissage, en association avec des réseaux de neurones. D'autre part, Hong *et al.* [1998] utilisent un modèle global basé sur des graphes étiquetés construits à partir de points de repère distribués sur le visage. Les noeuds de ces graphes sont formés par des vecteurs dont chaque élément est la réponse à un filtrage de *Gabor* extraite en un point donné de l'image. Finalement, Cootes *et al.* [2001] utilisent une représentation par modèle actif d'apparence (AAM) pour extraire automatiquement des paramètres caractérisant un visage.

Détection de signaux sociaux à partir des comportements vocaux
Le langage parlé a été largement étudié par la communauté du traitement automatique de la parole (TAP) résultant dans des avancés significatives dans le domaine de la reconnaissance de la parole, du locuteur et de la langue et plus récemment de la reconnaissance des émotions. Cependant, peu des travaux sont dédiés à l'étude de l'interaction sociale et de l'effet des comportements sur la qualité de l'interaction face à face. Les vocalisations non linguistiques sont généralement considérées comme des sources de bruits.

Les comportements vocaux sont généralement caractérisés par des descripteurs prosodiques : F0, énergie, durée [Crystal, 1969]. Dans le domaine de traitement de parole, la fréquence fondamentale caractérise les parties voisées du signal de parole et est liée à la sensation d'hauteur de la voix (aigue ou grave). Plusieurs méthodes existent pour l'estimation et l'évaluation de la fréquence fondamentale [Rabiner et Schafer, 1978] [Huang *et al.*, 2001] : les méthodes temporelles (autocorrélation, fonctions de différences moyennées (ASDF - Average Square Difference Function)), les méthodes d'estimation par maximum de vraisemblance, les méthodes reposant sur une analyse du cepstre, etc. [Calliope, 1997]. Les logiciels Praat [Boersma et Weenink, 2001] et Wavesurfer [Sjolander et Beskow, 2000] sont couramment utilisés pour l'extraction de la fréquence fondamentale. Un autre descripteur important est l'énergie du signal de parole qui permet de mesurer l'intensité sonore (faible ou forte).

Le rythme est une caractéristique intéressante et se traduit par des mesures diverses comme le nombre de voyelles par seconde [Pfau et Ruske, 1998] [Ringeval et Chetouani, 2008]. L'ensemble de ces descripteurs est souvent utilisé pour caractériser le signal de parole et particulièrement la partie voisée du signal.

Certains comportements para-linguistiques ont été étudiés dans le domaine du traitement des signaux de parole comme le rire [Truong et Leeuwen, 2005] [Truong et van Leeuwen, 2007] et les cris [Möller et Schönweiler, 1999] [Pal *et al.*, 2006]. Truong et van Leeuwen [2007] ont développé un système de détection automatique de rire à partir d'un signal audio, ils ont utilisés différents descripteurs prosodiques et acoustiques ainsi que différentes techniques d'apprentissage automatique comme les réseaux de neurones, machines à support vecteur (SVM) et les modèles de mélange des gaussiennes (GMM). Ils ont pu montrer que les meilleurs descripteurs pour la discrimination entre la parole normale et le rire se trouvent dans le domaine fréquentiel. Cependant, ils n'ont exploité que le flux audio pour la discrimination entre la parole et le rire. Ito *et al.* [2005] ont étudié la fusion statistique de deux systèmes de reconnaissance de rire : un système basé sur le flux audio et un autre basé sur l'information visuelle. Le système de détection visuelle de rire donne les meilleurs résultats. En revanche, la combinaison des deux systèmes permet de diminuer le taux de fausses alarmes.

Dans le cadre de l'analyse d'interaction sociale, Esposito *et al.* [2007] ont étudié les rôles des pauses et des silences dans l'interaction. Ils ont développé un système de détection automatique des pauses. Ils ont fait la différence entre les pauses vides (empty pauses) et les pauses remplies (filled pauses). Les pauses remplies correspondent aux pauses accompagnées de vocalisations non verbales. Ils ont montré que les individus utilisent largement des pauses vides durant leurs discours afin d'introduire de nouvelles idées. Ils utilisent également ces pauses pour la segmentation des phrases.

Enfin, la détection de silence a également été étudiée dans [Rabiner et Sambur, 1977] [Rabiner et Schafer, 1978], il existe maintenant des techniques robustes pour la détection de silences/pauses notamment dans le cas de paroles non bruitées.

1.5.1.4 Effet du contexte sur l'interprétation des signaux sociaux

La compréhension de l'interaction et particulièrement l'interprétation des différents signaux sociaux est un vaste domaine de recherche que plusieurs informaticiens, ergonomes, psychologues et sociologues ont abordé du point de vue de leur propre discipline. Malgré cette diversité, tous les chercheurs affirment que le contexte joue un rôle très important dans l'interaction et

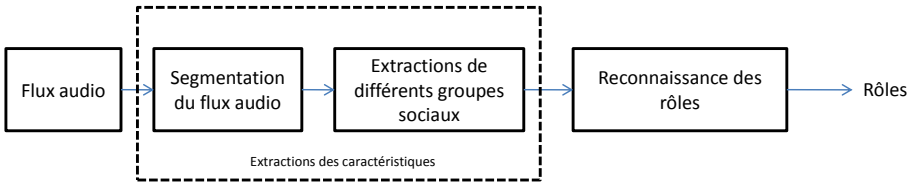


FIG. 1.7 – Reconnaissance automatique des rôles de locuteurs

la communication sociales. Tous les comportements produits par les différents interlocuteurs peuvent avoir des interprétations multiples en fonction du contexte interactionnel (le contexte social et culturel, le thème du discours, la nature des communications qui se développent, la participation des interlocuteurs et la synchronie qui s'établit, les évènements qui surviennent au cours des échanges, etc.). Par exemple, un sourire peut être interprété de manières différentes selon le contexte de l'interaction, il peut être un signe de politesse, de satisfaction ou de salutation.

Il existe divers signaux non verbaux (rire, expression de la douleur, etc.) avec à chaque fois des possibilités d'interprétations multiples. Le contexte semble pouvoir lever, du moins en partie, l'ambiguïté des interprétations. Afin d'illustrer les applications du traitement du signal social, nous avons choisis de présenter une application concrète : la reconnaissance de rôle des locuteurs dans une interaction [Favre *et al.*, 2008].

1.5.2 Application : Reconnaissance de rôle des locuteurs dans une interaction

La reconnaissance automatique des rôles a été étudiée dans deux différentes situations : des émissions radios [Barzilay *et al.*, 2000] [Favre *et al.*, 2008] [Vinciarelli, 2007] [Weng *et al.*, 2007] et des conversations de petits groupes d'interlocuteurs [Banerjee, 2004] [Dong *et al.*, 2007] [Garg *et al.*, 2008] [Zancanaro et Lepri, 2006]. Le travail de Garg *et al.* [2008] se décompose en deux parties : extraction des caractéristiques et reconnaissance des rôles, comme le montre la figure 1.7. L'extraction de caractéristiques consiste à d'abord segmenter automatiquement les locuteurs du flux audio, puis l'extraction des différents groupes sociaux (figure 1.8).

Le processus de segmentation automatique en locuteur transforme le flux audio en des séquences $S = \{s_i, \Delta t_i\}$, où $i \in \{i, ..., S\}$ et s_i est l'étiquette de l'interlocuteur qui a parlé au moment Δt_i. L'avantage de cette représentation est que chaque interlocuteur a_i est représenté par un vecteur

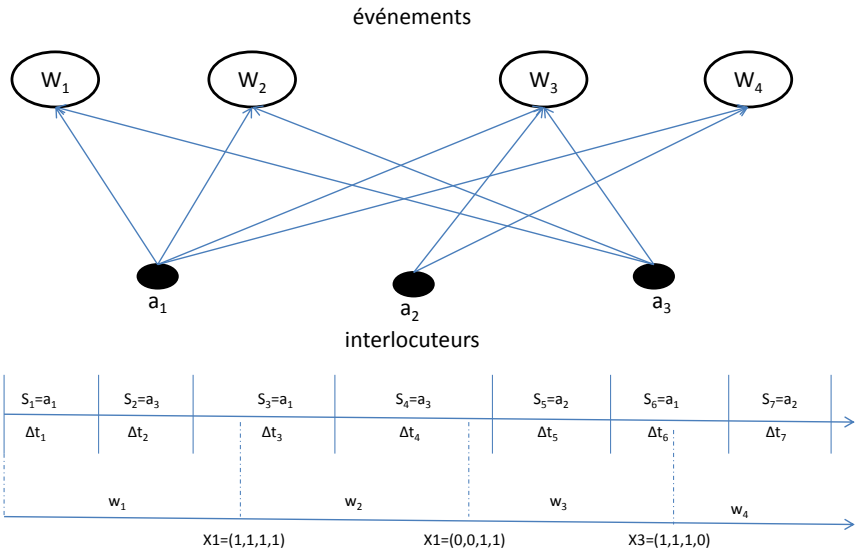


FIG. 1.8 – Extraction de caractéristiques pour la reconnaissance automatique des rôles [Favre *et al.*, 2008]

$\vec{x}_{ai}$= $(x_{ai1}, ..., x_{aiD})$, avec D le nombre de segments et a_{ij} le nombre de fois où l'interlocuteur a_i a participé à l'évènement j. Une autre variable de temps τ_i qui représente la durée totale du temps de parole l'interlocuteur a_i a été aussi prise en compte. Donc, le vecteur caractéristique devient $\vec{z}_{ai}$= $(\vec{x}_{ai}, \tau_i)$, de dimension $D+1$. Une Analyse en Composantes Principales ACP (Principal Component Analysis) [Bishop, 2006] est ensuite appliquée sur les vecteurs caractéristiques pour réduire leurs dimensions : projection du vecteur $\vec{z}_{ai}$ dans un espace de dimension $M < D+1$. Chaque enregistrement sera représenté par un vecteur $\vec{Y}$= $(\vec{y}_1, ..., \vec{y}_N)$, avec N est le nombre de tours de parole et $\vec{y}_i$ le vecteur caractéristique de l'interlocuteur qui a parlé durant le segment i.

La reconnaissance des rôles consiste à trouver la séquence des rôles qui permet d'optimiser la formule suivante :

$$R^* = \arg\max_{R \in R^N} P(Y|R)P(R) \tag{1.1}$$

avec $R = \{r_1, ..., r_N\}$ la séquence de N rôles et $r_i \in R$. R est un ensemble prédéfinis de rôles (invité, simple participant à la conversation, directeur, journaliste, animateur, etc).

TAB. 1.1 – Performances du système de reconnaissance de rôles évalué sur les $base_1$ et $base_2$ et les valeurs de l'écart type (σ)

	total(σ)	*AM*	*SA*	*GT*	*IP*	*HP*	*MW*
$base_1$							
HMM	74.2 (8.8)	97.8	9.5	65.9	38.0	61.1	63.0
HMM+1-gram	77.7 (11.6)	93.0	9.6	83.4	7.9	59.1	80.5
HMM+2-gram	79.7 (12.6)	95.5	12.2	82.6	25.7	63.5	80.4
HMM+3-gram	79.7 (9.3)	97.8	10.3	81.4	25.7	59.5	78.0
Bayes	82.5 (6.9)	98.0	3.6	91.8	9.0	64.5	79.7
$base_2$							
HMM	71.7 (7.2)	74.4	91.9	69.8	N	62.1	74.6
HMM+1-gram	83.7 (6.7)	74.5	92.0	90.3	N	58.4	19.0
HMM+2-gram	82.4 (7.4)	74.7	91.3	88.0	N	51.1	24.5
HMM+3-gram	86.1 (6.8)	74.4	91.9	92.0	N	72.8	30.5
Bayes	82.6 (6.9)	75.0	88.3	91.6	N	18.3	6.7

La probabilité à posteriori $P(Y|R)$ est estimée à l'aide d'un modèle de Markov caché (HMM) [Rabiner, 1989], où les rôles $r \in R$ sont les états du modèle. La probabilité à priori $P(R)$ est calculée à l'aide d'une modélisation statistique des données (*n-gram*) [Rosenfeld, 2000], $n \geq 1$:

$$p(R) = \prod_{k=1}^{N} p(r_k|r_{k-1}, r_{k-2}, ..., r_{k-n+1}) \tag{1.2}$$

Pour évaluer cette méthode, Favre *et al.* [2008] ont étudié trois types de bases de données : $base_1$ données issues de journaux télévisées (*news*), $base_2$ données issues d'émissions télévisées (*talk-show*) et $base_3$ données issues de réunions (*meeting*). Les individus dans la base $base_1$ peuvent prendre les rôles suivants : présentateur (AM), présentateur assistant (SA), invité (GT), participant à la conversation (IP), la personne qui gère la conversation (HP) et présentateur météo (WM). Les individus dans la $base_2$ peuvent jouer les même rôles que les acteurs de la $base_1$ à l'exception du rôle IP. Enfin, les individus de la $base_3$ jouent des rôles moins formels que les rôles existant dans la $base_1$ et la $base_2$: directeur de projet (PM), agent commercial (ME), expert en interface homme-machine (UI) et industriel (ID). Les tableaux 1.1 et 1.2 présentent les performances obtenues sur les $base_1$, $base_2$ et $base_3$.

Les résultats obtenus en utilisant un classifieur de type HMM et un classifieur bayésien ne sont pas significativement différents. Par contre, les deux classifieurs ne se comportent pas de la même façon. Ils ont des décisions différentes sur un nombre important d'exemples. Le modèle de langage (*n-gram*)

TAB. 1.2 – Performances de la reconnaissance de rôle évaluée sur la $base_3$ et les valeurs de l'écart type (σ)

	$total(\sigma)$	PM	ID	ME	UI
HMM	44.4 (26.8)	63.4	22.6	28.4	40.1
HMM+1-gram	40.7 (24.4)	70.6	12.4	16.1	32.0
HMM+2-gram	48.0 (25.9)	63.3	28.3	35.8	34.6
HMM+3-gram	46.9 (24.9)	61.8	25.0	39.4	33.8
Bayes	43.5 (23.9)	75.3	15.1	15.1	40.0

utilisé produit de meilleurs résultats sur les rôles les moins fréquents. Ce résultat suggère l'importance de la fusion des deux méthodes (*HMM + bayésien*).

1.6 La communication chez les bébés

Nous rappelons ici que l'objectif de notre thèse est la caractérisation automatique d'interaction parents-enfants. Dans ce cadre, il est important de préciser la nature de ces interactions.

Depuis son enfance, l'être humain développe continuellement des compétences de communication. De plus ces compétences diffèrent d'un individu à l'autre. Néanmoins, nous distinguons la communication de l'enfant qui est sans réelle intentionnalité communicative, lorsque les gestes et les vocalisations de l'enfant ne prennent sens que dans le cadre d'échanges avec l'adulte qui les interprète, de la communication de l'adulte qui est intentionnelle où l'individu utilise une conduite en sachant l'effet qu'elle va produire [Bates, 1979].

Les comportements non verbaux sont les principaux moyens de communication des enfants avant deux ans. En général, les enfants ne sont pas capables de produire de la parole compréhensible donc ils expriment leurs désirs à travers d'autres comportements non verbaux (gestes, postures, mimiques, cris, regard, etc). Les bébés disposent de mouvements expressifs porteurs de significations pour les parents, mais pas nécessairement pour eux car, au début de la vie, ils n'ont pas conscience d'émettre des signaux pour et vers leurs parents. Cependant, les parents vont essayer de les percevoir et de leurs donner un sens.

1.6.1 Compétences des bébés pour la communication

Les compétences des bébés sont assez limitées par rapport aux adultes mais leurs capacités d'apprentissage sont immenses [Thollon-Behar et Cohas, 2006].

Les bébés commencent par produire des réactions simples comme se tourner quand ils entendent un bruit ou une voix. De plus, il est aujourd'hui prouvé que les compétences cognitives des bébés sont très grandes. Par exemple, les bébés apprennent très vite que les objets de leur environnement ont des noms, qu'ils parviennent aisément à extraire des productions langagières des adultes qui les entourent.

Les aptitudes perceptives et motrices sont elles aussi exceptionnelles : les bébés de quelques secondes tournent les yeux à droite ou à gauche selon qu'on fait retentir un clic dans l'une ou l'autre direction dans la salle d'accouchement. Ce qui montre que, non seulement les bébés savent localiser les sons dans l'espace, mais qu'ils possèdent une coordination inter sensorielle. Ainsi, dès la première semaine, les enfants savent prévoir une trajectoire, se détournent et se protègent d'un objet avançant vers eux selon une trajectoire menaçante, mais ne se protègent pas d'un objet se déplaçant devant lui selon une autre trajectoire.

Du point de vue moteur, le bébé de 10 jours, tenu verticalement et le corps soutenu tend la main, touche les objets et éventuellement s'en saisit. Par contre, les gestes ne sont pas précis et si le bébé prend l'objet il n'en fait rien, néanmoins la capacité est présente. Cette capacité motrice du bébé lui permet d'exprimer ses désirs à partir de gestes simples comme le fait de tendre la main quand il veut quelque chose.

1.6.2 Evolution de la communication non verbale chez les enfants

Bower [1997] a montré que, dès sa naissance, le bébé réalise qu'il est un être humain et dispose de réponses spécifiques déclenchées exclusivement par d'autres êtres humains. Les bébés sont capables de produire des mouvements expressifs, et au cours du temps ces compétences vont évoluer. Les premières manifestations sont le regard, le sourire, la mimique et quelques expressions vocales.

1.6.2.1 Le regard

Le regard est fondamental car il n'y a pas de communication sans le contact visuel. L'effet que produit le regard sur l'adulte lors des premiers échanges est très important, c'est le signal le plus valorisé par les parents et particulièrement la mère car il donne le sentiment de s'adresser à une personne à part entière. De plus, il consolide l'attachement social. L'absence de regard est interprétée comme un évitement, souvent mal vécu par les parents, ou comme une rupture dans la communication.

1.6.2.2 Le sourire

Le sourire est considéré comme un signal de bien être (contrairement aux cris et aux pleurs). Plus tard, le sourire représente un signe de bonne socialisation (avec l'accueil des inconnus). Ce comportement est présent dès la naissance du bébé. Il y a deux types de sourires : sourire de contentement (sourire - reflexe) et sourire social (sourire - réponse). Le sourire reflexe est un signe du bien-être physique sans intention. Tandis que le sourire réponse a une dimension sociale, lorsque le bébé réplique réellement au sourire de sa mère, puis à celui de son père, et des autres. Le sourire social est un comportement dirigé et conscient. Il est souvent considéré comme le premier geste faisant de l'enfant un être social à part entière. Ensuite, au fil du temps, l'enfant apprend par imitation à nuancer ses sourires [Morris, 1994].

1.6.2.3 Les expressions vocales

Les expressions vocales sont présentes dès la naissance, et très rapidement les parents sont capables de les différencier et de les interpréter, notamment en fonction de leur intensité. Les cris ne sont pas identiques. Pendant les deux premiers mois, les vocalisations sont modulées et jouent un rôle important dans les proto-dialogues. Pendant cette période, les comportements non verbaux, particulièrement les cris, sont très importants dans l'engagement de la communication entre les parents et le bébé. Les cris peuvent être expressifs, par exemple des démonstrations spontanées d'émotions et d'états affectifs, parfois provoqués par les parents. Ils sont aussi utilisés pour appeler les parents, particulièrement la mère, pour attirer leur attention par exemple.

1.6.2.4 La mimique et la synchronie interactionnelle

Le répertoire de mimique chez les bébés est assez simple : tirer la langue, battre des cils ou ouvrir et fermer la bouche, etc. Ces mouvements semblent être des activités qui font plaisir à l'enfant. Il s'agit à ce stade d'un jeu social (activité préférée). La mimique est clairement dirigée vers les autres humains et témoigne de ce que le bébé se considère aussi comme un humain. D'une manière ou d'une autre, il sait que son visage est fait comme celui qui est en face de lui ; que sa bouche est semblable à celle qu'il voit devant lui.

Ces comportements de mimique sont très importants pour assurer la synchronie interactionnelle entre les parents et le bébé. La synchronie interactionnelle désigne une forme de comportement caractéristique de la communication humaine. Lorsque deux personnes appartenant au même groupe culturel s'engagent dans une conversation, une analyse détaillée de leurs mouvements montre que parallèlement, elles s'engagent dans une sorte de danse synchrone.

Pendant que l'une parle et fait certains mouvements : légers changements posturaux, changements de l'orientation de la tête, et l'autre fait les mouvements correspondants. En général, on n'est pas du tout conscient de cela, mais quand ce rituel est absent, il risque d'y avoir un défaut dans la communication. Une conversation réussie suppose une étroite synchronie des mouvements du corps. Les bébés s'engagent également dans des synchronies interactionnelles, quand les parents s'adressent à eux, ils bougent leurs corps selon des rythmes précis, parfaitement coordonnés aux unités de base du discours de l'adulte [Ruth et Arthur, 2004].

Cependant, seul le signal de parole déclenche ce comportement de synchronie interactionnelle chez l'enfant, et cela très précocement [Ruth et Arthur, 2004]. Bien sûr, cette communication est très émouvante pour l'adulte qui la partage (souvent la mère) et renforce le sentiment d'obtenir une réponse de son bébé. Par contre, il semble que certains troubles prénataux ou génétiques, comme l'autisme, puissent réduire ou supprimer l'aptitude du bébé à la synchronie interactionnelle. Ce trouble pourrait avoir des conséquences négatives à long terme, la mère se sentant rejetée par son bébé, le rejetterait à son tour. Cela pourrait entraîner une absence de communication prolongée qui retentit sur le développement ultérieur.

1.6.2.5 L'intersubjectivité

L'intersubjectivité ou partage du vécu traduit une relation dynamique entre deux sujets bien distincts en l'occurrence la mère et l'enfant. C'est l'aptitude de sentir les autres, différents de soi, tout en étant capable d'avoir ou de concevoir un état mental semblable au leur. Les enfants sont capables d'avoir ce mode de fonctionnement entre le septième et le neuvième mois de leur vie. Les comportements qui expriment cette propriété émotionnelle d'un affectif partagé sont désignés par les termes d'accordage affectif [Stern, 1997]. L'intersubjectivité est traduite par le passage de la dépendance absolue à l'indépendance relative de l'enfant par rapport ses parents. Cependant, le manque d'intersubjectivité est un signe pathologique du développement (retard du développement, autisme, etc.) [Merinfeld, 2005].

1.7 Autisme et enfant : aspect développemental

1.7.1 Le développement affectif de l'enfant

L'étude du développement affectif des enfants est un processus fondamental pour l'analyse des comportements et l'interaction sociale des enfants notamment dans le cadre des pathologies mentales et cognitives (langage, per-

ception, raisonnement, etc.). Le développement de l'enfant n'est pas un processus indépendant car l'environnement social et physique du bébé jouent un rôle important dans son développement. De plus, les compétences sociales des enfants se développent grâce aux communications avec son entourage et particulièrement avec ses parents. En complément, la découverte par les parents des compétences de leur enfant va faciliter la mise en place des interactions et de l'attachement précoce. La relation de l'enfant avec son entourage est conçue comme un processus bidirectionnel. L'enfant est à la fois soumis aux influences de ses parents et entraîne chez eux des modifications.

Le besoin d'attachement est un besoin primaire, inné chez l'homme. Ainsi, l'attachement est indispensable pour l'enfant pour survivre et pour se développer correctement. Il s'agit de la relation de l'enfant avec sa mère d'abord, puis avec ses parents et en général avec son entourage proche. Cet attachement se développe à partir de comportements innés : pleurs, succion, agrippement qui permettent de maintenir la proximité physique et l'accessibilité à la figure d'attachement privilégiée qui est le plus souvent représentée par la mère.

L'enfant veut toujours des réponses sécurisantes à ses désirs, aussi il a souvent besoin de réponses rapides, cohérentes, chaleureuses et prévisibles. Par conséquent, quand une sensation est désagréable, le bébé gémit, pleure, crie, hurle, et généralement quelqu'un, le plus souvent maman, vient et le soulage. Si les réponses de l'entourage sont adéquates au besoin d'attachement de l'enfant, celui-ci développera une base de sécurité et une image de lui-même positive. Par conséquent, à partir de cette base de sécurité, de nouvelles compétences apparaissent : la capacité de se séparer pour explorer l'environnement, la capacité d'attendre une réponse et plus tard de répondre à son tour aux besoins d'attachement.

Autrement, chez tout enfant se déroule un processus d'individuation et de séparation (psychique), qui permet le développement du sentiment de conscience de soi. Ce processus de séparation-individuation se manifeste par le développement des compétences de l'enfant, des compétences sensorielles et sociales, des compétences cognitives qui vont avoir un impact très important sur l'interaction des enfants avec leurs entourages.

1.7.1.1 Compétences sensorielles de l'enfant

On distingue les compétences perceptuelles (vision, audition et olfaction) et les compétences motrices qui sont généralement les premiers éléments développés l'enfant [Maury *et al.*, 2005]. Les bébés ont d'abord une préférence visuelle pour des modèles proches du visage humain par la forme et la taille. Par contre, leurs compétences visuelles restent limitées par rapport à celles des adultes. De plus, la perception visuelle de l'enfant est conditionnée par

l'ensemble de son développement sensori-moteur. Ainsi un déficit précoce de la fonction visuelle peut interférer avec le développement de l'enfant et influencer sur l'ensemble de ses compétences, qu'elles soient motrices, cognitives ou affectives, et avoir ainsi des répercussions sur ses performances interactionnelles puis sur son insertion sociale.

L'enfant, dès sa naissance, est très sensible aux événements sonores, le bébé tourne généralement sa tête ou ses yeux vers la direction de la source. Durant les premiers jours, le bébé préfère la voix humaine, et plus particulièrement celle de sa mère, aux autres sources sonores (bruits), cette préférence se manifeste par une augmentation des sourires et grand intérêt par tourner la tête en direction de la voix.

D'autres compétences, comme l'olfaction, sont assez développées et sont proches de celles de l'adulte. Très jeune, le bébé est capable de discriminer l'odeur du cou et/ou du sein de sa mère, de l'odeur d'une autre femme. Il est également capable de discriminer des saveurs : salé, sucré, amer, acide, avec une préférence pour le sucré.

Les compétences motrices et les mouvements corporels sont essentiellement réflexes : succion, grasping réflexe, réflexes archaïques. Cette motricité réflexe joue un rôle important dans l'interaction car elle présente un support aux parents pour forger leurs représentations de l'enfant. Par exemple quand un bébé serre le doigt de sa mère lors du grasping réflexe, sa mère éprouve une émotion qui, selon les cas, peut être rassurante (il tient à moi) ou inquiétante (il ne va pas me lâcher).

Toutes ces compétences vont avoir des conséquences positives sur l'interaction sociale de l'enfant avec son entourage, particulièrement avec ses parents.

1.7.1.2 L'interaction de l'enfant avec son entourage

L'interaction de l'enfant avec ses parents, particulièrement avec sa mère, est très importante et primordiale pour l'accompagner dans la découverte et la compréhension des différents dispositifs cognitifs. Les chercheurs en psychologie développementale ont montré que, à partir de quelques semaines après la naissance, la mère et le bébé échangent des signaux multimodaux : regards, quelques vocalisations, imitations, expressions faciales, puis chacun est influencé par le comportement de l'autre [Weinberg et Tronick, 1994] [Mayer et Tronick, 1985] [Beebe *et al.*, 1979]. On distingue trois niveaux d'interaction : comportemental, affectif et imaginaire.

Interaction comportementale L'interaction comportementale ou interaction réelle se définit par la manière dont les comportements de la mère vont influencer ceux du bébé et inversement. Elle est directement observable et elle

peut se situer dans les registres corporel, visuel et/ou vocal. En particulier, les vocalisations restent les moyens les plus utilisés et traduisent les besoins et les affects des bébés, permettant ainsi d'exprimer leurs désirs. La voix, particulièrement la voix maternelle, donne un tempo à l'interaction qui va permettre à l'enfant d'anticiper et de s'organiser. Les interactions vocales sont aussi importantes pour l'harmonisation de la relation entre la mère et l'enfant. Au début, les parents ont plus tendance à parler avec une voix harmonique avec le bébé, ils utilisent un registre particulier de parole appelé *motherese*, ce registre de parole joue un rôle très important dans le développement de l'enfant [Beebe et Lachmann, 1988] [Papousek et Papousek, 1977] [Papousek et Papousek, 1987] [Stern, 1993]. Le *motherese* est souvent utilisé par la mère dès les premiers instants de sa relation avec le bébé. A cette période néonatale, la motricité du bébé paraît parfois entraînée par la parole de la mère et synchronisée avec elle. De plus, l'interaction vocale est un processus fondamental pour l'acquisition et l'apprentissage du langage. Bruner [1968] affirme que l'acquisition du langage est basée sur les processus d'interaction mère-bébé. Concernant l'interaction corporelle, c'est la façon dont le bébé est tenu, porté, manipulé et touché, elle permet l'ajustement interactif entre l'enfant et sa mère. Ainsi, des études [Stack et Arnold, 1998] ont montré qu'avec seulement le toucher et les gestes, la mère peut engager facilement une interaction avec son bébé.

Interaction affective L'interaction affective représente l'influence réciproque de la vie émotionnelle du bébé et de celle de ses parents, particulièrement sa mère. Elle concerne le climat émotionnel ou affectif de l'interaction, le vécu agréable ou déplaisant de la communication. Les affects constituent l'objet même de la communication dans le jeu mère-bébé, surtout pendant les premières semaines. La tonalité affective globale des échanges entre les parents et l'enfant permet de dégager des sentiments de plaisir, bien-être, tristesse, ennui, indifférence, insécurité, excitation, voire de la haine appellé accordage affectif [Stern, 1985]. Il s'agit d'un acte d'intersubjectivité dans lequel le parent répond à une expression affective du bébé en la remaniant d'une autre façon et en la rejouant au bébé de sorte qu'il lui montre qu'il a partagé son expérience subjective interne.

Interaction imaginaire et fantasmatique L'interaction imaginaire fantasmatique représente l'influence réciproque du déroulement de la vie psychique de la mère et de celle de son bébé, aussi bien dans leurs aspects imaginaire, conscient ou inconscient que fantasmatique. Elle va ainsi donner sens à l'interaction comportementale. Chez le parent, la proximité de son bébé, va réactiver sa vie imaginaire et fantasmatique, il pourra ainsi parler de l'enfant

imaginaire de la grossesse. Quant au bébé, grâce à l'excitation des zones érogènes, il va halluciner la satisfaction et imaginer le plaisir qui vient de l'objet qu'il se représente. Par exemple : un fantasme maternel portant sur le danger de se séparer influence le comportement de la mère à l'égard de son bébé particulièrement dans une situation de séparation ; le comportement angoissé du bébé à ce moment là amène à penser qu'il partage avec sa mère une expérience émotionnelle qui contribue à la constitution de sa propre vie fantasmatique.

1.7.2 Autisme infantile

Il y a longtemps, la pathologie mentale de l'enfant, quelle qu'elle soit, était considérée comme l'expression d'une déficience du développement de l'intelligence. La psychose était à cette époque considérée comme exclusivement liée à l'adulte. Depuis, des similitudes cliniques ont peu à peu été mises à jour entre certaines psychoses décrites chez l'adulte et les manifestations de certains enfants, ce qui a fait supposer dès la fin du XIXème siècle la possibilité d'une éclosion très précoce de troubles psychotiques de la personnalité. En 1943, Léo Kanner, psychiatre américain d'origine autrichienne, décrivit pour la première fois l'autisme infantile précoce, à partir de l'observation d'un groupe d'enfants âgés de 2 ans et demi à 8 ans.

L'autisme est un trouble précoce, global et sévère du développement de l'enfant caractérisé par des perturbations dans les domaines des interactions sociales réciproques, de la communication (verbale et non verbale) et des comportements, un champ d'intérêts et des activités au caractère restreint, répétitif. Ces signes, lorsqu'ils sont associés, définissent alors le syndrome autistique. Celui-ci constitue un mode de fonctionnement pathologique très invalidant, caractérisé par une limitation intense des capacités d'échange et de communication, perturbant sévèrement les acquisitions, la socialisation et l'autonomie [Volkmar et Pauls, 2003]. Les difficultés de développement du langage, intriquées à celles du psychisme et de la vie affective, semblent au coeur du processus pathologique. Ainsi, le développement de la communication et celui des fonctions cognitives sont très perturbés, ce qui se traduit cliniquement par un isolement, une diminution des activités spontanées et des troubles du langage. Cependant, l'intensité de ces signes est très différente d'un enfant à l'autre et peut varier avec l'âge.

L'autisme est une pathologie neuro-développementale qui touche le développement et le fonctionnement du cerveau. Il ne peut être diagnostiqué avant l'âge de 2 ans, et les signes caractéristiques du syndrome autistique sont rarement complets avant l'âge de 3 ans [Schopler *et al.*, 1988] [Stone *et al.*, 1994] [Sullivan *et al.*, 1990].

Aujourd'hui, on distingue trois critères consensuels (apparus avant 36

mois) pour le diagnostic d'autisme : altération qualitative des interactions sociales, altération qualitative de la communication et du langage, comportements et activités restreints et stéréotypés :

- Altération qualitative des interactions sociales :
  - Fuite du regard
  - Impression de surdité au monde environnant
  - Absence d'intérêt envers autrui
  - Absence de jeux interactifs, de jeux de groupe, d'imitation ou d'imagination
  - Refus du contact corporel
- Altération qualitative de la communication et du langage :
  - Défaut d'expression, du contact oculaire
  - Défaut d'intention communicative (pointage, attention conjointe)
  - Absence du langage (50% des cas)
  - Quand le langage est présent il présente des particularités : modulation anormale de la voix et de la prosodie, inversion pronominale, peu de valeur de communication du langage
- Comportements et activités restreints et stéréotypés :
  - Stéréotypies, rituels
  - Comportements et activités répétitifs
  - Intolérance au changement
  - Hypo/hyperréactivité touchant toutes les modalités sensorielles

1.7.3 Premières manifestations cliniques de l'autisme

Bien avant que le diagnostic ne soit établi, les parents se rendent souvent compte des différences présentées par leurs enfants [Werner *et al.*, 2000]. L'ensemble des données recueillies, et notamment l'analyse systématique des données rapportées par les parents indique que dans 75 à 88% des cas des signes existent avant 2 ans et dans 31 à 55% avant 1 an.

Les parents d'enfants autistes ont rapporté que leurs enfants jouaient moins avec eux [Dahlgren et Gillberg, 1989] [Hoshino *et al.*, 1982], qu'ils préféraient être seuls et qu'ils réagissaient moins aux tentatives des parents de se joindre à leur jeu que les enfants qui ne souffraient pas de troubles autistiques [Dahlgren et Gillberg, 1989] [Hoshino *et al.*, 1982]. De plus, les parents ont rapporté que les enfants autistes souriaient moins aux autres [Hoshino *et al.*, 1982] [Wimpory *et al.*, 2000] et avaient plus tendance à afficher un visage sans expression que les enfants non autistes [Hoshino *et al.*, 1982].

De même, les données provenant des études longitudinales et prospectives portant sur les enfants à haut risque (frères et soeurs des enfants autistes)

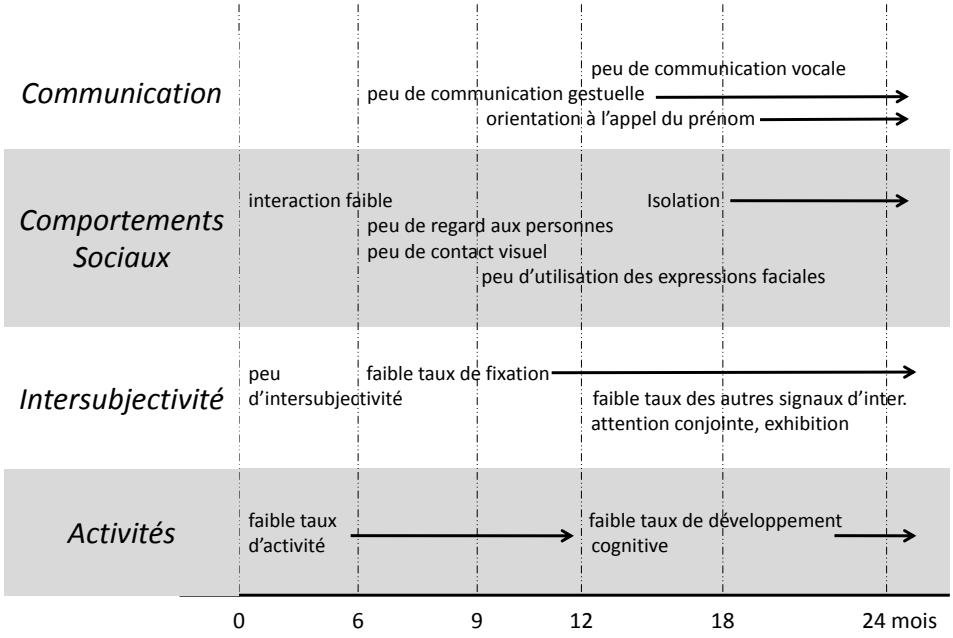


FIG. 1.9 – Premières manifestations de l'autisme

ont également donné des informations importantes sur les premiers signes comportementaux et développementaux de l'autisme. Ces études ont montré que les frères et soeurs diagnostiqués autistes plus tard présentaient plusieurs différences sociales comparativement aux enfants témoins à l'âge de 12 mois [Zwaigenbaum *et al.*, 2005].

Parmi les signes comportementaux, les enfants à devenir autistique passaient moins de temps à regarder les adultes pendant le jeu libre [Swettenham *et al.*, 1998], regardaient moins le visage d'un adulte feignant la détresse [Charman *et al.*, 1997], regardaient moins quand ils passaient des personnes aux objets [Swettenham *et al.*, 1998] [Charman *et al.*, 1997] et imitaient moins [Charman *et al.*, 1997] que les enfants typiques ou les enfants avec retard du développement.

Enfin, plusieurs études longitudinales de films familiaux (des vidéos tournées à la maison) des enfants à devenir autistique ont permis aux chercheurs d'extraire des comportements qui caractérisent le syndrome autistique [Saint-Georges *et al.*, 2010]. Pendant la première année, les enfants autistes ont tendance à manquer d'intérêt social, et interagir mal par rapport aux enfants avec développment normal [Maestro *et al.*, 2002] [Zakian *et al.*, 2000] [Adrien *et al.*, 1993], passent moins de temps à regarder les gens autour [Clifford et Dissanayake, 2007] [Maestro *et al.*, 2002], aussi passent moins de temps à vocaliser.

En général, les signes d'alerte de l'autisme apparaissent dès le premier semestre. La figure 1.9 résume les principaux signes d'alertes [Saint-Georges *et al.*, 2010]. Pendant les premiers six mois de vie, l'enfant autiste peut être caractérisé par le manque ou l'absence d'échange avec sa mère, l'absence d'intérêt pour les personnes : indifférence à la voix et au visage de la mère, absence d'échange du regard avec celle-ci. Puis, l'enfant autiste est aussi caractérisé par une indifférence au monde sonore. Du point de vue du comportement, l'enfant autiste est soit doté d'une sagesse excessive (enfant "trop calme"), soit au contraire, agitation désordonnée (enfant "trop excité"). Il présente aussi des troubles psychomoteurs qui se manifestent par des défauts d'ajustement posturaux et d'agrippement lors de la prise de l'enfant par l'adulte, aussi une absence d'attitude anticipatrice de l'enfant lorsque l'on ébauche le mouvement de le prendre dans les bras (l'enfant normal accompagne le mouvement en tendant les bras), des troubles graves et précoces du sommeil. Enfin, pendant les premiers six mois, l'enfant autiste est aussi caractérisé par une absence ou une pauvreté de vocalisations et une absence de sourire au visage humain, qui apparaît normalement vers le deuxième ou troisième mois et qui constitue un signe positif des capacités relationnelles de l'enfant.

Durant le deuxième semestre, les signes précédents se confirment (désintérêt pour les personnes, défaut d'ajustement postural, indifférence au monde sonore et visuel), mais d'autres signes apparaissent comme le défaut de contact (absence d'intérêt pour les personnes), absence d'angoisse de l'étranger et lors de la séparation d'avec les personnes qui s'occupent habituellement de lui. Normalement, l'angoisse de l'étranger apparaît vers 8 mois chez les enfants typiques : lorsqu'il est mis en présence d'un étranger en l'absence de sa mère, l'enfant montre, à cette période, des manifestations plus ou moins importantes d'angoisse. De plus, pendant cette période, l'enfant autiste n'imite pas, par exemple il ne participe pas à des activités comme "faire coucou" (bonjour).

Durant la deuxième année, les signes précédents deviennent plus explicites, notamment le désintérêt pour les personnes, une fascination trop vive pour les stimulations sensorielles [Saint-Georges *et al.*, 2010]. De plus, de nouvelles anomalies s'ajoutent comme l'absence de pointage, anomalies de la marche, phobies de certains bruits, auto-agressivité et stéréotypies gestuelles. Pendant cette période, les troubles de langage se confirment et la prononciation des premiers mots apparait en retard, après 18 mois.

1.7.4 Acquisition du langage et les interactions sociales

Les déficiences présentées depuis le très jeune âge par les enfants autistes invitent à nous interroger sur ce que nous savons de l'apprentissage du langage et des interactions sociales. Des travaux récents ont montré le rôle fondamen-

tal des expériences précoces relationnelles dans l'apprentissage du langage. Cet apprentissage se ferait de façon probabiliste : dans un premier temps, le bébé est capable de distinguer toutes les langues au niveau phonétique et d'identifier leurs caractéristiques prosodiques [Kuhl, 2003]. Puis il développe une stratégie d'apprentissage basée sur les signes et les caractéristiques de la langue d'entrée et l'exploitation de ses propriétés statistiques, conduisant à augmenter la perception de la langue maternelle et réduire celle des langues étrangères. Cependant, la simple exposition n'explique pas l'apprentissage du langage : la présence d'un être humain qui interagit avec le bébé a une forte influence dans l'apprentissage [Goldstein *et al.*, 2003].

Quelle que soit l'hypothèse neurobiologique ou génétique de l'autisme [Cohen *et al.*, 2005], nous devons garder en mémoire que la survie et le développement infantiles dépendent de l'interaction sociale avec un adulte pour entretenir les premiers besoins d'attachement émotionnel de l'enfant. En cas de privation précoce grave, les conséquences sur le développement de l'enfant sont parfois impressionnantes [Rutter *et al.*, 2007a] [Rutter *et al.*, 2007b] [Croft *et al.*, 2007]. Ainsi que la qualité de l'interaction sociale entre les parents et l'enfant dépend des deux partenaires car l'interaction parent-enfant est constituée de l'ensemble des processus bidirectionnels, où l'enfant est à l'origine de modifications chez les parents mais aussi il est soumis à l'influence du comportement des parents. Par conséquent, la découverte, par les parents, des compétences de leur enfant, va faciliter la mise en place des interactions et de l'attachement précoce.

D'autre part, des chercheurs étudiant l'acquisition du langage et d'autres étudiant les interactions sociales précoces ont trouvé une particularité commune très importante qui affecte le langage et le développement social des bébés. Le registre particulier de parole qui est adressée aux bébés et aux très jeunes enfants, appelé *motherese* pour la mère (ou fatherese pour le père), caractérisée par un *pitch* augmenté, un *tempo* plus lent et des contours intonatives exagérés [Fernald et Kuhl, 1987] semble jouer un rôle important dans l'interaction sociale et le développement du langage. Les études ont indiqué que cette prosodie particulière favorise l'attention des bébés, transmet des informations affectives et des informations phonologiques spécifiques [Kuhl, 2003]. De plus, quand on donne le choix aux bébés entre la parole normale et le *motherese*, ils préfèrent le *motherese* [Pegg *et al.*, 1992] [Cooper et Aslin, 1990] [Werker et McLeod, 1989] [Fernald, 1985] [Glenn et Cunningham, 1983].

1.8 Conclusions

Dans ce chapitre, nous avons défini la communication et les différentes typologies existantes dans la littérature. Ensuite, nous avons présenté les principaux types de signaux de communications appelés aussi signaux sociaux vis-à-vis de l'interaction sociale. Le support de l'interaction sociale est en partie la communication verbale et non-verbale. La première forme de communication considère essentiellement le langage alors que la communication non verbale exploite les expressions faciales, le contact visuel, les gestes, la posture, le langage corporel, etc. Dans le cadre d'une interaction sociale, les deux types de signaux sont complémentaires.

Dans notre étude, nous nous intéressons à la communication face à face, parent-enfant, qui regroupe les deux types de signaux verbaux et non verbaux. Cependant, dans la communication parent-enfant, les deux partenaires de communication ne partagent pas les mêmes signaux de communication car l'enfant n'est pas capable d'émettre les mêmes signaux que l'adulte et de produire de la parole compréhensible. Par conséquent, l'enfant exprime ses désirs à travers d'autres comportements non verbaux (gestes, postures, mimiques, cris, regard, vocalisations, etc.). Cette capacité de communication chez l'enfant se développe au cours du temps. Par contre ce processus de développement diffère d'un enfant à l'autre et dépend principalement de la qualité de l'interaction avec leurs parents. Ainsi que l'engagement des parents dans l'interaction dépend des réponses des enfants. Par conséquent, si l'enfant est atteint d'une pathologie développementale qui touche ses comportements, il risque d'avoir des problèmes de communication avec son entourage, particulièrement avec ses parents.

Les enfants autistes ont un développement déviant par rapport aux enfants normaux qui se reflète sur la qualité de l'interaction avec leurs parents. Afin d'étudier la qualité de l'interaction parent-enfant, des chercheurs ont identifié une particularité commune très importante qui affecte le langage et le développement social des enfants. Un registre particulier de parole qui est adressée aux bébés appelé *motherese* semble jouer un rôle modérateur dans l'interaction sociale et le développement du langage de l'enfant. Cette façon de parler favorise l'attention du bébé et permet de transmettre des informations affectives. Pour vérifier cette hypothèse, nous développons un algorithme de reconnaissance automatique de *motherese* dans les chapitres 2 et 3 et nous étudions son implication dans un modèle d'interaction computationnel (chapitre 4).

Chapitre 2

Emotion sociale : *Motherese*

Sommaire

2.1 Introduction

Le *motherese* a été identifié comme étant un support de l'interaction sociale entre la mère et l'enfant. Ce chapitre a pour objet de préciser la définition de ce registre de parole. Pour cela, nous proposons de caractériser les signaux émotionnels ainsi que leurs utilisations dans le contexte de l'interaction. Les particularités du *motherese* nous ont amenées à constituer un corpus de parole extrait de données réelles. Nous chercherons à qualifier les échanges des parents avec leur bébé dans le contexte des films familiaux (durée, qualité sonore). L'objectif principal de ce chapitre est de proposer un système automatique de détection du *motherese*. À cet effet les travaux portant sur la détection et la classification de signaux émotionnels seront passés en revue. Les systèmes de classification de l'état de l'art réquièrent plusieurs étapes : annotation des données, extraction de caractéristiques et classification. L'annotation permet d'obtenir une définition expérimentale des signaux étudiés. L'approche choisie pour la caractérisation et la détection consiste à exploiter à la fois des informations segmentales (acoustiques) et supra-segmentales (prosodie). Les performances du système sont évaluées sur la base de données constituées de parole spontanée et dans des contextes divers.

2.2 Signaux émotionnels

2.2.1 Définition

Les informations linguistiques sont nécessaires pour assurer une communication face à face, cependant elles ne forment qu'une partie du message communiqué. Lors d'une conversation, les interlocuteurs sont très sensibles aux informations extralinguistiques (état du locuteur, identité du locuteur, etc.) et aux informations paralinguistiques (intention, prosodie, etc.). Ainsi dans une conversation, les émotions du locuteur et sa manière de parler peuvent être plus importantes que les informations linguistiques. De plus, les émotions jouent un rôle très important dans la vie quotidienne et exercent une influence non négligeable sur les comportements humains. Elles oeuvrent pour transmettre plus de sens aux échanges et renseignent sur les intentions et motivations.

Cependant, il est difficile de définir une émotion, même si plusieurs travaux ont tenté de caractériser des états émotionnels [Cowie et Schröder, 2005] [Cowie et Cornelius, 2003]. En général, on distingue deux approches principales pour la description des émotions. Une première approche catégorielle où les émotions sont représentées par un état (tristesse, colère, etc.) à un instant donné. Les émotions sont alors considérées comme des caractéristiques

épisodiques basiques et universelles [Ekman, 1999c]. Une deuxième approche consiste à représenter l'état émotionnel par un point dans un espace multidimensionnel [Cowie et Cornelius, 2003] [Schröder, 2004].

Wundt [1913] a proposé une description dimensionnelle des états émotionnels sur trois dimensions : envie-aversion (positif-négatif), excitation-apaisement (actif-passif) et tension-soulagement. La dimension positive-négative reste la plus utilisée. Dans cette dimension, les états émotionnels sont organisés selon un axe négatif-positif (dégoût, colère, peur, surprise, excitation, joie). Dans la littérature, on distingue plusieurs représentations des états émotionnels se basant sur ces échelles. Whissel [1989] a proposé une liste d'émotions, que l'on retrouve également dans la représentation de Plutchik (cône des émotions), voir la figure 2.1. Galati et Sini [1995] ont fait une étude empirique pour répertorier les termes utilisés dans le langage courant francais.

Afin de faciliter la tâche de reconnaissance automatique des émotions, la majorité des études en informatique affective (affective computing) considère souvent l'approche catégorielle. Ils ne considèrent que les six émotions basiques proposées par Ekman [1999c] : colère, peur, tristesse, joie, dégoût et surprise. Le problème de l'approche catégorielle réside dans l'hypothèse discrète de l'état émotionnel. Du point de vue psychologique l'émotion est une notion floue et les émotions peuvent se chevaucher dans les interactions réelles, de plus l'expression d'un état émotionnel peut changer d'un individu à un autre [Picard, 2003]. Du fait de la difficulté de catégorisation et d'annotation des signaux émotionnels, plusieurs études se limitent à l'étude d'un nombre réduit d'émotions non prototypiques. Lee *et al.* [2001] ont développé une technique qui permet de discriminer les signaux positives des émotions négatives. Ainsi ces travaux se sont limités à la discrimination des parole émotionnelles et des paroles neutres [Ringeval *et al.*, 2010]. De plus, certains chercheurs ont étudié le problème de reconnaissance des émotions en tenant compte de leur application, comme les problèmes de la reconnaissance de la parole stressée vs. parole non stressée [Fernandez et Picard, 2003] [Narayanan, 2002], parole frustrée/contrariée vs. parole non frustrée [Ang *et al.*, 2002], la peur [Clavel *et al.*, 2008], l'empathie vs. neutre [Steidl *et al.*, 2005], le *motherese* [Mahdhaoui *et al.*, 2008] [Mahdhaoui et Chetouani, 2010].

2.2.2 Emotions et contexte

De nombreuses études sur la catégorisation des émotions ont été menées. Cependant, il n'existe actuellement pas de consensus sur la catégorisation des émotions tant elles sont dépendantes du contexte. Par conséquent, il est important d'étudier les émotions dans leurs contextes : culturel, social, linguistique, etc. Concernant les applications en vidéo surveillance, Clavel *et al.* [2008] ont

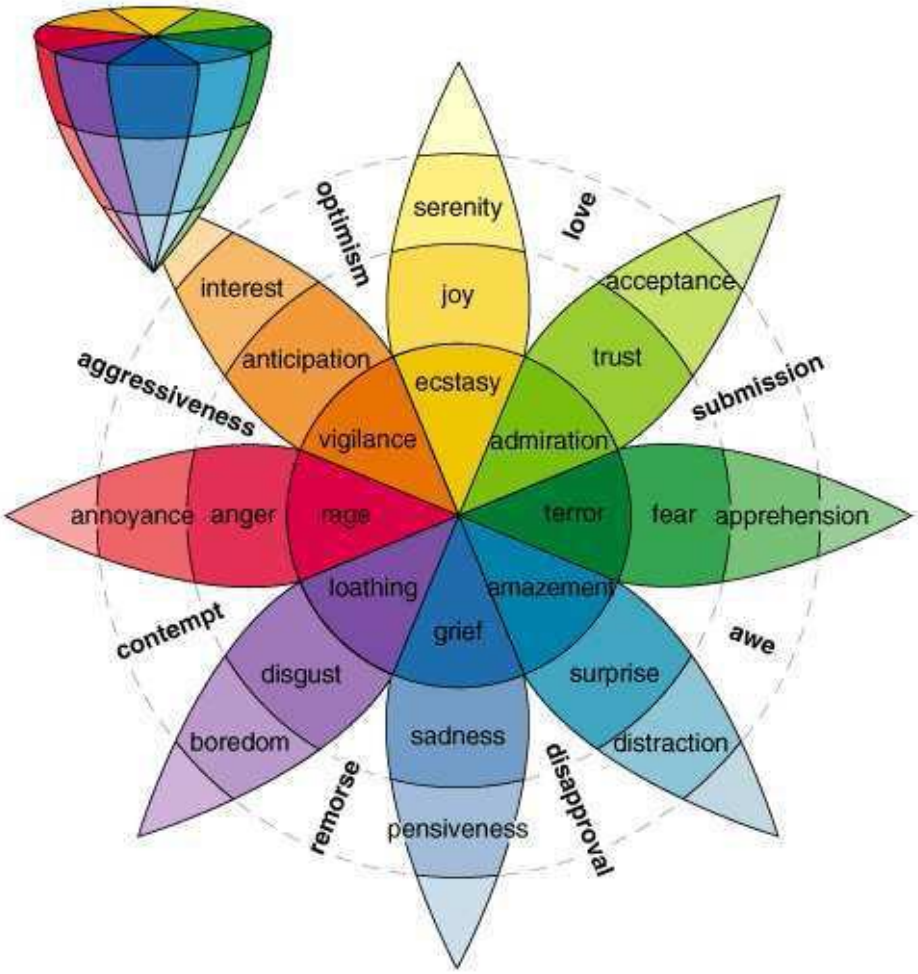


FIG. 2.1 – Cône des émotions de Plutchik [Plutchik, 1984]

étudié une émotion particulière : la peur vécue dans des situations anormales (événements imprévus constituant une menace pour la vie humaine). Il en résulte une définition expérimentale d'un ensemble d'états émotionnels regroupant non seulement la peur mais également l'inquiétude ou la panique. Dans le contexte social d'un centre d'appel, l'état émotionnel du client est une information importante pour évaluer son degré de satisfaction, mais la définition de ce dernier s'avère complexe.

Dans le contexte d'interaction parent-enfant, les comportements vocaux des parents comportent indéniablement une dimension affective. Les parents, ou tout autre interlocuteur (caregiver), adoptent spontanément et inconsciemment un langage particulier, caractérisé par un discours simplifié et une proso-

die particulière. Ce registre de parole appelé mamanais ou *motherese* présente des caractéristiques linguistiques particulières : énoncés plus courts avec plus de répétitions et d'expansion, meilleure articulation et complexité structurale réduite, ainsi des caractéristiques paralinguistiques particulières : tonalité globale plus haute, plus larges excursions tonales, des contours intonatifs plus distincts, un tempo ralenti et pauses plus longues. Ces caractéristiques se rapprochent de celles des émotions typiques et motivent l'approche suivie dans ce document consistant à exploiter des méthodologies de l'état de l'art en reconnaissance d'émotion [Schuller *et al.*, 2007]. En complément à cette caractérisation automatique, le contexte de production du *motherese* sera précisée par la modélisation des signaux d'interaction (cf. chapitre 4).

2.2.3 Signaux émotionnels et interaction

Le rôle des émotions dans les interactions sociales n'est plus à démontrer cependant il est pertinent de préciser son intervention. Par exemple, le traitement du signal social distingue les émotions individuelles des émotions sociales [Pantic *et al.*, 2009]. Les émotions individuelles correspondent le plus souvent aux émotions primaires comme la tristesse ou la joie. Elles sont considérées comme individuelles car elles ne sont pas dirigées vers autrui. Les émotions sociales, au contraire, sont produites dans l'intention de produire un effet chez l'autre comme l'admiration ou la compassion. Ces émotions sociales permettent également de réguler l'interaction et sont essentiellement formées par des émotions dites non-prototypiques.

Signalons cependant que la catégorisation prototypique / non-prototypique ne permet pas de discriminer entre les émotions individuelles et sociales. En effet, les émotions fournissent des informations sur la façon dont les interlocuteurs agissent en établissant et validant leur identité dans l'interaction sociale [Mackinnon, 1994]. Aussi, dans une conversation, les interlocuteurs se partagent les sentiments et les émotions par rapport au contexte. Par exemple quand une personne parle avec une voix triste, la personne en face change sa voix et essaie aussi de parler avec une voix triste. Au contraire quand une personne est en colère ou agressive, son interlocuteur peut parler avec une voix neutre pour la calmer ou au contraire se met en colère, en fonction du contexte de l'interaction et de la culture des interlocuteurs.

Les études sur les échanges entre un parent et son enfant montrent l'existence d'une adaptation mutuelle [Ruth et Arthur, 2004]. Le registre de parole *motherese* produit par les parents a clairement une vocation sociale et nous cherchons dans notre étude à qualifier son impact dans l'interaction et notamment chez les enfants à devenir autistique.

2.3 Parole adressée à l'enfant : *Motherese*

Nous cherchons à étudier une émotion sociale particulière dont le support est le *motherese* et émergeant dans un contexte spécifique : celui de l'interaction parent-enfant. Ce besoin est motivé par l'application visée par notre étude : la modélisation de l'interaction sociale parent-bébé et la compréhension de l'effet du *motherese* sur le développement des enfants, plus particulièrement les enfants autistes caractérisés par un développement déviant par rapport aux enfants typiques.

2.3.1 Définition

Motherese est un terme anglo-saxon utilisé pour désigner le langage modulé de la mère adressé à son enfant. D'après la définition tirée du dictionnaire d'orthophonie (2004), le *motherese* est considéré comme l'ensemble des modulations de la prosodie de la voix maternelle présentes dans le langage spécifiquement destiné au bébé, pendant la période d'acquisition langagière. Le *motherese* est un registre spécial de parole, il se caractérise par des modifications touchant chaque niveau du langage : les niveaux phonologique, lexical, morphosyntaxique et pragmatique [Rondal, 1983] [Garitte, 1998]

- Phonétiques et phonologiques : évolution de la hauteur tonale du discours, accentuation et allongement de la durée d'émission des substantifs et des verbes, ralentissement du rythme d'élocution et présence de pauses séparant clairement les énoncés.
- Lexicales : réduction de la diversité lexicale et utilisation de mots à référence concrète, tandis que les nouveaux mots sont généralement introduits dans les fins de phrases [Fernald et Mazzie, 1991].
- Morphosyntaxiques : grammaticalité et fluidité du discours, réduction de la longueur moyenne des productions verbales, phrases interrogatives plus nombreuses que les déclaratives, elles mêmes plus importantes que les exclamatives.
- Pragmatiques : répétitions, dominance des requêtes en demande d'information, feedbacks verbaux constitués d'approbation et de désapprobations verbales du discours, de corrections explicites, d'extensions et de répétitions des énoncées de l'enfant.

Aussi ces modifications altèrent le signal de parole résultant dans des caractéristiques prosodiques et acoustiques permettant de distinguer le *motherese* de la parole normale (ou le plus souvent celle adressée à un autre adulte).

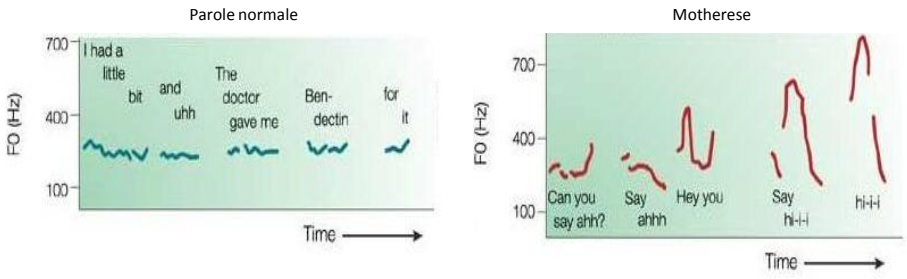


FIG. 2.2 – Le contour de pitch de *motherese* par rapport à la parole normale [Kuhl, 2004]

2.3.2 Caractéristiques prosodiques et acoustiques

Le *motherese* est caractérisé par un *pitch* augmenté, un tempo plus lent et des contours d'intonation exagérés [Gleitman *et al.*, 1984] [Grieser et Kuhl, 1988] [Fernald, 1992] (voir la figure 2.2). Les phonèmes et les voyelles sont également mieux articulés. Ainsi, les caractéristiques prosodiques, particulièrement l'exagération du contour du pitch facilite la discrimination des phonèmes par le bébé.

Toutes ces caractéristiques acoustiques semblent jouer un rôle important dans l'interaction sociale et le développement du langage [Rose *et al.*, 2003]. Il est difficile de modéliser ce registre de parole et de le détecter d'une manière automatique, voire d'une manière manuelle, comme c'est le cas pour un grand nombre de signaux de parole spontanée. En effet, la caractérisation de ces registres de parole (parole spontanée, parole affective) reste une question ouverte. Cependant, un grand nombre de paramètres acoustiques et prosodiques ont été proposés dans la littérature [Cowie *et al.*, 2001] [Schuller *et al.*, 2007], nous permettant d'envisager le développement d'un détecteur automatique.

2.4 Système état de l'art pour la détection de signaux émotionnels

2.4.1 Architecture générale du système

Un système de détection ou de reconnaissance des émotions, comme le montre la figure 2.3, se décompose en deux grandes parties. Tout d'abord, le signal émotionnel est caractérisé par des descripteurs. Puis les vecteurs de caractéristiques obtenus sont classifiés. Pour caractériser le signal de parole nous utilisons un certain nombre de descripteurs prosodiques et acoustiques. Ces

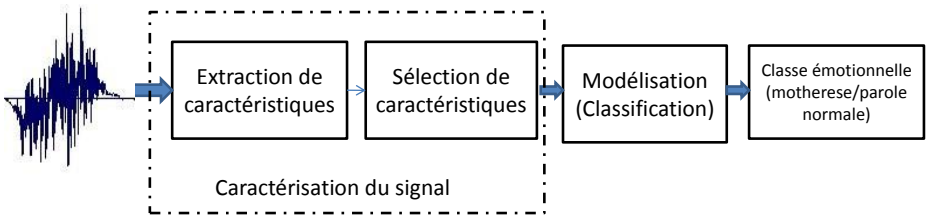


FIG. 2.3 – Système état de l'art pour la détection des signaux émotionnels

caractéristiques formeront l'entrée du système de classification/discrimination (*motherese / parole normale*).

2.4.2 Descripteurs bas niveaux

2.4.2.1 La qualité vocale

La qualité de la voix correspond généralement aux phénomènes acoustiques relatifs au timbre vocal. Elle permet de refléter l'identité, l'âge et le sexe du locuteur, mais aussi ses états affectifs. La qualité de la voix permet également de caractériser l'environnement du locuteur (bruit, vent, etc.). Dans une application de détection des émotions, la qualité vocale est principalement utilisée pour identifier les états affectifs. Les indices de la qualité vocale considérés sont aussi bien la prosodie que l'intonation, l'articulation, la hauteur, le rythme, l'intensité ou le timbre de l'émission vocale. Il s'agit d'une notion subjective.

Campbell et Mokhtari [2003] ont développé un algorithme pour la mesure de descripteurs de qualité vocale prenant en compte des modifications physiologiques dans la parole émotionnelle. Ils ont établi à partir d'une grande quantité de données, un lien entre la tendance à utiliser une qualité de voix soufflée et le degré d'attention porté à l'interlocuteur. Dans cet article, ils mettent en évidence le rôle des descripteurs modélisant la source glottale pour reconnaître les émotions. Montero *et al.* [1999] suggèrent ainsi que la qualité vocale est un facteur plus important que les autres paramètres prosodiques pour reconnaître/différencier certaines émotions.

2.4.2.2 Fréquence fondamentale

La fréquence fondamentale $F0$ correspond à la vitesse de vibration des cordes vocales et n'existe que pendant la phonation, elle n'est donc estimable que sur les parties voisées du signal de parole (structure harmonique). La $F0$

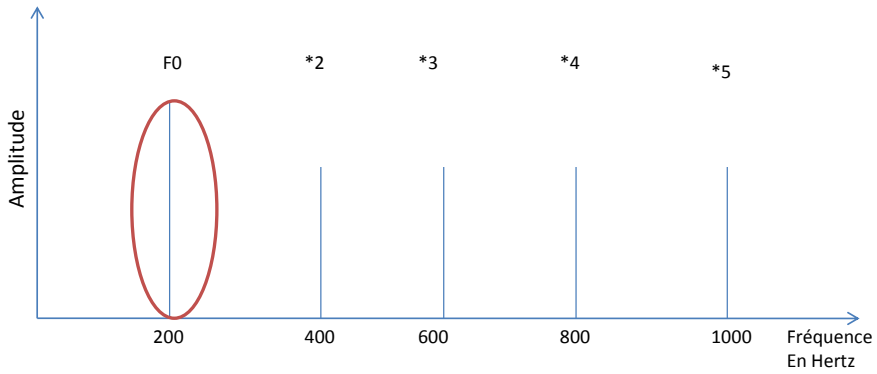


FIG. 2.4 – La fréquence fondamentale $F0$

est le corrélat physique de la notion perceptuelle de pitch.

Sachant que le signal de parole dit voisé est modélisé par un signal pseudo-périodique, la fréquence fondamentale est considérée comme l'inverse de la période T.

$$F0 = \frac{1}{T} \tag{2.1}$$

La fréquence fondamentale est aussi considérée comme la première harmonique du signal de parole comme le montre la figure 2.4. Dans les analyses acoustiques des signaux émotionnels, la $F0$ (ou le *pitch*) reste le paramètre acoustique le plus étudié du fait de sa représentativité des changements de la voix induits par l'état émotionnel du locuteur.

On distingue généralement deux modèles de perception du *pitch* [Scheirer, 2000] : le modèle de lieu et le modèle temporel. Le modèle de lieu suppose que la cochlée exécute une analyse spectrale et les pics des fréquences sont envoyés vers un processeur central, puis suivant la relation entre ces pics il estime la valeur du *pitch* [Goldstein, 1973]. Le modèle temporel suppose que la perception du *pitch* se fasse par une analyse temporelle de la périodicité dans chacune des bandes de fréquence obtenues par l'analyse spectrale de la cochlée [Slaney et Lyon, 1990].

Du point de vue des méthodes automatiques, plusieurs algorithmes d'estimation de la fréquence fondamentale ont été proposés. L'algorithme le plus utilisé est basé sur la fonction d'auto-corrélation du signal sonore dans le domaine temporel. Il consiste à rechercher des ressemblances entre des versions décalées du signal temporel $x(n)$:

$$\phi_T = \sum_{t=t0}^{t0+N} x(t)x(t+T) \tag{2.2}$$

La fréquence fondamentale est alors définie comme l'inverse de la période pour laquelle la fonction d'auto-corrélation est maximale :

$$F0 = \frac{1}{\arg\max(\phi_T)} \tag{2.3}$$

Enfin, il existe plusieurs outils d'extraction de la fréquence fondamentale. Dans notre travail, nous avons utilisé le logiciel libre d'analyse de la voix *Praat* [Boersma et Weenink, 2001].

2.4.2.3 Energie

L'énergie du signal $s(n)$ est le corrélat physique de l'intensité et est calculée à partir du signal temporel selon l'équation suivante :

$$E_s = \int_{-\infty}^{+\infty} |s(t)|^2 dt \tag{2.4}$$

L'énergie est généralement exprimée en décibels (perception auditive) :

$$E_s(dB) = 10 log_{10} E_s \tag{2.5}$$

L'évolution dans le temps de l'énergie peut déterminer le style d'intonation du locuteur. Ce paramètre est également utilisé pour la détection d'activité vocale. Dans le domaine de la détection des émotions, l'énergie d'un signal sonore (ou *loudness*) peut être utilisée pour caractériser le contenu émotionnel comme par exemple la surprise [Moncrieff *et al.*, 2001]. Cependant, au contraire de la fréquence fondamentale qui est différente selon le sexe (hommes : 60 - 100 Hz, femme/enfant : 200-300 Hz), les classes sonores homme et femme, peuvent avoir les mêmes caractéristiques énergétiques. En conséquence, ce paramètre est souvent exploité en complément d'autres paramètres.

2.4.2.4 Caractéristiques cepstrales

Les coefficients *MFCC* (Mel-Frequency Cepstral Coefficients) appartiennent à la famille des caractéristiques à court terme, largement utilisés en reconnaissance de la parole en raison de leur efficacité [Rabiner et Juang, 1993], ainsi qu'en reconnaissance des émotions [Truong et van Leeuwen, 2007] [Cowie *et al.*, 2001].

Les *MFCCs* sont basés sur une échelle de perception non linéaire appelée Mel. Celle-ci peut être définie par la relation suivante entre la fréquence en Hertz et sa correspondance en Mels :

$$mel(f) = x \times log_{10}(1 + \frac{f}{y}) \tag{2.6}$$

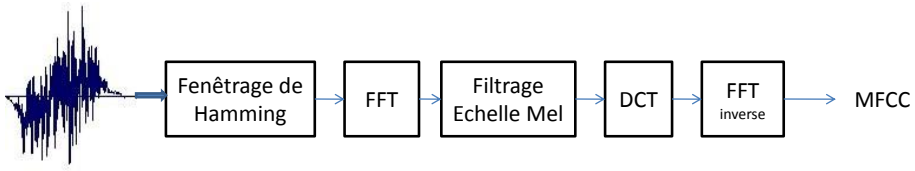


FIG. 2.5 – Processus de l'extraction des *MFCCs*

Plusieurs valeurs sont utilisées pour x et y. Les valeurs les plus couramment utilisées sont $x = 2595$ et $y = 700$.

Le processus d'extraction des *MFCCs* est composé des étapes décrites dans la figure 2.5 :

1. Fenêtrage du signal avec la fenêtre de Hamming.
2. Transformée de Fourier (FFT).
3. Filtrage par banc de filtres triangulaires espacés selon l'échelle Mel
4. Transformée en Cosinus Discrète
5. Transformé de Fourier inverse.

La première étape consiste à appliquer une pondération de Hamming sur chaque trame. Cette pondération ayant pour le but d'une part de concentrer la répartition de l'énergie sur les basses fréquences et d'autre part d'amoindrir les fortes variations du signal sur les bords de la fenêtre, variations qui entraînent une mauvaise estimation des coefficients du filtre si elles n'étaient pas atténuées. La deuxième étape consiste à calculer la FFT, cela permet d'obtenir une représentation fréquentielle du signal de parole en domaine fréquentiel. Dans la troisième étape, le spectre du signal est filtré par des filtres triangulaires (cf. figure 2.6) dont les bandes passantes sont équivalentes en fréquence exprimée en Mel. Enfin, les coefficients cepstraux peuvent être obtenus en appliquant une transformée de Fourier inverse à partir des coefficients en sorties des filtres. Cependant, le nombre de $MFCC$ est moins grand que le nombre de filtres, donc une transformée en cosinus discrète est généralement utilisée pour réduire la dimension du vecteur caractéristique. On obtient alors les coefficients $MFCC$ avec comme premier coefficient l'énergie du signal. En général, une dizaine de coefficients $MFCC$ est utilisée.

2.4.3 Techniques de classification

La classification automatique consiste à assigner des objets à des catégories ou classes. Dans notre cas, les objets sont des segments de signaux émotionnels

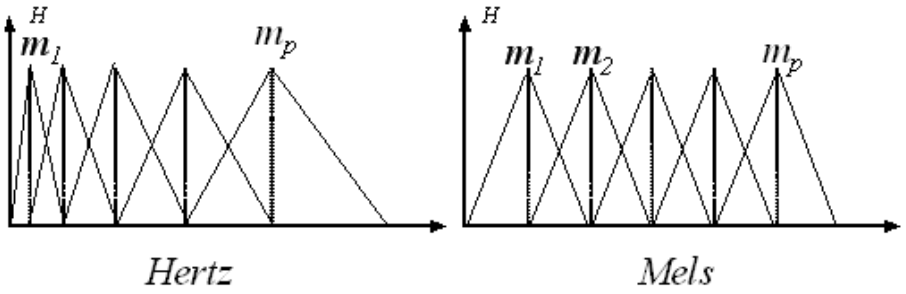


FIG. 2.6 – Filtres triangulaires pour l'extraction de coefficients *MFCC*

qu'il s'agit d'attribuer à des classes sonores définies (*motherese* et *parole normale*). La méthode d'apprentissage supervisé est appliquée lorsque l'ensemble des données et des étiquettes est fourni au système de classification. Dans le cas contraire on parlera d'apprentissage non supervisé.

Un grand nombre d'algorithmes de classification exploite l'apprentissage automatique afin d'en optimiser les paramètres et/ou de compiler les informations issues des bases de données. Le corpus joue alors un rôle primordial à la fois pour la modélisation et l'évaluation. Du fait de la diversité des caractéristiques et des classifieurs pertinents pour une tâche donnée, des méthodes de combinaison ont été proposées basées sur la fusion d'informations avec diverses approches (statistiques, vote, boosting). En général le système de classification automatique comporte deux étapes comme le montre la figure 2.7 :

- La phase d'apprentissage, qui consiste à estimer la description de l'espace des observations qui traduit le mieux l'association avec les classes correspondantes. Il s'agit donc de déterminer des fonctions de décision qui se basent sur les vecteurs caractéristiques précédemment calculés à partir de la base d'apprentissage. Ce sont ces fonctions de décision qui permettront d'associer une classe à un nouveau vecteur de caractéristiques.
- La phase de test, qui permet d'évaluer les performances du système de classification. Elle consiste à associer une classe au vecteur de caractéristiques par la fonction de décision apprise précédemment.

Plusieurs systèmes de classification supervisée ont été développés : arbre de décision, réseaux bayesiens, machines à vecteurs de support (SVM), plus proches voisins ($kppv$), mélange de gaussiennes (GMM) et réseaux neuronaux (RN), etc.

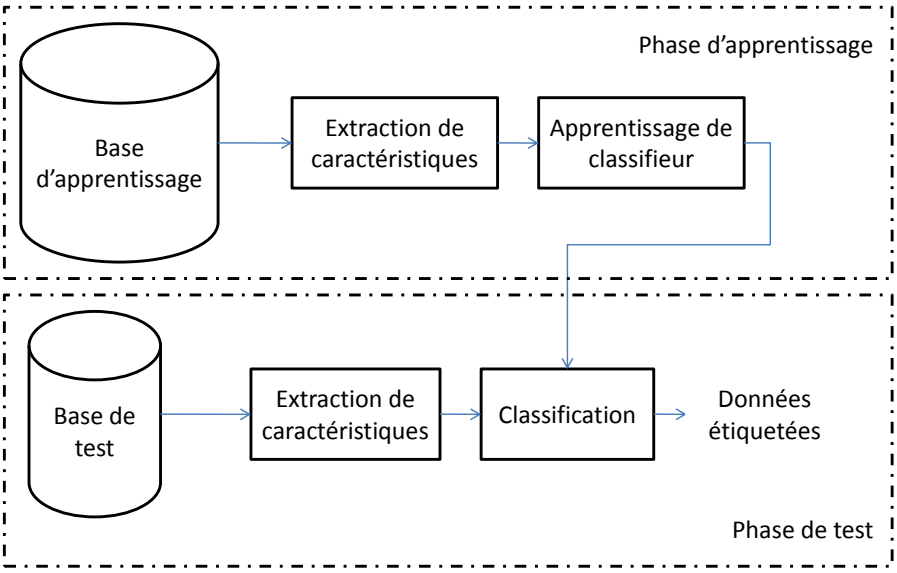


FIG. 2.7 – Système de classification automatique

2.4.4 K plus proches voisins : *kppv*

La méthode *Kppv* est une méthode non-paramétrique de classification : aucune estimation de paramètres n'est nécessaire [Duda *et al.*, 2000]. Cette méthode consiste à calculer la distance entre les exemples de la base de test et ceux de la base d'apprentissage. Ensuite, afin de définir l'étiquette de l'exemple de test, nous déterminons les k voisins les plus proches en exploitant les distances estimées.

Un modèle statistique peut-être dérivé du classifieur *kppv* en considérant les classes d'appartenance des voisins.

- Lorsque $k = 1$, l'estimation des probabilités exploite les distances aux classes :

$$P_{kppv}(C_n|x) = \frac{e^{-d_n}}{\sum_{i=1}^{N} e^{-d_i}} \tag{2.7}$$

  où C_n représente la catégorie (C_1 : *motherese* et C_2 : parole normale), x est le vecteur de caractéristique de la base du test, d_n est la distance minimale pour la classe C_n.
- Lorsque $k \neq 1$: nous estimons les k voisins les plus proches. Ensuite, les probabilités locales qui sont basées sur les classes des voisins sont définies de la manière suivante [Duda *et al.*, 2000] :

$$P_{kppv}(C_n|x) = \frac{k_n}{k} \tag{2.8}$$

où k_n représente le nombre des voisins de la classe C_n parmi les k voisins les plus proches.

L'avantage majeur du modèle statistique est que l'estimation de probabilité a posteriori permet d'intègrer et de combiner le classifieur *kppv* dans une méthodologie statistique.

2.4.5 Mélange de Gaussiennes : *GMM*

Le *GMM* (Gaussian Mixture Models) [Reynolds, 1995] est une méthode statistique utilisée pour modéliser la distribution des données de chaque classe dans un espace de descripteurs de dimension D. Cet espace est obtenu par la combinaison linéaire de plusieurs fonctions de densité de probabilité (Probability Density Functions *PDF*) de dimension D, ce qui permet d'approximer la forme globale de la collection de descripteurs.

La densité de mélange de gaussiennes est une somme pondérée de M composantes [Reynolds, 1995] :

$$P_{GMM}(x|C_n) = \sum_{i=1}^{M} w_i g_{(\mu_i,\Sigma_i)}(x) \tag{2.9}$$

où $P(x|C_n)$ est la densité de probabilité de la classe C_n évaluée à x, $g_{(\mu_i,\Sigma_i)}(x)$ sont les composantes de densités des gaussiennes et w_i sont les poids du mélange. Chaque composante de densité est définie par la matrice de covariance Σ_i et la moyenne μ_i, est calculé de la manière suivante :

$$g_{(\mu,\Sigma)}(x)\frac{1}{(2\pi^{d/2})\sqrt{det(\Sigma)}}e^{1/2(x-\mu)^T\Sigma^{-1}(x-\mu)} \tag{2.10}$$

Les poids du mélange w doivent satisfaire la contrainte suivante :

$$\sum_{i=1}^{M} w_i = 1 \tag{2.11}$$

L'apprentissage des *GMMs*, caractérisé par l'apprentissage des paramètres des modèles (μ,Σ), s'effectue par l'algorithme Expectation Maximisation (*EM*) qui garantit en théorie une convergence vers une solution optimale [Dempster *et al.*, 1977].

D'après la formule de Bayes, la probabilité à posteriori est obtenue grâce à la formule suivante :

$$P_{GMM}(C_n|x) = \frac{p(x|C_n)p(C_n)}{p(x)} = \frac{P(x|C_n)P(C_m)}{\sum_j P(x|C_j)P(C_j)} \tag{2.12}$$

où $p(C_n)$ est la probabilité à priori de la classe C_n et $p(x)$ est la fonction de densité de probabilité totale de la classe C_n évaluée à x.

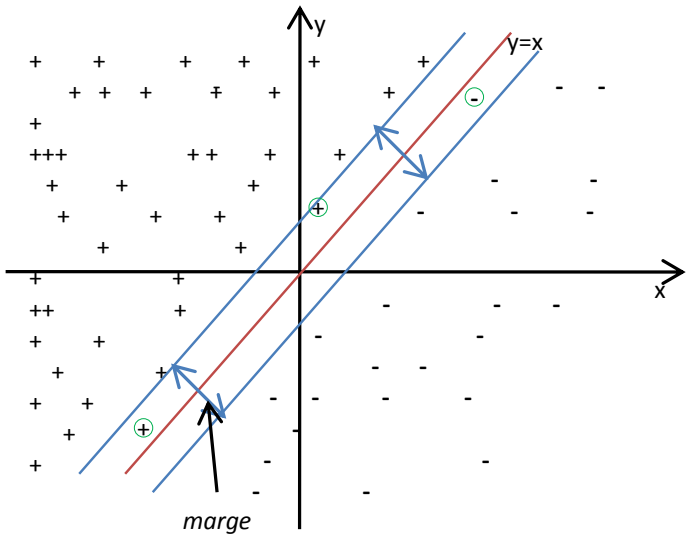


FIG. 2.8 – Exemple d'un problème de discrimination à deux classes, avec un séparateur linéaire

2.4.6 Machines à vecteurs de support : *SVM*

Cette technique, initiée par Vapnik [1995], consiste à estimer un hyperplan qui sépare les exemples positifs (*motherese*) des exemples négatifs (*parole normale*), en garantissant que la marge entre le plus proche des positifs et des négatifs soit maximale (cf. la figure 2.8).

Dans le cas linéaire (les données sont linéairement séparables), étant donné un ensemble des exemples $\{(x_1, y_1), ..., (x_l, y_l)\}$, avec x_i les vecteurs de caractéristiques de dimension D, $y_i \in \{-1, +1\}$ les étiquettes dans un problème à deux classes, $y_i = -1$ si l'exemple est négative et $y_i = +1$ si l'exemple est positive, l'algorithme estime l'hyperplan optimal permettant de maximiser la marge entre les observations correspondant aux deux classes (cf. la figure 2.8). L'hyperplan est donné par la solution du problème d'optimisation :

$$\min \frac{1}{2}||w||^2.w \tag{2.13}$$

avec les contraintes suivantes :

$$\forall i, y_i(w.x_i + b) \geq 1 \tag{2.14}$$

L'équation 2.13 est un problème d'optimisation sous contraintes qui est résolu

en introduisant les multiplicateurs de Lagrange $(\alpha_i)_{1\leq i\leq l}$ et un Lagrangien :

$$L(w,b,\alpha_i) = \frac{1}{2}||w||^2 - \sum_{i=1}^{l} \alpha_i \left[y_i(w.x_i + b) - 1\right] \tag{2.15}$$

où $\alpha = [\alpha_1 ... \alpha_l]^T$

Le Lagrangien L doit être minimisé par rapport aux variables dites primales w, b, et maximisé par rapport aux variables duales α_i : ce sont les conditions de Karush-Kuhn-Tucker (KKT) [Schoelkopf et Smola, 2002]. Après résolution, on retrouve que l'hyperplan optimal ne dépend que des n_s vecteurs supports du problème :

$$w = \sum_{i=1}^{n_s} \alpha_i y_i x_i \tag{2.16}$$

et la fonction de décision est définie par le signe de :

$$f(x) = w.x + b = \sum_{i=1}^{n_s} \alpha_i y_i (x.x_i) + b \tag{2.17}$$

Cependant, dans le cas où les données ne sont pas linéairement séparables, le problème ci-dessus n'a pas de solution discrète. Une solution consiste alors à rendre les contraintes moins rigides en introduisant des variables d'écart positives ξ_i ($\xi_i \geq 0$). Les contraintes deviennent alors [Vapnik, 1998] :

$$y_i(w.x_i + b) \geq 1 - \xi_i \tag{2.18}$$

Pour résoudre ce problème, on change la fonction d'optimisation :

$$\frac{1}{2}||w||^2 + C\sum_{i=1}^{l} \xi_i \tag{2.19}$$

Le paramètre $C > 0$ permet de régler le compromis entre la maximisation de la marge et la minimisation des erreurs de classification sur l'ensemble d'apprentissage. On parle alors de classificateurs à marge souple.

En pratique, quelques familles de fonctions à noyau paramétrable sont largement utilisées (polynomiale, gaussien, sigmoïde et laplacien) et il revient à l'utilisateur de SVM d'effectuer des tests pour déterminer celle qui convient le mieux à son application.

Le SVM standard ne permet pas d'aboutir directement à des probabilités à posteriori de classification, cependant pour garder une approche statistique commune aux autres classifieurs (fusion bayésienne), la sortie du SVM doit être représentée sous forme d'une probabilité.

Au lieu d'estimer les densités conditionnelles des classes $p(f|y)$, un modèle paramétrique est utilisé afin d'obtenir les probabilités a posteriori $p(y = 1|f)$ directement où f représente la valeur obtenue à la sortie du *SVM*. La toolbox *LIBSVM* [Chang et Lin, 2001] propose une méthode de calcul de ces probabilités à partir des paramètres obtenus par *SVM*. Ainsi les probabilités a posteriori sont déterminées à l'aide d'une fonction paramétrique sigmoïde par [Platt, 1999] :

$$P(y = 1|f) = \frac{1}{1 + exp(Bf + C)'} \tag{2.20}$$

dans laquelle les paramètres B et C sont estimés en utilisant la procédure du maximum de vraisemblance.

2.4.7 Réseaux de neurones : *RN*

Les réseaux de neurones sont composés d'éléments simples (ou neurones) fonctionnant en parallèle. Ces éléments ont été fortement inspirés par le système nerveux biologique. Le fonctionnement du réseau de neurones est influencé par les connections des éléments entre eux. On peut entraîner un réseau de neurones pour une tâche spécifique (détection des émotions par exemple) en ajustant les valeurs des connections (ou poids) entre les éléments (neurones). Plusieurs topologies de réseaux existent : les réseaux multicouches, les réseaux à connexions locales, les réseaux à connexion récurrentes et les réseaux à connexions complètes. Les réseaux multicouches (cf. la figure 2.9), sont les plus répandus, du fait de la simplicité du mécanisme d'apprentissage. Ils sont organisés en couches, chaque neurone prend généralement en entrée tous les neurones de la couche inférieure. Ils ne possèdent pas de cycles ni de connexions intra-classe. On définit alors une couche d'entrée, une couche de sortie, et n couches cachées.

Le neurone formel est l'unité élémentaire du réseau. Il effectue la somme pondérée de ses entrées, et la soumet à une fonction non linéaire dérivable. Pour un neurone formel possédant n entrées, la sortie est exprimée par :

$$y = \sum_{i=1}^{n} w_i x_i \tag{2.21}$$

puis active sa sortie grâce à une fonction non linéaire $z = f(y)$. Plusieurs fonctions sont utilisées pour l'activation comme la fonction sigmoïde :

$$g(a) = \frac{1}{1 + exp(-a)} \tag{2.22}$$

Dans notre travail, nous utilisons un réseau multicouche *MLP* [Bishop, 1995]. Considérons par exemple un réseau à une couche cachée (figure 2.9),

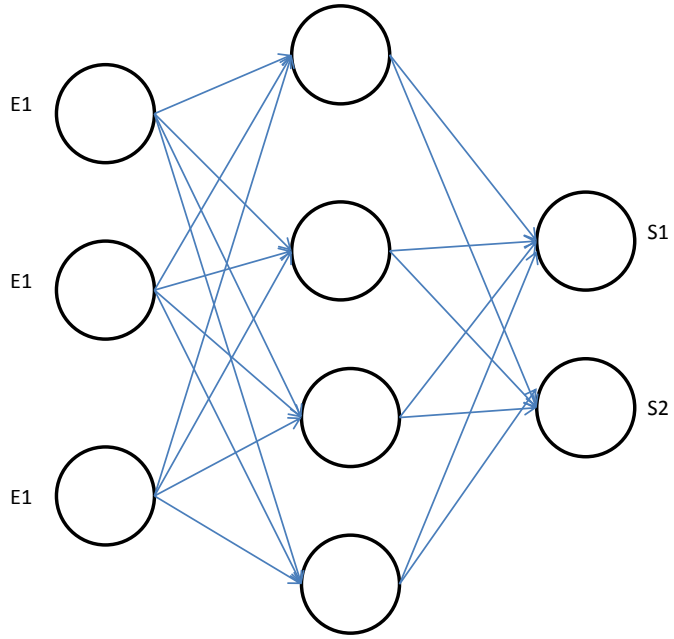


FIG. 2.9 – Réseau multicouche à une couche cachée, trois entrées et deux sorties.

possédant m cellules d'entrées $x_i = E_i$, une couche cachée à n neurones d'activation y_j et une couche de sortie à p neurones d'activation z_k, ce qui donne $n \times m$ connexions entre la couche d'entrée et la couche cachée (avec pondération v_{ij}) et $m \times p$ connexions entre la couche cachée et la couche de sortie (avec pondération w_{kj}). La première étape de l'apprentissage consiste à initialiser les poids des connexions (le plus souvent aléatoirement). Ensuite, pour chaque cellule d'entrée E_i, on propage les signaux vers la couche cachée :

$$y_j = f(\sum_{i=1}^{m} x_i v_{ij} + x_0) \tag{2.23}$$

Puis de la couche cachée vers la couche de sortie :

$$z_k = f(\sum_{j=1}^{n} y_j w_{kj} + y_0) \tag{2.24}$$

Les valeurs x_0 et y_0 sont des biais, des scalaires et non des sorties de la couche précédente. La troisième étape consiste à calculer les erreurs sur les couches de sortie pour chaque exemple de la base d'apprentissage appliqué en entrée du réseau : la différence entre la sortie désirée s_k et la sortie réelle/cible z_k :

$$E_k = z_k(1 - z_k)(s_k - z_k) \tag{2.25}$$

On propage cette erreur sur la couche cachée, l'erreur de chaque neurone de la couche cachée est donnée par :

$$F_j = y_j(1 - y_j) \sum_{k=1}^{p} w_{kj} E_k \tag{2.26}$$

Lors de la quatrième étape, il reste à modifier les poids des connexions, d'abord entre la couche d'entrée et la couche cachée

$$\begin{aligned} \Delta w_{kj} &= \eta y_j E_k \\ \Delta x_0 &= \eta E_k \end{aligned} \tag{2.27}$$

puis entre la couche cachée et la couche de sortie :

$$\begin{aligned} \Delta v_{ji} &= \eta x_i F_j \\ \Delta y_0 &= \eta F_k \end{aligned} \tag{2.28}$$

Finalement, il faut ré-itérer au niveau de la deuxième étape jusqu'à satisfaire un critère d'arrêt.

L'estimation des probabilités par le réseau est réalisée par la modification des fonctions d'activation de la couche de sortie. Bridle [1989] préconise l'utilisation de la fonction *Softmax* en imposant des contraintes sur les sorties :

$$\begin{aligned} &o_j > 0, \forall j \\ &\sum_{j=1}^{N} o_j = 1 \end{aligned} \tag{2.29}$$

où les o_j représentent une des N sorties du *MLP*. L'utilisation de la fonction *Softmax* exploite ces conditions :

$$o_j = \frac{e_j^I}{\sum_k e_k^I} \tag{2.30}$$

où les O_j sont les sorties, I sont les entrées.

2.5 Fusion de données

Il existe plusieurs manières de caractériser le signal de parole : descripteurs et/ou classifieurs différents avec à chaque fois des performances particulières. Le caractère optimal d'une solution n'est généralement pas garanti. La fusion d'informations ou de données offre alors un cadre théorique adapté pour la combinaison des modèles proposés. L'état émotionnel se traduisant par des

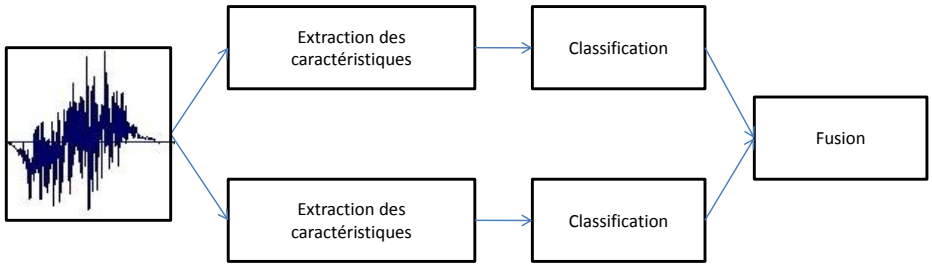


FIG. 2.10 – Fusion des données

modifications diverses du signal de parole, une caractérisation avec des descripteurs multiples est généralement suivie. Cependant, la fusion de données n'améliore pas toujours la classification, et une étude sur la complémentarité et/ou diversité des différents descripteurs et classifieurs est généralement nécessaire.

2.5.1 Méthodes de fusion

Dans la littérature, les méthodes se distinguent selon le niveau de la fusion : capteurs, espace des caractéristiques, opinion (ou score) ou bien encore décision [Monte-Moreno *et al.*, 2009]. Les deux principales approches sont la méthode *feature-level fusion* et la méthode *decision-level fusion* [Mahdhaoui *et al.*, 2008].

La fusion au niveau de l'espace des caractéristiques consiste à concaténer toutes les caractéristiques en un seul vecteur exploité ensuite par un classifieur unique. Par exemple, prenons deux vecteurs de caractéristique, un vecteur $X = (x_1, x_2, ..., x_n)$ de dimension n et un vecteur $Y = (y_1, y_2, ..., y_m)$ de dimension m, le résultat de la fusion est un vecteur $Z = (X, Y)$ de dimension $n+m$. La fusion au niveau de l'opinion des classifieurs consiste à prendre en compte toutes les prédictions établies par les différents classifieurs, puis à combiner les probabilités pour estimer la catégorie finale de l'objet à classifier [Gunatilaka et Baertlein, 2001], voir la figure 2.10. Une variété de méthodologies de fusion a été développée pour les applications en reconnaissance des formes [Varshney, 1997][Stanley *et al.*, 2002]. En reconnaissance de parole expressive, la fusion consiste à extraire des paramètres à des échelles temporelles différentes [Kim *et al.*, 2007] [Scherer *et al.*, 2007] [Chetouani *et al.*, 2009]. Kim *et al.* [2007] ont proposé une fusion de type *decision-level* pour un système temps réel de reconnaissance des émotions. Les auteurs ont combiné des caractéristiques cepstrales *MFCCs* et des caractéristiques prosodiques (*pitch* et *énergie*)

pour la discrimination de la parole neutre et l'émotion tristesse. Les auteurs ont montré expérimentalement que la fusion des caractéristiques donne de meilleurs résultats que le système à caractérisation unique, le taux d'erreur a été diminué de 27.3% à 33% selon différentes expériences, par rapport à la méthode conventionnelle. Scherer *et al.* [2007] ont combiné des caractéristiques spectrales *RASTA-PLP* [Hermansky *et al.*, 1991] et *l'énergie* pour développer un système de reconnaissance d'émotions. Les auteurs ont évalué plusieurs combinaisons de caractéristiques en utilisant un seul classifieur, ils ont montré que la fusion améliore significativement les performances du système de classification. Grâce à la fusion de caractéristiques, ils ont réussi à diminuer le taux d'erreur de classification d'environ 25% en combinant les descripteurs prosodiques et spectraux.

2.5.2 Méthodologie employée

Dans notre travail, nous avons utilisé une méthode simple pour la fusion des systèmes de reconnaissance (extraction de caractéristiques + classification), il s'agit d'une somme pondérée calculée à partir de la formule suivante :

$$P_{final} = \lambda \log P_1(C_n|x) + (1-\lambda) \log P_2(C_n|x) \tag{2.31}$$

Où P_1 et P_2 représentent les probabilités à posteriori respectivement du premier et deuxième système, x représente le segment à classifier et λ est le poids alloué à chacun des systèmes (variant de 0 à 1).

Certains classifieurs, comme le *kppv*, ne sont pas adaptés à cette approche car les probabilités estimées peuvent être nulles ($P_i = 0$). Nous avons mis à jour la formule précédente en introduisant l'exponentiel de probabilité initialement proposé par [Kim *et al.*, 2007] :

$$P_{final} = \lambda \log e^{P_1(C_n|x)} + (1-\lambda) \log e^{P_2(C_n|x)} \tag{2.32}$$

2.5.3 Méthodes de décision

Après la fusion des classifieurs, nous obtenons deux probabilités P_{C1} et P_{C2} qui correspondent respectivement à l'appartenance à la classe *motherese* et à la classe *non motherese*. Pour estimer la catégorie du segment, nous avons utilisé deux méthodes de décision :

1. Classification : L'hypothèse de classification correspond à la classe qui a la probabilité maximale. Si $P_{C1} > P_{C2}$, la classe du segment est *motherese*, tandis que si $P_{C1} < P_{C2}$, la classe du segment est *non motherese* :

$$C = \arg\max \{P_{fusion}(C_n|x)\} \tag{2.33}$$

2. Détection : L'hypothèse de classification est déterminée selon un seuil avec la formule suivante :

$$C = \begin{cases} C_1; \zeta < \theta \\ C_2; \zeta \geq \theta \end{cases} \tag{2.34}$$

où

$$\zeta = P_{fusion}(C_2|x) - P_{fusion}(C_1|x) \tag{2.35}$$

θ est le seuil de classification.

2.6 Bases de données

Les études sur le développement humain et les paroles émotionnelles requièrent des bases de données pour optimiser et évaluer les systèmes automatiques. On distingue deux types de bases de données : les bases de données actées et les bases de données vécues dans des contextes réels : "real-life" (comme les films familiaux).

2.6.1 Les bases de données actées

Les bases de données actées contiennent des émotions simulées, enregistrées dans des conditions artificielles, dans un laboratoire, studios ou salle acoustique (local insonorisé). Il s'agit de demander à un ensemble d'acteurs professionnels de simuler certaines émotions.

L'avantage de ce type de données est que le texte est contrôlé, car on demande au locuteur de lire des phrases avec une intonation donnée (permettant par la suite d'exploiter des informations lexicales par exemple [Schuller *et al.*, 2008]), de plus ces bases de données sont généralement exemptes de bruits ou d'hésitations.

La plupart des études sur la reconnaissance automatique des signaux émotionnels [Shami et Verhelst, 2007] [Ververidis et Kotropoulos, 2006] se basent sur des données actées.

2.6.2 Les bases de données d'émotions vécues en contexte réel

Les bases de données d'émotions "real-life" sont plus complexes à constituer [Devillers *et al.*, 2005]. La plupart des études utilisent des enregistrements et des extraits vidéo issus d'émissions de télévision [Devillers *et al.*, 2006]. Ainsi, pour traiter des émotions dans des applications réelles, Vidrascu et Devillers

[2005] ont étudié des données issues de conversations de centres d'appel. Clavel *et al.* [2006] ont développé le corpus *SAFE*, de type real-life. Le corpus *SAFE* contient 400 séquences audiovisuelles (durée moyenne de 8 secondes à 5 minutes). C'est une collection d'enregistrements vidéo extraits d'oeuvres cinématographiques récentes. Ce corpus présente des scènes d'actions normales et des scènes qui montrent des actions d'agression (menace physique, menace psychologique, enlèvement, prise d'otage, etc.). Le corpus est constitué de 7 heures d'enregistrements, dont 76% de parole dont 71% de scènes d'agressions.

Par ailleurs, dans le cadre d'interaction mère-enfant, quelques études ont montré que les mères à qui on demande de parler à un magnétophone comme si elles parlaient à leurs bébés, ne sont pas capables de produire la même qualité de parole, ni la même variété de courbes prosodiques du *motherese* qu'en présence du bébé dans des conditions réelles [Fernald, 1984].

Le nombre d'études sur la reconnaissance automatique de signaux émotionnels réels est assez limité par rapport aux nombreuses études sur les émotions actées. Cela ne diminue pourtant pas la nécessité ou l'importance de ce genre d'étude. Au contraire, le besoin de connaissance sur l'expression émotionnelle dans des occurrences naturelles devient de plus en plus grand. Les systèmes des reconnaissances d'émotions se doivent d'être robuste à la manière de produire les émotions (actée, spontanée) [Clavel *et al.*, 2008]. Une solution consiste à optimiser les systèmes sur des données réelles afin de s'assurer de leurs capacités de généralisation.

Corpus longitudinal Afin de permettre une étude longitudinale des états émotionnels et de leurs effets sur l'interaction face à face, il faut disposer d'un corpus longitudinal. Dans notre analyse de parole adressée à l'enfant (*motherese*), nous étudions un ensemble de films familiaux enregistrés dans des conditions réelles. Il s'agit d'enregistrements d'interactions spontanées entre des parents et des bébés, particulièrement entre mères et bébés sur une durée de plusieurs mois. Les films familiaux consistent en des enregistrements vidéo filmés par l'un des membres de la famille dans des conditions complètement réelles et dans des endroits différents (cuisine, salon, jardin, etc), ainsi que durant différents évènements (anniversaire, bain, repas, etc).

Les films familiaux permettent aux cliniciens d'accéder à des informations sur les comportements des enfants autistes avant que les diagnostics ne soient établis, sachant qu'il n'est pas possible de diagnostiquer la pathologie de l'autisme avant l'âge de deux ans.

Nous avons extrait aléatoirement des séquences audio de la base de données développée par Maestro *et al.* [2005], les séquences correspondent à 42 enfants (15 enfants avec développement normal, 12 enfants avec retard mental et 15 enfants autistes). Les différents segments sont des enregistrements audio de

mères italiennes s'adressant à leurs bébés généralement âgés de 0 à 2 ans. Plus des détails pour la base d'interaction parent-enfant sont donnés dans la section 4.4.1.

L'inconvénient majeur de l'utilisation de corpus "real-life" est l'absence d'étiquetage des données. L'annotation manuelle de tous les films est très coûteuse en temps, il n'est donc pas raisonnable d'annoter manuellement la totalité de la base de données. Par conséquent, il est nécessaire de développer une approche automatique. Nous allons chercher à définir expérimentalement le *motherese* par l'annotation d'une partie des données. Ensuite, ces données permettront de développer un détecteur automatique.

2.7 Annotation

2.7.1 Stratégie d'annotation

Le système état de l'art de détection des signaux émotionnels nécessite un grand nombre des données étiquetées pour l'apprentissage, par conséquent le processus d'annotation est une étape fondamentale pour le développement du système de classification.

Du fait de la complexité de la tâche d'annotation et afin de minimiser les erreurs d'annotation, nous avons demandé à deux psycholinguistes d'étiqueter les différents segments de parole (cf. voir la figure 2.11). D'abord le premier annotateur a segmenté les flux audio pour extraire des segments homogènes de parole : ne contenant que la voix de la mère. Chaque segment correspond à une phrase (un tour de parole ou une partie du tour de parole avec un contenu émotionnel homogène) dont la durée varie de 0.5 à 4 secondes. Ensuite, les segments sont catégorisés en deux classes : *motherese* et *parole normale* par les deux annotateurs séparément. Afin de s'assurer que les annotations fournies par les deux annotateurs offrent une bonne convergence, nous avons évalué la fiabilité des annotateurs en calculant la fidélité inter-juge. Enfin, afin de limiter la diversité des annotations, nous sélectionnons les segments annotés identiquement par les deux annotateurs.

2.7.2 Fidélité inter-juge

Du fait du caractère subjectif des émotions, les différentes annotations peuvent diverger. L'évaluation de l'accord entre les annotateurs est donc une étape nécessaire. Les méthodes statistiques sont souvent utilisées pour évaluer la similarité entre des annotations. Elles consistent à regrouper les différentes annotations dans une matrice, appelée matrice de contingence (cf. voir le tableau 2.1).

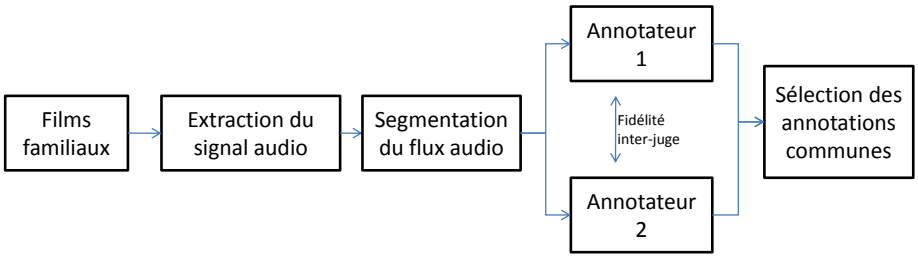


FIG. 2.11 – Stratégie d'annotation

TAB. 2.1 – Matrice de contingence

	Annotateur B			
Annotateur A	Etiquette	Motherese	Parole normale	Somme
	Motherese	N_{11}	N_{12}	$N_{1.}$
	Parole normale	N_{21}	N_{22}	$N_{2.}$
	Somme	$N_{.1}$	$N_{.2}$	N

Les nombres N_{ii} sont les nombres de fois où les deux annotateurs ont annotés de la même façon les même segments, tandis que les N_{ij} sont les nombres des fois où l'annotateur A a annoté un segment comme *motherese* alors que l'annotateur B l'a annoté comme *parole normale* ou l'inverse, et N est la somme de toutes les observations faites par les deux annotateurs.

Le test statistique *Kappa* est une méthode paramétrique largement utilisée pour calculer la fidélité inter-juge, elle a été proposée par Cohen [1960]. Elle permet de chiffrer l'intensité ou la qualité de l'accord réel entre des jugements qualitatifs appariés. Pour deux partitions à même nombre de classes, le coefficient *Kappa* mesure l'écart à la diagonale du tableau de contingence :

$$kappa = \frac{n \sum_{i=1}^{k} N_{ii} - \sum_{i=1}^{k} N_{i.} N_{.i}}{N^2 \sum_{i=1}^{k} N_{i.} N_{.i}} \tag{2.36}$$

La concordance observée P_o est la proportion d'individus classés dans les cases diagonales de concordance du tableau de contingence, soit la somme des effectifs diagonaux divisés par la taille de l'échantillon N :

$$P_o = \frac{1}{n} \sum_{i=1}^{k} N_{ii} \tag{2.37}$$

La concordance aléatoire P_e est égale à la somme des produits des effectifs marginaux divisés par le carré de la taille de l'échantillon :

$$P_e = \frac{1}{n^2} \sum_{i=1}^{k} N_{i.} N_{.i} \tag{2.38}$$

TAB. 2.2 – Interprétation de la valeur de *kappa*

kappa	Interprétation
0<	désaccord
0.0 - 0.20	accord très faible
0.21 - 0.40	accord faible
0.41 - 0.60	accord modéré
0.61 - 0.80	accord fort
0.81 - 1.00	accord presque parfait

TAB. 2.3 – Matrice de contingence : Exemple

	Annotateur B			
Annotateur A	Etiquette	Motherese	Parole normale	Somme
	Motherese	50	15	60
	Parole normale	20	20	40
	Somme	70	30	100

Le *kappa* exprime la différence relative entre la proportion d'accords observés P_o et la proportion d'accords aléatoires P_e qui est la valeur espérée, sous l'hypothèse nulle d'indépendance des variables, divisée par le complément à un de l'accord aléatoire.

$$kappa = \frac{P_o - P_e}{1 - P_e} \tag{2.39}$$

Le coefficient *kappa* est un nombre réel compris entre -1 et 1, trois situations sont généralement distinguées :

- $kappa = 1$: accord parfait entre les annotateurs ($P_o = 1$)
- $kappa = 0$: indépendance entre les annotateurs ($P_o = P_e$)
- $kappa = -1$: désaccord total entre les annotateurs ($P_o = 0$ et $P_e = 0.5$)

Ainsi, Landis et Koch [1977] ont proposé une interprétation de la valeur de *kappa* selon son ordre de grandeur (cf. tableau 2.2)

Exemple : Etant donné 100 segments de parole à annoter et deux annotateurs, avec deux classe *motherese/parole normale*. Supposons que la matrice de contingence entre les deux annotateurs soit comme présentée dans le tableau 2.3 : pour calculer la valeur de concordance, d'abord nous calculons les effectifs marginaux, $a = 70 \times 60/100 = 42$, puis $P_e = 42 + 12/100 = 0,54$. Enfin, la valeur de *kappa* est la suivante : $kappa = \frac{P_o - P_e}{1 - P_e} = \frac{0,7 - 0,54}{1 - 0,54} = 0,34$.

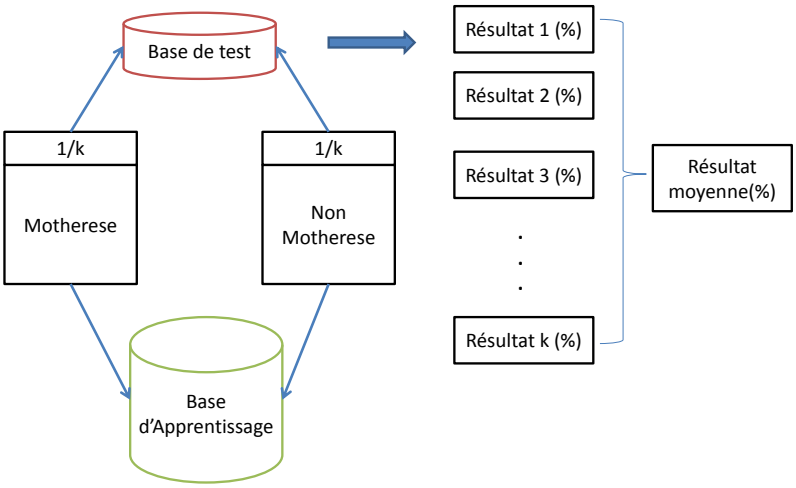


FIG. 2.12 – k-folds cross validation

2.8 Expérimentations

2.8.1 Protocole d'évaluation

2.8.1.1 Estimation de l'erreur de généralisation

Une procédure simple d'estimation de l'erreur de généralisation est la validation croisée dite *k-folds cross validation* (cf voir la figure 2.12). C'est une méthodologie statistique qui permet d'évaluer les performances d'un système de classification. Elle consiste à diviser l'ensemble des données en k sous-ensembles mutuellement exclusifs de taille approximativement égale. L'apprentissage de l'algorithme de classification est effectué en utilisant $k-1$ sous-ensembles et le test est effectué sur le sous-ensemble restant. Cette procédure est répétée k fois et chaque sous-ensemble est utilisé une fois pour le test. La moyenne des k taux d'erreur obtenus estime l'erreur de généralisation.

La technique *leave-one-out* représente le cas extrême de la méthode *k-folds corss validation*. Elle consiste à diviser l'ensemble des données en deux : une base d'apprentissage de $n-1$ exemples et une base de test ne contenant qu'un seul exemple.

2.8.1.2 Métriques d'évaluation

Le problème de détection de *motherese* est un problème de classification binaire. Nous avons utilisé la matrice de confusion pour évaluer la qualité de classification. Les lignes de la matrice de confusion représentent le nombre d'exemples d'une classe estimée automatiquement (hypothèse), tandis que les

TAB. 2.4 – Matrice de confusion

		Référence		
		Positive (C_1)	Negative(C_2)	
Hypothèse	Positive (C_1)	vrai positive (TP)	faux positive (FP)	$TP+FP$
	Négative (C_2)	faux négative (FN)	vrai négative (TN)	$FN+TN$
		$TP+FN$	$FP+TN$	

colonnes représentent le nombre d'exemples d'une classe réelle (référence) [Girard et Girard, 1999].

Le tableau 2.4 montre une matrice de confusion d'un problème de classification binaire. Cette matrice permet d'estimer les séquences de motherese correctement détectées (positifs) mais également d'affiner l'évaluation de la précision du système (faux positifs FP et faux négatifs FN).

1. TP est le nombre d'exemples positifs classés comme positifs
2. FP est le nombre d'exemples négatifs et classés comme positifs
3. TN est l'ensemble d'exemples négatifs classés comme négatifs
4. FN est l'ensemble d'exemples positifs et classés comme négatifs

Toujours à partir de la matrice de confusion (tableau 2.4), plusieurs métriques combinent les résultats précédents pour former l'exactitude (accuracy), la sensibilité et la spécificité :

$$Exactitude = \frac{TN+TP}{TN+FN+FP+TP} \tag{2.40}$$

$$Sensibilité = \frac{TP}{FN+TP} = TPR \tag{2.41}$$

$$Spécificité = \frac{TN}{FP+TN} = 1-FPR \tag{2.42}$$

L'exactitude est le nombre de prédictions justes rapportées au nombre total de prédictions, la spécificité représente le taux de classement corrects de la classe C_1 (*motherese*), tandis que la sensibilité représente le taux de classement corrects de la classe C_2 (*non motherese*).

2.8.1.3 Graphes *ROC*

Pour évaluer les performances du système, nous traçons également la courbe *ROC* (Receiver Operating Characteristic) [Duda *et al.*, 2000]. La courbe *ROC* représente le compromis entre le taux de vrais positifs (TPR) et le taux de faux positifs (FPR) dans un espace *ROC*. TPR et FPR caractérisent l'espace *ROC*.

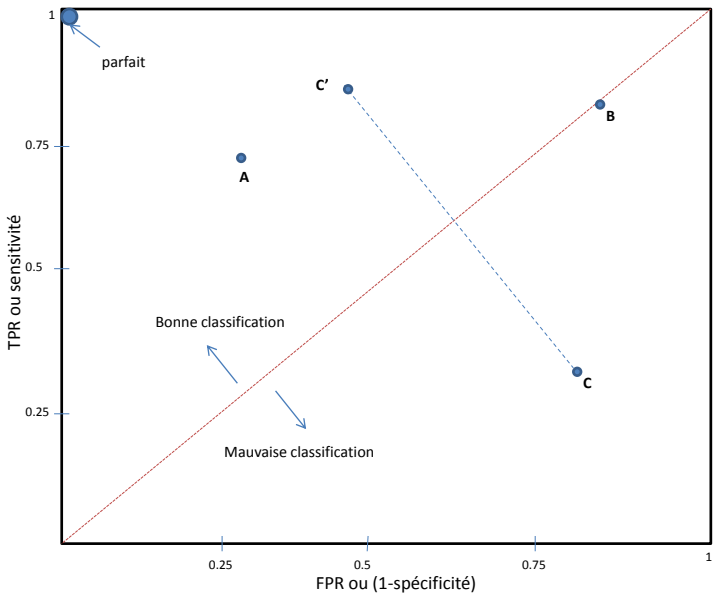


FIG. 2.13 – Espace *ROC*

Espace *ROC* Un espace *ROC* est défini par *FPR* (les abscisses) et *TPR* (les ordonnées), voir la figure 2.13. Puisque *TPR* est égale à sensibilité et *FPR* est égale à (1 - spécificité), l'espace *ROC* est représenté également par la sensibilité en fonction de (1 - spécificité).

La ligne diagonale divise l'espace *ROC* en deux zones. Les points au-dessus de la ligne diagonale indiquent des bons résultats de classification, tandis que les points au-dessous de la ligne indiquent les mauvais résultats.

Courbe *ROC* La courbe *ROC* est avant tout définie pour les problèmes à deux classes (les positifs et les négatifs), elle indique la capacité d'un classifieur à placer les positifs devant les négatifs. Sa construction s'appuie donc sur les probabilités d'être positif fournies par les classifieurs. Cependant, il n'est pas nécessaire que ce soit réellement une probabilité, une valeur quelconque dite "score" permettant d'ordonner les exemples suffit amplement. Ensuite, on fait varier la valeur du seuil de la mesure, pour chacune de ces variations on calcule la sensibilité et la spécificité du test au seuil fixé, chaque valeur du seuil est associé à un couple (sensibilité, spécificité) que l'on reporte dans l'espace *ROC* (cf. voir la figure 2.14).

Un classifieur avec un fort pouvoir discriminant occupera la partie supérieure gauche du graphique, voir la figure 2.15. Tandis qu'un classifieur avec un pouvoir discriminant moins puissant montrera une courbe *ROC* qui s'apla-

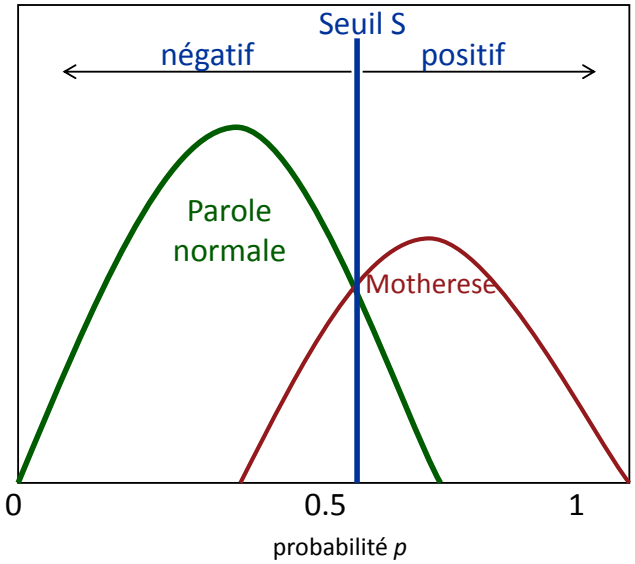


FIG. 2.14 – Détection du motherese : classes et seuil

tira vers la première diagonale du graphique. On peut utiliser la courbe pour décider quel est le seuil portant le meilleur compromis entre sensibilité et spécificité. Il s'agira du seuil où la courbe ROC montre un point d'inflexion (*C index*), voir la figure 2.15. On peut aussi comparer deux courbes ROC pour mettre en évidence quel est le test qui possède le meilleur pouvoir discriminant. Ce sera le test qui aura la courbe ROC la plus creuse, courbe verte dans la figure 2.15.

Aire sous courbe ROC (AUC) L'aire sous la courbe ROC est également une information pertinente et est notée AUC (Area Under ROC curve). Plus AUC est proche de 1 meilleur est le système. L'AUC peut être considerée comme la somme des aires de trapèzes dans l'espace ROC et se calcule de la manière suivante :

$$AUC = \frac{(Y(i) + Y(i+1)) \times (X(i+1) - X(i))}{2} \tag{2.43}$$

où X sont les abscisses et Y ordonnées, i représente l'indice des points et N est le nombre des points.

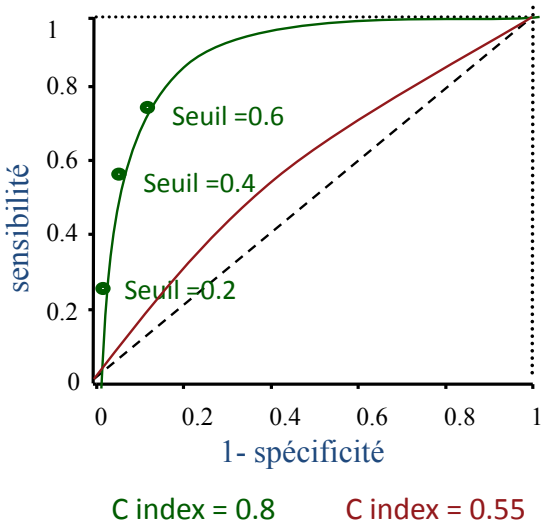


FIG. 2.15 – Courbe *ROC*

2.8.2 Résultats

2.8.2.1 Corpus d'étude

La base de données a été annotée par deux annotateurs selon la méthode décrite dans la section 2.7.1. La fidélité inter-juge entre les deux annotations est égale à Cohen's *kappa*=0.82. Selon le tableau 2.2, cela représente un accord fort entre les annotateurs. A partir de cette annotation, nous avons sélectionné 500 segments qui sont identiquement annoté par les deux annotateurs, 250 segments *motherese* et 250 segments *non motherese*, la durée moyenne des segments est de 0.5 à 4 secondes. La durée totale des segments de parole est de 15 minutes. La figure 2.16 montre la distribution des segments de *motherese* et de *non motherese*.

Les exemples sont extraits sur des périodes différentes. La figure 2.17 montre la distribution de données par semestre. Ces segments sont issus de 7 dyades mère-enfant. Le nombre de locuteurs est donc de 7. De plus, les productions de parole des mères sont assez variantes car les enregistrements sont faits durant 2 ans (étude longitudinale). Cela justifie l'indépendance en locuteur. Cependant, notre application est dépendante du genre, nous traitons uniquement la voix de la mère. Les performances des classifieurs sont estimées par la méthode de 10-folds cross validation.

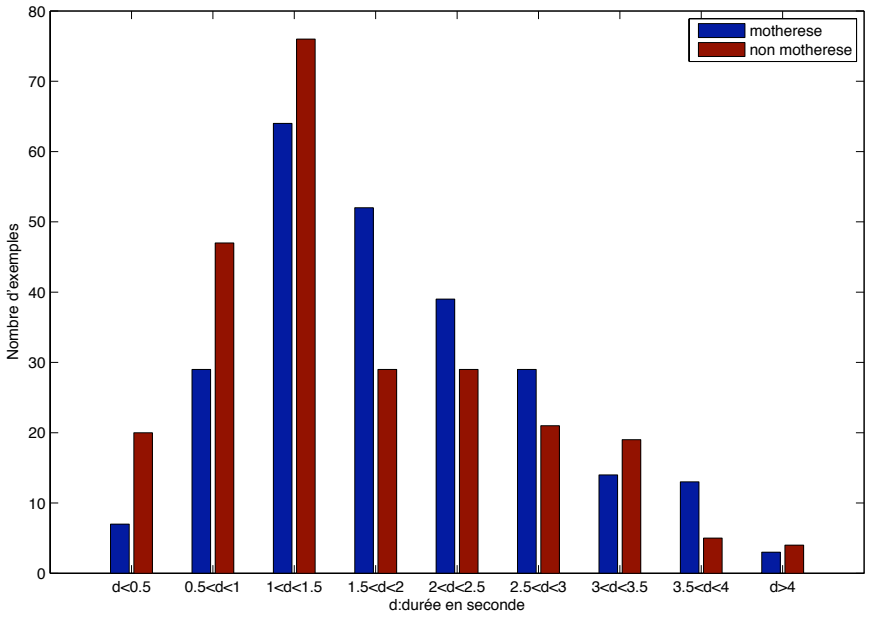


FIG. 2.16 – Distribution de données

2.8.2.2 Descripteurs acoustiques pour la caractérisation du *motherese*

Les différentes caractéristiques présentées ci-dessus ont été utilisées en tant que caractéristiques acoustiques du contenu émotionnel. D'abord, l'extraction de caractéristiques cepstrales ou segmentales est effectuée comme le montre la figure 2.18, la taille des trames est de 1024 échantillons (64 ms), l'entrelacement est de 512 échantillons (32 ms). La fréquence d'échantillonnage est de 16000Hz. La dimension du codage est définie à 16. Après la phase de codage, nous obtenons une matrice de dimension $16 \times M$, qui représente l'ensemble des caractéristiques segmentales, M étant le nombre de trames.

Pour analyser les caractéristiques supra-segmentales, il y a deux méthodes : l'approche *turn-level* et l'approche *segment-based*. La première méthode consiste à représenter le segment par un seul vecteur de caractéristique. On représente ainsi l'évolution générale des paramètres prosodiques sur la totalité du segment de parole. En utilisant cette méthode, nous obtenons un seul vecteur de caractéristique de dimension 6, où 6 est le nombre de statistiques appliquées sur les valeurs du pitch et d'énergie du segment étudié (la moyenne, la variance et l'étendue). La deuxième méthode considère uniquement les zones voisées du signal de parole. Supposons qu'un exemple x est

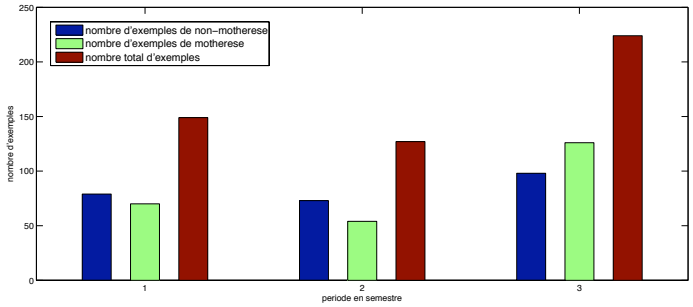


FIG. 2.17 – Distribution de données par semestre

segmenté en plusieurs zones voisées F_x. Pour chaque partie, la caractéristique supra-segmentale est composée par les six mesures statistiques du pitch et de l'énergie suivant : la moyenne, la variance et l'étendue. Trois mesures sont utilisées pour décrire le contour du pitch, et trois pour l'énergie. Enfin, nous obtenons une matrice de dimension $6 \times N$, N est le nombre de zones voisées. Aussi, la dimension de la matrice dépend du nombre de zones voisées. La classification des caractéristiques supra-segmentales, basées uniquement sur les zones voisées, suit l'approche *SBA* (segment based approach) proposée par Shami et Kamel [2005], voir la figure 2.19. Pour chaque segment de parole x, la probabilité à posteriori est estimée à partir des probabilités locales des N zones voisées F_x :

$$P_{supra}(C_n|x) = \sum_{i=1}^{N} P_{supra}(C_n|F_{xi}) \times length(F_{xi}) \tag{2.44}$$

où $length(F_{xi})$ représente la durée de chaque zone voisée. Ensuite, la probabilité globale est normalisée par la durée des zones voisées :

$$P_{supra}(C_n|x) = \frac{1}{\sum_{i=1}^{N} length(F_{xi})} P_{supra}(C_n|x) \tag{2.45}$$

Les descripteurs précédents sont normalisés (moyenne nulle et écart-type unitaire) selon la formule suivante (*z-normalisation*) [Truong et van Leeuwen, 2007] :

$$\hat{F}_x = (F_x - \mu)/\sigma \tag{2.46}$$

F_x est un vecteur caractéristique segmental ou suprasegmental, μ est la moyenne de tous les codes et σ est l'écart-type.

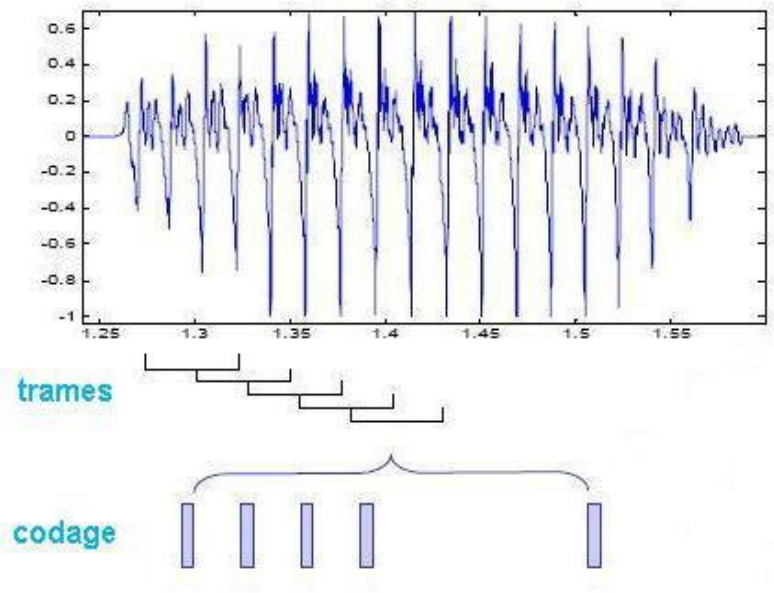


FIG. 2.18 – Codage MFCC de la parole

TAB. 2.5 – Processus d'optimisation du classifieur *GMM* avec les caractéristiques segmentales et les caractéristiques supra-segmentales

Descripteurs	M : nombre de gaussiennes														
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Segmentaux	59	53.5	60.5	58	55	51	57.5	60	64	62	70.5	67.5	62	63.5	72.75
Supra-segmentaux	60	52	58.5	61	55	52	59.5	60	65	62	65.5	64	61.5	64	69.5

2.8.2.3 Résultats des systèmes de classification

D'abord, nous avons commencé par optimiser les paramètres des classifieurs. Cette étape consiste à trouver le nombre optimal de gaussiennes M pour le classifieur *GMM* (cf. le tableau 2.5) et le nombre optimal de voisins (k) pour le classifieur *kppv* (cf. le tableau 2.6), ainsi que le kernel optimal utilisé avec le classifieur *SVM* (cf. le tableau 2.7), en utilisant les deux caractéristiques : segmentale et supra-segmentale. Pour optimiser le classifieur *SVM*, nous avons testé quatre types de noyau : linéaire, polynomial, gaussien et laplacien, leurs performances sont comparables.

Les tableaux 2.8 et 2.9 regroupent respectivement les performances optimales des classifieurs *GMM* et *kppv*, chacun des deux classifieurs considère deux types de caractéristiques : les caractéristiques segmentales (*MFCC*) et les caractéristiques supra-segmentales. En comparant les quatres classifieurs, concernant le *GMM*, la meilleure configuration est obtenue avec les caracté-

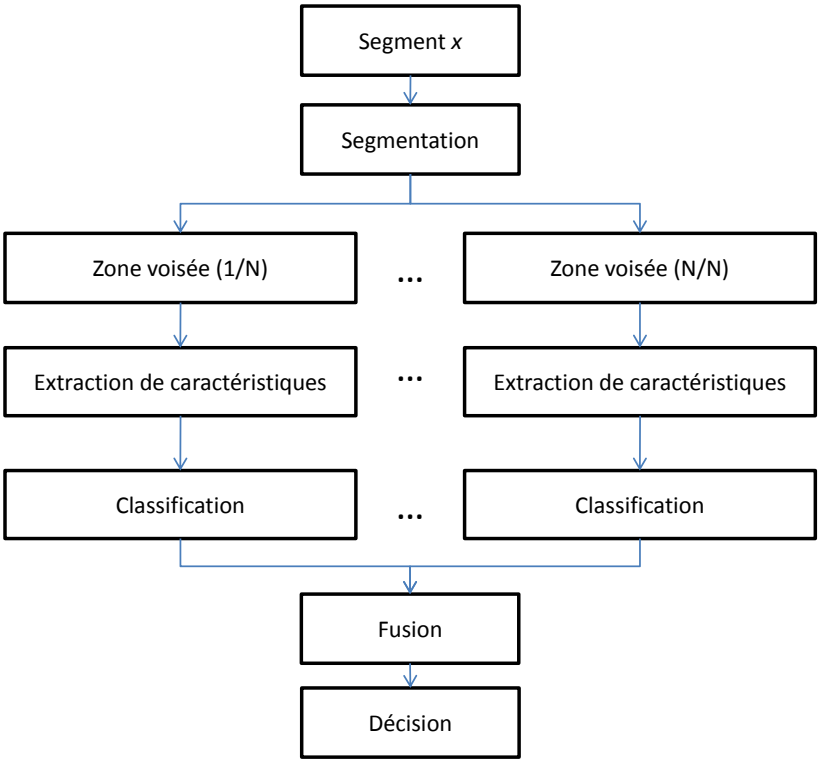


FIG. 2.19 – Approche SBA : *Segment-based approach*

ristiques segmentales (72.75% avec *MFCC* vs. 69.5% avec les caractéristiques supra-segmentales). Les meilleures performances sont obtenues en utilisant 16 gaussiennes. Concernant le classifieur *kppv* la meilleure configuration est également obtenue avec les caractéristiques segmentales (70.25% avec *MFCC* vs. 65.5% avec les caractéristiques supra-segmentales). Les meilleures performances sont obtenues en utilisant respectivement $k = 11$ et $k = 7$ voisins.

Le classifieur *SVM* ne donne pas de performances satisfaisantes, les meilleurs résultats obtenus en utilisant un noyau linéaire, 59.75% avec les caractéristiques segmentales et 54.75% avec les caractéristiques supra-segmentales. Une explication est le manque de données qui sont souvent nécessaires pour une bonne généralisation. Enfin, le classifieur *MLP* donne des résultats satisfaisants en utilisant les caractéristiques segmentales (61.5%), tandis que le résultat n'est pas satisfaisant en utilisant les caractéristiques supra-segmentales (50.25%).

Du point de vue caractéristiques, nous avons cherché à exploiter la définition phonétique du motherese (cf. section 2.3) qui est généralement qualifié par une modulation de la prosodie (i.e. caractéristiques supra-segmentales).

TAB. 2.6 – Processus d'optimisation du classifieur *kppv* avec les caractéristiques segmentales et les caractéristiques supra-segmentales

Descripteurs	k : nombre de plus proches voisins										
	1	3	5	7	9	11	13	15	17	19	21
Segmentaux	62.5	65	67	67.25	69	70.25	63	65.25	66	67.5	69
Supra-segmentaux	61	64.5	64	65.5	62	59.25	64.5	65	61.25	62	63.25

TAB. 2.7 – Performances du classifieur *SVM* en utilisant différents types de noyaux

	Caractéristiques segmentales	Caractéristiques supra-segmentales
linéaire	59.75%	54.75%
polynomial	58.5 %	54 %
gaussien	59%	53.75%
laplacien	57%	54 %

Le taux de classification obtenu par le classifieur basé sur les caractéristiques supra-segmentales est de 69.5% alors que celui exploitant les *MFCCs* obtient 72.75%. Ce résultat montre la difficulté de développer un système de classification de signaux émotionnels en se basant uniquement sur les définitions phonétiques et ceci pour plusieurs raisons. Une première raison est que, dans notre cas, il s'agit d'émotions spontanées et non-prototypiques et surtout il s'agit d'une définition expérimentale obtenue par annotation. Une seconde raison possible réside dans la variabilité des données. Notre base de données est constituée de voix de mères avec des intervalles d'enregistrement importants : les films retracent des interactions de 18 mois. De plus, l'environnement sonore est très variable : microphones différents et en mouvement, bruit des activités connexes (bain, cuisine, famille, parc, etc.).

En résumé, les résultats expérimentaux montrent la supériorité du classifieur *GMM* exploitant des *MFCC* (72.75%). Une fois encore cette approche est pertinente en reconnaissance automatique des émotions.

Afin de bénéficier des caractéristiques court et long termes, nous avons combiné les caractéristiques segmentales et supra-segmentales. Les résultats expérimentaux de fusion, en utilisant la somme pondérée, montrent que la fusion des caractéristiques a amélioré les performances par rapport aux résultats obtenus avec les caractérisations uniques. La combinaison des caractéristiques cepstrales *MFCC* et caractéristiques prosodiques (*pitch* et *énergie*) avec un classifieur *GMM* donne de meilleurs résultats par rapport au meilleur système obtenu en utilisant uniquement le *GMM* et les *MFCC*, ainsi que la courbe *ROC* représentée dans la figure 2.20 le montre. La fusion de classifieurs améliore la qualité de classification par rapport aux systèmes avec une caractéristique unique. $\lambda = 0$ correspond aux caractéristiques cepstrales *MFCC*, $\lambda = 1$ cor-

TAB. 2.8 – Performance du système de classification de *motherese* avec le classifieur *GMM* (en %)

	Segmentaux	Supra-segmentaux
$Taux$	72.75%	69.5%
M	16	16

TAB. 2.9 – Performance du système de classification de *motherese* avec le classifieur *kppv* (en %)

	Segmentaux	Supra-segmentaux
$Taux$	70.25%	65.5%
k	11	7

respond aux caractéristiques prosodiques et $\lambda = 0.7$ correspond à la fusion de deux systèmes (meilleure fusion). La figure 2.20 montre également la contribution plus importante du *MFCC* ($\lambda = 0.7$) par rapport aux caractéristiques prosodiques ($1 - \lambda = 0.7$).

2.9 Conclusions

Cette partie a été dédiée à la description de notre système de détection du *motherese*. Dans un premier temps nous avons définis le phénomène émotionnel et son utilisation par rapport au contexte social d'interaction. Ensuite nous avons étudié les particularités de la parole adressée à l'enfant, le *motherese*. Ce registre de parole, qui est défini par sa prosodie, est caractérisé par un *pitch* augmenté, un tempo plus lent et des contours d'intonation exagérés [Fernald et Kuhl, 1987]. Ce registre de parole joue un rôle important dans l'interaction sociale parent-enfant et le développement du langage de l'enfant. Dans un deuxième temps, nous avons présenté l'architecture générale d'un système de classification automatique de signaux émotionnels. Nous avons commencé par présenter les descripteurs bas niveaux que nous avons utilisé pour caractériser le signal de *motherese*. Ensuite, nous avons présenté les différentes techniques de classification exploitées pour la classification des vecteurs caractéristiques. A la fin de cette étape, nous avons développé un système de classification statistique supervisée de *motherese*. Ce système est composé de deux étapes, une première étape consiste à apprendre un modèle de classification sur des données étiquetées, et une deuxième étape consiste à catégoriser les données de test en utilisant le modèle généré pendant la phase d'apprentissage. Les différents systèmes de classification développés dans ce chapitre ont été appliqués sur une base de données réelles, cette base de données est une collection de

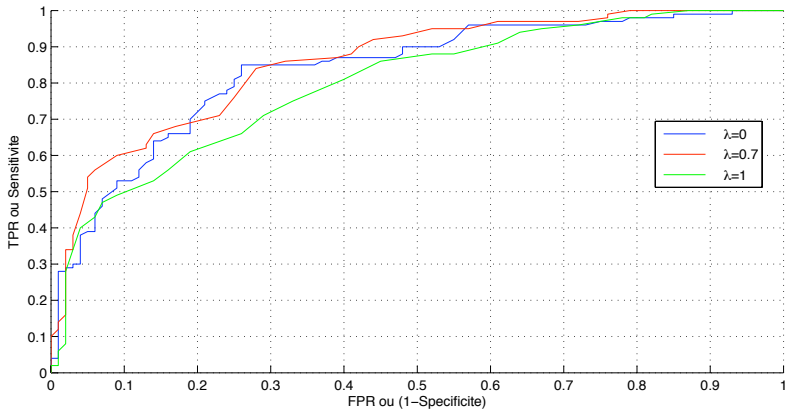


FIG. 2.20 – Courbe ROC : fusion des caractéristiques cepstrales *MFCC* et prosodiques en utilisant le classifieur *GMM*

séquences d'interaction parent-enfant naturelles et spontanées. Ces systèmes offrent des résultats satisfaisants, l'inconvénient majeur est qu'ils nécessitent beaucoup de données d'apprentissage pour construire le modèle de classification.

Pour améliorer les performances des systèmes de classification supervisée, nous avons combiné les caractéristiques et les classifieurs en utilisant une méthode de fusion basée sur une simple somme pondérée (cf. les équations 2.31 et 2.32).

Malgré la définition théorique du *motherese*, défini par la prosodie, les résultats expérimentaux montrent que le système composé d'un classifieur *GMM* et des descripteurs acoustiques *MFCC* offre les meilleures performances de classification. Ce résultat montre encore une fois la pertinence de l'approche *GMM+MFCC*. Ensuite, le résultat obtenu a été légèrement amélioré par la fusion des caractéristiques prosodiques et cepstrales. Malgré les résultats obtenus, nous sommes assez loin d'une reconnaissance parfaite. Plusieurs directions de recherche sont envisageables pour améliorer la qualité de classification. Une meilleure caractérisation du *motherese* dans le domaine spectrale peut conduire à de meilleures performances. De plus, l'ensemble des systèmes développés sont complètement supervisés, ils nécessitent beaucoup des données étiquetées pour l'apprentissage, alors que dans le cas d'analyse des données réelles, l'annotation des données n'est pas toujours évidente. Par conséquent, nous avons développé une nouvelle approche de classification, dite classification semi-supervisée, qui consiste à combiner les données étiquetées et non étiquetées pour l'apprentissage (cf. chapitre 3).

Chapitre 3

Classification semi-supervisée de signaux émotionnels

Sommaire

3.1 Introduction

La classification supervisée du *motherese* requiert la constitution et l'annotation d'une base de données. De cette base dépendent les performances du système. Le nombre important de films à traiter (1200 enregistrements), les situations diverses associées ainsi que la nécessité de préciser expérimentalement la définition du *motherese*, nous impose de trouver des solutions automatiques adaptées à ces problématiques. L'apprentissage semi-supervisé propose un cadre formel permettant de renforcer les règles de catégorisations

des systèmes de classification supervisée en combinant apprentissage et prédiction sur des données étiquetées et non étiquetées. Ce chapitre a pour objectif la description d'un algorithme de co-apprentissage de classification du *motherese*.

Dans un premier temps, nous rappelons les principales approches d'apprentissage automatique : apprentissage supervisé et non supervisé ainsi que l'apprentissage semi-supervisé. Ensuite, nous présentons les approches état de l'art en classification semi-supervisée et en particulier, nous nous focaliserons sur l'algorithme de co-apprentissage initié par Blum et Mitchell [1998] et Nigam *et al.* [2000] car offrant un mécanisme pertinent de renforcement des règles de catégorisation. Le co-apprentissage permet l'exploitation d'un grand nombre de classifieurs et par cette occasion nous proposons de mettre en oeuvre des méthodes d'extraction de caractéristiques différentes de celles déjà utilisées dans le chapitre 2. Elles sont basées sur des fonctionnelles de descripteurs bas niveaux du spectre inspirées de l'état de l'art en reconnaissance des émotions. Dans ce contexte, nous proposons une nouvelle méthode de co-apprentissage automatique pour la classification de signaux émotionnels. Il s'agit d'un algorithme itératif qui permet de fusionner plusieurs caractéristiques et classifieurs. La fin de ce chapitre est dédiée à la comparaison de notre approche avec celles de l'état de l'art.

3.2 Apprentissage automatique

L'apprentissage automatique en classification consiste à estimer des règles/fonctions de catégorisation f à partir d'observations : données apprentissage ici désignée par D. L'ensemble D peut regrouper les données X et leurs étiquettes Y et l'on parle alors d'apprentissage supervisé dans le cas contraire, il s'agit d'apprentissage non-supervisé.

3.2.1 Apprentissage supervisé

La classification supervisée suppose qu'il existe déjà une catégorisation des objets. Son principe est de rattacher les objets X à leurs classes d'appartenance Y, qui sont choisies parmi un ensemble de classes connues. Il s'agit ainsi d'établir des règles ou fonctions de catégorisation permettant par la suite de prédire la classe de nouveaux objets.

Pour cette approche, on dispose d'un ensemble de données étiquetées :

$$D = \{(x_t, y_t) | x_t \in X, y_t \in Y\}_{t=1}^{n} \tag{3.1}$$

où n représente le nombre de paires d'entrées : avec x_t le vecteur de caractéristique et y_t l'étiquette. Y correspond à l'ensemble fini de classes auxquelles

peuvent appartenir les différentes entrées possibles $x \in X$. Dans le cas où x est une séquence, Y peut aussi être un ensemble de séquences de classes.

Un apprentissage complètement supervisé nécessite une investigation manuelle pour l'annotation des données. Il s'agit alors de demander à un ensemble d'experts d'associer les objets à des classes prédéfinies comme nous l'avons fait pour la constitution de la base de données *motherese-parole normale* (cf. chapitre 2). L'algorithme consiste à optimiser les paramètres du modèle, ici un classifieur, afin de pouvoir reproduire automatiquement le processus d'étiquetage réalisé par un humain. Pour une entrée x_t quelconque, l'algorithme permet de prédire la valeur de la cible y_t qui aurait normalement été donnée par un humain. Deux stratégies sont généralement distinguées pour la résolution d'un problème d'apprentissage :

- Approche générative qui modélise la distribution jointe $P(X, Y)$ des entrées et des sorties. On compte parmi les méthodes génératives : l'analyse discriminante linéaire ou bien encore les mélanges de modèles gaussiens (GMM).
- Approche discriminante qui modélise directement la règle de classification $P(X|Y)$ comme les *SVMs* ou *kppv*.

A noter qu'il n'est pas toujours possible d'avoir accès à des données étiquetées pour des raisons diverses (annotation manuelle très coûteuse). Il est aussi nécessaire d'avoir un expert spécialiste pour étiqueter les données (psycholinguiste, musicien, etc.) ce qui n'est pas toujours possible.

Par conséquent, des approches alternatives ont été proposées pour pallier le manque de données étiquetées. L'apprentissage non supervisé offre un cadre théorique pertinent pour l'étude de données non étiquetées.

3.2.2 Apprentissage non supervisé

La classification de type non supervisé consiste à diviser un groupe hétérogène de données D en sous-groupes (*clusters*) de manière à ce que les données considérées comme les plus similaires soient associées au sein d'un groupe homogène et qu'au contraire, les données considérées comme différentes se retrouvent dans des groupes distincts. L'objectif étant de permettre une extraction de connaissance organisée à partir de ces données. Dans cette méthode, l'ensemble des exemples ne contient que des vecteurs caractéristiques (sans étiquettes) :

$$D = \{x_t | x_t \in X\}_{t=1}^{n} \tag{3.2}$$

où n représente le nombre d'entrées x_t.

Cette volonté de division est bien sûr ambiguë et le plus souvent formalisée par l'objectif de définir des groupes d'exemples (la distance entre les exemples

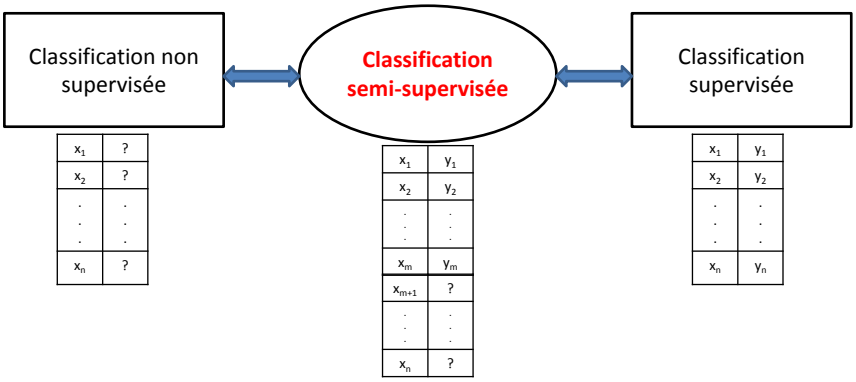


FIG. 3.1 – Apprentissage semi-supervisé

d'un même groupe est minimale et la distance entre les groupes est maximale). Plusieurs techniques d'apprentissage non supervisé ont été proposées. Citons les méthodes de regroupement hiérarchiques [Fisher, 1987], les méthodes de regroupement basées sur les *k-moyennes* [Zhang *et al.*, 1997] ou celles basées sur la théorie des graphes [Hinnerburg et Keim, 1999]. L'algorithme le plus répandu est certainement celui des *k-moyennes* [Zhang *et al.*, 1997] basé sur la quantification vectorielle. Le principe des *k-moyennes* est le suivant : d'abord on dispose d'observations que l'on souhaite rassembler en classes ou clusters, sans que l'on dispose de connaissance à priori des propriétés particulières sur ces clusters ; seul leur nombre n est fixé à priori. Il s'agit d'un processus itératif où chaque itération est composée de deux étapes :

- Recherche, pour chaque point d'observation, de son meilleur représentant parmi n référents, où chaque référent représente une classe ;
- Optimisation de chacun de ces référents pour qu'ils représentent au mieux les points d'observations en n classes.

Les approches non-supervisée sont génératives dans le sens où elles exploitent la structure intrinsèque des données. Elles supposent une variable latente (la classe de l'objet). Cependant, dans le cas de notre problématique de détection du *motherese* dans les films familiaux, l'approche non supervisée n'est pas appropriée. La variable latente associée à la classe ne l'est plus car partiellement observée pour les données étiquetées par les psychologues. L'approche semi-supevisée offre la possibilité de renforcer le modèle de classification en exploitant les données non étiquetées.

3.3 Apprentissage semi-supervisé

3.3.1 Définition et état de l'art

Dans l'objectif d'améliorer notre détecteur de parole émotionnelle, nous exploitons une approche de classification semi-supervisée qui permet de combiner des données étiquetées et des données non étiquetées pour l'apprentissage. La classification semi-supervisée se situe entre la classification supervisée et la classification non supervisée (cf. la figure 3.1) [Chapelle *et al.*, 2006]. On peut ainsi chercher à catégoriser les données non étiquetées en contraignant le nombre de classes ou se basant sur le prédicteur. On peut également chercher à améliorer les performances de la règle de catégorisation apprise. L'apprentissage semi-supervisé intervient lorsqu'on dispose à la fois d'un ensemble de données étiquetées $< (x_1, y_1), ..., (x_m, y_m) >$ et non étiquetées $< (x_{m+1}, ?), ..., (x_n, ?) >$ (cf. figure 3.1).

Depuis les années 1970, de nombreux travaux ont été proposés autour de l'apprentissage semi-supervisé, parmi les premiers nous citons les travaux de O'Neill [1978] et les travaux de Ganesalingam et McLachlan [1978]. Ces travaux reposent sur le concept de données manquantes. Dans la même période, Dempster *et al.* [1977] ont proposé un algorithme d'estimation du maximum de vraisemblance par itération successive en deux étapes. Un peu plus tard, vers les années 1990, grâce au développement des dispositifs d'acquisition de données qui ont permis d'avoir beaucoup de données disponibles, plusieurs chercheurs se sont intéressés à ce domaine de recherche. La plupart des travaux étaient consacrés à la classification de documents et de pages web. Nigam *et al.* [2000] ont proposé une méthode partiellement supervisée pour la classification automatique de textes. Ils ont présenté un des tous premiers algorithmes semi-supervisé pour apprendre les paramètres du modèle génératif Naïve Bayes pour la classification de textes. Enfin, Chapelle *et al.* [2006] ont publié un livre sur ce thème de recherche très actif, ce qui constitue une preuve de l'intérêt actuel suscité par ce sujet.

3.3.2 Techniques d'apprentissage semi-supervisé

Les solutions proposées au problème de l'apprentissage semi-supervisé dans la littérature s'appuient sur des hypothèses et sur des modèles différents. Nous trouvons les méthodes génératives et discriminatives, ainsi que d'autres méthodes ne reposant pas sur un modèle probabiliste. Parmi les méthodes les plus utilisées nous citons les méthodes génératives, les méthodes prédictives en particulier les méthodes transductives, la méthode *self-training* et les méthodes de co-apprentissage automatique (*co-training*) [Zhu, 2006].

3.3.2.1 Les méthodes génératives

Quand le problème de l'utilisation de données partiellement étiquetées a été posé pour la première fois, les modèles génératifs ont été les premiers à être adaptés pour combiner des données étiquetées et non étiquetées pour l'apprentissage automatique [Hosmer, 1973] [McLachlan, 1977]. Les premiers travaux de Hosmer [1973] ont consisté à chercher la proportion de mâles et de femelles d'une population des flétans (grand poisson plat des mers froides) à partir d'une base de données partiellement étiquetées. L'utilisation d'échantillons étiquetés avec une grande quantité d'échantillons non étiquetés a permis à Hosmer d'estimer plus précisément des paramètres d'apprentissage. D'un point de vue théorique, Hosmer [1973] a réalisé son travail par maximum de vraisemblance grâce à un algorithme itératif, qui été développé par la suite par Dempster *et al.* [1977] en un algorithme, *EM*, connu et assez utilisé dans les problèmes de données partiellement étiquetées. Cet algorithme est mieux adapté pour traiter les problèmes de données manquantes ce qui est le cas des données non-étiquetées pour lesquelles les étiquettes constituent les données manquantes. Les modèles considérés consistent principalement en des mélanges de distributions de gaussiennes et des mélanges de distributions multinomiales [Cooper et Freeman, 1970].

Principe de l'algorithme *EM* Comme nous l'avons dit dans la section 3.2.2, les méthodes de classification non supervisée supposent une variable latente. Ainsi, dans le cas de données manquantes, ces variables latentes ne sont pas observées lors de la phase d'estimation des paramètres de classification. Pour résoudre ce problème, l'algorithme *EM* a été proposé afin de pallier le problème de données partiellement étiquetées. L'algorithme *EM* consiste d'abord à reconstituer les données manquantes. Il construit une distribution de probabilité sur les valeurs pouvant être prises par les variables latentes et se sert de celles-ci pour mettre à jour les paramètres utilisés dans les expressions de la densité marginale. L'algorithme *EM* est un processus itératif constitué de deux étapes :

- étape E : lors de cette étape, l'information actuellement disponible, les données observées mais aussi l'estimé courant des paramètres, est utilisée pour construire une distribution de probabilité sur les valeurs pouvant être prises par les variables latentes pour chacun des exemples $(P(X, Y))$.
- étape M : lors de cette étape, les paramètres du modèle sont estimés en utilisant les distributions de probabilité définies à l'étape précédente. Cette étape peut généralement être traitée en ayant recours aux méthodes d'estimation utilisées lorsque toutes les données sont observées,

comme le maximum de vraisemblance.

3.3.2.2 Les méthodes transductives

La méthode transductive est un cas particulier de la classification prédictive. Un exemple en est le *SVM* transductif [Joachims, 1999b]. Les machines à vecteurs supports tranductives utilisent la notion de marge appliquée aux exemples non étiquetés pour déplacer la frontière de décision vers des zones où peu d'exemples de l'ensemble d'apprentissage peuvent être trouvés. Au contraire du *SVM* supervisé qui consiste à maximiser la marge uniquement à partir de données étiquetées, le *SVM* transductif propose de maximiser la marge sur l'ensemble des données étiquetées et non étiquetées, en choisissant l'étiquetage des données non étiquetées le plus favorable. Ce qui revient à minimiser l'équation suivante :

$$\frac{1}{2}||w||^2 \tag{3.3}$$

sujet à :

$$\begin{cases} \forall i \in \{1, ..., n_l\} : y_i[w^{'} x_i + b] \geq 1, \\ \forall j \in \{n_l + 1, ..., n\} : y_j^*[w^{'} x_j + b] \geq 1, \theta \end{cases} \tag{3.4}$$

où n_l est le nombre de données étiquetées parmi n exemples.

D'un point de vue théorique, le but du *SVM* transductif est de minimiser une borne sur l'erreur commise sur les données non étiquetées [Vapnik, 1995]. Cependant, cette technique n'est pas suffisante et ne permet pas de résoudre correctement le problème surtout quand l'ensemble des données non étiquetées dépassent la centaine. Par conséquent, des approches heuristiques permettant de faire face à ce problème sont apparues. Joachims [1999a] a proposé la technique SVM^{light} qui nécessitent en pratique de fixer la proportion d'exemples étiquetés positivement et négativement afin d'éviter l'obtention de solutions dégénérées. Les *SVM* transductifs ont été appliqué à plusieurs domaines et ils ont aboutis à de bonnes performances notamment pour la classification de données textuelles [Joachims, 1999b].

3.3.2.3 Self-training

La technique d'apprentissage *self-training* est largement utilisée et étudiée par la communauté de l'apprentissage semi-supervisé du fait de sa facilité d'implémentation mais aussi de ses bonnes performances. Cette méthode comme la plupart des algorithmes d'apprentissage semi-supervisé consiste à entraîner un classifieur avec des données étiquetées et des données non étiquetées. Elle utilise ses propres prédictions sur des exemples non étiquetées

TAB. 3.1 – Algorithme Self-training

Entrée :
Ensemble L des données étiquetées
Ensemble U des données non étiquetées
Un nombre n des données à ajouter à l'ensemble L à chaque itération
Tant que $U \neq \emptyset$
Apprendre un classifieur h sur l'ensemble L
Etiqueter l'ensemble U en utilisant le classifieur h
T est l'ensemble des n exemples de U prédits avec la meilleure confiance par h
Ajouter T à L
Supprimer T de U
Fin

afin d'élargir son ensemble d'apprentissage. L'algorithme de *self-training* a été proposée par Zhu *et al.* [2003]. Le principe de cette méthode est présenté dans le tableau 3.1 : h est un algorithme de classification traditionnel qu'on entraîne d'abord avec les données d'apprentissage étiquetées L, puis on l'utilise pour classifier les données non étiquetées. A chaque itération nous sélectionnons un certain nombre d'exemples n de la base U, les n exemples sélectionnés sont les exemples prédits avec la meilleure confiance (prédits avec des probabilités supérieures à un seuil fixé). Ces exemples sont considérés maintenant comme des données d'apprentissage étiquetées. Ce processus est répété jusqu'à que l'ensemble des données non étiquetées devient vide.

L'algorithme *self-training* a été appliqué à de nombreux problèmes de classification, par exemple dans la désambiguïsons de sens [Yarowsky, 1995], l'analyse syntaxique et la segmentation automatique de la parole [Guz *et al.*, 2010].

3.3.2.4 Co-apprentissage automatique (*Co-training*)

La méthode de co-apprentissage automatique suppose que chaque exemple x_t peut être divisé en au moins deux parties x_t^1 et x_t^2 contenant chacune suffisamment d'information pour bien prédire l'étiquette y_t. Cette méthode consiste à entraîner plusieurs classifieurs, chacun basé sur un descripteur donné, puis à les améliorer entre eux à l'aide d'une masse importante de données non étiquetées. En entraînant deux modèles différents sur ces deux représentations, on peut alors utiliser leurs prédictions sur des entrées non-étiquetées afin qu'ils élargissent mutuellement leurs ensembles d'entraînement. La régularisation implicite de cette approche force donc les sorties de chaque modèle à être similaires sur les entrées non étiquetées. Cette régularisation est aussi explicitement utilisée par des approches qui sont dites du *multi-view learning* que nous détaillerons par la suite.

TAB. 3.2 – Algorithme de co-apprentissage

Entrée
Ensemble L des données étiquetées
Ensemble U des données non étiquetées
Tant que $U \neq \emptyset$
Apprendre un classifieur h_1 sur l'ensemble L
Apprendre un classifieur h_2 sur l'ensemble L
Etiqueter aléatoirement un nombre p des exemples de l'ensemble U en utilisant le classifieur h_1
Etiqueter aléatoirement un nombre p des exemples de l'ensemble U en utilisant le classifieur h_2
Ajouter l'ensemble T des exemples labelisées par h_1 et h_2 à l'ensemble L
Supprimer T de U
Fin

La version standard de cette approche a été proposée par Blum et Mitchell [1998] (cf. le tableau 3.2), elle peut aussi être vue comme un algorithme de *self-training* croisé à deux classifieurs.

Blum et Mitchell [1998] ont appliqué l'algorithme standard de co-apprentissage automatique à la classification automatique de pages web. D'abord, ils ont supposé qu'une page web x_t est caractérisée par deux descripteurs différents et indépendants x_t^1 et x_t^2, le premier descripteur x_t^1 correspond au nombre de mots dans la page x_t et x_t^2 correspond au nombre de mots soulignés qui représente le nombre de liens hypertextes vers la page x_t depuis les autres pages web de la base de données. Ils ont supposé que les deux descripteurs, x_t^1 et x_t^2, sont suffisants pour caractériser une page web. Ensuite, ils ont entrainé un algorithme de Bayes naïf (naive Bayes algorithm) en parallèle sur chacun des deux descripteurs x_t^1 et x_t^2. Expérimentalement, ils ont montré que le taux d'erreur est diminué de moitié en utilisant l'algorithme de co-apprentissage avec *multi-vue*, il est ainsi passé de 11% à 5%.

L'algorithme de co-apprentissage a été appliqué avec réussite à divers problèmes de reconnaissance de formes comme dans le traitement automatique de langue [David et Cardie, 2001] [Sarkar, 2001] et la reconnaissance automatique de la parole [Virginia et Ballard, 1998].

3.4 Apprentissage automatique et caractérisation multiple

L'apprentissage automatique a été étudié à partir de deux aspects différents : caractérisation unique et caractérisation multiple. Les algorithmes classiques de classification, notamment la classification semi-supervisée, se basent sur une caractérisation unique ; un seul descripteur par objet. Dans des cas critiques, un seul descripteur n'est pas fiable pour caractériser un objet. Par conséquent, il est indispensable d'utiliser plus qu'un descripteur afin de béné-

ficier des capacités de discrimination de chacun d'entre eux.

3.4.1 Caractérisation unique

Traditionnellement, un algorithme de classification ne considère qu'une seule forme de caractérisation de l'objet à classifier. Malgré la contrainte de caractérisation unique de l'objet, les systèmes n'utilisant qu'un seul descripteur peuvent aboutir à de bonnes performances de classification. L'algorithme d'apprentissage semi-supervisé de type *self-training* est une forme particulière de cette méthode. Elle ne considère qu'un seul descripteur pour décrire un objet. Par contre, pour avoir de bonnes performances, le descripteur utilisé doit être adapté au problème de classification et il doit avoir un grand pouvoir discriminant.

L'utilisation d'un seul descripteur n'empêche pas d'avoir plusieurs prédictions. En utilisant divers classifieurs, nous obtenons différentes prédictions possibles. Dans ce contexte, Zhou et Goldman [2004] ont développé un système robuste de classification dit démocratique qui se base sur un système de vote entre les différents classifieurs en utilisant un seul descripteur. Cette méthode est très efficace pour un problème à caractérisation unique.

Dans le domaine de la reconnaissance des formes, s'il existe plusieurs manières de caractérisation adéquate pour résoudre le problème de classification, il est fortement conseillé d'utiliser parallèlement toutes ces caractéristiques afin de mieux prédire les catégories [Muslea *et al.*, 2000] (caractérisation multiple).

3.4.2 Caractérisation multiple

Une caractéristique utilisée toute seule peut ne pas être suffisante pour bien caractériser l'objet. En général, un objet est caractérisé par différentes caractéristiques, chaque caractéristique permettant de le décrire par rapport à un point de vue donné (cf. figure 3.2). Intuitivement, dans un système à caractérisation multiple, un exemple x_t peut être décrit de différentes manières. Par exemple, pour un signal de parole nous disposons de plusieurs descripteurs : *pitch*, *énergie*, *MFCC*, etc. Dans le cadre de la classification de signaux émotionnels, plusieurs études ont montré que l'utilisation parallèle des caractéristiques long-termes et des caractéristiques court-termes permet d'améliorer la classification [Kim *et al.*, 2007], d'où l'intérêt d'utiliser différentes caractérisations. De plus, l'avantage de cette représentation est que les différents descripteurs peuvent être complémentaires entre eux. Afin de bénéficier de la complémentarité des caractéristiques, la meilleure solution consiste à les combiner ou à les utiliser en parallèle.

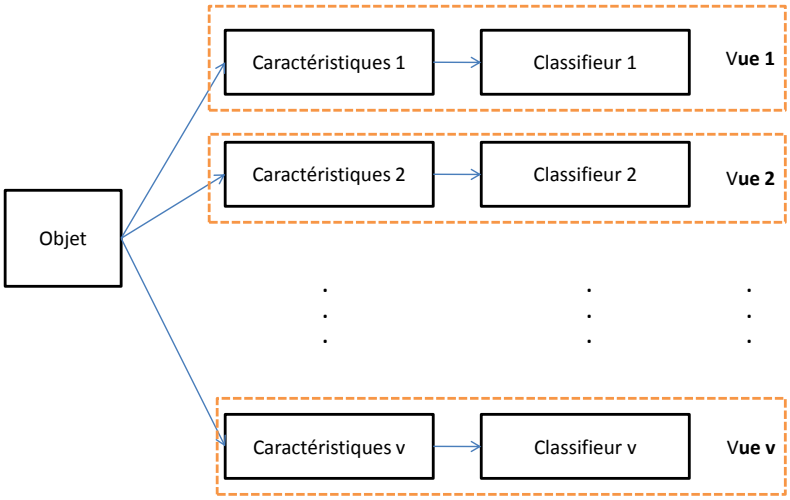


FIG. 3.2 – Caractérisation multiple : approche *multivue*

Les algorithmes de classification basés sur une caractérisation multiple (*multi-vue*) consistent à utiliser parallèlement différents descripteurs afin de bénéficier des prédictions de chacun et donner une prédiction unique par objet, qui représente la combinaison des différentes prédictions. L'algorithme de co-apprentissage automatique est une forme particulière de cette méthode. Il suppose qu'il y ait au moins deux caractérisations indépendantes possibles pour décrire un objet (cf. figure 3.2).

Les méthodes *multi-vue* ont donné de bonnes performances dans de nombreuses applications de classification, comme dans la reconnaissance de la parole [Sa et Ballard, 1998] et la classification de pages web [Blum et Mitchell, 1998] [Zhang et Sun, 2010]. Récemment, Zhang et Sun [2010] ont développé une technique d'apprentissage basée sur une caractérisation multiple en utilisant différentes techniques de réseaux de neurones pour chaque descripteur. La méthode proposée consiste à classifier des pages web en deux catégories : pages de cours et pages de non cours. Pour caractériser l'ensemble des pages web, l'algorithme considère deux descripteurs : le nombre de mots (*View1*) et le nombre de liens hypertextes vers la page (*View2*). Les auteurs ont montré que l'approche *multi-vue* donne de meilleurs résultats que les méthodes utilisant une seule forme de caractérisation (cf. figure 3.3).

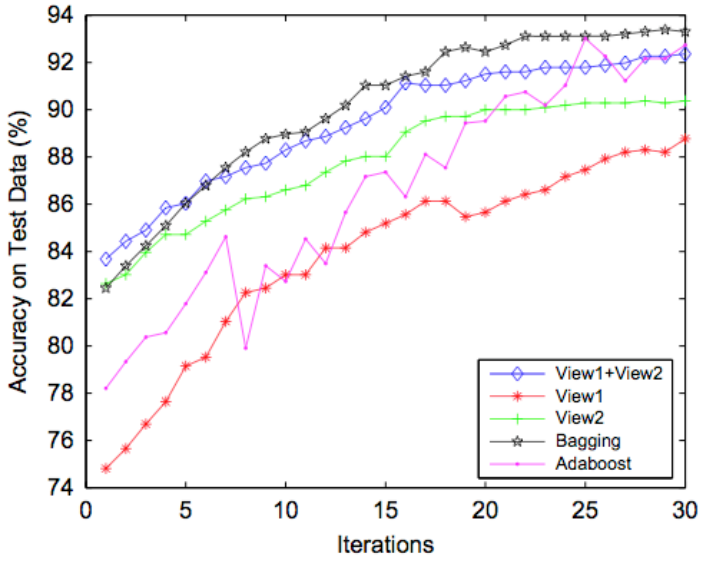


FIG. 3.3 – Comparaison des performances de l'approche *multi-vue* *View1*+*View2*) et l'approche à caractérisation unique *View1* et *View2* [Zhang et Sun, 2010]

3.5 Algorithme de co-apprentissage avec caractérisation multiple

L'algorithme de co-apprentissage automatique (*co-training*) permet de bénéficier des avantages de la méthode d'apprentissage semi-supervisé, ainsi des avantages de l'approche *multi-vue*. Il consiste à utiliser plusieurs caractéristiques et intégrer toutes les informations disponibles auprès des différents points de vue (descripteur + classifieur) pour prédire plus précisément les catégories des objets à classifier. Par contre, cette méthode ne permet pas la fusion des caractéristiques. Par conséquent, afin de renforcer la confiance de classification, nous ajoutons un module de fusion statistique (probabilité à posteriori) qui permet de prendre en compte les différents points de vue et d'avoir une prédiction unique par objet.

3.5.1 Méthode statistique d'apprentissage semi-supervisé

L'algorithme d'apprentissage *multi-vue* est une procédure itérative, comme le montre la figure 3.4. Etant donné un ensemble de données étiquetées L et un ensemble de données non étiquetées U, ainsi qu'un ensemble de descripteurs,

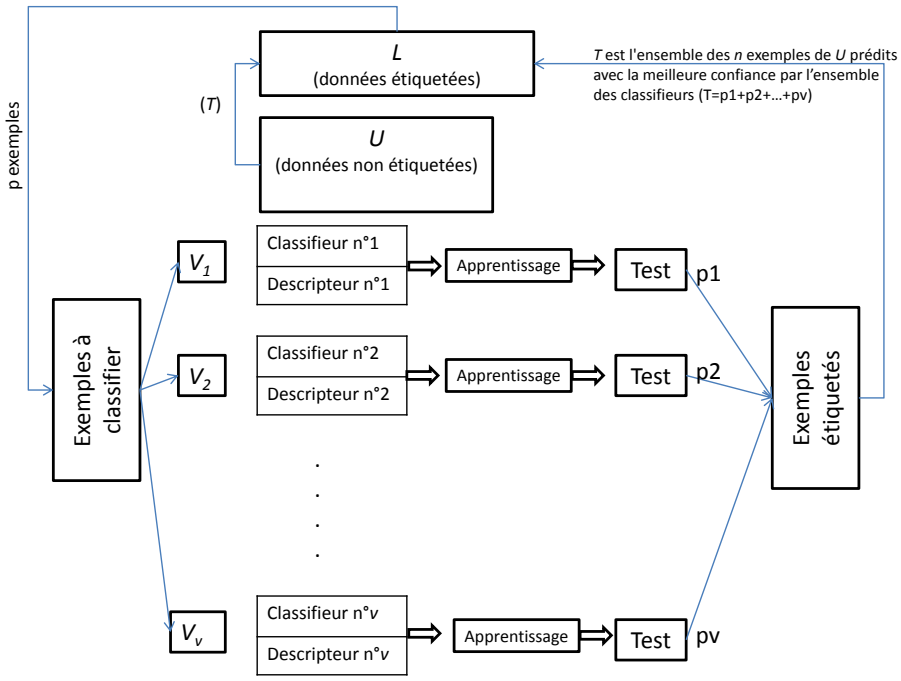


FIG. 3.4 – Architecture du système de classification semi-supervisée de type co-apprentissage

l'algorithme consiste utiliser différents vues. Ensuite il prend T exemples qui sont prédits avec la meilleure confiance à partir de la base de test U et les met dans la base d'apprentissage L. Chaque classifieur permet d'étiqueter un ensemble d'exemple P_i.

$$T = \sum_{i=1}^{v} P_i \tag{3.5}$$

Blum et Mitchell [1998] ont proposé le premier algorithme de ce genre. L'algorithme de Blum et Mitchell [1998] suppose qu'il y ait au moins deux descripteurs indépendants possibles pour caractériser un objet, ce qui donne deux points de vues différents. Supposons que l'on dispose de v descripteurs différents, nous avons en conséquence v points de vue différents $\{V_1, V_2, ..., V_v\}$. Dans ce cas, un objet x_t est défini par $v+1$ *uplets* : $\langle x_t^1, x_t^2, ..., x_t^v, y_t\rangle$, où y_t est l'étiquette de l'objet x_t et $\{x_t^1, x_t^2, ..., x_t^v\}$ sont les différents points de vue d'un problème à v points de vue différents. Cependant, un exemple de test est défini par $\langle x_t^1, x_t^2, ..., x_t^v, ?\rangle$, dans ce cas l'étiquette y_t est inconnue.

L'inconvénient de cette méthode est qu'elle prend séparément les différentes prédictions obtenues par chaque vue. La méthode standard de co-

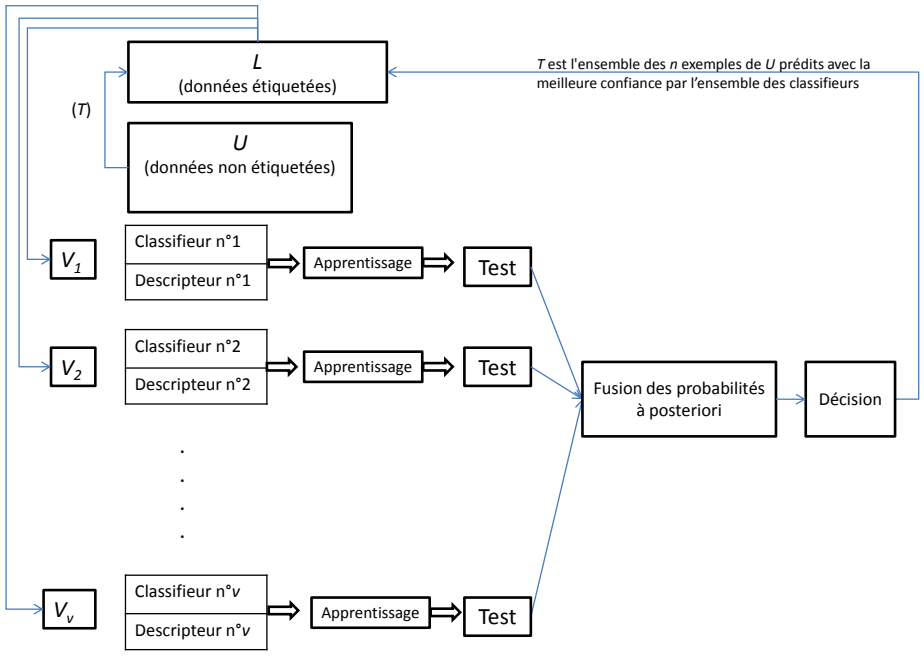


FIG. 3.5 – Architecture du système de co-apprentissage avec fusion de données

apprentissage ne permet pas la fusion de caractéristiques, les différents classifieurs sont indépendants. Par conséquent, afin de bénéficier de la complémentarité de différents descripteurs et classifieurs, nous proposons un algorithme qui permet de prendre en compte l'ensemble des points de vue et d'obtenir une prédiction unique par objet.

3.5.2 Co-apprentissage automatique pour la classification du *motherese*

Dans le cadre du co-apprentissage automatique de type *multi-vue*, nous avons proposé une nouvelle méthode de co-apprentissage (cf. figure 3.5). Il s'agit d'un algorithme de classification qui consiste à combiner les prédictions issues de différents classifieurs (les probabilités à posteriori) afin d'obtenir une prédiction unique pour chaque exemple de test. La méthode proposée est une nouvelle forme de co-apprentissage automatique, elle est plus appropriée aux problèmes impliquant à la fois la classification semi-supervisée et la fusion de données. Cet algorithme est conçu pour améliorer les performances de classification grâce à la combinaison de données non étiquetées.

3.5.2.1 Algorithme

TAB. 3.3 – Algorithme de co-apprentissage automatique pour la classification du *motherese*

Entrée

Ensemble L de m exemples étiquetés

L=$\{(l_1^1, ..., l_1^v, y_1), ..., (l_m^1, ..., l_m^v, y_m)\}$ avec $y_i = \{1, 2\}$ (problème bi-classe)

Ensemble U de n exemples non étiquetés

U=$\{(x_1^1, ..., x_1^v), ..., (x_n^1, ..., x_n^v)\}$

v = nombre de *vue*

Initialisation

ω_k (poids des classifieurs)= $1/v$ pour tous les classifieurs

Tant que $U \neq \emptyset$

A. Classifier tous les exemples de la base de test U

Pour $k = 1, 2, ..., v$

1. Apprendre le classifieur h_k sur l'ensemble L
2. Etiqueter les exemples de l'ensemble U en utilisant le classifieur h_k
3. Estimer la probabilité de classification de chaque exemple x_i de U, $p(C_j|x_i) = \sum_{k=1}^{v} \omega_k \times h_k(C_j|x_i^k)$
4. $Labels(x_i) = decision(p(C_j|x_i))$

Fin pour

B. Mettre à jour la base d'apprentissage L **et la base de test** U

$U_j = \{z_1, ..., z_{nj}\}$ sont les ensembles d'exemples classifiés dans C_j

Pour $i = 1, 2, ..., nj$

$prob(z_i, j) = \frac{\sum_{k=1}^{v} \omega_k \times h_k(C_j|z_i^k)}{\sum_{k=1}^{v} \omega_k}$

Fin pour

$marge_j = moyenne(prob(z_i, j))$

Prendre T_j exemples de U_j qui sont prédis dans C_j avec une probabilité supérieur à $marge_j$

$T = \sum_j T_j$

Ajouter l'ensemble T à L et le supprimer de U

C. Mettre à jour les poids

$\omega_k = \frac{\sum_{i=1}^{size(T)} h_k(z_i^k)}{\sum_{k=1}^{v} \sum_{i=1}^{size(T)} h_k(z_i^k)}$

Fin Tant que

L'algorithme de co-apprentissage proposé est présenté dans le tableau 3.3. Il prend en entré un ensemble L=$\{(l_1^1, .., l_1^v, y_1), ..., (l_m^1, ..., l_m^v, y_m)\}$ de données étiquetées et un ensemble U=$\{(x_1^1, .., x_1^v), ..., (x_n^1, ..., x_n^v)\}$ de données non étiquetées. Chaque exemple de la base de données étiquetées est représenté de la façon suivante $(l_i^1, l_i^2, ..., l_i^v, y_i)$: par v descripteurs et y_i représente la catégorie de l'exemple. Dans le cas d'un problème bi-classe $y_i \in \{-1, +1\}$. Les

exemples de la base de données non étiquetées sont représentés de la manière suivante $(x_i^1, x_i^2, ..., x_i^v)$. Les labels associés aux exemples non étiquetés ne sont pas disponibles. L'algorithme considère un ensemble de classifieurs et de descripteurs qu'on appelle des *multi-vue*. Les différentes *vues* servent chacune à modéliser les segments de parole, elles doivent être suffisantes pour la tâche de discrimination.

La méthode proposée a une forte similitude avec celle de la fusion de données qui consiste en l'intégration de différents descripteurs et classifieurs dans un seul algorithme. D'abord pour initialiser l'algorithme, les poids des classifieurs sont initialisés à $1/v$, où v représente le nombre de *vues*, ainsi chaque ensemble de classifieur et descripteur contribue avec le même poids dans la classification. Ensuite l'algorithme procède itérativement en la classification des exemples non étiquetés de l'ensemble U. Intuitivement, l'algorithme se décompose en trois processus : une première étape consiste à l'estimation des probabilités à posteriori des exemples non étiquetés, une deuxième étape consiste à mettre à jour les bases de données et une dernière étape consiste mettre à jour les poids des classifieurs.

Estimation des probabilités à posteriori : L'algorithme de co-apprentissage proposé est une forme de fusion statistique bénéficiant de différents classifieurs statistiques h_k que nous avons adapté afin de produire une probabilité à posteriori pour chaque exemple non étiqueté (voir le chapitre 2). Les classifieurs utilisés estiment des probabilités à posteriori en sortie ; pour chaque exemple x_i, chaque classifieur h_k estime une probabilité à posteriori $h_k(C_j|x_i^k)$. Par conséquent, nous obtenons une probabilité à posteriori par classifieur, d'où plusieurs prédictions possible, autant le nombre des *vues* v. Afin d'avoir une prédiction unique par exemple de test x_i, nous combinons les différents résultats (probabilités) de la manière suivante :

$$p(C_j|x_i) = \sum_{k=1}^{v} \omega_k \times h_k(C_j|x_i^k) \tag{3.6}$$

où ω_k représente les poids des classifieurs et h_k est le prédicteur associé à la *vue* k. Ensuite, pour estimer la classe C_j de l'exemple x_i, nous appliquons une méthode de décision [Duda *et al.*, 2000], dans notre exemple nous prenons en compte le maximum à posteriori.

Mettre à jour la base d'apprentissage L et la base de test U : Cette étape consiste à prendre en compte les résultats de l'estimation des probabilités à posteriori pour classifier les exemples prédits avec des probabilités supérieures à un seuil dynamique à définir à chaque itération. D'abord, nous

considérons que U_j est l'ensemble des données classifiées dans la catégorie C_j, dans notre problème bi-classe U_1 représente l'ensemble des données *motherese* et U_2 représente l'ensemble des données *non motherese*. Ensuite, nous calculons une valeur de confiance de classification que nous appelons aussi la marge de classification :

$$prob(z_i, j) = \frac{\sum_{k=1}^{v} \omega_k \times h_k(C_j|z_i^k)}{\sum_{k=1}^{v} \omega_k} \tag{3.7}$$

$$marge_j = moyenne(prob(z_i, j)) \tag{3.8}$$

où $marge_j \in \{0, 1\}$. Cette valeur de marge est inspirée des techniques de classification à vaste marge (SVM) qui consistent à optimiser une marge de classification qui permet de définir un seuil discriminatif. Elle peut être interprétée comme une mesure de confiance de classification [Schapire et Singer, 1999]. Ensuite, nous définissons un ensemble T contenant les ensembles T_j correspondant aux exemples classifiés dans les classes associées C_j avec des probabilités supérieures à la valeur de la marge $marge_j$. Dans notre exemple, nous avons deux ensembles T_1 et T_2 et deux valeurs de marge $marge_1$ et $marge_2$ correspondant successivement à la classe *motherese* et la classe *non motherese*. Par conséquent, nous obtenant un ensemble $T = \{T_1, T_2\}$. Enfin, nous ajoutons l'ensemble T à la base de données étiquetées L et on le supprime de la base de données non étiquetées U.

Mettre à jour les poids des classifieurs : Les poids sont mis à jour à la fin de chaque itération, un poids w_k est estimé pour chaque classifieur h_k. La nouvelle valeur du poids w_k est proportionnelle à la contribution du classifieur correspondant h_k à la classification de l'ensemble T. Ensuite, afin de garder une estimation probabiliste, les poids sont normalisés pour que leur somme soit égale à 1 ($\sum_v w_k = 1$). La formule de mise à jour les différents poids est la suivante :

$$\omega_k = \frac{\sum_{i=1}^{size(T)} h_k(z_i^k)}{\sum_{k=1}^{v} \sum_{i=1}^{size(T)} h_k(z_i^k)} \tag{3.9}$$

Pour valider les différentes *vues*, supposons que nous disposons d'un ensemble d de descripteurs et c de classifieurs, le travail d'optimisation des *vues* consiste à prendre en compte les d descripteurs, puis à choisir parmi les c classifieurs, ceux qui sont les plus adaptés à chaque descripteur. Dans notre étude, les *vues* sont interprétés à partir de la phase d'apprentissage supervisé, en prenant en compte différentes caractéristiques conventionnelles et non conventionnelles associées avec des différents classifieurs. Dans notre travail, nous utilisons quatre techniques de classification différentes.

3.5.2.2 Descripteurs conventionnels et non conventionnels pour la construction des différents points de vue

Dans la littérature, nous distinguons différentes caractérisations de signaux émotionnels [Shami et Verhelst, 2007] [Clavel *et al.*, 2008] [Schuller *et al.*, 2009] [Schuller *et al.*, 2007], nous en avons présenté et utilisé quelques une dans le chapitre 2. Par contre, la caractérisation de paroles affectives et spontanées reste un défi pour la communauté du traitement de la parole, particulièrement la parole émotionnelle spontanée.

Schuller *et al.* [2009] ont récemment organisé un *workshop-challenge* sous forme d'une compétition internationale dédiée à l'évaluation et la comparaison des systèmes de reconnaissance d'émotions, y compris la reconnaissance du *motherese*. Cette compétition traite également des différentes approches d'extraction de caractéristiques pour le développement de systèmes robustes de reconnaissance d'émotions. Cependant, il n'existe pas de caractéristiques génériques idéales pour la détection des émotions, ce qui rend cette tâche dépendante et sensible à l'émotion recherchée.

Théoriquement, le *motherese* est caractérisé par sa prosodie mais aussi par d'autres caractéristiques acoustiques et spectrales. Le signal du *motherese* est perceptuellement différent de la parole adressée aux adultes. Pour caractériser spectralement le signal, nous avons adapté un modèle basé sur le système de la perception humaine. Cette approche est une analyse en bancs de filtres avec une distribution en fréquences imitant la répartition et la forme des filtres de la cochlée. Cette répartition est non-linéaire et plusieurs implémentations sont possibles. Dans notre étude, nous utilisons l'échelle non linéaire de *Bark* [Zwicker, 1961] [Zwicker et Terhardt, 1980]. L'échelle de *Bark* [Zwicker, 1961] correspond à une répartition dans le spectre des fréquences d'un ensemble de bandes de fréquences critiques. Ces bandes sont des regroupements des excitations sonores ayant des fréquences voisines et perceptivement proches au sein de certaines bandes fréquentielles. Cette échelle en bande critique est une échelle perceptive (correspondant à la tonie), dont l'unité est le *Bark*, regroupe un ensemble variable de fréquences.

Les bandes critiques montrent que notre oreille fonctionne de facon sélective en fonction des fréquences. La largeur d'une bande critique, quelle que soit sa fréquence centrale, est appelée un *Bark*. Jusqu'à 1 kHz, cette largeur de bande est linéaire et est égale à 100 Hz. Au-delà de 1 kHz, elle varie dans une proportion logarithmique avec la fréquence. Dans l'échelle des *Barks*, un accroissement de 1 *Bark* correspond à une augmentation en fréquence de 1 bande critique. Par conséquent, la relation *Bark*/*Hertz* est quasi-linéaire jusqu'à 500 Hz ; au-delà, elle est quasi-logarithmique.

Comme le décrit la figure 3.6, le filtre *Bark* attribue des larges bandes

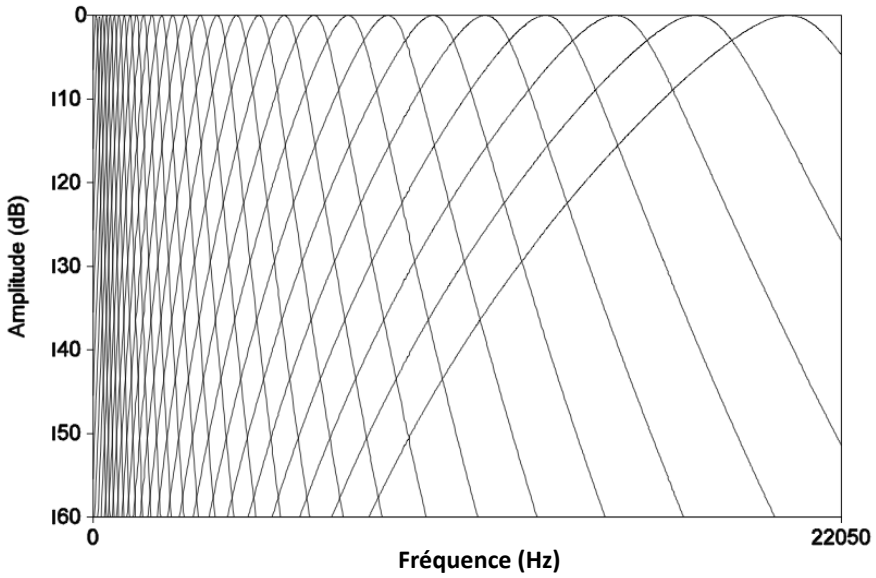
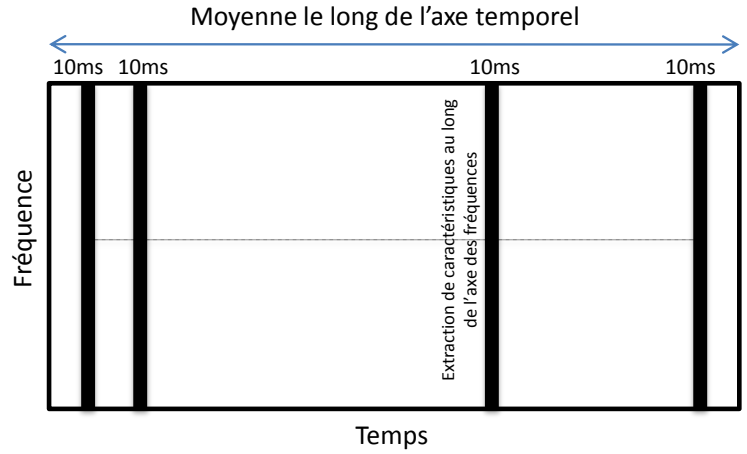


FIG. 3.6 – Filtre *Bark* utilisé pour l'extraction des caractéristiques

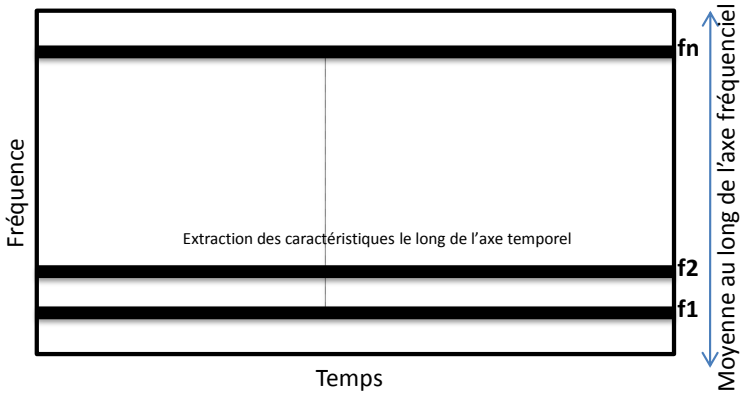
(en Hz) pour les hautes fréquences et il maintient une bonne résolution pour les basses fréquences. Ensuite, nous appliquons F statistiques sur la répresentation spectrale obtenue. Les F statistiques peuvent être extraites sur l'axe temporel ou l'axe de fréquenciel. Nous avons utilisé trois stratégies différentes pour l'extraction des caractéristiques.

- **Approche *TL*** (axe temporel) figure 3.7.a. : extraire les F statistiques à partir du vecteur spectral (pour chaque fenêtre du temps (trame)), puis calculer la moyenne au long de l'axe temporel pour obtenir les premiers F caractéristiques.
- **Approche *SL*** (axe fréquenciel) figure 3.7.b. : extraire les F statistiques au long de l'axe du temps (pour chaque bande de fréquence), puis calculer la moyenne au long de l'axe fréquenciel pour obtenir les F caractéristiques.
- **Approche *MV*** (valeurs moyennes) est la valeur moyenne des deux approches précedentes.

En résumé, afin de caractériser le signal de parole, nous utilisons des caractéristiques conventionnelles et non conventionnelles. Le comportement de chaque descripteur est en effet dépendant de la fenêtre temporelle sur laquelle il est considéré. Afin de modéliser d'une manière globale chaque caractéristique, nous utilisons des statistiques ou ce qu'on appelle des fonctionnels.



(a)



(b)

FIG. 3.7 – Extraction des caractéristiques *Bark*

Cela consiste à appliquer la fonctionnelle F sur les descripteurs bas niveaux (*pitch*, *énergie*, *Bark*) (cf. tableau 3.4) :

$$F : X \rightarrow x \in R \tag{3.10}$$

Dans notre travail, nous utilisons 32 statistiques (fonctionnelles) pour modéliser l'évolution du spectre le long des axes temporels et fréquenciels.

TAB. 3.4 – Les descripteurs bas niveaux et les fonctionnels appliqués

Descripteurs bas niveaux	**Fonctionnelles (32)**
Bark *pitch* *energie*	maximum, minimum, moyenne, écart-type, variance, skewness, kurtosis, interquartile, étendue, écart type moyen (mean absolute deviation :MAD), MAD basée sur les médianes, i.e. median(abs(x-median(x))), premier et deuxième coefficients de la régression linéaire , premier, deuxième et troisième coefficients de la régression quadratique, 9 quantiles correspondant aux valeurs de probabilités suivantes : 0.025, 0.125, 0.25, 0.375, 0.50, 0.625, 0.75, 0.875, 0.975, 2 quantiles correspondant aux 0.1 et 0.9, la valeur de l'étendue de l'interquantile entre ce deux valeurs, la valeur absolue et le signe de l'intervalle entre l'apparence maximale et minimale.

TAB. 3.5 – Différents descripteurs

$f1$	16 *MFCCs*
$f2$	*pitch* (la moyenne, la variance et l'étendue) + *énergie* (la moyenne, la variance et l'étendue)
$f3$	*pitch* avec 35 statistiques
$f4$	*énergie* avec 35 statistiques
$f5$	*pitch* avec 35 statistiques + *énergie* avec 35 statistiques
$f6$	*Bark* en utilisant les approches *TL* + *SL* + *MV* (96 statistiques)
$f7$	*Bark* en utilisant l'approche *TL* (32 statistiques)
$f8$	*Bark* en utilisant l'approche *SL* (32 statistiques)
$f9$	*Bark* en utilisant l'approche *MV* (32 statistiques)

3.5.3 Expérimentations

3.5.3.1 Différents points de vue pour la classification du *motherese*

Grâce aux différents classifieurs et descripteurs, nous obtenons plusieurs combinaisons (descripteur+classifieur) possibles pour la classification. Cependant ces combinaisons ne sont pas tous efficaces pour la classification du *motherese*. Par conséquent, nous avons pris en compte tous les descripteurs, puis nous avons recherché le classifieur correspondant le plus adapté à chaque descripteur. A l'issue de la phase d'extraction de caractéristiques nous avons obtenus 9 descripteurs qui sont présentés dans le tableau 3.5. Les 9 descripteurs correspondent aux caractéristiques cepstrales 16 *MFCCs* $f1$, des caractéristiques prosodiques $f2$ (*pitch*+*énergie*) avec des statistiques classiques comme présenté dans le chapitre 2. Aussi nous appliquons d'autres statistiques sur les caractéristiques prosodiques pour obtenir des vecteurs de plus grandes dimensions, $f3$, $f4$ et $f5$. Enfin les descripteurs $f6$, $f7$, $f8$ et $f9$ représentent les caractéristiques perceptives.

Nous avons testé les différentes techniques sur la même base de données présentée dans le chapitre 2, contenant 500 segments de parole distribués équitablement entre *motherese* et *non motherese*. Les différents résultats sont représentés dans le tableau 3.6. Le tableau 3.6 montre que le système

TAB. 3.6 – Performance (accuracy) des différents classifieurs supervisés (en %)

	Caractéristiques	*GMM*	*kppv*	*SVM*	*MLP*
Descripteur cepstral	$f1$	**72.75**	70.25	59.75	61. 5
Descripteurs prosodiques	$f2$	**69.5**	65.5	54.75	50.25
	$f3$	54.75	**55**	50	50
	$f4$	67	**68.5**	65.5	58.5
	$f5$	62.25	65.5	**65.5**	54.5
Descripteurs perceptives	$f6$	**61**	50.5	49	54.5
	$f7$	55.5	51	52	**58.5**
	$f8$	**65**	52	50.5	55.5
	$f9$	58.75	50.5	50.5	**64**

TAB. 3.7 – Différents *vues*

	Descripteur	Classifieur
V_1	$f1$	*GMM*
V_2	$f2$	*GMM*
V_3	$f3$	*kppv*
V_4	$f4$	*kppv*
V_5	$f5$	*SVM*
V_6	$f6$	*GMM*
V_7	$f7$	*MLP*
V_8	$f8$	*GMM*
V_9	$f9$	*MLP*

GMM+MFCC donne les meilleures performances avec un taux de reconnaissance de 72.75 %, ce qui confirme l'efficacité du descripteur cepstral pour la caractérisation de signaux émotionnels, particulièrement le *motherese*. Pour chaque descripteur nous prenons en compte le classifieur optimal, nous obtenons les 9 combinaisons présentées dans le tableau 3.7.

3.5.3.2 Résultats

Dans la phase de construction des différentes *vues*, nous avons obtenus 9 classifieurs (voir le tableau 3.7). Ensuite, nous utilisons les différentes *vues* ensemble dans l'algorithme de co-apprentissage que nous avons proposé. La figure 3.8 montre l'architecture générale de la méthode statistique de classification semi-supervisée du *motherese*. En entrée, nous avons un ensemble de données étiquetées L et un ensemble de données non étiquetées U. Pour construire l'ensemble L, nous avons sélectionné aléatoirement à partir de la

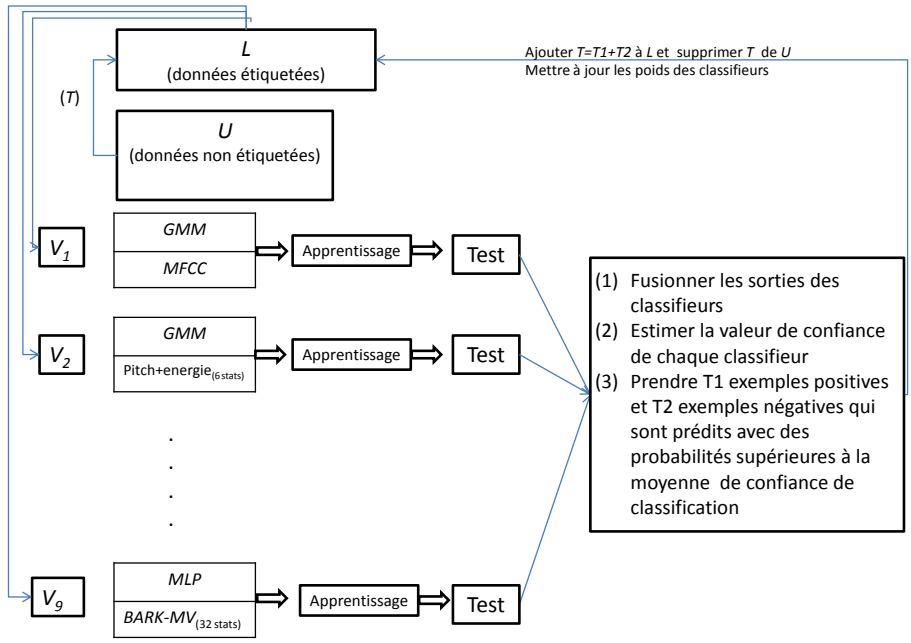


FIG. 3.8 – Méthode statistique pour la classification semi-supervisée du *motherese*

base de données contenant au total 500 segments, 100 segments de parole (50 segments du *motherese* et 50 segments du *non motherese*). Les 400 segments qui restent formeront la base U. Aussi nous définissons les poids des différentes *vues* $\{V_1,V_2,...,V_9\}$, le poids de chacun est initialisé à $1/9$.

Pour estimer la dépendance entre la quantité de données étiquetées pour l'apprentissage et les performances de classification, nous exécutons l'algorithme de co-apprentissage en utilisant différentes quantités de données. La figure 3.9 présente les performances du système en utilisant différentes quantités d'annotations de 10% à 100% de la base d'apprentissage L (de 10 à 100 segments). Nous constatons que les performances de classification évoluent avec la quantité des annotations disponibles. De plus, la figure 3.9 montre que notre système de co-apprentissage donne des résultats satisfaisants même lorsque nous exploitons que très peu de données étiquetées. Quand nous utilisons 10 segments étiquetés (5 segments du *motherese* et 5 segments du *non motherese*) nous obtenons un score de reconnaissance de 66,75%, tandis que les performances sont de l'ordre de 75,75% en utilisant la totalité de la base L pour l'apprentissage (estimés sur 400 segments).

Afin de montrer les avantages de notre approche par rapport aux autres

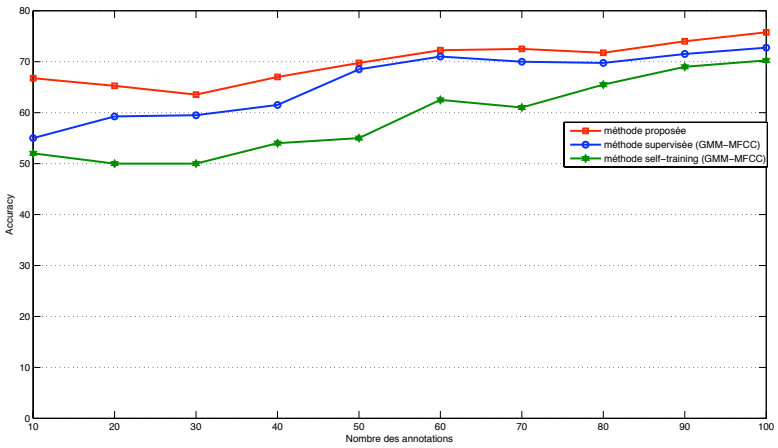


FIG. 3.9 – Performance de classification avec différente quantité de données étiquetées d'apprentissage

méthodes de classification semi-supervisée, nous comparons les résultats obtenus par notre système avec ceux obtenus par la méthode *self-training*. La méthode d'apprentissage semi-supervisé *self-training* en utilisant le meilleur classifieur obtenu, *GMM* et *MFCC*. La figure 3.9 montre que notre méthode donne de meilleurs résultats par rapport à la méthode *self-training*, 75.75% vs 63% en utilisant la totalité de données étiquetées L. Ce résultat est expliqué par la sensibilité de la méthode *self-training* aux erreurs de classification générées lors des premières itérations. De plus, la différence entre les deux méthodes est importante dans le cas où très peu de données d'apprentissage sont disponibles, 66,75% vs 52% lorsque nous utilisons 10 annotations (cf. tableau 3.8).

Pour évaluer les performances de fusion de données dans le cadre du co-apprentissage, nous comparons les performances de la méthode proposée avec l'algorithme de co-apprentissage standard. Le tableau 3.8 montre que la méthode de co-apprentissage proposée donne de meilleurs résultats. L'approche *multi-vue*, ainsi que la fusion de classifieurs permettent de renforcer les règles de catégorisation. Les performances inferieures de la méthode de co-apprentissage standard est du aux contributions des classifieurs qui génèrent des erreurs de classification indépendamment des autres, cela se propagent sur le reste des itérations et affectent le processus d'apprentissage.

Pour montrer l'apport de la méthode de co-apprentissage proposée par rapport à la méthode de classification supervisée, nous comparons aussi les performances de notre système avec le meilleur système de classification su-

TAB. 3.8 – Performance de classification avec différente quantité de données étiquetées d'apprentissage

Nombre des annotations	10	20	30	40	50	60	70	80	90	100
Méthode proposée	66.75	65.25	63.5	67	69.75	72.25	72.5	71.75	74	75.75
Co-training standard	63.5	62.5	62	64.5	68.5	69.75	71.25	69.5	71	71.5
Self-trainig GMM-MFCC	52	50	50	54	55	62.5	61	65	69	70.25
Self-training GMM-prosodie	54	52.5	53	52	53.5	58	59	62	64.5	67.75
Méthode supervisée	55	59.25	59.5	61.5	68.5	71	70	69.75	71.5	72.75

pervisée qui est *GMM+MFCC*. La figure 3.9 montre que notre méthode donne de meilleurs résultats, surtout lorsque nous disposons de peu de données étiquetées, 66,75% vs 55% en utilisant 10 segments étiquetés (cf. tableau 3.8). La différence entre les performances des deux systèmes se réduit lorsque nous augmentons la quantité de données étiquetées. Ceci s'explique par l'aptitude du système de classification supervisée à mieux modéliser les données lorsque le nombre d'exemples étiquetés est plus important.

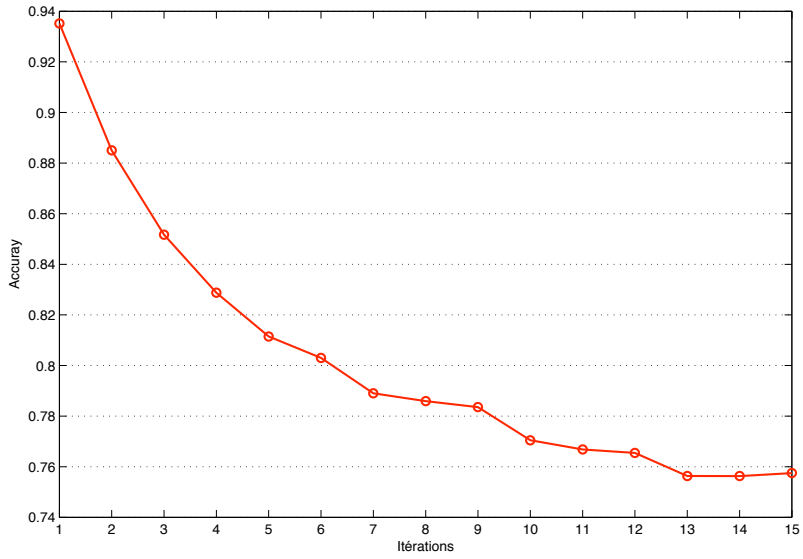


FIG. 3.10 – Performance par itération

La figure 3.10 montre que les performances de classification de l'algorithme de co-apprentissage sont meilleures lors des premières itérations. Dans la première itération, la performance est de 93,52%, tandis qu'à la dernière itération, la performance finale est de 75,75%. Ce résultat est totalement raisonnable, comme le montre la figure 3.11, le système ne fait pas beaucoup des erreurs. Dans la première itération le système a classifié 108 exemples dont 101 sont

correctement classifiés (vrai positif + vrai négatif) et seulement 7 exemples qui sont mal classifiés (faux positifs + faux négatifs) (cf. figure 3.11).

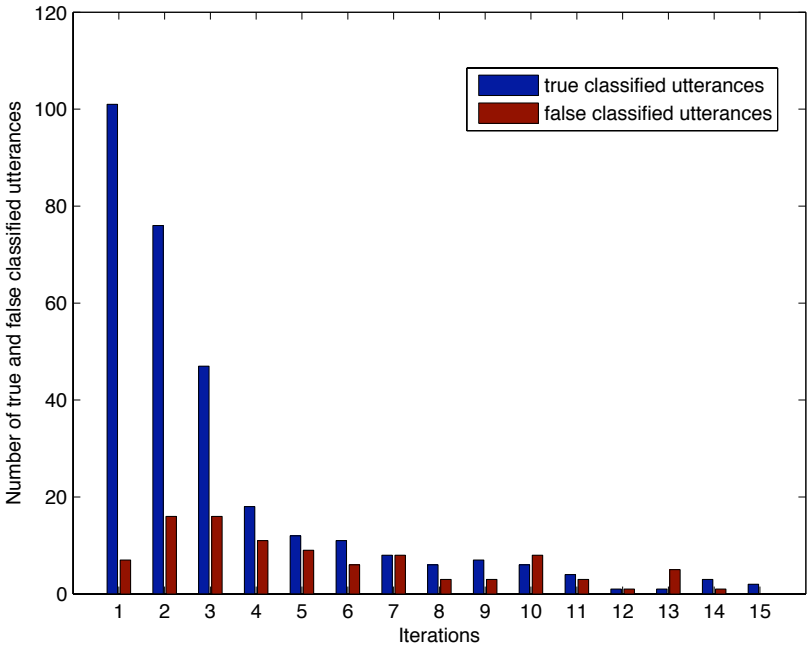


FIG. 3.11 – Nombre des segments classifiés correctement par itération

3.6 Conclusion

Ce chapitre nous a permis de présenter une solution innovante au problème de l'apprentissage lorsque l'ensemble de données étiquetées servant à l'apprentissage de la règle de classification est partielle. Cette solution repose sur l'extension de l'algorithme de co-apprentissage automatique proposé par Blum et Mitchell [1998]. La méthode proposée se situe entre le problème de fusion de données et l'apprentissage semi-supervisé. Dans un premier temps, nous avons introduit les différentes méthodes d'apprentissage semi-supervisé. Ensuite nous avons étudié l'impact de la caractérisation multiple du signal émotionnel par rapport à la caractérisation unique. Puis nous avons montré l'architecture du système de co-apprentissage proposé. La méthode proposée est une variante de l'approche standard de co-apprentissage initialement proposée par Blum et Mitchell [1998]. Elle s'inspire également de l'algorithme *Adaboost* pour le calcul d'erreur de chaque classifieur. L'originalité de notre

méthode réside dans le fait d'utiliser plusieurs classifieurs et descripteurs pour la classification de signaux émotionnels et de pénaliser d'une manière dynamique chaque classifieur par rapport à son erreur de classification (degrés de confiance).

Enfin, des expérimentations sur de données réelles nous ont permis de valider l'apport de cette solution pour traiter des situations ou on dispose de très peu de données d'apprentissage. Notre méthode donne de meilleurs résultats par rapport aux méthodes état de l'art de classification semi-supervisée. De plus, en se comparant avec les systèmes supervisés, l'approche semi-supervisée améliore nettement les résultats, en particulier lorsque le nombre de données étiquetées est faible.

Chapitre 4

Modélisation de l'interaction

Sommaire

4.1 Introduction

Dans le chapitre 1, nous avons exposé les différentes formes de communication ainsi que leurs supports avec une attention particulière pour l'interaction parent-enfant. Dans cette étude, nous nous sommes focalisés sur l'analyse des productions verbales et affectives de la mère (*motherese*). Nous allons dans ce

chapitre étudier l'interaction avec une perspective multi-modale (communication verbale et non-verbale) et développementale. Cette étude se situe dans le cadre du traitement du signal social. Nous évoquerons de ce fait l'analyse automatique de comportements.

Ensuite, nous analyserons l'interaction parent-enfant en comparant trois groupes d'enfant : avec développement typique, à devenir autistique et avec retard mental. Enfin, nous présentons une méthode de regroupement pour la classification des comportements en se basant sur une représentation statistique des données d'interactions. La fin de ce chapitre est dédiée à l'évaluation et l'analyse des résultats obtenus.

4.2 Analyse de comportements et d'interaction humains

4.2.1 État de l'art

L'analyse et la compréhension de l'interaction apportent une nouvelle perception aux environnements intelligents pour la reconnaissance de l'activité humaine. La prise de conscience d'un comportement implique au préalable de différencier les protagonistes et leurs interdépendances. L'analyse des comportements humains regroupe plusieurs disciplines de recherche : psychologie, sociologie, informatique, science du langage, etc. Dans la littérature, on distingue plusieurs travaux se portant sur ce thème de recherche. Parmi les applications les plus connues, nous citons la reconnaissance du rôle des locuteurs dans une conversation [Favre *et al.*, 2008] (voir le paragraphe 1.5.2), la reconnaissance d'activité humaine [Turaga *et al.*, 2008] [Oikonomopoulos *et al.*, 2009], les agents conversationnels [Mancini *et al.*, 2007], etc.

Récemment, des chercheurs en traitement du signal social[1] se sont intéressés à l'identification des personnes dominantes dans une conversation ou dans un débat [Poggi et D'Errico, 2010] [Aran et Gatica-Perez, 2010]. Aran et Gatica-Perez [2010] ont proposé une technique qui permet de combiner les comportements non verbaux vocaux et visuels afin d'identifier la personne dominante dans une conversation face à face. Ils se sont basés sur l'hypothèse que la personne qui domine une conversation, est celle qui parle majoritairement et utilise fréquemment des gestes [Hall *et al.*, 2005]. Pour estimer la personne dominante dans une conversation i, en utilisant une seule caractéristique non verbale (un comportement) f, les auteurs ont utilisé la méthode

[1] http ://sspnet.eu/

suivante (*rule-bases estimator*) :

$$MD_i = \arg\max_p(f_p^i), p \in \{1, 2, ..., P\} \tag{4.1}$$

où p est l'identifiant du locuteur étudié, f_p^i correspond au comportement du locuteur dans la conversation i et P représente le nombre d'interlocuteurs dans la conversation. Cette méthode ne nécessite pas de données étiquetées pour réaliser l'apprentissage. Cependant son inconvénient est qu'elle ne prend en compte qu'un seul descripteur à la fois, elle ne peut pas bénéficier de la fusion de différentes caractéristiques. Les résultats expérimentaux montrent que la durée totale des tours de parole est le meilleur indice pour l'identification de la personne dominante de la conversation. Cette information est parfois complétée par des informations visuelles [Jayagopi *et al.*, 2009]. La fusion des activités vocales et visuelles peut améliorer nettement la reconnaissance. Comme la méthode proposée ne prend en compte qu'une caractéristique à la fois, Aran et Gatica-Perez [2010] ont défini un modèle pour chaque descripteur, puis ils ont combiné les différents résultats obtenus (*decision-level fusion*). Enfin, pour estimer l'importance de la domination des différents interlocuteurs, la méthode consiste à trier d'abord les interlocuteurs en utilisant des descripteurs uniques. Puis, pour l'estimation globale en tenant compte de tous les descripteurs, il faut combiner les résultats obtenus par l'utilisation des descripteurs uniques. Intuitivement, pour une conversation i, en utilisant une combinaison de descripteurs C, le rang de l'interlocuteur dominant est définit de la manière suivante :

$$R_iC = \arg\max_p(\sum_{f \in C} r_{fp}^i), C \subseteq F \tag{4.2}$$

où r_{fp}^i est le rang de l'interlocuteur p, en utilisant le descripteur f, dans la conversation i, et F est l'ensemble de descripteurs. Les résultats expérimentaux montrent la nécessité de combiner des caractéristiques visuelles et auditives, l'approche visuelle ne donnant pas satisfaction. Concernant la tâche de reconnaissance de la personne la plus dominante (MD), le meilleur résultat (85.07%) est obtenu grâce aux descripteurs audio. Pour la reconnaissance de la personne la moins dominante (LD), le meilleur résultat (85.92%) est obtenu également grâce aux descripteurs audio. Ensuite, grâce à la fusion des deux descripteurs (audio et visuels), le score de reconnaissance de MD est amélioré de 3% et celui de LD de 7%.

L'analyse de comportements humains a été également appliquée à l'indentification des conflits dans un débat, particulièrement dans un débat politique [Vinciarelli *et al.*, 2009c] [Vinciarelli *et al.*, 2009a] [Poggi *et al.*, 2010]. Poggi *et al.* [2010] ont défini la notion d'accord entre les différents interlocuteurs

dans un débat télévisé en se basant sur des signaux verbaux et non verbaux. Pour une meilleure caractérisation des signes d'accord et désaccord, Bousmalis *et al.* [2009] ont défini la notion d'accord entre deux interlocuteurs et ses différentes formes dans le cadre d'une interaction face à face (mouvement de la tête, sourire, mouvement du regard, imitation, etc). Les applications visées sont l'enrichissement d'informations : un manager pourra ainsi connaître l'évolution et la structuration de la réunion.

Dans le cadre de l'analyse d'interactions dans les réseaux sociaux, Wu *et al.* [2008] ont proposé une méthode qui permet d'identifier l'interlocuteur qui monopolise et dirige le forum, mais aussi l'identification des principaux sujets de discussion, ce travail est détaillé dans le paragraphe 4.6.2.2.

Malgré les nombreux travaux qui existent pour l'analyse de comportements humains, la représentation ou le codage de données issues des interactions réelles reste un problème ouvert et est un des premiers défis à résoudre dans la tâche d'analyse de signaux sociaux. Cette représentation doit être en adéquation avec la qualité des données et les processus de compréhension et de modélisation.

4.2.2 Représentation de données d'interactions

La représentation de données d'interactions est une étape nécessaire pour l'analyse et la compréhension de comportements humains. La méthode naïve de représentation de comportements humains consiste à représenter quantitativement chaque comportement par sa fréquence d'apparition. Malgré sa simplicité, cette forme de représentation a été largement utilisée [Nicolini *et al.*, 2010] [Aran et Gatica-Perez, 2010]. Cette simple représentation peut se transformer en une représentation matricielle. Par exemple dans le cadre de l'analyse de l'interdépendance des locuteurs dans une conversation, les séquences d'interaction sont représentées par une matrice de distances symétrique X de dimension $n \times n$, où n est le nombre de personnes enregistrées dans la discussion. Pour simplifier le problème, une interaction implique une relation symétrique entre deux individus. Lorsque le locuteur i parle à son interlocuteur j et que j écoute i, alors i et j interagissent. La valeur X_{ij} représente le nombre d'interaction entre deux locuteurs i et j. Pour l'analyse des réseaux sociaux, voir la figure 4.1, Wu *et al.* [2008] ont proposé une représentation matricielle d'interactions. Les différentes connexions entre les pages web et les internautes qui les utilisent sont représentées par une matrice X de dimension $n \times m$. La matrice X représente le n conversations associées à m individus.

La représentation d'interaction peut être également donnée sous forme d'un graphe. C'est une variante de la représentation matricielle. Dans le but d'identifier des structures dans un match de football (ex : stratégies de jeux),

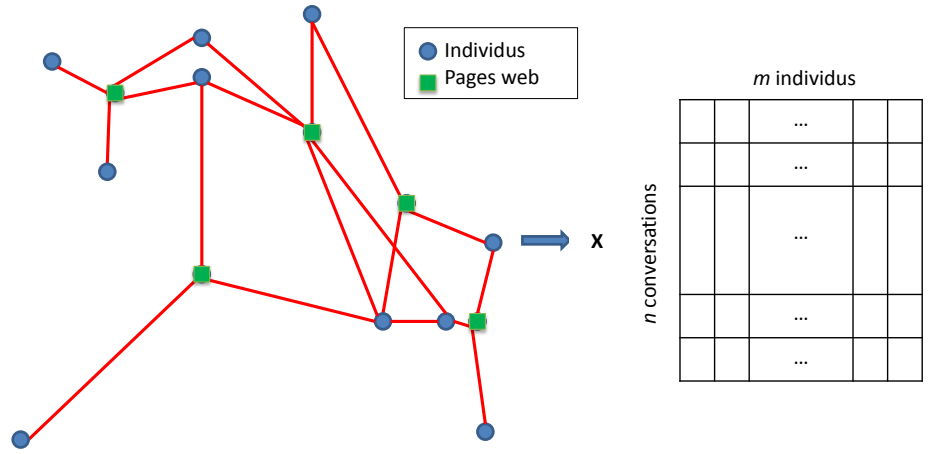


FIG. 4.1 – Différentes connexions entre les pages web et les internautes qui les utilisent [Wu *et al.*, 2008]

Park et Yilmaz [2010] ont représenté le nombre de jeux (passage du ballon) entre les joueurs sous la forme d'un graphe d'interactions G. Les joueurs sont les sommets du graphe. Etant donné un graphe d'interaction G, la matrice d'adjacence correspondant X est définie de la manière suivante :

$$X_{ij} = \begin{cases} w \quad si\ i\ et\ j\ ont\ joué\ ensemble,\ w \geq 1, \\ 0 \quad sinon \end{cases} \tag{4.3}$$

Avec w le nombre de passages du ballon du joueur i vers le joueur j.

La deuxième famille de représentation d'interaction sociale repose sur le modèle de Markov caché [Rabiner, 1989]. Cette méthode consiste à estimer la probabilité d'apparition de séquences de comportements (parole, geste, etc) de personnes [Maisonnasse et Brdiczka, 2005] [Favre *et al.*, 2008]. Le résultat de cette représentation est un ensemble de *n+1 uplets*, de séquences d'actions $< a_1, a_2, ..., a_n, p >$, où a_i est l'action, n est la longueur de la séquence d'interaction et p est la probabilité d'apparition de cette séquence. L'ensemble des séquences de dimension n forment le modèle $n-gram$ d'interaction. Dans le cas où $n = 1$, les séquences représentent simplement les probabilités d'apparition de chaque comportement au cours de l'interaction.

Enfin, Favre *et al.* [2008] ont proposé une méthode intelligente pour la représentation de données d'interaction. Cette méthode repose sur les tours de parole. Comme nous l'avons expliqué dans le paragpraphe 1.5.2, chaque séquence d'interaction est décomposée en plusieurs tours de parole grâce à un algorithme de segmentation en locuteur. Ensuite, chaque séquence d'interactions est représentée par un vecteur $X = \{x_1, x_2, ..., x_n\}$ de dimension n, où n

représente le nombre de locuteur participant, x_i est un vecteur de dimension m, où m est le nombre de tours de parole. Par exemple, $x_1 = (1, 1, 0, 1, 1)$ signifie d'abord qu'il y a 5 locuteurs dans la conversation étudiée x_1, et que les locuteurs 1, 2, 4 et 5 ont participé (pris un tour de parole) dans la séquence x_1 mais pas le locuteur 3 ($x_1(3) = 0$).

4.3 Interaction parent-enfant

4.3.1 Définition

L'interaction parent-enfant est un cas particulier de l'interaction sociale. Cependant, comme nous l'avons montré dans le chapitre 1, les enfants et les adultes ne partagent pas les mêmes signaux de communication. Les interactions parent-enfant évoluent de façon considérable au cours du temps : les premiers échanges s'établissent dans une position face à face, étayée par un engagement de regards mutuels avant l'installation des autres compétences sensorielles et sensori-motrices chez l'enfant. L'interaction parent-enfant a été largement étudiée par le communauté de la psychologie de l'enfant. Par contre il n'existe que très peu de travaux qui traitent ce problème du point de vue computationnel. Dans la littérature, les travaux, qui se sont intéressés à l'interaction parent-enfant, se sont particulièrement basés sur l'analyse d'interaction en étudiant les expressions faciales [Jaffe *et al.*, 2001], le regard ou les comportements vocaux [Crown *et al.*, 2002].

L'analyse d'interaction parent-enfant permet d'identifier certaines règles de communication. Ainsi elle offre la possibilité d'identifier les signaux de communication, produits par le parent, primordiaux dans l'engagement et la régulation de l'interaction. Cette tâche se révèle complexe du fait de la stochasticité de l'interaction : variations inter et intra couple interactif (parent-enfant) mais également la coordination ou adaptation mutuelle de part la nature bidirectionnelle de l'interaction. Il est possible que l'enfant, également le parent, arrive à prédire le comportement de son interlocuteur (sourire, vocalisation) grâce à certaines règles naturelles d'interaction. Par exemple, lorsque le parent fait un sourire à l'enfant, l'enfant ne va pas forcément faire un sourire. Par contre lorsque l'enfant fait un sourire, le parent doit forcément faire de même et continue à sourire jusqu'à que l'enfant s'arrête [Symons et Moran, 1994]. Par conséquent, la plupart des études sur l'interaction parent-enfant estime que les comportements des parents sont dépendants des actions précédentes de l'enfant. Par contre, il est difficile de prédire les comportements de l'enfant. Il n'existe pas de règles nous permettant de prédire la réponse de l'enfant vis-à-vis d'un comportement ou d'une série de comportements réalisés

par le parent [Symons et Moran, 1994] [Messinger *et al.*, 2010]. Pour étudier ce phénomène, des travaux se basant sur des analyses statistiques d'interactions parent-enfant [Muratori *et al.*, 2010] ont fait appel à des méthodes automatiques qui permettent d'extraire les comportements des parents qui ont plus de pouvoir d'engager une interaction avec l'enfant et évaluer la dépendance de comportement de l'enfant et de parent au cours du temps [Messinger *et al.*, 2010]. La plupart de ces travaux se basent sur l'analyse des films familiaux.

Les méthodes d'apprentissage automatique ont été largement utilisées pour modéliser les comportements humains au cours de l'interaction [Butko et Movellan, 2008] [Schulz et Gopnik, 2004]. Afin d'identifier les signaux pertinents dans la communication intentionnelle de l'enfant, Messinger *et al.* [2010] ont appliqué des méthodes des reconnaissance de formes afin de prédire les comportements des enfants et de leurs mères au cours de l'interaction. Les auteurs ont utilisé une base de données longitudinale : collection des scènes d'interactions mère-bébé filmées dans un laboratoire. Ils se sont intéressés à trois comportements : le sourire de l'enfant, le sourire de la mère et l'orientation du regard de l'enfant. Afin de vérifier que l'interaction mère-enfant devient de plus en plus stable au cours du temps, les auteurs ont comparé les scènes d'interactions enregistrées sur des intervalles de temps différents. D'abord, ils ont modélisé les interactions mère-enfant. Ce modèle décrit les couples de comportements. Ils ont utilisé le critère de maximum de vraisemblance pour estimer la distribution des probabilités des transitions entre les différents états d'interaction :

$$p(i_t, m_t, i_{t-1}, m_{t-1}) \tag{4.4}$$

où i est le comportement de l'enfant, m est le comportement de la mère, i_t représente l'état de l'enfant à l'instant t et i_{t-1} est l'état de l'enfant à l'instant $t-1$. Puis, la similarité entre les différentes scènes est calculée en utilisant les coefficients de Bhattacharyya :

$$f(p_1, p_2) = \int \sqrt{p_1(s)} \times \sqrt{p_2(s)} ds \tag{4.5}$$

où p_1 et p_2 sont les distributions des interactions dans deux scènes consécutives. Les résultats expérimentaux montrent que la similarité entre les scènes consécutives augmente au cours du temps. Les enfants et leur mère deviennent de plus en plus susceptibles d'interagir et surtout avec une certaine stabilité. Par conséquent les comportements des enfants deviennent de plus en plus faciles à prédire au cours du temps.

4.3.2 Etude de films familiaux pour l'analyse d'interaction parent-enfant

Les films familiaux offrent un cadre expérimental pertinent pour l'étude de l'interaction parent-enfant [Saint-Georges *et al.*, 2010] [Muratori *et al.*, 2010]. L'analyse de films familiaux nous permet d'avoir des observations longitudinales de l'interaction sociale parent-enfant. Elle permet également d'avoir une observation qualitative du développement cognitif et neuro-moteur de l'enfant. Les films familiaux offre un cadre expérimental riche et adapté aux troubles du développement.

Dans le cadre de l'analyse de l'interaction parent-enfant, en particulier les enfants autistes, et dans l'objectif d'identifier des signes infantiles de l'autisme et les premiers symptômes, nous étudions des bases d'observation directe des enfants en situation d'interaction avec leurs parents (les films familiaux). L'analyse des films familiaux permettent d'examiner les interactions sociales précoces [Condon et Sander, 1974] [Stern *et al.*, 1975] et les comportements des parents et des enfants pendant la période néonatale [Brazelton *et al.*, 1975]. Ainsi, l'utilisation de films familiaux permet de préciser certains signes précoces de pathologies développementales de l'enfant. Même si un grand nombre de travaux porte sur des films familiaux, très peu se consacrent à l'étude du développement des enfants avant l'âge de deux ans. Massie [1975] est le premier a avoir publié des études basées sur les films familiaux représentant des enfants à devenir autistique, il s'est intéressé à l'analyse de l'interaction mère-bébé et l'étude des premiers symptômes autistiques. Il a mis en évidence les difficultés des enfants autistes dans l'interaction et l'attachement à ses parents.

D'autres équipes de recherche ont abordé ce problème, Receveur *et al.* [2005] ont ainsi développé des outils pour la reconnaissance de signes précoces de l'autisme (geste, imitation, attention). Ils ont étudié les perturbations précoces de communication chez les enfants autistes. Ils ont montré que l'interaction sociale et l'attention conjointe sont déjà défectueuses dès les premiers mois. Dans le même contexte, les travaux de l'équipe de *Pise* sur le thème des signes précoces de l'autisme et ses conséquences sur l'interaction parent-enfant [Maestro *et al.*, 2001] [Maestro *et al.*, 2002] [Maestro *et al.*, 2006] offrent une description détaillée du développement de l'enfant. Ces auteurs ont étudié les premières apparitions des symptômes autistiques (par exemple attention aux personnes ou aux objets, imitations, etc) en analysant des films familiaux avec différentes conditions et états pathologiques. L'objectif de ces travaux est d'identifier les caractéristiques comportementales des enfants autistes et leurs capacités d'interaction sociale et de communication.

TAB. 4.1 – Nombre et durée (en minute) des différentes scènes des trois groupes d'enfant distribuées sur trois semestres

	AD (15)	*MR* (12)	*TD* (15)	*t-test* (valeur de *p*)		
				AD vs *MR*	*AD* vs *TD*	*MR* vs *TD*
Nombre $S1$	162	100	164	.116	.163	.138
Durée $S1$	258.5	148.5	237.3	.142	.759	.105
Nombre $S2$	164	111	201	.639	.088	.079
Durée $S2$	258.1	176.3	347.5	.863	.148	.278
Nombre $S3$	127	90	120	.917	.747	.701
Durée $S3$	212	194.5	194.5	.202	.474	.371

4.4 Bases de données des interactions parent-enfant

4.4.1 Bases de données longitudinale : films familiaux

Dans notre travail, nous exploitons la base de données conçue par l'équipe de Muratori *et al.* [2010] [Maestro *et al.*, 2005]. Cette base de données est composée de plusieurs séquences d'interactions issues de 42 films familiaux. Les films familiaux représentent trois groupes d'enfant (*AD*, *MR* et *TD*) filmés pendant leurs premiers 18 mois. Le premier groupe *AD* (Autistic Disorder) est composé de 15 enfants (10 garçons et 5 filles), ces enfants sont diagnostiqués autiste plus tard. Le deuxième groupe *MR* (Mental Retardation) est composé de 12 enfants (7 garçons et 5 filles), ces enfants ont un retard de développement. Enfin, le troisième groupe *TD* (Typical Development) est composé de 15 enfants (9 garçons et 6 filles), ces enfants ont un développement typique. De plus, pour effectuer une étude longitudinale, les séquences vidéos sont organisées en trois intervalles de temps, premier semestre $S1$, deuxième semestre $S2$ et troisième semestre $S3$, voir le tableau 4.1. Le *t-test* a été utilisé pour vérifier l'homogénéité des trois groupes d'enfants, vis-à-vis de la moyenne d'âge des enfants, le nombre et la longueur des scènes d'interactions.

Cette base de données est conçue pour les recherches longitudinales sur l'interaction parent-enfant permettant ainsi d'étudier l'évolution du développement des enfants au cours du temps. De plus, cette base de données permet potentiellement d'identifier les signes précoces de l'autisme et les différents symptômes de cette pathologie développementale. Elle permet également de comparer la qualité de l'interaction des enfants pathologiques à celles des enfants avec développement normal, ainsi d'identifier les conséquences de l'altération des interactions sur les comportements des parents et des enfants.

Tab. 4.2 – Comportements de parent

Comportements	E/S	Description
Regulation up/down	*S*	Le parent agit sur l'humeur du bébé. Le parent peut calmer ou exciter et provoquer le bébé
Toucher	*E*	Le parent stimule le bébé par le toucher
Vocalisations	*E*	Le parent stimule le bébé par l'appel du prénom ou par des autres formes de vocalisation
Geste/montrer quelque chose	*E*	Le parent stimule le bébé par des gestes ou par le fait de lui montrer quelque chose

4.4.2 Annotation de l'interaction face à face

4.4.2.1 Codage de données

Après la collection et l'organisation de données d'interaction, la deuxième étape consiste à annoter toutes les séquences d'interaction parent-enfant. Le logiciel professionnel *Observer* 4.0 a été utilisé pour organiser et analyser les données. Il a été configuré pour étiqueter manuellement les comportements des bébés et des parents sous une échelle donnée *ICBS* (Infant and Caregiver Behavior Scale) qui défini l'ensemble des comportements des parents et des enfants. Les tableaux 4.2 et 4.3 présentent respectivement l'ensemble des étiquettes correspondant aux comportements des parents et les réponses des bébés dans le cadre de l'interaction sociale. Ces étiquettes sont divisées en deux groupes, d'abord les étiquettes des comportements des parents *CB* (Caregiver Behavior), puis les étiquettes des comportements des enfants *IB* (Infant Behavior). Les comportements sont caractérisés sous la forme d'évènements (E) ou d'états (S). Un comportement de type état est un comportement qui se déroule sur un intervalle de temps, et les comportements de type évènement sont représentés par leur fréquence d'apparition (cf. tableaux 4.2 et 4.3).

4.4.2.2 Annotation

Quatre annotateurs ont participé à l'annotation d'interactions, deux pédopsychiatres et deux psychologues (pour enfants). La première étape consiste à coder les comportements sous formes d'étiquettes, le tableau 4.4 présente les différentes étiquettes associées aux comportements des parents et des enfants. Ensuite, les annotateurs ont été entrainés à annoter les cas pathologiques et typiques en utilisant le logiciel *Observer* et en exploitant la grille des comportements présentée dans le tableau 4.4. Après la phase d'entraînement, les annotateurs ont annoté la totalité de la base d'interaction. L'intervalle de tolérance entre deux annotations d'un même comportement est fixé à 1 seconde pour les comportements de type évènement et de 3 secondes pour les comportements de type état. L'étiquetage enregistré en dehors de ces fenêtres a été

TAB. 4.3 – Comportements des enfants

Comportements	E/S	Description
Comportements envers des objets		
S'orienter vers un objet	*E*	Le bébé s'oriente vers un objet (regard spontané)
Admiration d'un objet	*E*	Le bébé déplace son regard pour suivre la trajectoire d'un objet
Explorer un objet	*S*	L'enfant touche un objet
Regarder un objet/autour	*S*	Le bébé regarde un objet ou simplement regarder autour de lui
Sourire à un objet	*E*	Le bébé sourit intentionnellement à un objet
Jouer avec un objet	*S*	Le bébé trouve du plaisir et de la satisfaction après avoir un contact physique et|ou visuel avec un objet
Chercher contact avec un objet	*E*	Le bébé fait des mouvements spontanés et intentionnelles pour entrer en contact avec un objet
Comportements expressifs		
Vocalisation simple	*E*	Le bébé produit des sons envers une personne ou un objet
Crier	*S*	Le bébé commence à crier suite à un évèvement spécifique
Comportements d'exploration		
S'orienter vers une personne	*E*	Le bébé s'oriente vers une personne
Admiration d'une personne	*E*	Le bébé déplace son regard pour suivre la trajectoire d'une personne
Explorer une personne	*S*	L'enfant touche une personne
Comportements réceptifs		
Regarder une personne	*S*	Le bébé regarde une personne
Sourire à une personne	*E*	Le bébé sourit intentionnellement à une personne
Jouer avec une personne	*S*	Le bébé trouve du plaisir et de la satisfaction après avoir un contact physique et|ou visuel avec une personne
Syntonie	*S*	Le bébé montre des signes d'expressions congrues envers des sollicitations affectives du parent
Comportements actifs		
Chercher un contact avec une personne	*E*	Le bébé fait des mouvements spontanés et intentionnelles pour entrer en contact avec une personne
Solliciter	*E*	Le bébé présente une action verbale, vocale ou tactile pour attirer l'attention du partenaire ou d'obtenir une autre type de réponse
Comportements d'intersubjectivité		
Anticipation de l'intention des autres	*E*	Le bébé fait des mouvements d'anticipation pour prédire l'action des autres
Gestes communicatifs	*E*	Le bébé communique avec des gestes sociaux
Regard référentiel	*E*	Le bébé déplace son regard vers une situation spécifique
Accepter l'invitation	*E*	Le comportement de l'enfant est communiqué 3 secondes après la sollicitation du parent
S'orienter à l'appel du prénom	*E*	Le bébé s'oriente vers la personne qui l'a appelé par son prénom
Imitation	*E*	Le bébé répète le geste ou l'action de l'autre dans un intervalle de temps très court
Pointer compréhension/déclarative	*E*	Le bébé déplace son regard vers la direction indiquée par le parent en utilisant son doigt pour indiquer quelque chose dans le but de partager un état émotionnel
Le maintien de l'engagement social	*S*	Le bébé prend un rôle actif au sein d'une interaction bidirectionnelle Le bébé interagit, vocalise et maintient son tour de parole
Vocalisation significative	*E*	Le bébé produit et communique intentionnellement des sons dont ils ont des sens

signalé comme une erreur d'annotation et était considéré comme un désaccord. A l'issue de cette annotation manuelle, la valeur de fidélité inter-juge, *Cohen's Kappa*, entre les différents annotateurs est égale à 0.77 pour les comportements de type état et de à 0.75 pour les comportements de type évènement, ce qui représente un accord fort entre les annotateurs (cf. la section 2.7.2). La

TAB. 4.4 – Les étiquettes représentant les comportements de parent et d'enfants

Enfant Comp. basiques	**étiq.**	**Enfant Compo. complexes**	**étiq.**	**Enfant Vocalisations**	**étiq.**	**Parent sollicitation**	**étiq.**
Regarder des personnes	B_lkp	Jouer*	C_ejx	simple	v_sim	REGULATION*,UP	G_reu
Regarder des objets	B_lko	Jouer avec une personne	C_ejp	significative	v_mea	REGULATION*,DOWN	G_red
Regarder autour	B_lka	Jouer avec un objet	C_ejo	cries	V_cry	Demander	g_req
Exploration	B_exo	Syntonie	C_snt			Appel du prénom	g_nam
Contacter une personne	b_ctp	Engagement social	C_seg			Toucher	g_tou
Contacter un objet	b_cto	Anticipation	c_ant			Vocalisation	g_voc
S'orienter XXX	b_orx	Gestes communicatifs	c_com			Geste	g_ges
S'orienter vers une personne	b_orp	Regard référentiel	c_ref			Montrer un objet	g_sho
S'orienter vers un Objet	b_oro	Sollicitation	c_sol			NULL (cg)	G_nul
S'orienter à l'appel du prénom	b_orn	Acc invitation	c_acc				
Suivie du regard d'une personne	b_gfp	S'offerir	c_off				
Regard de suivie d'un objet	b_gfo	Imitation	c_imi				
Sourire à une personne	b_smp	Pointer	c_poi				
Sourire à un objet	b_smo						

TAB. 4.5 – Fréquences et durées des différents comportements de parent et des enfants pour le trois groupes *AD*, *MR* et *TD* durant le 3 premiers semestres $S1$, $S2$ et $S3$

	Groupe *AD*			Groupe *MR*			Groupe *TD*		
Etiquette (fréquence/durée)	$S1$	$S2$	$S3$	$S1$	$S2$	$S3$	$S1$	$S2$	$S3$
S'orienter à l'appel du prénom	0.2(0.2)	0.3(0.4)	0.3(0.4)	0.1(0.3)	1.1(1.1)	0.7(0.7)	0.1(0.2)	0.4(0.4)	0.7(0.6)
Sourire à une personne	2.2(3.9)	4.6(6.2)	1.3(1.3)	3.6(2.8)	4.4(3.8)	1.8(1.8)	1.4(1.6)	3(3.2)	3.3(2.3)
Jouer avec une personne (fréq.)	0.1(0.2)	0.3(0.7)	0.1(0.1)	0.1(0.2)	0.2(0.3)		0.0(0.1)	0.3(0.1)	0.4(0.5)
Jouer avec une personne (dur.)	0.6(1.6)	2(4.2)	0.1(0.4)	1.5(3.4)	2.2(3.4)			1.8(5.2)	1.6(2.4)
Syntonie (fréq.)	0.1(0.1)	0.1(0.2)	0.1(0.1)	0.3(0.6)	0.2(0.2)	0.1(0.1)	0.03(0.1)	0.1(0.2)	0.04(0.1)
Syntonie (dur.)	0.2(1)	1.7(5.1)	0.1(0.4)	2.8(5.3)	0.6(0.8)	0.1(0.3)		1.2(2.4)	0.5(1.4)
Regard référentiel		0.1(0.2)		0.04(0.1)	0.1(0.1)	0.4(0.6)		0.1(0.2)	
Solliciter	0.1(0.2)	0.7(1.2)	0.5(0.7)	0.4(0.8)	1.6(1.3)	1.9(2.1)	0.1(0.2)	1.1(1.4)	0.7(0.6)
Accepter invitation		0.5(0.5)	0.6(0.7)	0.2(0.5)	1.2(1.1)	2.3(2.4)		0.2(0.4)	0.1(0.7)
Engagement social (fréq.)	0.1(0.4)	0.1(0.2)	0.1(0.2)	0.2(0.2)	0.5(0.4)	0.5(0.6)	0.1(0.3)	0(0.06)	0.2(0.3)
Engagement social (dur.)	2.1(6.3)	0.7(1.6)	2.6(4.9)	2.9(4.7)	5.1(6)	6.1(7.5)	1.1(3.6)		5.4(11)
Vocalisations simples	8.2(9.3)	12.2(12.4)	6.2(6.3)	14.9(14.1)	17.3(10.8)	13.1(6.3)	7.2(7.7)	7(7)	14.7(11.4)
Vocalisations significatives	0.03(0.1)	0.04(0.1)	0.5(1.4)		0.2(0.5)	2.7(3.8)	0(0.03)	0.03(0.1)	0.3(0.5)
Regulation down (fréq.)	0.1(0.2)	0.03(0.1)		0.2(0.4)	0.1(0.3)	0.1(0.3)	0.4(0.5)	0.03(0.1)	0.3(0.3)
Regulation down (fréq.)	0.7(1.6)	0.02(0.1)		1.5(3.9)	0.9(2.8)	0.6(1.3)	2.8(3.8)	0.04(0.1)	1.5(2.8)
Sollicitation de parent	5.8(12.6)	6.3(13.6)	4.4(4.3)	7.6(12.9)	8.5(13.5)	6.9(10.8)	6.1(13.6)	4.5(7.2)	6.4(9.3)
Par l'appel du prénom	1.7(1.5)	3.2(2.5)	1.8(1.5)	1.2(1.2)	3.3(2.2)	2.6(1.9)	1.8(2)	2.9(2.5)	4.2(3.5)
Par les gestes	0.4(0.9)	1.4(2.3)	1.1(1.5)	0.8(1.5)	2.4(2.3)	1.5(1.6)	0.3(0.5)	1(1.1)	2.4(4.4)
Par le toucher	7.6(11.7)	3.7(4.8)	2.4(2.8)	7.1(6.5)	3.2(3.1)	1.5(1.6)	8.2(7.7)	2.8(3.1)	2.5(1.8)
Par le "request behavior"	2.2(3.1)	5.1(5.7)	5.2(5.8)	3.4(2.2)	7.9(5.6)	8.6(6.4)	2(1.9)	4.04(3.5)	6.8(2.8)
Par montrer un objet	0.7(1.3)	1.3(1.2)	1.1(1.1)	1.5(1.7)	2.6(2.4)	1.8(1.8)	0.7(1.1)	0.9(1.2)	1.2(1.8)
par la vocalisation	22.3(21.8)	22.9(27.1)	14.7(13.6)	31.3(16.1)	31.7(19.9)	25.7(14.1)	23.6(25.7)	15.5(11.6)	21.5(14)

figure 4.2 donne une visualisation d'annotation d'interactions parent-enfant.

4.4.2.3 Analyse de données

Les fréquences des comportements (valeur moyenne d'apparition d'un comportement par minute) et les durées moyennes des comportements de type état ont été estimées. Le tableau 4.5 présente les fréquences, les durées moyennes et les valeur de l'écart-type des différents comportements. Les cases vides correspondent aux signaux introuvables ou existant mais avec une fréquence très faible. Ensuite, une analyse préliminaire réalisée par un test *Anova* [*temps(3)* $\times$ *groupe(3)*], présentée dans le tableau 4.6, a été réalisé pour comparer le trois groupes par rapport aux fréquences et durées des comportements le long des

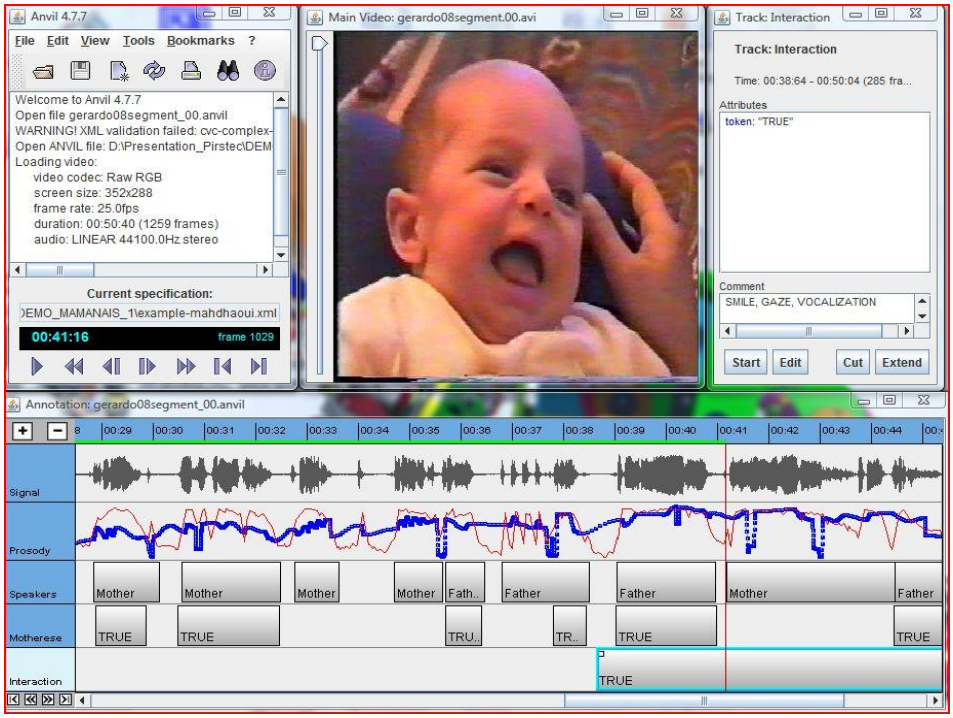


FIG. 4.2 – Exemple d'annotation

trois semestres. Dans le tableau 4.6, les valeurs significatives de p sont présentées par des chiffres positifs inférieur à 1, sachant que plus la valeur de p est petite plus elle est significative, ns indique que la valeur de p est non significative. Afin d'identifier les comportements les plus significatifs pour chaque groupe d'enfant, un test *post-hoc* a été utilisé pour chaque semestre (cf. tableau 4.7). Le test *post-hoc* permet de comparer le degré de significativité d'un tel comportement pour chaque groupe d'enfants. Enfin, une analyse de comportements discriminants a été utilisée pour déterminer la puissance de discrimination de chaque comportement entre le groupe AD et le groupe TD ou le groupe MR durant les trois semestres $S1$, $S2$ et $S3$, voir le tableau 4.8.

Le tableau 4.7 regroupe les effets significatifs obtenus par le test *Anova* et le test *post-hoc* des comportements sortant significatifs en appliquant le test *Anova* à deux dimension (tableau 4.6). Durant $S1$, nous constatons qu'il y a une différence significative entre les différents groupes en terme de durée de la 'syntonie'. De plus le test *post-hoc* montre que cette différence est due à la valeur de durée élevée de 'syntonie' pour le groupe TD par rapport aux groupes AD et MR. Par contre, il n'y a pas de différence significative entre le groupe AD et MR durant la période $S1$. Durant la période $S2$, le test *Anova* montre qu'il y a de différences significatives au niveau des comportements

TAB. 4.6 – Effet significative du *temps*, *temps* × *groupe* et *groupe* (*Anova* [*temps(3)* × *groupe(3)*])

Comportements	Fréquence/durée	*temps*		*temps* × *groupe*		*groupe*	
		$F(2.36)$	p	$F(4.36)$	p	$F(2.36)$	p
Comportement des enfants							
Jouer avec une personne	fréquence	2.69	0.07	2.28	0.07		ns
Jouer avec une personne	durée		ns	2.21	0.08		ns
S'orienter à l'appel du prénom	fréquence	7.99	0.00	3.72	0.01	3.25	0.05
Sourire à une personne	fréquence	3.29	0.04		ns		ns
Syntonie	fréquence		ns		ns	3.87	0.03
Syntonie	durée		ns	2.12	0.09		ns
Regard référentiel	fréquence	2.59	0.08	4.08	0.00	5.09	0.01
Solliciter	fréquence	6.11	0.00		ns	8.94	0.00
Accepter invitation	fréquence	12.66	0.00	2.43	0.06	9.06	0.00
Engagement social	fréquence		ns		ns	8.51	0.00
Engagement social	durée	2.45	0.09		ns	3.40	0.04
Vocalisations simples	fréquence		ns	2.08	0.09	3.57	0.04
Vocalisations significatives	durée	7.20	0.00	3.75	0.01	4.39	0.02
Comportement du parent							
Regulation down	fréquence	3.88	0.03		ns	7.87	0.00
Regulation down	durée	3.95	0.02		ns	3.56	0.04
Sollicitation	fréquence	11.17	0.00		ns	4.84	0.01

suivants : 's'orienter à l'appel du prénom', 'engagement social' (fréquence et durée) et 'accepter invitation'. De plus, le test *post-hoc* montre que le groupe *TD* a des scores plus élevés que les groupes *AD* et *MR* au niveau de ces comportements, à l'exception du comportement 's'orienter à l'appel du prénom' qui est uniquement plus élevé que le groupe *AD*. Par ailleurs, il n'y a pas de différences significatives entre les groupes *MR* et *AD* durant la période $S2$. Durant la période $S3$, le test *Anova* montre aussi d'autres différences significatives. Le test *post-hoc* montre que le groupe *TD* a des scores significativement plus importants que le groupe *AD* au niveau des comportements 'regard référentiel', 'solliciter', 'accepter invitation' et 'vocalisations significatives'. Au niveau de tous ces comportements il n'existe pas de différences significatives entre les groupes *AD* et *MR*. Cependant, le comportement 'engagement social' est le seul qui ne présente pas de différence entre les groupes *MR* et *TD*. Pour les autres comportements, 'sourire à une personne' , 'jouer avec une personne' (enjoying with people) (fréquence et durée), 'vocalisations simples' et 'regulation-down', le groupe *AD* présente des scores plus importants que le groupe *MR* (des scores inférieurs).

Dans le tableau 4.8, nous constatons que durant $S1$, le comportement 'regulation-down' est significativement discriminant entre le groupe *AD* (80%) et le groupe *MR* (55%). Durant $S2$, les comportements 's'orienter à l'appel du prénom' et 'engagement social' sont également significativement discriminants entre le groupe *AD* (73%) et le groupe *TD* (46% et 60%). Durant $S3$,

TAB. 4.7 – Test *Anova* et *post-hoc* pour les comportements significatifs

Comportements	Classe	fréquence/durée	$F(2,39)$	p	*Post hoc* (*<.099 ;**<.05)		
					AD vs TD	AD vs MR	MR vs TD
$S1$							
Syntonie	IB	durée	334	.04	$AD < TD*$	ns	$MR<TD$
$S2$							
S'orienter à l'appel du prénom	IB	fréquence	503	.01	$AD < TD**$	ns	ns
Engagement social	IB	fréquence	1004	.00	$AD < TD**$	ns	$MR < TD**$
		durée	791	.00	$AD < TD**$	ns	$MR < TD**$
Accepter invitation		fréquence	498	.01	$AD < TD*$	ns	$MR < TD**$
$S3$							
Sourire à une personne	IB	fréquence	355	.04	ns	$AD < MR**$	ns
Jouer avec une personne		fréquence	573	.00	ns	$AD < MR**$	$MR < TD**$
		durée	588	.00	ns	$AD < MR**$	$MR < TD**$
Engagement social	IB	fréquence	376	.03	$AD < TD**$	ns	ns
Regard référentiel			562	.00	$AD < TD**$	ns	$MR < TD*$
Solliciter			419	.02	$AD < TD**$	ns	$MR < TD*$
Accepter invitation	IB	fréquence	501	.01	$AD < TD**$	ns	$MR < TD*$
Vocalisations simples	IB	fréquence	41	.02	ns	$AD < MR**$	ns
Vocalisations significatives	IB	fréquence	406	.02	$AD < TD**$	ns	$MR < TD*$
Regulation down	CB	fréquence	313	.05	ns	$AD < MR**$	ns

TAB. 4.8 – Analyse des comportements discriminants

Comportements	Classe	fréquence/durée	Entre les groupes	Wilks Lambda	sig.	Pourcentage des cas bien identifier		
						AD	TD	MR
$S1$								
Regulation down	CB	fréquence	$AD<MR$	828	.03	80%		55%
$S2$								
S'orienter à l'appel du prénom	IB	fréquence	$AD<TD$	810	.01	73%	46%	
Engagement social	IB	fréquence	$AD<TD$	782	.00	73%	60%	
$S3$								
S'orienter à l'appel du prénom	IB	fréquence	$AD<MR$	318	.00	93%		77%
Sourire à une personne	IB	fréquence	$AD<MR$	318	.00	93%		77%
Accepter invitation	IB	fréquence	$AD<TD$	790	.01	93%	46%	
Jouer avec une personne	IB	durée	$AD<MR$	803	.03	93%		33%
Vocalisations simples	IB	fréquence	$AD<TD$	760	.00	73%	73%	
			$AD<MR$	814	.03	73%		50%
Regulation down	CB	fréquence	$AD<MR$	730	.00	100%		66%

différent comportements 's'orienter à l'appel du prénom', 'sourire à une personne', 'jouer avec une personne', 'vocalisations simples' et 'regulation-down' permettent aussi de différencier les groupes *AD* et *MR*. Enfin, les comportements 'accepter invitation' et 'vocalisations simples' permettent de discriminer le groupe *AD* et le groupe *TD*.

4.5 Analyse de l'interaction parent-enfant

Dans le paragraphe 4.3, les comportements des enfants étaient étudiés sans prendre en compte le processus d'interaction avec le parent. Dans cette partie, nous étudions les comportements de l'enfant dans les situations interactives. Le schéma 4.3 montre les différentes étapes de ce travail. D'abord,

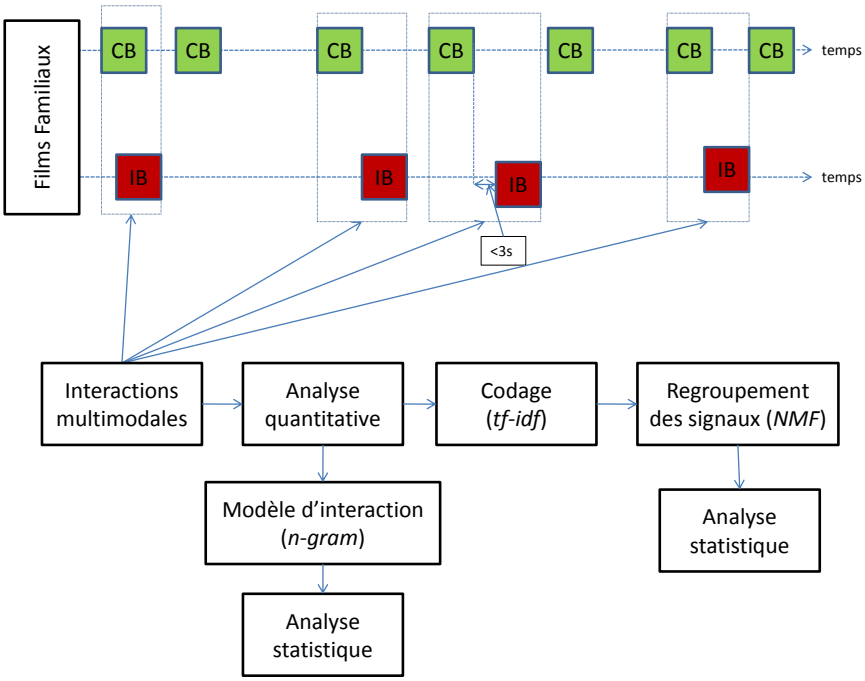


FIG. 4.3 – Shéma d'analyse de l'interaction parent-enfant

nous construisons une base d'interaction multimodale qui regroupe les séquences d'interactions parent-enfant. A l'issue de cette analyse quantitative, nous construisons un modèle d'interaction (*n-gram*). Ensuite, nous poursuivons par faire une simple analyse quantitative des données d'interactions. Enfin, dans le but d'identifier les différents groupes d'interaction par type d'enfant, nous proposons une méthode de codage des données d'interaction qui fera l'objet d'un processus de regroupement.

4.5.1 Création de la base d'interaction multimodale

A l'issue de l'annotation de la base d'interaction, nous obtenons un ensemble de données étiquetées. Cependant, ces données brutes ne sont pas exploitables directement pour l'analyse de l'interaction parent-enfant, elles ne sont pas prises dans un contexte interactif. Par conséquent, afin d'étudier les comportements de l'enfant vis-à-vis les sollicitations du parent, nous avons construit une base d'interaction multimodale à partir des données issues de la phase d'annotation. Comme l'indique la figure 4.3, nous avons commencé par extraire les séquences d'interactions parent-enfant. C'est l'ensemble des comportements de parent suivis par des réponses de l'enfant dans une fenêtre de

temps inférieure à 3 secondes. Par ailleurs, les comportements du type état des enfants (avec des durées), la 'syntonie', le 'maintien de l'engagement social' et 'jouer avec une personne ou un objet' (enjoying with people or objects), sont inclus directement dans la base d'interaction.

Cette base d'interaction fera l'objet d'analyses quantitatives et statistiques afin de comprendre et étudier les interactions des différents groupes d'enfants. Les figures 4.4, 4.5 et 4.6 montrent les 20 couples d'interaction les plus fréquents respectivement durant le premier, deuxième et troisième semestre ($S1$, $S2$ et $S3$). Le couple d'interaction '$G\_reu\#V\_sim$' signifie que l'un des parents a communiqué un signal de *regulation-up* et l'enfant a répondu par une vocalisation simple dans une fenêtre de 3 secondes.

Ensuite, nous avons construit un modèle d'interaction *n-gram*. Le modèle d'interaction est inspiré des modèles de langage[2]. C'est une méthode souvent utilisée en traitement automatique de parole. Elle consiste à modéliser les séquences d'interaction, afin d'estimer la probabilité d'une suite de comportements. Ainsi, avec un modèle *3-gram*, la probabilité d'une séquence de comportements SEQ contenant n comportements est :

$$P(SEQ) = \prod_{i=1}^{n} P(comp_i | comp_{i-2}, comp_{i-1}) \tag{4.6}$$

L'estimation des probabilités des suites de comportements est calculée à partir de la base d'interaction. Pour cela, nous avons utilisé le *Maximum de vraisemblance*. Par exemple pour un modèle *3-gram* :

$$P_{mv} = \frac{tf(comp_i comp_j comp)}{tf(comp_i comp_j)} \tag{4.7}$$

où tf correspond au décompte des interactions (*term-frequency*).

La qualité d'un modèle d'interaction dépend de sa capacité à représenter le maximum des séquences de comportement. La mesure la plus couramment utilisée pour évaluer le modèle *n-gram* est la perplexité. La perplexité d'un modèle *n-gram* correspond à sa capacité de prédiction. Plus la valeur de perplexité est petite, plus le modèle d'interaction possède des capacités de prédiction. Pour un modèle *n-gram*, la perplexité se définit ainsi :

$$PP = 2^{\frac{-1}{n} \sum_{t=1}^{n} \log_2 P(comp_t | h)} \tag{4.8}$$

où h correspond à l'historique du comportement $comp_t$.

L'estimation du modèle est dépendant de la quantité de données. Nous utilisons un modèle *2-gram* qui donne une meilleure représentation des données. La valeur de perplexité est égale à 1.9 pour le modèle typique, 2.2 pour

[2]http ://www.w3.org/TR/ngram-spec/

le modèle des enfants autistes et 2.3 pour le modèle des enfants avec retard mental.

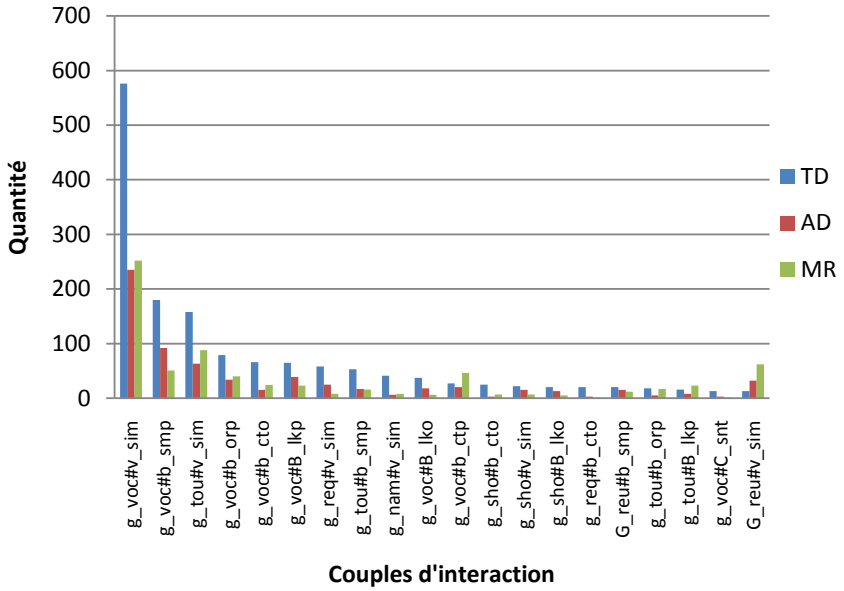


FIG. 4.4 – Les couples d'interaction les plus fréquents durant le premier semestre

4.5.2 Caractérisation quantitative de l'interaction parent-enfant

Les figures 4.7, 4.8 et 4.9 représentent respectivement les modèles probabilistes des signaux d'interaction parent-enfant pour les enfants avec développement typique, les enfants autistes et les enfants avec retard mental. Chaque figure est composée de trois sous figures correspondant respectivement aux premier $S1$, deuxième $S2$ et troisième $S3$ semestres. Les signaux d'interaction sont regroupés (cf. les tableaux 4.3, 4.2). Du coté de l'enfant, nous avons défini six groupes de comportement. Le premier groupe correspond aux comportements d'intersubjectivité ($b\_intersubjectivity$). Cela comprend les comportements suivants : vocalisations significatives, accepter invitation, s'orienter à l'appel du prénom, regard référentiel, des gestes communicatifs, imitation, le maintien de l'engagement social et l'anticipation de l'intention des autres. Le deuxième groupe correspond aux comportements envers les objets (b_vers_objet). Il regroupe les signaux suivants : s'orienter vers un

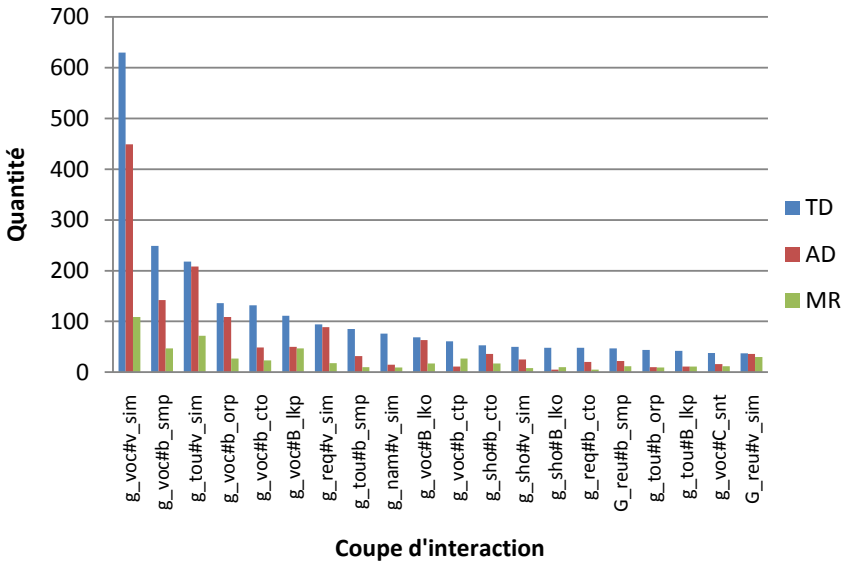


FIG. 4.5 – Les couples d'interaction les plus fréquents durant le deuxième semestre

objet, admiration d'un objet, explorer un objet, regarder un objet ou regarder autour, sourire à un objet, jouer avec un objet et chercher contact avec un objet. Le troisième groupe correspond aux comportements expressifs ($b\_sim\_expressif$) : vocalisations simples et cris. Le quatrième groupe correspond aux comportements d'exploration ($b\_sim\_exploratoire$) : s'orienter vers une personne, admiration d'une personne et explorer une personne. Le cinquième groupe correspond aux comportements réceptifs ($b\_sim\_receptif$), cela regroupe les comportements suivants : regarder une personne, sourire à une personne, jouer avec une personne et la syntonie. Le sixième groupe correspond aux comportements actifs (b_sim_actif) : chercher contact avec une personne et solliciter une personne. Du coté des parents, nous avons défini 3 groupes : les vocalisations ($G\_vocalisation$), regulation-up/down (G_reu) et Geste/montrer quelque chose (G_ges_show). Le groupe $G\_vocalisation$ correspond aux vocalisations normales et l'appel du prénom, ainsi que d'autres formes de vocalisation.

Les connexions colorées sur les diagrammes d'interaction représentent les combinaisons les plus fréquentes. Nous avons sélectionné les couples d'interactions dont la probabilité est supérieure à 1%.

Les diagrammes d'interaction montrent que la communication vocale est très importante dans l'interaction parent-enfant. D'après l'ensemble des fi-

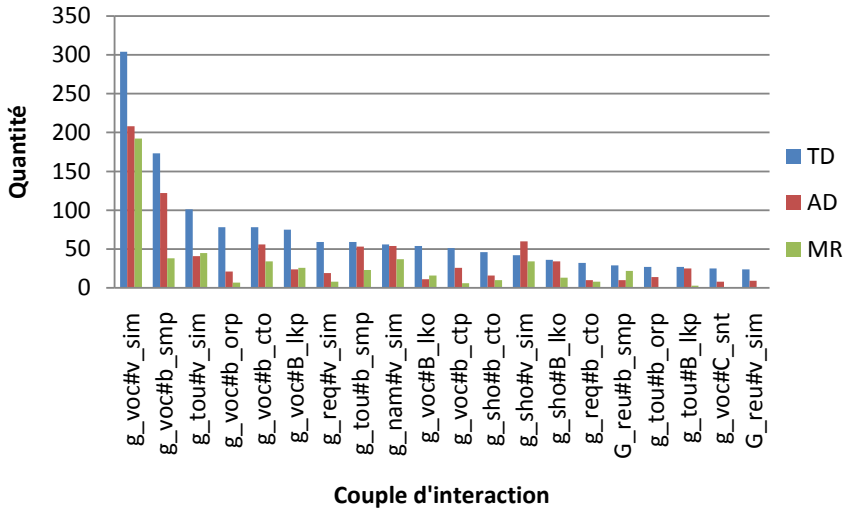


FIG. 4.6 – Les couples d'interaction les plus fréquents durant le troisième semestre

gures, nous constatons que dans les trois groupes, durant les trois semestres, la réponse vocale de l'enfant aux vocalisations de parent est le type d'interaction le plus fréquent, près de 70% des comportements de l'enfant sont en réponses à des vocalisations. Cela montre que l'interaction vocale est très importante dans les trois groupes et durant les trois semestres, ce qui aide à prédire le comportement de l'enfant après une vocalisation de la part du parent. Ainsi, afin d'étudier l'effet du *motherese* sur l'interaction, nous avons étudié l'interaction d'un enfant autiste (*Diego*), nous avons trouvé qu'à peu près 80% des vocalisations sont des *motherese* (cf. figure 4.10). Cela évolue au cours du temps, du premier jusqu'au troisième semestre. La figure 4.10 montre également qu'à peu près 72% des comportements d'enfant sont initiés par le *motherese*.

Les diagrammes d'interaction montrent également que le signal *regulation-up* (*G_reu*) apparaît seulement durant le premier semestre chez les enfants typiques alors que les parents des enfants autistes et les enfants avec retard mental continuent à utiliser cette stimulation pour agir sur l'interaction de l'enfant. Ce résultat s'explique par le fait que les parents des enfants autistes stimulent beaucoup plus leurs enfants pour essayer d'attirer leur attention et les encourager à mieux interagir. Cela confirme l'hypothèse que les enfants autistes portent moins d'attention à leur environnement et qu'ils souffrent de

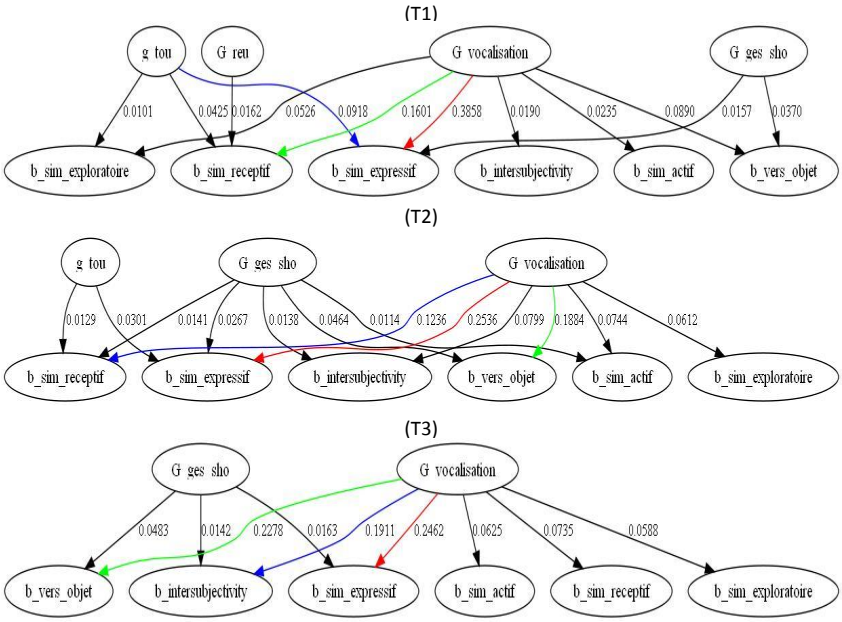


FIG. 4.7 – Modèle probabiliste des signaux d'interaction parent-enfant pour les enfants avec développement typique

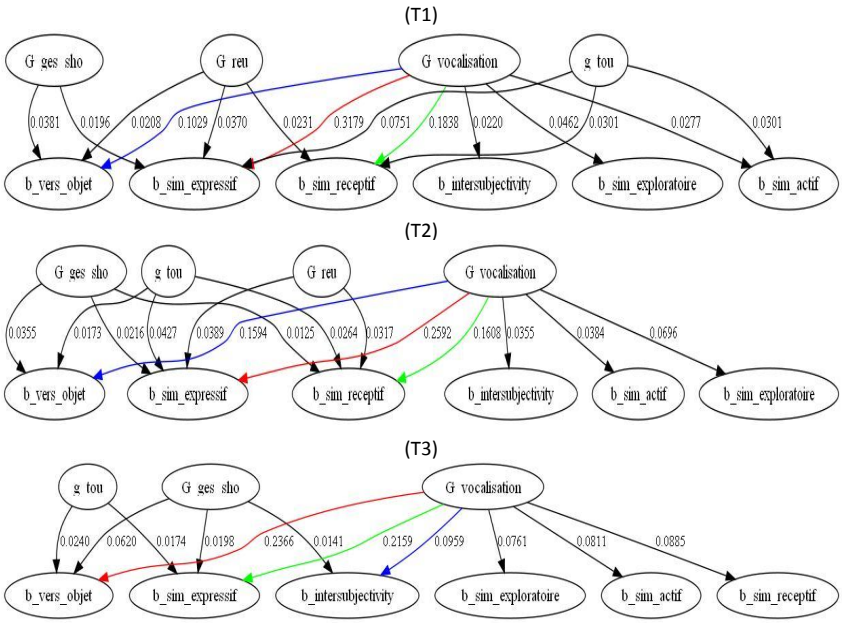


FIG. 4.8 – Modèle probabiliste des signaux d'interaction parent-enfant pour les enfants à devenir autistique

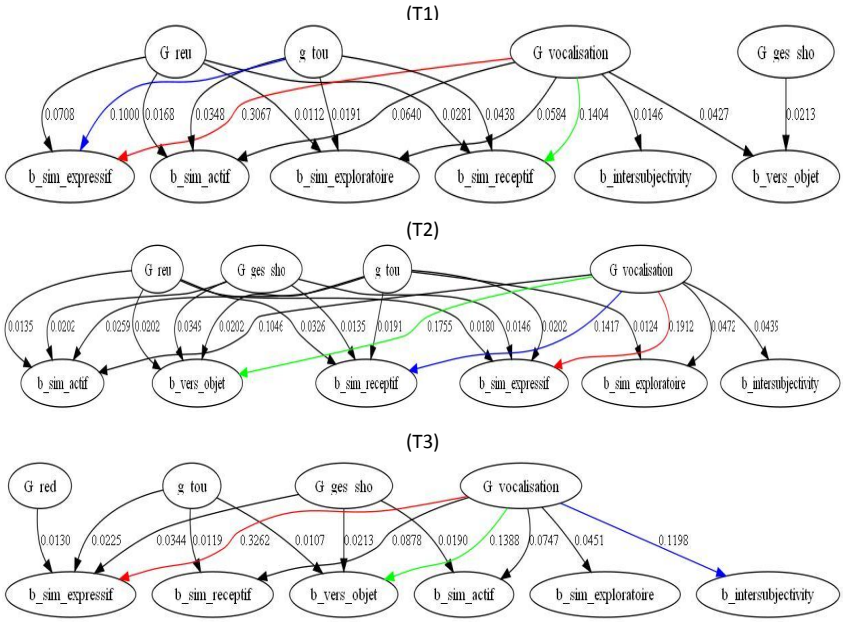


FIG. 4.9 – Modèle probabiliste des signaux d'interaction parent-enfant pour les enfants avec un retard mental

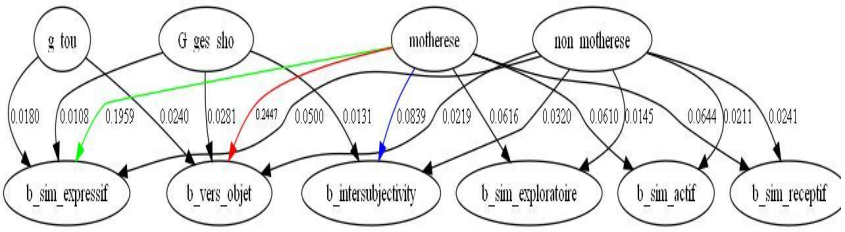


FIG. 4.10 – Modèle probabiliste des signaux d'interaction parent-enfant pour un enfant autiste

problèmes d'interaction sociale ce qui oblige leurs parents à les stimuler plus. De plus, pour les interactions initiées par des gestes ou toucher, il y a une simplification des registres chez les enfants typiques : au troisième semestre, les enfants répondent principalement à la stimulation vocale de leurs parents, alors que les parents des autistes et des retards mentaux ont aussi recours au toucher et à la communication gestuelle pour obtenir une réponse du bébé.

4.5.3 Analyse statistique de l'interaction parent-enfant

Pour comparer l'importance de chaque couple d'interaction par groupe et par semestre, en collaborant avec des cliniciens, nous avons utilisé un modèle linéaire généralisé à effet sujet aléatoire *GLMER* [Lee *et al.*, 2006]. En travaillant à l'aide du logiciel *R*, nous avons donc effectué une régression linéaire à effets aléatoires.

Nous avons vérifié la distribution des items et des regroupements à l'aide d'histogrammes. La plupart des groupes d'interactions, également les signaux de comportement des parents et des enfants (*IB* et *IG*) suivent une loi de quasi-Poisson, voir la figure 4.11 pour la distribution des signaux d'intersubjectivité. Les items les moins fréquents, ne sont pas analysables à cause de leur distribution qui ne suit aucune loi de probabilité précise. Cependant, ces signaux seront analysés dans un cadre interactif. Tous les signaux *IB* et *IG* et les groupes d'interactions satisfaisant la loi de quasi-Poisson ont été intégrés dans le modèle *GLMER*.

D'abord, nous avons effectué deux analyses uni-variées. La première analyse est effectuée avec le diagnostic comme variable explicative (elle évalue l'effet groupe) et la deuxième analyse propose le temps comme variable explicative (elle évalue l'effet semestre). Ensuite, une analyse multi-variée a été effectuée, prenant en compte simultanément l'effet semestre et l'effet groupe.

D'après ces analyses, nous avons justifié les hypothèses cliniques sur le développement de l'enfant annoncées dans le chapitre 1. Pour les enfants de type *TD*, nous constatons que les parents utilisent de plus en plus des vocalisations au cours du temps. Durant $S1$, 77% des comportements de parent sont des vocalisations, durant $S2$ le taux de vocalisation est égale à 83.1% et durant $S3$ le taux de vocalisation est égale à 91.3%. Cependant l'évolution de l'utilisation des vocalisations chez le parent des enfants normaux n'est pas significative. De plus la communication par le toucher décroît significativement au cours du temps, cela est remplacé par d'autres signaux de communication, en particulier des vocalisations. Ainsi que les comportements de 'regulation-up' et 'regulation-down' décroissent également significativement chez les parents des enfants normaux. Du coté des enfants, nous constatons que les vocalisations apparaîssent significatives durant le deuxième semestre et le troisième semestre. De plus, les comportements d'intersubjectivités croissent d'une manière significative au cours du temps. En comparant le modèle du développement des enfants typiques avec les enfants pathologiques, nous constatons que les parents des enfants autistes utilisent beaucoup de 'regulation-up', au contraire des parents des enfants typiques. La quantité d'utilisation de 'regulation-up' devient de plus en plus importante au cours du temps. D'autres comportements restent importants durant le deuxième et

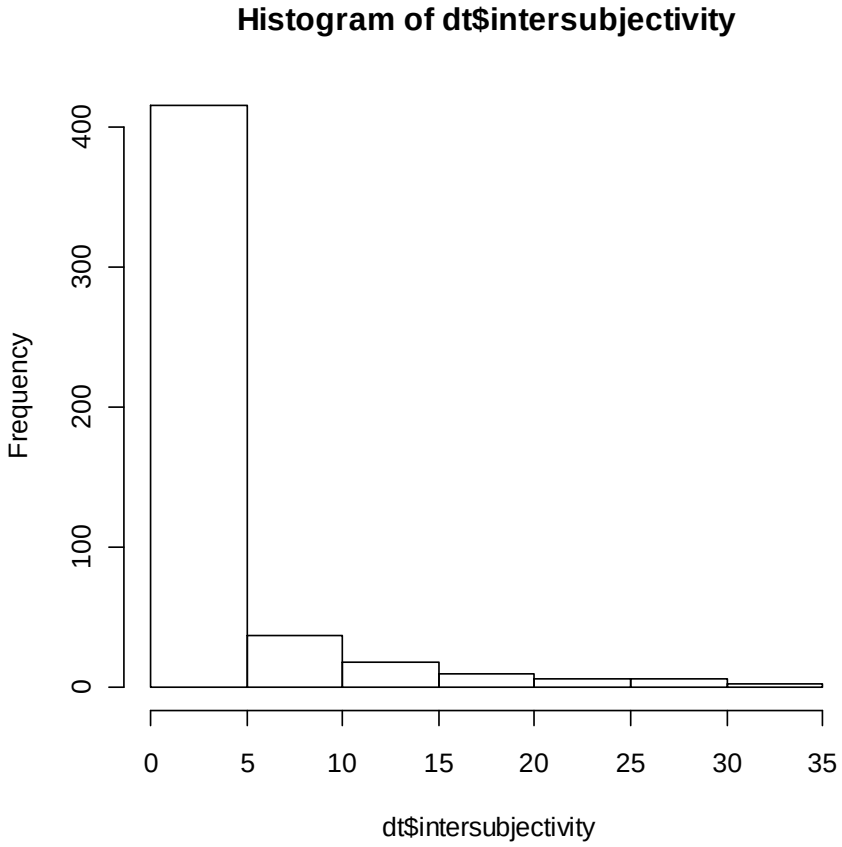


FIG. 4.11 – Factorisation en matrices non-négatives

troisième semestres comme le toucher et d'autres comportements non verbaux qui remplacent les messages verbaux.

4.6 Compréhension de l'interaction parent-enfant

Dans le but d'identifier les comportements les plus pertinents discriminants les trois groupes d'enfants, nous avons développé une technique de regroupement automatique de signaux qui permet d'extraire les différents patterns interactifs par type d'enfant. La première étape de cette technique consiste à adapter une représentation particulière des données d'interactions, cette représentation doit être consistante avec la méthode d'apprentissage utilisée pour

le regroupement. Ensuite, une méthode de classification non supervisée (*clustering*) a été utilisée pour distinguer les différents groupes des comportements.

4.6.1 Caractérisation statistique de l'interaction

Pour la représentation des signaux d'interaction, nous avons utilisé un modèle d'espace vectoriel. Dans cette représentation, chaque scène de la base d'interaction est représentée par un vecteur à N dimensions, où N correspond au nombre de caractéristiques décrivant l'ensemble des données. Dans notre étude, les caractéristiques sont les couples des comportements. Un vecteur représentant une scène contient le poids de chaque couple de comportement dans cette scène. Une valeur fréquemment utilisée pour ce poids est le *tf-idf* (*term frequency-inverse document frequency*) [Salton et Buckley, 1988]. Nous avons adapté cette méthode *tf-idf*, souvent utilisée en classification de documents et en recherche d'information, pour représenter les données d'interactions.

La méthode *tf-idf* est une fonction de pondération statistique qui permet d'évaluer l'importance d'un couple de comportement spécifique pour une scène d'interaction dans un corpus. Le poids dépend du nombre d'occurrences des comportements (*tf*) dans la scène. Il varie également en fonction de la fréquence des comportements dans le corpus (*idf*). La fréquence du terme t_i est simplement le nombre d'occurrences de ce terme dans la scène considérée. Cette somme est en général normalisée pour éviter les biais liés à la longueur de la scène. Soit la scène d_j et le comportement t_i, alors la fréquence du comportement dans la scène est calculée de la manière suivante :

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \tag{4.9}$$

où n_{ij} est le nombre d'occurrences du comportement t_i dans la scène d_j. Le dénominateur est le nombre d'occurrences de tous les termes dans la scène d_j. La fréquence inverse (inverse document frequency) est une mesure de l'importance du comportement dans l'ensemble du corpus. Dans le schéma *tf-idf*, elle vise à donner un poids plus important aux comportements les moins fréquents, considérés comme plus discriminants. Elle consiste à calculer le logarithme de l'inverse de la proportion de scènes du corpus qui contiennent le terme :

$$idf_i = log \frac{|D|}{|\{d_j : t_i \in d_j\}|} \tag{4.10}$$

Où $|D|$ représente le nombre total de scènes dans le corpus et $|\{d_j : t_i \in d_j\}|$ représente le nombre de scènes où le comportement t_i apparaît. Finalement, le poids s'obtient en multipliant les deux mesures :

$$(tf - idf)_{ij} = tf_{ij} \times idf_i \tag{4.11}$$

4.6.2 Regroupement non supervisé des signaux d'interaction

L'objectif d'un algorithme de clustering est de renvoyer une partition des données, dans laquelle les objets à l'intérieur de chaque cluster sont aussi proches que possible les uns des autres et aussi éloignés que possible des objets des autres clusters. Dans la littérature, nous distinguons plusieurs méthodes de clustering comme l'algorithme des *k-moyennes* [Forgy, 1965] et les méthodes hiérarchiques. Une technique plus récente dite de factorisation en matrices non-négatives (*NMF*) a été adaptée pour résoudre notre problème de regroupement. Ces techniques de regroupement ont été largement utilisées pour résoudre les problèmes de réduction de dimensionnalité et de structuration des documents. Ainsi, la factorisation en matrices non-négatives a été particulièrement appliquée à de très nombreux domaines : représentation d'images de visages [Lee et Seung, 1999], classification d'images [Guillamet *et al.*, 2001], surveillance de messages électroniques [Berry et Browne, 2005], extraction de caractéristiques sémantiques dans des textes [Lee et Seung, 1999], regroupement de gênes impliqués dans le cancer [Liu et Yuan, 2008], et détection de l'activité neuronale pour les interfaces cerveau-machine [Kim *et al.*, 2005].

4.6.2.1 Algorithme de *k-moyennes*

Les *k-moyennes* [Forgy, 1965] est l'algorithme de clustering le plus connu et le plus utilisé, du fait de sa simplicité de mise en oeuvre. Il consiste à partitionner les données en K clusters. Contrairement à d'autres méthodes dites hiérarchiques, qui créent une structure en "arbre de clusters" pour décrire les groupements, les *k-moyennes* ne créent qu'un seul niveau de clusters. Chaque cluster de la partition est défini par ses objets et son centroïde. Ensuite, l'algorithme minimise itérativement la somme des distances entre chaque objet et le centroïde de cluster, sachant que la position initiale des centroïdes conditionne le résultat final, de sorte que les centroïdes doivent être initialement placés le plus loin possible les uns des autres de façon à optimiser l'algorithme. Puis, l'algorithme met à jour les objets des clusters jusqu'à ce que la somme ne puisse plus diminuer. Enfin, le résultat est un ensemble de clusters compacts et clairement séparés, sous réserve qu'on ait choisi la bonne valeur K (nombre de clusters). Les principales étapes de l'algorithme des *k-moyennes* sont :

1. Choix aléatoire de la position initiale des K clusters.
2. Affecter les objets à un cluster suivant un critère de minimisation des distances (généralement selon une mesure de distance euclidienne).
3. Une fois tous les objets placés, recalculer les K centroïdes.

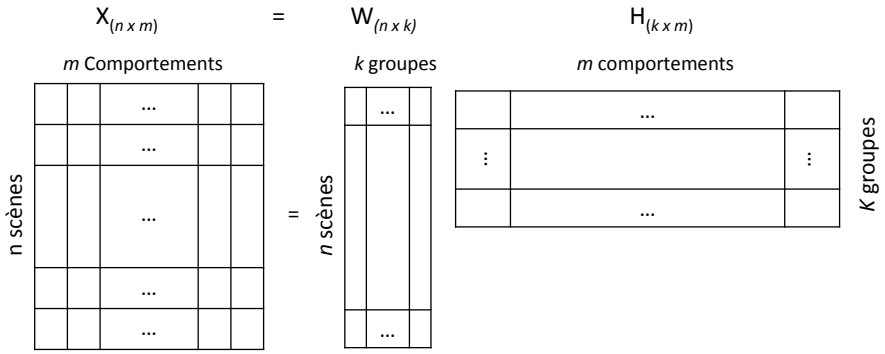


FIG. 4.12 – Factorisation en matrices non-négatives

4. Réitérer les étapes 2 et 3 jusqu'à ce que plus aucune ré-affectation ne soit faite.

4.6.2.2 Factorisation en matrices non négatives

La factorisation en matrices non négatives (NMF) [Lee et Seung, 2000] est une méthode générale de décomposition matricielle, introduite par Lee et Seung [1999]. Elle permet d'approximer toute matrice X de taille $(n \times m)$ et dont les éléments sont tous positifs, grâce à une décomposition de la forme suivante (figure 4.12) :

$$X = WH \tag{4.12}$$

où W et H sont des matrices de taille $(n \times k)$ et $(k \times m)$. La matrice X contient les vecteurs réels de dimension m, la matrice W contient les vecteurs correspondants dans un espace de dimension $k < m$, et la matrice de passage H contient les vecteurs de base. Déterminer les matrices W et H revient à minimiser la distance entre la matrice initiale X et le produit WH ; plus précisément, il faut minimiser la norme de Frobenius :

$$||X - WH||^2 = \sum_{i,j}(X_{ij} - (WH)_{ij})^2 \tag{4.13}$$

sous les contraintes de non-négativité. C'est un problème d'optimisation non trivial, que [Lee et Seung, 2000] proposent de résoudre en initialisant W et H aléatoirement, puis en alternant les mises à jour suivantes :

$$H_{ij} \leftarrow H_{ij}\frac{(W^T X)}{(W^T WH)_{ij}} \tag{4.14}$$

$$W_{ij} \leftarrow W_{ij} \frac{(XH^T)_{ij}}{(WHH^T)_{ij}} \tag{4.15}$$

Lee et Seung [2000] montrent que leur algorithme converge vers un minimum local. Sa complexité est en *O(nmkt)*, où t est le nombre maximal d'itérations.

La méthode *NMF* a été appliquée avec succès, notamment en reconnaissance des visages [Lee et Seung, 1999], en classification de documents textuels [Xu *et al.*, 2003] [Shanaz *et al.*, 2006], en recherche d'informations [Lee et Seung, 2000]. Récemment elle a été utilisée pour l'analyse de signaux sociaux par Wu *et al.* [2008] pour découvrir simultanément les membres les plus impliqués dans un réseau social en ligne (discussion en ligne, forum) et les thèmes de discussion latents dans le forum. Dans le cadre de l'analyse des interactions en ligne (*forum*), les degrés de manifestations des internautes avec les sujets des discussions et leurs relations entre eux en communiquant leurs opinions sont deux éléments nécessaires pour extraire les différents sujets de discussion et l'intérêt que donnent les internautes à ces sujets. Pour résoudre ce problème Wu *et al.* [2008] ont d'abord proposé de mesurer la fréquence de participation de chaque internaute à chaque sujet de discussion, puis de construire une matrice X de dimension $(n \times m)$ qui contient les valeurs des fréquences, avec n et m respectivement le nombre d'internautes et le nombre de sujets de discussion. Ensuite, pour détecter les k groupes de discussion les plus importants, la matrice X est décomposée en deux matrices non négatives W et H grâce à la méthode *NMF*. Après factorisation, les deux matrices résultantes indiquent l'ensemble des discussions décrivant chaque cluster c_i. Un cluster c_i correspondant à la i^{eme} discussion est obtenu simplement grâce à la formule suivante :

$$c_i = \arg\max_j W_{ij} \tag{4.16}$$

où j est l'indice du sujet de discussion latent.

4.7 Expérimentations

4.7.1 Mesure de la performance des résultats de clustering

En apprentissage non supervisé comme en apprentissage supervisé, l'évaluation des résultats d'une méthode donnée, de même que la comparaison de différentes méthodes, est une problématique importante. Mais si la validation croisée est une méthodologie reconnue pour l'évaluation en classification supervisée, l'évaluation de la pertinence des groupes formés en classification

non supervisée reste un problème ouvert. La difficulté vient principalement du fait que l'évaluation des résultats des algorithmes de clustering est subjective par nature car il existe souvent différents regroupements pertinents possibles pour un même jeu de données. Cependant, il existe quatre méthodes principales pour mesurer la qualité des résultats d'algorithmes de clustering, mais chacune souffre de différentes limitations :

1. Utiliser des données artificielles pour lesquelles le regroupement attendu est connu. Mais les méthodes sont alors évaluées uniquement sur les distributions spécifiques correspondantes, et les résultats sur des données artificielles ne peuvent pas être généralisés aux données réelles.
2. Utiliser des jeux de données étiquetées et observer si la méthode retrouve les classes initiales. Mais les classes d'un problème supervisé ne correspondent pas nécessairement aux groupes qu'il est pertinent d'identifier pour une méthode non supervisée, d'autres regroupements pouvant être également pertinents.
3. Utiliser des critères numériques, tels que l'inertie intra-cluster et/ou la séparation inter-clusters. Mais de tels critères ont une part importante de subjectivité car ils utilisent des notions prédéfinies de ce qu'est un bon clustering. Par exemple, la séparation des clusters n'est pas toujours un bon critère à utiliser car parfois, des clusters qui se chevauchent peuvent être plus pertinents.
4. Utiliser un expert pour évaluer le sens d'un clustering donné dans un champ d'application spécifique. Mais s'il est possible à un expert de dire si un regroupement donné a du sens, il est beaucoup plus problématique de quantifier son intérêt ou de dire si un regroupement est meilleur qu'un autre. De plus, l'intérêt de la méthode ne peut être généralisé à différents types de jeux de données.

Avec nos types de données il est très difficile d'avoir une classification de référence pour évaluer la performance des algorithmes. Par conséquent, nous calculons les indices de qualité basés sur des calculs de distance dont les plus connus sont l'inertie inter-classes et intra-classes (*Homogeneity-Separation*) [Lebart *et al.*, 1982] :

– L'inertie Intra-classes permet de mesurer le degré d'homogénéité entre les données associées à une classe. Elle calcule leurs distances par rapport au point représentant le profil de la classe.

$$Intra = \frac{1}{n} \sum_{C \in P} \frac{1}{2n} \sum_{i \in C} \sum_{j \in C} d(i,j)^2 \tag{4.17}$$

– L'inertie Inter-classes mesure le degré d'hétérogénéité entre les classes. Elle calcule les distances entre les points représentant les profils des

TAB. 4.9 – Nombre des clusters obtenus par groupe d'enfant

	Méthode	*TD*	*AD*	*MR*
*S*1	*NMF*	11	12	5
	K-moyennes	12	8	7
*S*2	*NMF*	14	8	11
	K-moyennes	14	11	11
*S*3	*NMF*	9	10	14
	K-moyennes	12	12	10

différentes classes de la partition.

$$Inter = \frac{1}{n} \sum_{C \in P} n_C d^2(c, c_G) \tag{4.18}$$

où c est le centre de la classe C et c_G est le centre du nuage de points matérialisant les données.

4.7.2 Résultats de clustering

Nous avons utilisé la méthode de factorisation en matrices non-négatives *NMF* pour regrouper les signaux d'interaction. De plus, pour étudier l'évolution des trois groupes d'enfant durant les trois semestres, nous calculons l'information mutuelle normalisée (*NMI*) proposée par Strehl et Ghosh [2002]. Nous considérons le modèle d'enfants typiques comme un modèle de référence. L'information mutuelle normalisée entre deux résultats de clustering différents $\widehat{y}^1$ et $\widehat{y}^2$ permet de mesurer la corrélation entre deux regroupements.

$$NMI(\widehat{y}^1, \widehat{y}^2) = \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} n_{i,j}^{1,2} log\left(\frac{n \times n_{i,j}^{1,2}}{n_i^1 \times n_j^2}\right)}{\sqrt{\left(\sum_{i=1}^{k} n_i^1 log\frac{n_i^1}{n}\right)\left(\sum_{j=1}^{k} n_j^2 log\frac{n_j^2}{n}\right)}} \tag{4.19}$$

où n_i^1 est le nombre de comportements catégorisés dans le cluster c_i en utilisant le clustering $\widehat{y}^1$, n_j^2 est le nombre de comportements catégorisés dans le cluster c_j en utilisant le clustering $\widehat{y}^2$ et $n_{i,j}^{1,2}$ est le nombre de comportements catégorisé dans le cluster c_i en utilisant le clustering $\widehat{y}^1$ et dans le cluster c_j en utilisant le clustering $\widehat{y}^2$.

Afin d'évaluer les résultats de la méthode de clustering proposée, nous avons testé en parallèle la méthode des *k-moyennes*. Le tableau 4.9 montre les meilleures solutions de regroupement en se basant sur les mesures de l'homogénéité et la séparation entre les groupes. Ainsi le tableau 4.9 montre que

TAB. 4.10 – Les valeurs d'information mutuelle normalisée entre les groupes TD/AD, TD/MR et AD/MR

	Méthode	TD/AD	TD/MR	AD/MR
$S1$	*NMF*	0.482	0.491	0.473
	K-moyennes	0.343	0.301	0.485
$S2$	*NMF*	0.438	0.520	0.506
	K-moyennes	0.240	0.262	0.399
$S3$	*NMF*	0.372	0.456	0.465
	K-moyennes	0.273	0.219	0.416

le nombre des clusters croît au cours du temps (5-7 clusters au premier semestre vs. 14-10 clusters au troisième semestre). Ce résultat s'explique par la pauvreté des situations d'interactions (peu de diversité) au premier semestre.

Le tableau 4.10 présente les valeurs de l'information mutuelle normalisée entre les enfants typiques et les enfants autistes, les enfants typiques et les enfants avec retard mental, les enfants autistes et les enfants avec retard mental. Durant le premier semestre, nous constatons que la similarité, estimée à l'aide de l'information mutuelle normalisée, entre les enfants typiques et les enfants autistes, et la similarité entre les enfants typiques et les enfants avec retard mental sont quasi-égales pour les deux méthodes de clustering. Cependant, la méthode *NMF* donne de meilleurs résultats que l'algorithme des *k-moyennes*. Nous avons obtenu une valeur de similarité égale à 0.482 entre les groupes TD/AD durant le premier semestre et 0.438 durant le deuxième semestre, contre une valeur égale à 0.343 obtenue avec la méthode des *k-moyennes* durant le premier semestre et 0.240 durant le deuxième semestre. Les résultats obtenus par la méthode *NMF* sont en accord avec les observations cliniques du développement des enfants avec ou sans pathologie.

Avec la méthode *NMF*, nous constatons que les groupes TD et MR sont plus corrélés que les groupes TD et AD durant les deuxième et troisième semestres. Le retard mental est généralement considéré comme un "simple retard" dans le développement. La méthodologie proposée ici permet de comparer les semestres entre eux. Nous avons ainsi comparé le semestre 2 du groupe TD et le semestre 3 du groupe MR. L'information mutuelle de 0.522 indique une plus forte ressemblance entre ces deux semestres (comparé à $NMI(TD_{S3}, MR_{S3})$. La similarité entre les groupes TD et AD décroît au cours du temps. Le développement déviant (et non retardé) du groupe AD explique en partie ce résultat. Le caractère déviant de développement est la principale différence entre les enfants avec retard mental et les enfants autistes [Mays et Gillon, 1993].

Plusieurs études basées sur l'analyse des films familiaux [Baranek, 1999] ont montré que durant la première année de vie, les enfants avec retard mental et les enfants autistes sont caractérisés par un handicap d'interaction social par rapport aux enfants avec développement typique. Ce handicap se manifeste par un manque de réponse à l'appel du prénom, tendance à s'intéresser moins aux personnes, etc. Ces signes précoces de l'autisme deviennent plus évidents chez les enfants autistes. Cette hypothèse est confirmée par la méthode *NMF*. Dans le tableau 4.10, nous constatons que durant le premier et le deuxième semestres, les groupes AD et MR sont quasi-similaires ($S1$: NMI=0.473 et $S2$: NMI=0.506). De plus, en analysant les contenus des clusters obtenus par la méthode *NMF*, nous constatons que la plupart des couples des comportements initiés par un signal de regulation-down (G_reu) sont souvent groupés ensemble chez les enfants autistes. Ce résultat est justifié par le manque d'interaction des enfants autistes, ce que oblige le parent à communiquer avec des signaux de 'regulation-up' afin d'encourager l'enfant à interagir plus et à attirer son attention.

Le tableau 4.11 montre le résultat du regroupement des signaux d'interaction des enfants autistes durant le troisième semestre, en utilisant la méthode de clustering NMF. Nous constatons que la plupart des signaux de régulation sont regroupés dans un même cluster (cluster 2). Ainsi dans le cluster 6, toutes les dyades d'interaction sont initiées par des vocalisations de la part de parent. Dans le cluster 7, nous trouvons toutes les dyades d'interaction où l'enfant répond par des comportements d'intersubjectivité.

Pour étudier le regroupement en tenant en compte du *motherese*, nous avons étudié un cas d'enfant autiste (*Diego*) (cf. tableau 4.12). Le tableau 4.12 montre que le signal du *motherese* est souvent regroupé avec les signaux de 'regulation-up'. Ce résultat s'explique par le rôle commun de régulation d'interaction qui joue les signaux du *motherese* et de 'régulation-up'.

4.8 Conclusion

Ce chapitre traite le problème de l'analyse et la compréhension de l'interaction humaine, en particulier l'interaction parent-enfant. Dans un premier temps, nous avons exposé quelques travaux sur l'analyse de comportements et d'interaction humains et différentes formes de représentation de données d'interaction. Ensuite, nous avons passé en revue quelques méthodes portant sur l'analyse de l'interaction parent-enfant en se basant sur l'étude de films familiaux. Nous avons étudié une base de données longitudinale d'interactions parent-enfant. Cette base de données a été soigneusement annotée par des experts du développement de l'enfant.

TAB. 4.11 – Résultat de regroupement des signaux d'interaction (enfant autiste durant le troisième semestre)

Cluster 1	'g_ges#B_exo', 'g_ges#b_cto'
Cluster 2	'G_reu#B_lkp', 'G_reu#B_exo', 'G_reu#B_lka', 'G_reu#b_cto' 'G_reu#b_orp', 'G_reu#v_sim'
Cluster 3	'g_nam#b_cto', 'g_voc#B_lko'
Cluster 4	'g_ges#b_ctp', 'g_ges#b_smp', 'g_ges#v_sim', 'g_nam#c_imi', 'g_req#b_ctp', 'g_req#b_orn', 'g_req#b_oro', 'g_sho#b_oro', 'g_tou#b_smp', 'g_tou#c_acc', 'g_tou#v_sim', 'g_voc#b_smp', 'g_voc#c_off'
Cluster 5	'g_ges#B_lkp', 'g_ges#b_orp', 'g_ges#c_sol', 'g_nam#B_lko', 'g_req#B_lko', 'g_req#b_orp', 'g_sho#B_lkp', 'g_sho#b_orp', 'g_tou#B_exo', 'g_tou#B_lka', 'g_tou#B_lkp', 'g_tou#b_orp'
Cluster 6	'g_nam#b_ctp', 'g_nam#b_orp', 'g_nam#b_smp', 'g_nam#v_sim', 'g_req#v_sim', 'g_tou#b_ctp', 'g_voc#B_exo', 'g_voc#B_lka', 'g_voc#b_ctp', 'g_voc#c_sol'
Cluster 7	'g_ges#c_acc', 'g_ges#c_com', 'g_nam#B_lkp', 'g_req#c_acc', 'g_req#c_com', 'g_voc#b_gfp', 'g_voc#c_acc', 'g_voc#c_com'
Cluster 8	'G_reu#b_ctp', 'G_reu#b_smp', 'G_reu#c_acc', 'g_ges#c_poi', 'g_nam#b_orn', 'g_nam#c_acc', 'g_req#B_lkp', 'g_req#b_smp', 'g_sho#B_lko', 'g_sho#b_smp', 'g_tou#b_cto', 'g_voc#b_orn', 'g_voc#b_orp', 'g_voc#v_mea'
Cluster 9	'g_ges#b_gfo', 'g_ges#c_imi', 'g_ges#v_mea', 'g_nam#b_gfo', 'g_req#B_exo', 'g_req#b_cto', 'g_req#b_gfo', 'g_req#c_imi', 'g_req#c_poi', 'g_req#v_mea', 'g_sho#b_gfo', 'g_tou#B_lko', 'g_voc#B_lkp', 'g_voc#C_seg', 'g_voc#b_cto', 'g_voc#b_gfo', 'g_voc#b_oro', 'g_voc#c_imi', 'g_voc#c_poi'
Cluster 10	'g_nam#c_sol', 'g_sho#B_exo', 'g_sho#b_cto', 'g_sho#b_smo', 'g_sho#c_acc', 'g_sho#v_sim', 'g_tou#c_sol', 'g_voc#C_ejo', 'g_voc#b_smo'

Pour étudier l'évolution des signaux d'interactions les plus importants (significatifs) pour chaque type d'enfant au cours du temps, nous avons effectué des analyses statistiques qui ont confirmées les résultats obtenus par les cliniciens, mais cette fois-ci dans le cadre d'interaction parent-enfant. Ensuite, afin d'étudier l'interaction parent-enfant, ainsi que pour identifier les signes caractéristiques du développemental de l'enfant, nous avons adapté une méthode de regroupement non supervisée de signaux d'interaction. Cette méthode repose sur la factorisation de matrices non négatives *NMF*. La méthode de classification est couplée avec un système de codage, *tf-idf*, souvent utilisé dans le domaine du regroupement de document. C'est une forme de représentation statistique qui prend en compte à la fois l'importance d'un signal spécifique pour une scène d'interaction dans le corpus et la fréquence générale des signaux dans le corpus. Les résultats obtenus par cette méthode sont en accord avec les observations cliniques.

TAB. 4.12 – Résultat de regroupement des signaux d'interaction en tenant en compte le signal du *motherese* (cas d'un enfant autiste)

Cluster 1	'G_reu#B_lkp', 'G_reu#B_exo', 'G_reu#B_lka', 'G_reu#b_cto', 'G_reu#b_orp', 'G_reu#v_sim', 'g_motherese#b_cto', g_motherese#B_lko', 'g_motherese#b_ctp', 'g_motherese#b_orp', 'g_motherese#b_smp', 'g_motherese#c_imi', 'g_motherese#v_mea', 'g_motherese#c_sol', 'g_motherese#c_com'
Cluster 2	'g_ges#b_ctp', 'g_ges#b_smp', 'g_ges#v_sim', 'g_motherese#c_imi', 'g_non_motherese#b_ctp', 'g_req#b_orn', 'g_sho#b_oro','g_tou#v_sim', 'g_tou#b_smp', 'g_tou#c_acc', 'g_motherese#c_off', 'g_ges#b_cto', 'g_ges#B_exo'
Cluster 3	'g_ges#B_lkp', 'g_ges#b_orp', 'g_ges#c_sol', 'g_non_motherese#B_lko', 'g_req#B_lko', 'g_non_motherese#b_orp', 'g_sho#B_lkp', 'g_sho#b_orp', 'g_tou#B_exo', 'g_tou#B_lka', 'g_tou#B_lkp', 'g_tou#b_orp'
Cluster 4	'g_motherese#v_sim', 'g_motherese#c_sol', 'g_motherese#v_sim', 'g_tou#b_ctp', 'g_motherese#B_exo', 'g_motherese#B_lka',
Cluster 5	'g_ges#c_acc', 'g_ges#c_com', 'g_motherese#B_lkp', 'g_motherese#c_acc', 'g_motherese#c_com', 'g_motherese#b_gfp', 'g_motherese#c_acc'
Cluster 6	'G_reu#b_smp', 'G_reu#c_acc', 'g_ges#c_poi', 'g_motherese#b_orn', 'g_motherese#b_orn', 'g_motherese#c_acc', 'g_motherese#B_lkp', 'g_sho#B_lko', 'g_sho#b_smp', 'g_non_motherese#b_cto', 'g_non_motherese#b_orp', 'g_non_motherese#b_smp'
Cluster 7	'g_ges#b_gfo', 'g_ges#c_imi', 'g_ges#v_mea', 'g_motherese#b_gfo', 'g_motherese#B_exo', 'g_motherese#b_gfo', 'g_motherese#c_imi', 'g_motherese#c_poi', 'g_motherese#v_mea', 'g_sho#b_gfo', 'g_tou#B_lko', 'g_motherese#B_lkp', 'g_motherese#C_seg', 'g_non_motherese#b_cto', 'g_motherese#b_oro', 'g_motherese#c_poi','g_non_motherese#b_gfo',
Cluster 8	'g_sho#B_exo', 'g_sho#b_cto', 'g_sho#b_smo', 'g_non_motherese#c_acc' 'g_sho#c_acc', 'g_sho#v_sim', 'g_tou#c_sol', 'g_non_motherese#C_ejo', 'g_non_motherese#b_smo', 'g_non_motherese#B_lkp', 'g_non_motherese#B_lka', 'g_non_motherese#b_gfp, 'g_non_motherese#B_exo', 'g_non_motherese#c_imi'

Conclusion et perspectives

Conclusions générales

Dans cette thèse, nous avons abordé le problème de l'analyse de signaux sociaux et en particulier l'interaction parent-enfant. Dans le chapitre 1, nous avons défini le concept de communication et d'interaction sociales. Nous avons également exploré le domaine de l'analyse de signaux sociaux et leurs interprétations dans le cadre de l'interaction face à face. Ensuite, nous avons étudié la communication du point de vue développemental, en nous focalisant sur les capacités d'interaction des enfants. Cette étude a permis d'identifier le rôle potentiel du *motherese* dans l'interaction. Le *motherese* est le support d'une émotion sociale non prototypique, qui semble jouer un rôle important dans l'interaction parent-enfant. Nous étudions ainsi l'impact de cette forme d'émotion et ses influences sur les comportements de l'enfant, dans le cadre des interactions réelles extraites à partir des films familiaux. Les films familiaux sont des collections de scènes d'interaction parent-enfant, enregistrées par des parents dans des contextes réels (repas, bain, anniversaire, etc.). Cependant, cette analyse requiert l'annotation de la base d'interaction. Par conséquent, nous avons développé un système de détection automatique du *motherese*. Nous avons développé un système de classification supervisée. Cela passe par trois étapes : acquisition et annotation des données, extraction de caractéristiques et classification. Nous avons également proposé des méthodes de classification plus avancées.

En théorie, le *motherese* est défini par sa prosodie : un pitch augmenté, un tempo plus lent et des contours d'intonation exagérés (comparaison avec la "parole normale"). Dans le chapitre 2, au contraire de la définition théorique, nous avons montré que la définition expérimentale du *motherese* se base essentiellement sur des caractéristiques spectrales. Les meilleures performances de classification sont obtenues en utilisant des descripteurs cepstraux *MFCC*. Malgré les bonnes performances obtenues, le système proposé présente beaucoup de limites. Il est complètement supervisé, il nécessite beaucoup de données d'apprentissage pour atteindre une qualité de classification satisfaisante. Aussi, ce type d'apprentissage (supervisé) est assez sensible à la diversité des données. Nous nous sommes alors tourné vers une approche semi-supervisée qui nécessite beaucoup moins de données d'apprentissage.

Dans le chapitre 3, nous avons proposé un algorithme de classification semi-supervisée de type co-apprentissage automatique. Cet algorithme consiste à combiner des exemples étiquetés et non étiquetés pour l'apprentissage. Le prin-

cipal avantage de cette méthode est qu'elle permet de renforcer les foctions de catégorisation. Elle permet également d'offrir une définition expérimentale plus générale. De plus, cette méthode permet d'utiliser différentes caractérisations en parallèle. L'algorithme proposé est une extension de la méthode standard de co-apprentissage automatique proposée par [Blum et Mitchell, 1998]. Il considère un ensemble de classifieurs et de descripteurs qu'on appelle des *vues* qui servent chacune à modéliser les segments de parole. Le but de l'algorithme est de combiner les différentes *vues* afin d'obtenir une prédiction unique par exemple du test. Ainsi, la contribution de chaque *vue* dans la décision finale est définie par son degré de confiance dynamique, estimée à chaque itération. Pour valider cette méthode, nous nous comparons avec des méthodes état de l'art d'apprentissage semi-supervisée, ainsi que la méthode supervisée. Notre méthode donne de bonnes performances, en particulier dans le cas où très peu de données d'apprentissage sont disponibles.

Afin de justifier l'hypothèse clinique sur l'importance du *motherese* dans le développement des enfants et en particulier les enfants autistes, nous avons proposé, dans le chapitre 4, une modélisation statistique de l'interaction. Elle permet d'identifier le rôle du *motherese* dans l'interaction, ainsi que les signaux pertinent pour l'interaction suivant le type de développement d'enfant, nous étudions trois types d'enfants : enfant avec développent typique, enfants autistes et enfants avec retard mental. De plus, dans cette partie nous avons étudié l'interdépendance des signaux de communication par la construction d'un modèle statistique. Enfin, pour identifier les groupes de comportements les plus pertinents discriminants les trois groupes d'enfants, nous avons développé une technique de regroupement automatique de signaux qui permet d'extraire les groupes d'interactions par type d'enfant. Pour cela, nous avons adapté une méthode de représentation de données *tf-idf*, souvent utilisée pour la représentation de documents. Ensuite, nous avons utilisé une méthode de regroupement non supervisée, cette méthode est basée sur la factorisation de matrices non négatives. La technique de regroupement proposée a été validée expérimentalement.

Perspectives

Plusieurs perspectives nous paraissent intéressantes à explorer. Une des premières perspectives concerne l'amélioration de la caractérisation du *motherese* pouvant s'appuyer sur la proposition de nouvelles caractéristiques acoustiques du signal. La segmentation du *motherese* reste ainsi un problème majeur du fait de la qualité diverse des films familiaux. Etudier la complémentarité des différents descripteurs permet d'avoir une meilleure fusion de caractéris-

tiques et de bénéficier des avantages de chacun. Afin d'améliorer le système de reconnaissance, l'utilisation d'un module de sélection de caractéristiques peut être également envisagée.

Le système de classification semi-supervisée peut être amélioré en tenant compte de la complémentarité des différents classifieurs et descripteurs. Nous proposons de développer une méthode non supervisée pour estimer l'erreur de classification à chaque itération. Le système proposé peut être considéré comme un apprentissage actif permettant de s'adapter à l'évolution des données.

Pour la reconnaissance du *motherese*, nous avons exploité uniquement le flux audio. Il est intéressant de définir cette émotion à partir des expressions faciales ou des gestes. La multimodalité joue un rôle important dans la communication. Nous pourrions ainsi étudier le *motionese* (modification de la gestuelle).

Le modèle d'interaction sociale que nous proposons est suffisant pour l'extraction de signaux pertinents d'interaction et également pour l'étude de l'interdépendance entre les signaux sociaux. Une approche complémentaire consisterait à développer un système de prédiction des réponses de l'enfant. Au-delà de l'intérêt pour la compréhension de l'interaction, il s'agirait d'exploiter des nouvelles techniques d'apprentissage et de classification (Les champs aléatoires conditionnels). Les champs aléatoires conditionnels (ou *Conditional Rondom Fields* : *CRF*) se situent dans un cadre probabiliste et sont basés sur une approche conditionnelle pour étiqueter les séquences de données. Ces modèle sont mieux adaptés pour la prédiction des comportements. Ils considèrent la probabilité conditionnelle $p(x|y)$.

Enfin, l'analyse de signaux sociaux reste un domaine émergent dans l'analyse des comportements humains. Ce domaine permet de transmettre, à la machine ou au robot, les compétences humaines en interaction sociale. La plupart des systèmes d'interactions existants s'appuient essentiellement sur des modèles à base des règles. Il est intéressant de développer des méthodes d'apprentissage actif et dynamique de comportement, en tenant compte de l'historique de l'interaction et des différents états affectifs en faisant appel à des compétences diverses : informatique, traitement de signal, sociologie, psychologie, médecine, etc.

Publications

Articles dans des revues internationales avec comité de lecture

[1] **Mahdhaoui, A.** and Chetouani, M. (en révision). Supervised and semi-supervised infant-directed speech classification for parent-infant interaction analysis. Speech Communication Journal.

[2] **Mahdhaoui, A.** and Chetouani, M. and Cassel, R.S. and Saint-Georges, C. and Parlato, E. and Laznik, M.C. and Apicella, F. and Muratori, F. and Maestro, S. and Cohen, D. (2010). Computerized home video detection for motherese may help to study impaired interaction between infants who become autistic and their parents. International Journal of Methods in Psychiatric Research.

[3] Chetouani, M. and **Mahdhaoui, A.** and Ringeval, F. (2009). Time-Scale Feature Extractions for Emotional Speech Characterization. Cognitive Computation, Springer, publisher. Vol 1 No 2 Pages 194-201.

Chapitres d'ouvrages scientifiques

[4] **Mahdhaoui, A.** and Chetouani, M. and Zong, C. and Cassel, R.S. and Saint-Georges, C. and Laznik, M-C. and Maestro, S. and Apicella, F. and Muratori, F. and Cohen, D. (2009). Automatic Motherese Detection for Face-to-Face Interaction Analysis. Multimodal Signals : Cognitive and Algorithmic Issues, Springer Verlag, publisher. Vol LNAI 5398 Pages 248-255.

Communications internationales avec actes

[5] **Mahdhaoui, A.** and Chetouani, M. (2010). Emotional Speech Classification Based On Multi View Characterization. IAPR International Conference on Pattern Recognition, ICPR 2010.

[6] AL Moubayed, S. and Baklouti, M. and Chetouani, M. and Dutoit, T. and **Mahdhaoui, A.** and Martin, J-C and Ondas, S. and Pelachaud, C. and Urbain, J. and Yilmaz, M. (2009). Generating Robot/Agent Backchannels

During a Storytelling Experiment. IEEE International Conference on Robotics and Automation (ICRA'09), Japan, 2009..

[7] **Mahdhaoui, A.** and Chetouani, M. and Kessous, L. (2009). Time-Frequency Features Extraction for Infant Directed Speech Discrimination. ISCA Tutorial and Research Workshop on Non-Linear Speech Processing (NOLISP09) .

[8] **Mahdhaoui, A.** and Chetouani, M. (2009). A new approach for motherese detection using a semi-supervised algorithm. IEEE Workshop on Machine Learning for Signal Processing (MLSP'09).

[9] **Mahdhaoui, A.** and Chetouani, M. and Cassel, R.S. and Saint-Georges, C. and Laznik, M-C. and Apicella, F. and Muratori, F. and Maestro, S. and Cohen, D. (2009). Home video segmentation for motherese may help to detect impaired interaction between infants . Innovative Research In Autism (IRIA2009).

[10] **Mahdhaoui, A.** and Ringeval, .F and Chetouani, M. (2009). Emotional Speech Characterization Based on Multi-Features Fusion for Face-to-Face Interaction . International Conference on Signals, Circuits and Systems (SCS09).

[11] **Mahdhaoui, A.** and Chetouani, M. (2008). Automatic motherese detection for Parent-Infant Interaction. Speech and face to face communication, workshop dedicated to the memory of Christian Benoit. Grenoble, France.

[12] **Mahdhaoui, A.** and Chetouani, M. and Zong, C. (2008). Motherese Detection Based On Segmental and Supra-Segmental Features. IAPR International Conference on Pattern Recognition, ICPR 2008. Tampa Florida, USA.

[13] AL Moubayed, S. and Baklouti, M. and Chetouani, M. and Dutoit, T. and **Mahdhaoui, A.** and Martin, J-C and Ondas, S. and Pelachaud, C. and Urbain, J. and Yilmaz, M. (2008). Multimodal Feedback from Robots and Agents in a Storytelling Experiment. eINTERFACE'08 Proceedings of the 4th International Summer on Multi-Modal Interfaces, August 4-29,2008, Paris-Orsay, France. Pages 43-55.

Bibliographie

P. Aarabi, D. Hughes, K. Mohajer et M. Emami : The automatic measurement of facial beauty. *In Proceedings of IEEE International Conference on Systems*, pages 2644–2647, 2001. 29

J-C. Abric : *Psychologie de la communication : théories et méthodes*. Armand Colin, 1996. 13

JL. Adrien, P. Lenoir, J. Martineau, A. Perrot, L. Hameury, C. Larmande et D. Sauvage : Blind ratings of early symptoms of autism based upon family home movies. *Journal of the American Academy of Child and Adolescent Psychiatry*, 32:617–626, 1993. 45

S. AL Moubayed, M. Baklouti, M. Chetouani, T. Dutoit, A. Mahdhaoui, J-C Martin, S. Ondas, C. Pelachaud, J. Urbain et M. Yilmaz : Generating robot/agent backchannels during a storytelling experiment. *In IEEE International Conference on Robotics and Automation (ICRA)*, 2009. 3

N. Ambady et R. Rosenthal : Thin slices of expressive behavior as predictors of interpersonal consequences : a meta-analysis. *Psychological Bulletin*, 111(2):256–274, 1992. 19

J. Ang, R. Dhillon, A. Krupski, E. Shriberg et A. Stolcke : Prosody-based automatic detection of annoyance and frustration in human-computer dialog. *In International Conference on Spoken Language Processing*, pages 2037–2040, 2002. 51

O. Aran et D. Gatica-Perez : Fusing audio-visual nonverbal cues to detect dominant people in small group conversation. *In International conference on pattern recognition*, 2010. 3, 116, 117, 118

M. Argyle : *The Psychology of Interpersonal Behaviour*. Penguin, 1967. 2

M. Argyle : *Bodily communication*. Methuen, London, 1975. 30

K. Bailly et M. Milgram : Boosting feature selection for neural network based regression. *Neural Networks*, 22:748–756, 2009. 27

A.I. Banerjee, S. Rudnicky : Using simple speech based features to detect the state of a meeting and the roles of the meeting participants. *In Proceedings of International Conference on Spoken Language Processing*, pages 2189–2192, 2004. 33

G. Baranek : Autism during infancy : a retrospective video analysis of sensory-motor and social behaviors at 9-12 months of age. *Autism and Developmental Disorders*, 29:213–224, 1999. 146

R. Barzilay, M. Collins, J. Hirschberg et S. Whittaker : The rules behind the roles : identifying speaker roles in radio broadcasts. *In Proceedings of American Association of Artificial Intelligence Symposium*, pages 679–684, 2000. 33

E. Bates : *The emergence of symbols : cognition and communication in infancy*. Academic press, New York, 1979. 36

B. Beebe et F.M. Lachmann : The contribution of mother-infant mutual influence to the origins of self- and object-representations. *Psychoanalytic Psychology*, 5:305–337, 1988. 42

B. Beebe, D. Stern et J. Jaffe : *Of speech and time : Temporal speech patterns in interpersonal contexts*, chapitre The kinesic rhythm of mother-infant interactions, pages 23–34. Hillsdale, NJ : Lawrence Erlbaum., 1979. 41

C. Ben Abdelkader et Y. Yacoob : *Computational Forensics*, chapitre Statistical estimation of human anthropometry from a single uncalibrated image. Springer-Verlag, Berlin, 2009. 29

M.W. Berry et M. Browne : Email surveillance using non-negative matrix factorization . *Computational & Mathematical Organization Theory*, 11: 249–264, 2005. 140

C. Bishop : *Pattern Recognition and Machine Learning*. Springer Verlag, 2006. 34

C.M. Bishop : *Neural Networks for Pattern Recognition*. OXFORD University Press, 1995. 65

M.J. Black et Y. Yacoob : Recognizing facial expressions in image sequences using local parametrized models of image motion. *International Journal of Computer Vision*, 25(1):23–48, 1997. 31

A. Blum et T. Mitchell : Combining labeled and unlabeled data with co-training. *In COLT : Proceedings of the Workshop on Computational Learning Theory*, pages 92–100, 1998. 7, 88, 95, 97, 99, 112, 150

P. Boersma et D. Weenink : Praat, a system for doing phonetics by computer. *Glot international*, 5(9):341–345, 2001. 31, 58

K. Bousmalis, M. Mehu et M. Pantic : Spotting agreement and disagreement : A survey of nonverbal audiovisual cues and tools. *In Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, 2009. 118

T.G.R. Bower : *Le développement psychologique de la première enfance*. Margada, San Francisco, 1997. 37

E. Boyle, A.H. Anderson et A. Newlands : The effects of visibility on dialogue in a cooperative problem solving task. *Language and Speech*, 37:1–20, 1994. 30

T B. Brazelton, E. Tronick, L. Adamson, H Als et S. Wise : Early mother-infant reciprocity. *Parent-infant interaction*, 33:137–154, 1975. 122

L. Bretzner, I. Laptev et T. Lindeberg : Hand gesture recognition using multi-scale colour feature, hierarchical models and particle filtering. *In International Conference on Automatic Face and Gesture Recognition*, 2002. 29

JS. Bridle : *Neurocomputing - Algorithms, Architectures and Applications*, chapitre Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition, pages 227–236. Springer-Verlag, Berlin, 1989. 67

T.J Bruneau et A. Francine : Le silence dans la communication. *Communication et langages*, 20:5–14, 1973. 21

J.S. Bruner : *Processes of cognitive growth : Infancy* . Heinz Werner Lectures, University Press with Barri Publishers, 1968. 42

P.M Brunet, G. McKeown, R. Cowie, H. Donnan et E. Douglas-Cowie : Social signal processing : What are the relevant variables ? and in what ways do they relate ? *In IEEE Intl. Workshop on Social Signal Processing*, 2009. 12

N.J. Butko et J.R. Movellan : An infomax model of eye movement. *In IEEE International Conference on Development and Learning*, 2008. 121

Calliope : *La parole et son traitement automatique*. Dunod, 1997. 31

N. Campbell et P. Mokhtari : Voice quality : the 4th prosodic dimension. *In 15th International Congress of Phonetic Sciences, Barcelone, Espagne*, pages 2417–2420, 2003. 56

G. Castellano, L. Kessous et G. Caridakis : Multimodal emotion recognition from expressive faces, body gestures and speech. *In International Conference on Affective Computing and Intelligent Interaction*, 2007. 30

C.C. Chang et C.J. Lin : Libsvm : a library for support vector machines. Rapport technique, Departement of computer science and information engineering, Taiwan University, 2001. 65

O. Chapelle, B. Schölkopf et A. Zien : *Semi-Supervised Learning*. Cambridge, MA : MIT Press, 2006. 91

T. Charman, J. Swettenham, S. Baron-Cohen, A. Cox, G. Baird et A. Drew : Infants with autism : An investigation of empathy, pretend play, joint attention, and imitation. *Developmental Psychology*, 33(5):781–789, 1997. 45

M. Chetouani, A. Mahdhaoui et F. Ringeval : Time-scale feature extractions for emotional speech characterization. *Cognitive Computation, Springer, publisher*, 1(2):194–201, 2009. 68

C. Clavel, L. Devillers, T. Ehrette et C. Sedogbo : The safe corpus : illustratin extreme emotion in dynamic situations. *In LREC*, 2006. 71

C. Clavel, I. Vasilescu, L. Devillers, Richard et T. G. Ehrette : Fear-type emotion recognition for future audio-based surveillance systems. *Speech Communication*, 50(6):487–503, 2008. 51, 71, 104

S. Clifford et C. Dissanayake : The early development of joint attention in infants with autistic disorder using home video observations and parental interview. *Journal of Autism and Developmental Disorders*, 38:791–805, 2007. 45

D. Cohen, N. Pichard, S. Tordjman, C. Baumann, L. Burglen, S. Excoffier, G. Lazar, P. Mazet, C. Pinquier, A. Verloes et D. Heron : Specific genetic disorders and autism : clinical contribution towards identification. *Journal of Autism and Developmental Disorder*, 35:103–116, 2005. 47

J. Cohen : A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960. 73

J. Cohn, A. Zlochpwer, J.J. Lien et T. Kanade : Feature-point tracking by optical flow discriminates subtle differences in facial expression. *In International Conference on Automatic Face and Gesture Recognition*, pages 396–401, 1998. 31

J.F. Cohn : Foundations of human computing : facial expression and emotion. *In ACM International Conference on Multimodal Interfaces*, 2006. 19

WS. Condon et LW. Sander : Neonate movement is synchronized with adult speech : Interactional participation and language acquisition. *Science*, 183:99–101, 1974. 122

D.B. Cooper et J.H. Freeman : On the asymptotic improvement in the outcome of supervised learning provided by additional nonsupervised learning. *IEEE Transactions on Computers*, 19(11):1055–1063, 1970. 92

RP. Cooper et RN. Aslin : Preference for infant-directed speech in the first month after birth. *Child Development*, 61:1584–1595, 1990. 47

T.F. Cootes, G.J. Edwards et C.J. Taylor : Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6): 681–685, 2001. 31

S. Cormier : *La communication et la gestion*. PU Québec, 2000. 16

M. Costa, W. Dinsbach, ASR. Manstead et P.E.R. Bitti : Social presence, embarrassment, and nonverbal behavior. *Journal of Nonverbal Behavior*, 25(4):225–240, 2001. 22

M. Coulson : Attributing emotion to static body postures : recognition accuracy, confusions, and viewpoint dependence. *Journal of Nonverbal Behavior*, 28(2):117–139, 2004. 22, 30

R. Cowie et R.R. Cornelius : Describing the emotional states that are expressed in speech. *Speech Communication*, 40:5–32, 2003. 50, 51

R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz et J.G. Taylor : Emotion recognition in human-computer interaction. *IEEE Signal Processing magazine*, 18(1):32–80, 2001. 55, 58

R. Cowie et M. Schröder : *Machine Learning for Multimodal Interaction*, chapitre Piecing together the emotion jigsaw, pages 305–317. Springer Verlag, 2005. 50

C. Croft, C. Beckett, M. Rutter, J. Castle, E. Colvert, C. Groothues, A. Hawkins, J. Kreppner, SE. Stevens et EJS Sonuga-Barke : Early adolescent outcomes of institutionally-deprived and non-deprived adoptees. ii : language as a protective factor and a vulnerable outcome. *Journal of Child Psychology and Psychiatry*, 48(1):31–44, 2007. 47

CL. Crown, S. Feldstein, M. Jasnow, B. Beebe et J. Jaffe : The cross-modal coordination of interpersonal timing : Six-week-olds infants gaze with adults vocal behavior. *Journal of Psycholinguistic Research*, 31:1–23, 2002. 120

D. Crystal : Prosodic systems and intonation in english. *In Cambridge University Press, Cambridge*, 1969. 31

SO. Dahlgren et C. Gillberg : Symptoms in the first two years of life : A preliminary population study of infantile autism. *European Archives of Psychiatry and Neurological Sciences*, 238(3):169–174, 1989. 44

N. Dalal et B. Triggs : Histograms of oriented gradients for human detection. *In IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005. 28

T. Darrell, G. Gordon, M. Harville et J. Woodfill : Integrated person tracking using stereo, color, and pattern detection. *International Journal of Computer Vision*, 37(2):175–185, 2000. 29

P. David et C Cardie : Limitations of co-training for natural language from large datasets. *In Proceeding of Empirical Methods in Natural Language Processing*, pages 1–10, 2001. 95

M. De Bonis : *Connaitre les émotions humaines.* Margada, 1996. 4

A.P. Dempster, N.M. Laird et D.B. Rubin : Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977. 62, 91, 92

L. Devillers, R. Cowie, J. Martin, E. Douglas-Cowie, S. Abrilian et M. McRorie : Real life emotions in french and english tv video clips : an integrated annotation protocol combining continuous and discrete approaches. *In Language Resources and Evaluation (LREC)*, 2006. 70

L. Devillers, L. Vidrascu et L. Lamel : Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18:407–422, 2005. 70

W. Dong, B. Lepri, A. Cappelletti, A.S. Pentland, F. Pianesi et M. Zancanaro : Using the influence model to recognize functional roles in meetings. *In Proceedings of the International Conference on Multimodal Interfaces*, pages 271–278, 2007. 33

R. Duda, P. Hart et D. Stork : *Pattern Classification, second edition.* Wiley-Interscience, 2000. 61, 76, 102

N. Eagle et A. Pentland : Reality mining : sensing complex social signals. *Journal of Personal and Ubiquitous Computing*, 10(4):255–268, 2006. 26

Y. Eisenthal, G. Dror et E. Ruppin : Facial attractiveness : beauty and the machine. *Neural Computation*, 18(1):119–142, 2005. 29

P. Ekman : Facial expressions, the handbook of cognition and emotion. *In T. Dalgleish and M. Power, John Wiley & Sons Ltd*, 1999a. 31

P. Ekman : *Gesture, Speech and Signals*, chapitre Emotional And Conversational Nonverbal Signals, pages 45–55. London : Oxford University Press, 1999b. 16

P. Ekman : *Handbook of Cognition and Emotion*, chapitre Basic Emotions, pages 301–320. John Wiley, New York, 1999c. 51

P. Ekman et W.V. Friesen : *Facial action coding system : A technique for the measurement of facial movement.* Consulting Psychologists Press, 1978. 19

P. Ekman et E.L. Rosenberg : What the face reveals : Basic and applied studies of spontaneous expression using the facial action coding system (facs). *In Oxford University Press*, 2005. 22

E. Esposito, V. Stejskal, Z. Smékal et N. Bourbakis : The significance of empty speech pauses : Cognitive and algorithmic issues. *In International conference on Advances in brain, vision and artificial intelligence*, pages 542–554, 2007. 32

S. Favre, J. Salamin, H. abd Dines et A. Vinciarelli : Role recognition in multiparty recordings using social affiliation networks and discrete distributions. *In Proceedings of the ACM International Conference on Multimodal Interfaces*, pages 29–36, 2008. v, 33, 34, 35, 116, 119

A. Fernald : Intonation and reference. *In meeting of the British Psychological Society, Developmental Psychology Section*, Lancaster, England, 1984. 71

A. Fernald : Four-month-old infants prefer to listen to motherese. *Infant Behavior and Development*, 8:181–195, 1985. 47

A. Fernald : *The adapted mind : Evolutionary psychology and the generation of culture*, chapitre Human maternal vocalizations to infants as biologically relevant signals : An evolutionary perspective, pages 391–428. London : Oxford University Press, 1992. 55

A. Fernald et P. Kuhl : Acoustic determinants of infant preference for motherese speech. *Infant Behavior and Development*, 10:279–293, 1987. 47, 85

A. Fernald et C. Mazzie : Prosody and focus in speech to infants and adults. *Developmental Psychology*, 27:209–221, 1991. 54

R. Fernandez et R.W. Picard : Modeling drivers speech under stress speech. *Speech Communication*, 40:145–159, 2003. 51

D. Fisher : Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2(2):139–172, 1987. 90

E. Forgy : Cluster analysis of multivariate data : Efficiency vs. interpretability of classifications. *Biometrics*, 21:768, 1965. 140

W.T. Freeman et M. Roth : Orientation histograms for hand gesture recognition. Rapport technique, Mitsubishi Electric Research Laboratories, Cambridge Research centre, 1995. 29

Y. Freund et R. Schapire : A decision-theoretic generalization of on-line learning and an application to boosting. *Journal on Computer and System Sciences*, 55(1):119–139, 1997. 28

D. Galati et B. Sini : *Les émotions dans les interactions*, chapitre Les structures sémantiques du lexique français des émotions. Plantin, A., 1995. 51

T. Gandhi et M.M. Trivedi : Pedestrian protection systems : issues, survey, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 8(3):413–430, 2007. 30

S. Ganesalingam et G.J. McLachlan : The efficiency of a linear discriminant function based on unclassified initial samples. *Biometrika*, 65(3):658–665, 1978. 91

C. Garcia et G. Tziritas : Face detection using quantized skin color regions merging and wavelet packet analysis. *IEEE Transactions on Multimedia*, 1 (3):264–277, 1999. 28

N. Garg, S. Favre, H. Salamin, D. Hakkani-Tür et A. Vinciarelli : Role recognition for meeting participants : an approach based on lexical information and social network analysis. *In Proceedings of the ACM International Conference on Multimedia*, pages 693–696, 2008. 33

C. Garitte : *Le développement de la conversation chez l'enfant.* DeBoek Université, 1998. 54

D.M. Gavrila : Visual analysis of human movement : a survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999. 27

M.C. Girard et C.M. Girard : *Traitement des données de télédétection.* DUNOD Ed. Paris, 1999. 76

L.R. Gleitman, E.L. Newport et H. Gleitman : The current status of the motherese hypothesis. *Journal of Child Language*, 11:43–79, 1984. 55

SM. Glenn et CC. Cunningham : What do babies listen to most ? a developmental study of auditory preferences in nonhandicapped infants and infants with down's syndrome. *Developmental Psychology*, 19:332–337, 1983. 47

J.L. Goldstein : An optimum processor theory for central formation of the pitch of complex tones. *Journal of the Acoustical Society of America*, 54(6):1496–1516, 1973. 57

M. Goldstein, A. King et M. West : Social interaction shapes babbling : testing parallels between birdsong and speech. *In Proceedings of the National Academy of Sciences*, volume 100, pages 8030–8035, 2003. 47

J.E. Grahe et F.J. Bernieri : The importance of nonverbal cues in judging rapport. *Journal of Nonverbal Behavior*, 23(4):253–269, 1999. 19

M.R. Greenwald et A.G. Banaji : Implicit social cognition : attitudes, self-esteem, and stereotypes. *Psychological Review*, 102(1):4–27, 1995. 25

D.L. Grieser et P.K. Kuhl : Maternal speech to infants in a tonal language : Support for universal prosodic features in motherese. *Developmental Psychology*, 24:14–20, 1988. 55

D. Guillamet, B. Schiele et J. Vitrià : Color histogram classification using nmf. rapport technique 057. Rapport technique, Centre de Visió Per Computador, Université autonome de Barcelone, 2001. 140

A.H. Gunatilaka et B.A. Baertlein : Feature-level and decision-level fusion of noncoincidently sampled sensors for land mine detection. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 23(6):577–589, 2001. 68

H. Gunes, M. Piccardi et M. Pantic : *Affective Computing : Focus on Emotion Expression, Synthesis, and Recognition*, chapitre From the lab to the real world : Affect recognition using multiple cues and modalities, pages 185–218. Vienna, Austria : I-Tech Edu, 2008a. 26

H. Gunes, M. Piccardi et M. Pantic : From the lab to the real world : affect recognition using multiple cues and modalities. *In Affective Computing : Focus on Emotion Expression, Synthesis, and Recognition*, pages 185–218, 2008b. 30

U. Guz, S. Cuendet, D. Hakkani-Tur et G. Tur : Multi-view semi-supervised learning for dialog act segmentation of speech. *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, 18(2):320–329, 2010. 94

E.T. Hall : *The Silent Language*. Doubleday, 1959. 23

J.A. Hall, E.J. Coats et L.S. Lebeau : Nonverbal behavior and the vertical dimension of social relations : a meta-analysis. *Psychological Bulletin*, 1313(6):898–924, 2005. 116

H. Hermansky, N. Morgan, A. Bayya et P. kohn : Rasta-plp speech analysis. Rapport technique, ICSI Technical Report TR-91-069, 1991. 69

A. Hinnerburg et D.A. Keim : Cluster discovery methods for large data bases - from the past to the future. *In ACM SIGMOD Int. Conf. on Management of Data. Tutorial Session*, 1999. 90

J. Hirschberg : Pitch accent in context : predicting intonational prominence from text. *Artificial Intelligence*, 63(1-2):305–340, 1993. 20

J. Hirschberg et B. Grosz : Intonational features of local and global discourse structure. *In Proceedings of the Speech and Natural Language Workshop*, pages 441–446, 1992. 20

H. Hong, H. Neven et C. von der Malsburg : Online facial expression recognition based on personalized gallery. *In Conference on Automatic Face and Gesture Recognition*, pages 354–359, 1998. 31

Y. Hoshino, H. Kumashiro, Y. Yashima, R. Tachibana, M. Watanabe et H. Furukawa : Early symptoms of autistic children and its diagnostic significance. *Folia Psychiatrica et Neurologica Japonica*, 36(4):367–374, 1982. 44

D.W. Hosmer : A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different thypes of sample. *Biometrics*, 29:761–770, 1973. 92

C.R.L. Hsu, M. Abdel-Mottaleb et A.K. Jain : Face detection in colour images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):696–706, 2002. 28

X. Huang, A. Acero et H. W.Hon : *Spoken Language Processing*. Prentice-Hall, Englewood Cliffs, NJ, 2001. 31

A. Ito, X. Wang, M. Suzuki et S. Makino : Smile and laughter recognition using speech processing and face recognition from conversation video. *In Proceedings of the International Conference on Cyberworlds*, pages 437–444, 2005. 32

J. Jaffe, B. Beebe, S. Feldstein, CL. Crown, et MD Jasnow : Rhythms of dialogue in infancy : Coordinated timing in development. *Monographs of the Society for Research in Child Development*, 66(2):131, 2001. 120

D.B. Jayagopi, H. Hung, C. Yeo et D. Gatica-Perez : Modeling dominance in group conversations from non verbal activity cues. *IEEE Transaction on Audio, Speech, and language processing*, 17(3):501–513, 2009. 117

T. Joachims : Making large-scale support vector machine learning practical. *In MIT Press, Cambridge, MA, USA*, pages 169–184, 1999a. 93

T. Joachims : Transductive inference for text classification using support vector machines. *In Proc. 16th International Conf. on Machine Learning*, pages 200–209, 1999b. 93

S.E. Johnson et P.C. Woodland : Speaker clustering using direct maximisation of the mllr-adapted likelihood. *In International Conference on Spoken Language Processing*, 1998. 27

D. Keltner et P. Ekman : *Handbook of Emotions*, chapitre Facial expression of emotion, pages 236–249. Guilford Press, 2000. 19

D. Keltner et J. Haidt : Social functions of emotions at four levels of analysis. *Cognition and Emotion*, 13(5):505–521, 1999. 21

K. Kim, P.G. Georgiou, S. Lee et N. Narayanan : Real-time emotion detection system using speech : Multi-modal fusion of different timescale features. *In IEEE 9th Workshop on Multimedia Signal Processing*, pages 48–51, 2007. 68, 69, 96

S-P. Kim, Y.N. Rao, D. Ergodmus, J.C. Sanchez, M.A.L. Nicolelis et J.C. Principe : Determining patterns in neural activity for reaching movements using non-negative matrix factorization. *EURASIP Journal on Applied Sig. Special Issue of Trends in Brain Computer Interfaces*, 19:3113–3121, 2005. 140

P.R. Kleinginna et A.M. Kleinginna : A categorized list of emotion definitions with suggestions for a consensual definition. *Motivation and Emotion*, 5(4):345–379, 1981. 4

W.W. Kong et S. Ranganath : Automatic hand trajectory segmentation and phoneme transcription for sign language. *In Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, 2008. 29

P. Kuhl : Human speech and birdsong : Communication and the social brain. *In The Proceedings of the National Academy of Sciences*, volume 100(17), pages 9645–9646, 2003. 47

PK. Kuhl : Early language acquisition : Cracking the speech code. *Nature Reviews Neuroscience*, 5(11):831–843, 2004. v, 55

J.R. Landis et G.G. Koch : The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174, 1977. 74

H.C. Lasswell : *The Comunication of Ideas*, chapitre The Structure and Function of Communication in Society, page 37. New YOrk, Harper and Brothers, 1948. 14

MC. Laznik, S. Maestro, F. Muratori et E. Parlato : Les interactions sonores entre les bébés devenus autistes et leur parents. *In* G Konopczynski, éditeur : *Au commencement tait la voix*, pages 171–81, 2005. 5

L. Lebart, A. Maurineau et M. Piron : *Traitement des données statistiques*. Dunod, Paris, 1982. 143

C.M. Lee, S. Narayanan et R. Pieraccini : Recognition of negative emotions from the speech signal. *In IEEE Automatic Speech Recognition and Understanding*, 2001. 51

D. D. Lee et H.S. Seung : Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999. 140, 141, 142

D. D. Lee et H.S. Seung : Algorithms for non-negative matrix factorization. *In NIPS*, pages 556–562, 2000. 141, 142

Y. Lee, J.A. Nelder et Y. Pawitan : *Generalized Linear Models with Random Effects : Unified Analysis via H-likelihood*. Chapman & Hall/RC, 2006. 137

W. Liu et K. Yuan : Sparse p-norm nonnegative matrix factorization for clustering gene expression data. *International Journal of Data Mining and Bioinformatics*, 2(3):236–249, 2008. 140

N.J. Mackinnon : Symbolic interactionism as affect control theory. *In State University of New York Press, New York*, 1994. 53

S. Maestro, F. Muratori, F. Barbieri, C. Casella, V. Cattaneo, M.C. Cavallaro, A. CESARI, A. MILONE, L. RIZZO, V. VIGLIONE, D. STERN et F. PALACIO-ESPASA : Early behavioral development in autistic children : The first 2 years of life through home movies. *Psychopathology*, 34:147–152, 2001. 122

S. Maestro, F. Muratori, M.C. Cavallaro, C. Pecini, A. Cesari, A. Paziente, D. Stern, B. Golse et F. Palacio-Espasa : How young children treat objects and people : an empirical study of the first year of life in autism. *Child psychiatry and Human Development*, 35(4):383–396, 2005. 71, 123

S. Maestro, F. Muratori, MC. Cavallaro, F. Pei, D. Stern, B. Golse et F. Palacio-Espasa : Attentional skills during the first 6 months of age in autism spectrum disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, 41:1239–1245, 2002. 45, 122

S. Maestro, F. Muratori, A. Cesari, C. Pecini, F. Apicella et D. Stern : A view to regressive autism through home movies. is early development really normal. *Acta Psychiatrica Scandinavica*, 113:68–72, 2006. 122

A. Mahdhaoui et M. Chetouani : Emotional speech classification based on multi view characterization. *In IAPR International Conference on Pattern Recognition, ICPR*, 2010. 51

A. Mahdhaoui, M. Chetouani et C. Zong : Motherese detection based on segmental and supra-segmental features. *In IAPR International Conference on Pattern Recognition, ICPR*, 2008. 51, 68

J. Maisonnasse et O. Brdiczka : Détection automatique des groupes d'interactions. *In 2nd French-speaking conference on Mobility and ubiquity computing*, 2005. 119

M. Mancini, R. Bresin et C. Pelachaud : An expressive virtual agent head driven by music performance. *IEEE Transactions on Audio, Speech and Language Processing*, 15(6):1833–1841, 2007. 116

H. Massie : The early natural history of childhood psychosis. *Journal of the American Academy of Child Psychiatry*, 14:683–707, 1975. 122

M. Maury, A. Lazartigues, D. Sauvage, JP. Visier et JP. Raynaud : *MATURATION ET VULNERABILITE*, chapitre Développement affectif du nourrisson. L'installation précoce de la relation mère-enfant et son importance, pages 46–53. Service de pédopsychiatrie CHU Anger, 2005. 40

N Mayer et E.Z. Tronick : *Social perception in infants*, chapitre Mothers' turn-giving signals and infant turn-taking in mother-infant interaction, pages 199–216. Norwood, NJ : Ablex, 1985. 41

R. Mays et J. Gillon : Autism in young children : an update. *Pediatric Health Care*, 7:17–23, 1993. 145

I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner et H. Bourlard : Modeling human interaction in meetings. *In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 748–751, 2003. 25

G.J. McLachlan : Estimating the linear discriminant function from initial samples containing a small number of unclassified observations. *Journal of the American statistical association*, 72(358):403–406, 1977. 92

D. McNeill : Hand and mind : What gestures reveal about thought. *In University Of Chicago Press*, 1996. 22

A. Mehrabian et S.R. Ferris : Inference of attitude from nonverbal communication in two channels. *Journal of Counseling Psychology*, 31(3):248–252, 1967. 19

A. Mehrabian's : *Nonverbal communication.* Aldine transaction, 1972. 15, 16

E.G. Merinfeld : Attachement et intersubjectivité : premiers liens de l'enfant introduction. *Cahiers critiques de thérapie familiale et de pratiques de réseaux*, 35:166, 2005. 39

D.M. Messinger, P. Ruvolo, PV. Ekas et A. Fogel : Applying machine learning to infant interaction : The development is in the details (in press). *Neural Networks*, 2010. 121

S. Möller et R. Schönweiler : Analysis of infant cries for the early detection of hearing impairment. *Speech Communication*, 28(3):175–193, 1999. 32

T.B. Moeslund et E. Granum : A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231–268, 2001. 27

S. Moncrieff, C. Dorai et S. Venkatesh : Affect computing in film through sound energy dynamics. *In Proceedings of ACM*, 2001. 58

E. Monte-Moreno, M. Chetouani, M. Faundez-Zanuy et J. Sole-Casals : Maximum likelihood linear programming data fusion for speaker recognition. *Speech Communication*, 51(9):820–830, 2009. 68

J. Montero, J. Gutierrez-Arriola, J. Colas et JM. Pardo : Analysis and modelling of emotional speech in spanish. *In Proceedings of ICPhS, San Francisco, U.S.A*, 1999. 56

D. Moraru : Segmentation en locuteurs. Rapport technique, CLIPS-IMAG, Université Joseph Fourrier, 2002. 27

D. Morris : *Bébé révélé*. Calmann-Lévy, 1994. 38

F. Muratori, F. Apicella, P. Muratori et S. Maestro : Intersubjective disruptions and caregiver-infant interaction in early autistic disorder. *Research in Autism Spectrum Disorders*, In Press, Corrected Proof:10, 2010. 121, 122, 123

F. Muratori et S. Maestro : Autism as a downstream effect of primary difficulties in intersubjectivity interacting with abnormal development of brain connectivity. *International Journal for Dialogical Science*, 2(1):93–118, 2007. 5

I. Muslea, S. Minton et C. Knoblock : Selective sampling with redundant views. *In Proceedings of Association for the Advancement of Artificial Intelligence*, pages 621–626, 2000. 96

S. Narayanan : Towards modeling user behavior in human-machine interactions : Effect of errors and emotions. *In International Standards for Language Engineering workshop*, 2002. 51

C. Nicolini, B. Lepri, S. Teso et A. Passerini : From on-going to complete activity recognition exloiting related activities. *In Human Behavior Understanding*, 2010. 118

K. Nigam, A.K. McCallum, S. Thrun et T. Mitchell : Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39(2-3):103–134, 2000. 88, 91

A. Oikonomopoulos, I. Patras et M. Pantic : Implicit spatiotemporal shape model for human activity localisation and recognition. *In Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 27–33, 2009. 116

T. O'Neill : Normal discrimination with unclassified observations. *Journal of the American Statistical Association*, 73(364):821–826, 1978. 91

C. Padgett, G. Cottrel et R. Adolphs : Categorical perception in facial emotion classification. *In ACM Siggraph*, pages 75–84, 1996. 31

P. Pal, A.N. Iyer et R.E. Yantorno : Emotion detection from infant facial expressions and cries. *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006. 32

M. Pantic, A. Pentland, A. Nijholt et T. Huang : Human computing and machine understanding of human behavior : a survey. *In ICMI'06 : Proceedings of the 8th international conference on Multimodal interfaces*, 2006. ISBN 1-59593-541-X. 18

M. Pantic, A. Pentland, A. Nijholt et T. Huang : Human-centred intelligent human-computer interaction (hci2) : how far are we from attaining it ? *International Journal of Autonomous and Adaptive Communications Systems*, 1(2):168–187, 2008. 18

M. Pantic, I. Poggi, F. D'Errico, M. Schroder, C. Pelachaud, D. Heylen et V. Vinciarelli : D2.2 : First draft of the agenda on ssp research. Rapport technique, SSPNet Social Signal Processing Network., 2009. 2, 53

M. Pantic et L.J.M. Rothkrantz : Automatic analysis of facial expressions : the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000. 27

C. Papageorgiou, M. Oren et T. Poggio : A general framework for object detection. *In IEEE International Conference on Computer Vision*, pages 555–562, 1999. 28

H. Papousek et M. Papousek : *Studies in mother-infant interaction : The Loch Lomond Symposium*, chapitre Mothering and cognitive head start : Psychobiological considerations, pages 63–85. London : Academic Press, 1977. 42

H. Papousek et M Papousek : *Handbook of infant development (2nd ed)*, chapitre Intuitive parenting : A dialectic counterpart to the infant's integrative competence, pages 669–720. New York : Wiley, 1987. 42

K-J. Park et A. Yilmaz : Social network approach to analysis of soccer game. *In International Conference on Pattern Recognition*, 2010. 119

V. Pavlovic, R. Sharma et T. Huang : Visual interpretation of hand gestures for human-computer interaction : A review. *PAMI*, 19 (7):677–695, 1997. 29

JE. Pegg, JF. Werker et PJ. McLeod : Preference for infant-directed over adult-directed speech : Evidence from 7-week-old infants. *Infant Behavior and Development*, 15:325–345, 1992. 47

A. Pentland : Social signal processing. *IEEE Signal Processing Magazine*, 24(4):108–111, 2007. 2, 12

T. Pfau et G. Ruske : Estimating the speaking rate by vowel detection. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 945–948, 1998. 32

W.R. Picard : Affective computing : Challenges. mit media laboratory cambridge usa. *International Journal of Human-Computer Studies*, 59(1-2):55–64, 2003. 51

J.C. Platt : *Advances in Large Margin Classifiers*, chapitre Probabilistic Outputs for SVM and comparison to regularized likelihood methods. MIT Press, Cambridge, MA, 1999. 65

R. Plutchik : *Approaches to Emotion*, chapitre A General Psychoevolutionary Theory. Erlbaum, 1984. v, 52

I. Poggi, F. D'Ericco et L. Vincze : Agreement and its multimodal communication in debates : A qualitative analysis. *Cognitive Computation*, 3:1–14, 2010. 117

I. Poggi et F. D'Errico : Dominance signals in debates. *In Human Behavior Understanding*, 2010. 116

L. Rabiner et M. Sambur : Voiced-unvoiced-silence detection using the itakura lpc distance measure. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 323–326, 1977. 32

L.R. Rabiner : A tutorial on hidden markov models and selected applications in speech recognition. *IEEE*, 77:257–286, 1989. 35, 119

L.R. Rabiner et B.H. Juang : *Fundamentals of Speech Recognition*. Prentice-Hall, 1993. 58

L.R. Rabiner et R.W. Schafer : *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, NJ, 1978. 31, 32

C. Receveur, P. Lenoir, H. Desombre, S. Roux, C. Barthelemy et J. Malvy : Interaction and imitation deficits from infancy to 4 years of age in children with autism : A pilot study based on videotapes. *Autism*, 9:69–82, 2005. 122

D.A. Reynolds : Speaker identification and verification using gaussian mixture speaker models. *Speech communication*, 17:91–108, 1995. 62

D.A. Reynolds, R.B. Dunm et J.J. Laughlin : The lincoln speaker recognition system : Nist eval2000. *In Proceedings of International Conference on Spoken Language Processing (ICSLP 2000)*, 2000. 27

D.A. Reynolds, E. Singer, B. A. Carson, G.C. O'Leary, J.J. McLaughlin et M.A. Zissman : Blind clustering of speech utterances based on speaker and language characteristics. *In International Conference on Spoken Language Processing*, 1998. 27

J.W. Riley et M.W. Riley : *Sociologie de l'information*, chapitre La communication de masse et le système social, page 79. Paris, Larousse, 1973. 14

L. Rimé, B. et Schiratura : *Fundamentals of non verbal behaviour*, chapitre Gesture and speech, pages 239–281. Cambridge, Cambridge University Press, 1991. 18

F. Ringeval et M. Chetouani : A vowel based approach for acted emotion recognition. *In Interspeech*, 2008. 32

F. Ringeval, J. Demouy, G. Szaszak, M. Chetouani, L. Robel, J. Xavier, D. Cohen et M. Plaza : Automatic intonation recognition for the prosodic assessment of language impaired children. *IEEE Transactions on Audio, Speech and Language Processing*, in press, 2010. 51

J.A Rondal : *L'interaction adulte-enfant et la construction du langage*. Bruxelles, Mardaga, 1983. 54

S.A. Rose, J.F. Feldman et J.J. Jankowski : Infant visual recognition memory : Independent contributions of speed and attention. *Developmental Psychology*, 39:563–571, 2003. 55

R. Rosenfeld : Two decades of statistical language modeling : Where do we go from here. *IEEE*, 88:1270–1278, 2000. 35

H.A. Rowley, S. Baluja et T. Kanade : Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998. 27

F. Ruth et E. Arthur : Parent-infant synchrony and the social-emotional development of triplets. *Developmental psychology*, 40(6):1133–1147, 2004. 39, 53

M. Rutter, E. Colvert, J. Kreppner, C. Beckett, J. Castle, C. Groothues, A. Hawkins, SE. Stevens et EJS. Sonuga-Barke : Early adolescent outcomes for institutionally-deprived and non-deprived adoptees. i : Disinhibited attachment. *Journal of Child Psychology and Psychiatry*, 48(1):17–30, 2007a. 47

M. Rutter, J. Kreppner, C. Croft, M. Murin, E. Colvert, C. Beckett, J. Castle et EJS. Sonuga-Barke : Early adolescent outcomes of institutionally deprived and non-deprived adoptees. iii. quasi-autism. *The Journal of Child Psychology and Psychiatry*, 48(12):1200–1207, 2007b. 47

V.R. Sa et D. Ballard : Category learning through multi-modality sensing. *Neural Computation*, 10(5):24, 1998. 97

C. Saint-Georges, R.S. Cassel, D. Cohen, M. Chetouani, M.C. Laznik, S. Maestro et F. Muratori : What studies of family home movies can teach us about autistic infants : A literature review. *Research in Autism Spectrum Disorders*, 4(3):355–366, 2010. 45, 46, 122

G. Salton et C Buckley : Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988. 139

A. Sarkar : Applying co-training methods to statistical parsing. *In Proceeding of the 2nd annual mééting of the north americain chapter of the association for computational linguistics*, pages 175–182, 2001. 95

G.P. sathas : *Conversation Analysis - The Study of Talk-in-Interaction*. Sage Publications, 1995. 21

R.E. Schapire et Y. Singer : Improved boosting algorithms using confidence-rated predictions. *Machine Learning, Springer Netherlands*, 37(3):297–336, 1999. 103

E.D. Scheirer : *Music Listening Systems*. Thèse de doctorat, Massachusetts Institute of Technology, 2000. 57

S. Scherer, F. Schwenker et G. Palm : Classifier fusion for emotion recognition from speech. *In IEEE Intelligent Environments*, pages 152–155, 2007. 68, 69

FW. Schneider, JA. Gruman et LM. Coutts : *Applied Social Psychology : Understanding and Addressing Social and Practical Problems*. Sage, 2005. 24

B. Schoelkopf et A.J. Smola : Learning with kernels. *In The MIT Press, Cambridge*, 2002. 64

E. Schopler, RJ. Reichler et BR. Renner : The childhood autism rating scale. *In Los Angeles : Western Psychological Services*, 1988. 43

M. Schröder : Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions. *In Proc. Workshop on Affective Dialogue Systems Kloster Irsee*, pages 209–220, 2004. 51

B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous et V. Aharonson : The relevance of feature type for the automatic classification of emotional user states : low level descriptors and functionals. *In Proceedings of Interspeech*, pages 2253–2256, 2007. 53, 55, 104

B. Schuller, S. Steidl et A. Batliner : The interspeech 2009 emotion challenge. *In Interspeech*, pages 312–315, 2009. 104

B. Schuller, B. Vlasenko, D. Arsic, G Rigoll et A. Wendemuth : Combining speech recognition and acoustic word emotion models for robust text-independent emotion recognition. *In IEEE International Conference on Multimedia & Expo, ICME*, 2008. 70

LE. Schulz et A. Gopnik : Causal learning across domains., . *Developmental Psychology*, 40(2):162–167, 2004. 121

A. Sepheri, Y. Yacoob et L. Davis : Employing the hand as an interface device. *Journal of Multimedia*, 1(7):18–29, 2006. 29

M. Shami et W. Verhelst : An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. *Speech Communication*, 49:201–212, 2007. 70, 104

M.T. Shami et M.S. Kamel : Segment-based approach to the recognition of emotions in speech. *In IEEE Multimedia and Expo*, 2005. 81

F. Shanaz, M. Berry, P. Pauca et R. Plemmons : Document clustering using nonnegative matrix factorization. *Information Processing and Management*, 42(2):373–386, 2006. 142

C. Shannon et W. Warren : *La Théorie mathématique de la communication (trad. française)*. Paris, Retz-CEPL, 1975. 13

P.E. Shrout et D.W. Fiske : Nonverbal behaviors and social evaluation. *Journal of Personality*, 49(2):115–128, 1981. 21

M-H. Siu, G. Yu et H. Gish : Segregation of speakers for speech recognition and speaker identification. *In Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 91)*, 1991. 27

K. Sjolander et J. Beskow : Wavesurfer - an open source speech tool. *In International Conference on Spoken Language Processing,*, pages 464–467, 2000. 31

M. Slaney et R.F. Lyon : A perceptual pitch detector. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP*, pages 357–360, 1990. 57

D.M. Stack et S.L. Arnold : Changes in mothers touch and hand gestures influence infant behaviour during face-to-face interchanges. *Infant Behavior and Development*, 21:799–806, 1998. 42

R.J. Stanley, P.D. Gader et K.C. Ho : Feature and decision level sensor fusion of electromagnetic induction and ground penetrating radar sensors for landmine detection with hand-held units. *Information fusion*, 3:215–223, 2002. 68

S. Steidl, M. Levit, A. Batliner, E. Noth et E. Niemann : Off all things the measure is man automatic classification of emotions and interlabeler consistency. *In IEEE international conference on acoustics, speech, and signal processing*, 2005. 51

D.N. Stern : *The interpersonal world of the infant : A view from psychoanalysis and development psychology.* Basic Books, 1985. 42

D.N. Stern : *The perceived self : Ecological and interpersonal sources of the self-knowledge*, chapitre The role of feelings for an interpersonal self, pages 205–215. New York : Cambridge University Press, 1993. 42

D.N. Stern : *La constellation maternelle.* Calmann-Lévy, Paris, 1997. 39

DN. Stern, J. Jaffe, B. Beebe et SL. Bennett : Vocalization in unison and alternation : Two modes of communication within the mother-infant dyad. *Annals of the New York Academy of Science*, 263:89–100, 1975. 122

WL. Stone, E L. Hoffman, SE. Lewis et OY. Ousley : Early recognition of autism : Parental reports vs. clinical observation. *In Archives of Paediatrics and Adolescent Medicine 148*, 1994. 43

A. Strehl et J. Ghosh : Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Machine Learning Research*, 3:583–617, 2002. 144

A. Sullivan, J. Kelso et M. Stewart : Mothers' views on the ages of onset for four childhood disorders. *Child Psychiatry Human Development*, 20(4):269–278, 1990. 21, 43

K.K. Sung et T. Poggio : Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):39–51, 1998. 27

J. Swettenham, S. Baron-Cohen, T. Charman, A. Cox, G. Baird, A. Drew, L. Rees et S. Wheelwright : The frequency and distribution of spontaneous attention shifts between social and nonsocial stimuli in autistic, typically developing, and nonautistic developmentally delayed infants. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 39(5):747–753, 1998. 45

D.K Symons et G. Moran : Responsiveness and dependency are different aspects of social contingencies : An example from mother and infant smiles. *Infant Behavior and Development*, 17(2):209–214, 1994. 120, 121

MP. Thollon-Behar et C. Cohas : Children between 20 and 30 months : communicative skill within relationship with adult and ability to interact with peers. *Pratiques psychologiques*, 12:59–72, 2006. 36

E.L. Thorndike : A constant error on psychological ratings. *Journal of Applied Psychology*, 4:25–29, 1920. 24

S.E. Tranter et D.A. Reynolds : An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1557–1565, 2006. 27

K.P ; Truong et D.A. Leeuwen : Automatic detection of laughter. *In Proceedings of the European Conference on Speech Communication and Technology*, pages 485–488, 2005. 32

K.P. Truong et D.A. van Leeuwen : Automatic discrimination between laughter and speech. *Speech Communication*, 49(2):144–158, 2007. 32, 58, 81

P. Turaga, R. Chellappa, V.S. Subrahmanian et O. Udrea : Machine recognition of human activities : A survey. *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY*, 18(11):1473–1488, 2008. 116

J. Van den Stock, R. Righart et Gelder B.de : Body expressions influence recognition of emotions in the face and voice. *Emotion*, 7(3):487–494, 2007. 22, 30

V. Vapnik : *Statistical Learning Theory*. New York : Wiley, 1998. 64

V.N. Vapnik : *The Nature of Statistical Learning Theory*. Springer, New York, 1995. 63, 93

P.K. Varshney : *Distributed Detection and Data Fusion (Signal Processing and Data Fusion)*. Springer-Verlag, 1997. 68

D. Ververidis et C. Kotropoulos : Emotional speech recognition : Resources, features, and methods. *Speech Communication*, 48:1162–1181, 2006. 70

L. Vidrascu et L. Devillers : Real-life emotion representation and detection in call centers data. *In Affective computing and intelligent interaction*, 2005. 70

A. Vinciarelli : Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. *IEEE Transactions on Multimedia*, 9(9):1215–1226, 2007. 33

A. Vinciarelli, A. Dielmann, S. Favre et H. Salamin : Canal9 : a database of political debates for analysis of social interactions. *In International conference on affective computing and intelligent interaction*, 2009a. 117

A. Vinciarelli et J.M. Odobez : Application of information retrieval technologies to presentation slides. *IEEE Transactions on Multimedia*, 8(5):981–995, 2006. 25

A. Vinciarelli, M. Pantic et H.and Pentland A. Bourlard : Social signal processing : Survey of an emerging domain. *Image and Vison Computing*, 27(12):1743–1759, 2009b. 2, 12

A. Vinciarelli, H. Salamin et M. Pantic : Social signal processing : Understanding social interaction through nonverbal behavior analysis. *In International Workshop on Computer Vision and Pattern Recognition for Human Behavior*, 2009c. v, 2, 30, 117

P. Viola et M.J. Jones : Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004. 28, 30

D.S. Virginia et D. Ballard : Category learning from multi-modality. *Neural Computation*, 10(5):1097–1117, 1998. 95

FR. Volkmar et D. Pauls : Autism. *Lucet*, 362:1133–1141, 2003. 43

A. Waibel, T. Schultz, M. Bett, M. Denecke, R. Malkin, I. Rogina et R. Stiefelhagen : Smart : the smart meeting room task at isl. *In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 752–755, 2003. 25

L. Wang et T. Hu, W. Tan : Recent developments in human motion analysis. *Pattern Recognition*, 36(3):585–601, 2003. 30

R.M. Warner et Sugarman D.B. : Attributions of personality based on physical appearance, speech, and handwriting. *Journal of Personality and Social Psychology*, 50(4):792–799, 1986. 25

M.K. Weinberg et E.Z. Tronick : Beyond the face : An empirical study of infant affective configurations of facial, vocal, gestural and regulatory behaviors. *Child Development*, 65:1503–1515, 1994. 41

C.Y. Weng, W.T. Chu et J.L. Wu : Movie analysis based on roles social network. *In Proceedings of IEEE International Conference on Multimedia and Expo*, pages 1403–1406, 2007. 33

JF. Werker et PJ. McLeod : Infant preference for both male and female infant-directed talk : A developmental study of attentional affective responsiveness. *Canadian Journal of Psychology*, 43:230–246, 1989. 47

E. Werner, G. Dawson, J. Osterling et N. Dinno : Recognition of autism spectrum disorder before one year of age : A retrospective study based on home videotapes. *Journal of Autism and Developmental Disorders*, 30:157–162, 2000. 44

C.M. Whissel : The dictionary of affect in language. emotion : Theory, research, and experience. *In New York : Academic Press*, 1989. 51

N. Wiener : *God and Golem.* Cambridge, Massassuchets Institute of Technology, MIT, 1964. 13

DC. Wimpory, RP. Hobson, JMG. Williams et S. Nash : Are infants with autism socially engaged ? a study of recent retrospective parental reports. *Journal of Autism and Developmental Disorders*, 30(6):525–536, 2000. 44

B. Wu et R. Nevatia : Tracking of multiple, partially occluded humans based on static body part detection. *In IEEE Proc. of the conference on Computer Vision and Pattern Recognition*, pages 951–958, 2006. 29

Y. Wu et Huang. T.S. : Vision-based gesture recognition : a review. *In Proceedings of the International Gesture Workshop*, pages 103–109, 1999. 29

Z.L. Wu, C.W. Cheng et C.H. Li : Social and semantics analysis via non-negative matrix factorization. *In Proceeding of the 17th international conference on World Wide Web*, pages 1245–1246, 2008. vi, 118, 119, 142

W. Wundt : *Grundriss der Psychologie.* Leipzig, 1913. 51

W. Xu, X. Liu et Y. Gong : Document clustering based on non-negative matrix factorization. *In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003. 142

M.H. Yang, D. Kriegman et N. Ahuja : Detecting faces in images : a survey. *IEEE transactions on pattern analysis and machine intelligence*, 24(1):34–58, 2002. 30

D. Yarowsky : Unsupervised word sense disambiguation rivaling supervised methods. *In Dans 33rd Annual Meeting of the ACL, Cambridge, MA*, pages 189–196, 1995. 94

A. Zakian, J. Malvy, H. Desombre, S. Roux et P. Lenoir : Signes précoces de l'autisme et films familiaux : une nouvelle étude par cotateurs informés et non informés du diagnostic. *L'Encéphale*, 26:38–44, 2000. 45

M. Zancanaro et F. Lepri, B. Pianesi : Automatic detection of group functional roles in face to face interactions. *In Proceedings of the International Conference on Multimodal Interfaces*, pages 28–34, 2006. 33

Q Zhang et S. Sun : Multiple-view multiple-learner active learning. *Pattern Recognition*, 43(9):3113–3119, 2010. vi, 97, 98

T. Zhang, R. Ramakrishnan et M. Livny : Birch : A new data clustering algorithm and its applications. *Data Mining and Knowledge Discovery*, 1(2):141–182, 1997. 90

Q. Zhao, J. Kang, H. Tao et W. Hua : Part based human tracking in a multiple cues fusion framework . *In International Conference on Pattern Recognition*, pages 450–455, 2006. 28

W. Zhao, R. Chellappa, A. Rosenfeld et P.J. Phillips : Face recognition : A literature survey. *In CVL Technical Report, University of Maryland*, 2000. 30

Y. Zhou et S. Goldman : Democraticco-learning. *In Proceedings of the16th IEEE International Conference on Tools with Artificial Intelligence*, pages 594–602, 2004. 96

X. Zhu : Semi-supervised learning literature survey. Rapport technique, Univ. of Wisconsin, Madison, 2006. 91

X. Zhu, Z. Ghahramani et J. Lafferty : Semi-supervised learning using gaussian fields and harmonic functions. *In ICML*, 2003. 94

L. Zwaigenbaum, S. Bryson, T. Rogers, W. Roberts, J. Brian et P. Szatmari : Behavioral manifestations of autism in the first year of life. *International Journal of Developmental Neuroscience*, 23(2-3):143–152, 2005. 45

E. Zwicker : Subdivision of the audible frequency range into critical bands. *Journal of the Acoustical Society of America*, 33(2):248–248, 1961. 104

E. Zwicker et E. Terhardt : Analytical expressions for crtitical-band rate and critical bandwith as a function of frequency. *JASA*, 68(5):1523–1525, 1980. 104