Représentation de comportements emotionnels multimodaux spontanés : perception, annotation et synthèse

Sarkis Abrilian

Université Paris-Sud 11
Faculté des Sciences d'Orsay
91405 ORSAY CEDEX

LIMSI-CNRS
BP 133
F-91403 ORSAY CEDEX

Thèse pour obtenir le grade de Docteur de l'Université Paris 11
Discipline : Informatique

Présentée et soutenue publiquement par

Sarkis ABRILIAN

Représentation de Comportements Emotionnels Multimodaux Spontanés : Perception, Annotation et Synthèse

le 7 septembre 2007, devant le jury composé de :

Joseph MARIANI	Directeur
Jean-Claude MARTIN	Encadrant
Laurence DEVILLERS	Encadrant
Lola CANAMERO	Rapporteur
Anne NICOLLE	Rapporteur

Remerciements

Je voudrais tout d'abord remercier mes rapporteurs, Lola Canamero et Anne Nicole, pour le temps qu'ils m'ont accordé et leurs remarques pertinentes m'ayant permis d'améliorer mon manuscrit.
De chaleureux remerciements à mes coencadrants de thèse, Jean-Claude Martin et Laurence Devillers, pour m'avoir aidé scientifiquement et soutenu moralement tout au long de ma thèse. Merci particulièrement à Jean-Claude qui m'a fait confiance depuis le tout début en 2002 lors de mon stage de master, puis pendant un an sur le projet européen NICE, avant de me proposer un financement de thèse dans le cadre du projet HUMAINE sur les émotions.

Merci à Joseph Mariani d'avoir accepté de diriger ma thèse.

Merci à Marianne Najm et Fabien Bajeot pour le travail d'annotation qu'ils ont effectué, merci également à tous ceux qui ont accepté de participer à mes expériences !

Je remercie vivement les personnes avec qui j'ai partagé le bureau pendant ces années, Stéphanie Buisine, Guillaume Pitel, Nicolas Sabouret, Ouriel Grynspan, Haïfa Zargayouna, David Leray, Aurélie Zara, François Bouchet. Merci particulièrement à Stéphanie Buisine pour ses analyses pertinentes lors des travaux que nous avons effectué en collaboration.

Merci à Jean Paul Sansonnet pour avoir donné un second souffle à l'agent Lea sur lequel j'ai travaillé avant de commencer ma thèse. Merci à Patrick Paroubek pour ses conseils.

Merci à Julien Poudade, avec qui j'ai effectué tout mon parcours universitaire : DEUG, Licence, et Maitrise à Paris 6, puis DEA et Thèse à Orsay. Merci également à Lionel Landwerlin, son accolyte sur les AIBOs.

Un grand merci à Catherine Pelachaud et Maurizio Mancini avec qui j'ai collaboré lors de ma thèse sur l'agent Greta développé à Paris 8.

Merci à Ellen Douglas-Cowie et Roddy Cowie pour leur collaboration dans le cadre du projet HUMAINE, merci également aux autres membres du projet HUMAINE avec qui j'ai eu l'occasion de travailler.

Merci à mes amis, pour leur présence et leur soutien.

Merci enfin à ma famille, qui m'a toujours encouragé à atteindre mon but.

Résumé

L'objectif de cette thèse est de représenter les émotions spontanées et les signes multimodaux associés pour contribuer à la conception des futurs systèmes affectifs interactifs. Les prototypes actuels sont généralement limités à la détection et à la génération de quelques émotions simples et se fondent sur des données audio ou vidéo jouées par des acteurs et récoltées en laboratoire. Afin de pouvoir modéliser les relations complexes entre les émotions spontanées et leurs expressions dans différentes modalités, une approche exploratoire est nécessaire. L'approche exploratoire que nous avons choisie dans cette thèse pour l'étude de ces émotions spontanées consiste à collecter et annoter un corpus vidéo d'interviews télévisées. Ce type de corpus comporte des émotions plus complexes que les 6 émotions de base (colère, peur, joie, tristesse, surprise, dégoût). On observe en effet dans les comportements émotionnels spontanés des superpositions, des masquages, des conflits entre émotions positives et négatives.

Nous rapportons plusieurs expérimentations ayant permis d'apporter des éléments de réponses aux questions suivantes : Comment annoter de manière fiable et représenter des émotions complexes spontanées ? A quels niveaux d'abstraction et à quel niveau temporel ? Comment représenter le contexte d'une émotion ? Quels sont les paramètres des comportements multimodaux pertinents pour la perception des émotions ? Quelles sont les relations entre les émotions et les comportements multimodaux associés ? Comment appliquer ce type de connaissance à la spécification d'agents animés expressifs montrant des émotions complexes ? Comment ces expressions multimodales d'émotions complexes issues des vidéos d'origines ou générées dans des animations d'agent sont-elles perçues par différents utilisateurs ?

Ces expériences ont permis la définition de plusieurs niveaux de représentation des émotions et des paramètres comportementaux multimodaux apportant des informations pertinentes pour la perception de ces émotions complexes spontanées. En perspective, les outils développés durant cette thèse (schémas d'annotation, programmes de mesures, protocoles d'annotation) pourront être utilisés pour la conception de modèles qui permettront à des systèmes interactifs affectifs de détecter/synthétiser des expressions multimodales d'émotions spontanées.

TABLE DES MATIERES

Partie 1) Introduction

1.1 *Contexte*

A la suite d'un Master Recherche en Sciences Cognitives obtenu à l'Université Paris-Sud, j'ai travaillé pendant un an sur la spécification d'agents conversationnels animés (ACAs) dans le cadre du projet européen NICE (développement d'un jeu interactif éducatif). Les comportements étudiés se limitaient à des références à des objets via des gestes déictiques. J'ai eu ensuite l'opportunité d'intégrer le projet européen HUMAINE pour effectuer une thèse sur la représentation de comportements plus riches que ceux que j'avais étudié jusqu'alors, les émotions spontanées.

Dans des perspectives à long terme de conception de systèmes interactifs "intuitifs", les interfaces homme-machine (IHM) multimodales visent à permettre à l'utilisateur d'interagir naturellement avec un système via plusieurs modalités de communication comme la parole, les expressions faciales, les gestes de la main, et les mouvements du corps. En effet, l'être humain combine spontanément ces différentes modalités lors de comportements communicatifs ou émotionnels. Il peut également détecter les états émotionnels de ses semblables à partir de ces indices multimodaux. Il n'existe pas ou très peu d'interfaces homme-machine à l'heure actuelle qui gèrent les états émotionnels. Ce domaine est en émergence, comme en témoigne le nombre de conférences, projets qui s'intéresse à ce nouveau domaine.

Les systèmes interactifs intuitifs de demain devront donc être capables de détecter ces signes multimodaux et émotionnels émis par les utilisateurs mais aussi de générer de tels comportements via des personnages virtuels dotés de capacités communicatives (ou "Agents Conversationnels Animés"). De telles interfaces ont déjà montré des résultats d'évaluation positifs par exemple dans des prototypes d'applications éducatives ou ludiques dans lesquelles la dimension émotionnelle de l'interaction est particulièrement pertinente. Ceci nécessite de pouvoir représenter informatiquement les relations entre les différents niveaux impliqués : émotions, comportements multimodaux, contexte, etc. D'autres applications sont envisageables dans le domaine des robots compagnons (Cañamero et al. 2006), des communications, de la sécurité, de la santé ou encore du jeu vidéo et des arts interactifs.

1.2 *Objectifs de la thèse*

1.2.1 Problématique

Les travaux existants sur l'expression spontanée et multimodale des émotions sont rares. La majorité des études sur les comportements émotionnels est limitée aux émotions de base (colère, joie, tristesse, dégoût, peur, surprise (Ekman, 2003)), utilise des données actées (jouées par des acteurs) collectées en laboratoire dans des contextes restreints, et se limite à une modalité (parole, expressions faciales ou mouvements du corps). Ainsi, une des équipes de psychologues les plus avancées dans le domaine (Swiss National Research Center in Affective Sciences : http://www.unige.ch/fapse/emotion/members/scherer/scherer.html) continue à utiliser des données actées pour tester des hypothèses théoriques sur les émotions et leur expression (Scherer et Ellgring, 2007) (Bänziger et al., 2006). Analyser le comportement spontané de l'utilisateur, ou générer des personnages réalistes dans des applications narratives nécessite des connaissances en comportement spontané. Il existe encore peu de travaux sur les relations entre les émotions et la coordination de leurs expressions dans plusieurs modalités, et peu de travaux sur les émotions non actées. L'équipe de psychologues qui a coordonné le réseau HUMAINE (Roddy Cowie et Ellen Douglas-Cowie), et avec qui nous avons collaboré aucours de cette thèse, commence à travailler sur des données réelles après avoir principalement travaillé sur des données induites.

1.2.2 Objectifs

L'objectif de cette thèse est de contribuer à la définition de représentations pour les émotions spontanées et les comportements multimodaux associés. Ces représentations peuvent servir à l'étude expérimentale des comportements émotionnels, à la spécification d'agents conversationnels animés (ACAs), et à long terme à la détection automatique des émotions.

L'approche choisie consiste à collecter et annoter un corpus vidéo exploratoire d'interviews télévisées. Ce type de corpus comporte des émotions plus complexes que les 6 émotions de base : on y observe des superpositions, masquages, conflits entre plusieurs émotions.
Nous cherchons donc à identifier les niveaux de représentation et les paramètres comportementaux apportant des informations pertinentes pour la perception de ces émotions complexes.
Cette étude soulève plusieurs questions : Comment collecter, annoter et représenter des émotions complexes ? A quels niveaux d'abstraction et à quel niveau temporel ? Comment

représenter le contexte d'une émotion ? Quels sont les indices pertinents dans les comportements expressifs ? Quels sont les paramètres des comportements multimodaux pertinents pour la perception des émotions ? Quelles sont les relations entre les émotions et les comportements multimodaux associés ? Comment appliquer ce type de connaissance à la spécification d'agents animés expressifs montrant des émotions complexes ? Comment ces expressions multimodales d'émotions complexes issues des vidéos d'origines ou générées dans des animations d'agent sont-elles perçues par différents utilisateurs ?

Nous détaillons ci-dessous les différents objectifs conceptuels et méthodologiques que nous nous sommes fixés pour apporter des réponses à ces questions via une approche pluridisciplinaire.

Objectifs conceptuels

- ***Définir des schémas de représentation multi niveaux des comportements émotionnels multimodaux spontanés***

La représentation de comportements émotionnels multimodaux spontanés nécessite la définition d'un schéma multi niveaux car différents niveaux de représentation se révèlent nécessaires. Les théories classiquement utilisées pour représenter les émotions sont les catégories verbales (*labels* / *étiquettes* d'émotions joie, colère, tristesse…), les dimensions abstraites (activation, valence, intensité…) ou encore les variables d'évaluation des événements déclenchant l'émotion. Les deux premières théories sont classiquement utilisées pour l'annotation des émotions. Des schémas combinant ces deux représentations ont déjà été mis en œuvre. Les variables d'évaluation des événements causant l'émotion font l'objet de nombreuses études théoriques, mais n'ont pas encore fait l'objet d'annotations manuelles. Nous avons adopté, comme nous le détaillerons plus loin dans ce document, une approche exploratoire. Nous ne nous sommes pas limités à une seule de ces théories mais avons opté pour l'exploration de trois approches théoriques des émotions (catégories, dimensions, appraisal). Pour l'annotation des comportements multimodaux, les paramètres expressifs des différentes modalités susceptibles d'apporter une information quant à l'émotion produite doivent être pris en compte : par exemple les paramètres expressifs des gestes et les expressions faciales. Qu'en est-il pour des comportements spontanés observés dans des vidéos d'interviews télévisées ? Quelles dimensions sont pertinentes ? Utilisables ? Fiables ?

- ***Permettre le calcul des relations entre les niveaux de représentation***

Quelles analyses peut-on faire à partir de ces annotations pour atteindre nos objectifs? L'analyse temporelle des annotations à différents niveaux permettra par exemple de trouver différentes relations: relations entre étiquettes émotionnelles et dimensions (abstraites ou d'évaluation), relations entre étiquettes émotionnelles et comportements multimodaux, relations inter vidéos (comparaison des dimensions et des comportements multimodaux entre étiquettes identiques dans 2 vidéos différentes)… Quelles sont les mesures les plus pertinentes pour des émotions complexes observables dans des interviews télévisées ?

Au niveau de la génération de comportement émotionnel multimodal, l'animation d'un ACA à partir de ces relations, et des tests perceptifs sur ces animations pourront permettre de comparer les annotations effectuées sur des vidéos réelles et celles effectuées sur des animations afin de valider les annotations.

Objectifs méthodologiques

- ***Intégrer différentes sources de connaissances sur les émotions : annotation manuelle, annotation automatique, tests perceptifs sur les vidéos originales, tests perceptifs sur des animations d'ACA***

Deux techniques de traitement de données sont généralement utilisées dans l'étude de corpus de comportements multimodaux observés dans des vidéos : l'annotation manuelle et la détection automatique (audio, traitement d'image). Le choix de l'une ou l'autre dépend de l'objectif de recherche, de la taille du corpus traité, ainsi que de la finesse d'annotation recherchée.

L'annotation manuelle des émotions est basée sur la perception subjective, par un sujet, des comportements expressifs d'un individu. Sachant que les annotations manuelles doivent être validées pour être utilisables, quels sont les modes de validation possibles ? Les résultats d'annotation manuelle peuvent-ils être validés par des résultats extraits de la détection automatique ? Et vice-versa ? Des tests perceptifs effectués sur des agents animés expressifs qui rejouent les comportements observés dans des vidéos peuvent-ils valider les annotations de ces vidéos ?

Deux méthodes d'annotations manuelles sont généralement utilisées, l'annotation continue ou l'annotation discrète. L'annotation continue fournit des informations fines sur les variations temporelles des paramètres annotés. L'annotation discrète permet d'apporter des informations

au niveau de segments temporels définis par l'annotateur, ou encore au niveau global d'une vidéo.

L'annotation manuelle est coûteuse en temps d'annotation. Les systèmes de détection automatique ont l'avantage de pouvoir traiter rapidement de gros corpus mais sont limités dans le sens où ils doivent se focaliser sur très peu de paramètres, par exemple la localisation du visage, la mesure de la quantité de mouvement dans une vidéo, ou encore l'extraction de paramètres prosodiques à partir d'un fichier audio. Ils sont généralement appliqués à des données actées récoltées en laboratoire avec une haute résolution vidéo et audio. Leur application à des interviews télévisées de basse résolution mais plus spontanée reste donc à explorer. Les systèmes de détection automatiques utilisent des algorithmes (par exemple pour la parole, des algorithmes et outils permettent de détecter la fréquence fondamentale maximum « F0 Max » ou moyenne « F0 Range », qui déterminent la hauteur du son).

Des tests perceptifs peuvent être effectués, tant au niveau des vidéos originales qu'au niveau d'animations avec un ACA, pour valider les annotations, ou comparer la perception de différents groupes d'individus. Ces groupes servant à étudier les différences individuelles peuvent être définis selon différents critères : sexe, âge, trait de personnalité, profession, expertise dans le domaine des émotions.

- ***Définir un cadre méthodologique et logiciel : plateformes, protocoles, outils***

Pour pouvoir effectuer les annotations, il est nécessaire de définir un schéma de représentation adapté au corpus à traiter. Un protocole d'annotation doit également être défini: choix des sujets (combien, tranche d'âge, niveau d'étude…), choix de/des outils d'annotation (le schéma de représentation étant implémenté en fonction de l'outil utilisé : Praat (Boersma & Weenink), Anvil (Kipp 2001), FeelTrace (Cowie et al. 2000). Ces outils peuvent être combinés. Anvil permet d'importer des informations audio provenant de Praat (transcriptions, intensité, fréquence fondamentale…). Le traitement des données collectées à partir de ces différents outils peut être assuré par programme informatique.

1.2.3 Revue des différentes questions sur un exemple illustratif

Dans une des interviews que nous avons collectées (extrait n°3), une femme réagit à une décision prise par la justice, et qu'elle considère injuste (son père et son frère ont été emprisonnés). Voici une transcription de cet extrait vidéo montrant les élisions phonétiques et des exemples d' « affect burst » ([breath], [sil]) :

> *« qu'on me r'donne mon père et qu'on m'relache mon ptit frère [sil] parce que c't'un gamin on lui brise sa jeunesse car on est en train d'lui briser sa jeunesse [breath] et mon père mon père qui a travaillé pendant plus d'trente ans [breath] été comme hiver dehors [breath] se r'trouver là comme un malpropre c'est quand même scandaleux moi j'veux qu'on m'rende mon père moi [breath] on m'l'a pris mon père l'jour qu'on m'a pris mon père on m'a pris moi en même temps parce que j'étais toujours avec mon père [breath] tout l'temps [breath] et c'est ce qui me fait plus mal j'aurai préféré qu'on [breath] qu'on m'prenne moi qu'on m'prenne mon père j'aurai préféré être à sa place aujourd'hui [breath] [sil][sigh] »*

Cette vidéo contient des comportements émotionnels complexes (superposition de colère et désespoir) et des indices multimodaux variés (pleurs, gestes répétés, expressions faciales, etc.). Le premier problème rencontré concerne la segmentation : à quel moment commence et finit un segment émotionnel ? Comment s'articulent les segments successifs? L'étude de cet extrait vidéo soulève également des questions telles que : Comment prendre en compte plusieurs émotions qui se chevauchent, les variations d'intensité d'une émotion au cours du temps ? Comment étiqueter les comportements ? Quelles dimensions du comportement sont pertinentes ? Comment annoter les mouvements expressifs pertinents du corps humain ? Quelles sont les informations récupérables par traitement automatiquement de la voix et de l'image ? La personne n'est pas statique devant la caméra, elle pivote, se retourne, et continue à effectuer des gestes. Le schéma de représentation doit prendre en compte ces différences par rapport à des comportements plus contrôlés collectés en laboratoire.

En dehors des difficultés d'annotation, le problème de l'éthique est également à prendre en compte, les personnes impliquées dans le corpus doivent rester anonymes, et la diffusion du corpus n'est pas possible.

L'animation à l'aide d'un ACA, à partir des annotations émotions et comportements multimodaux sur cette vidéo, se doit d'être « proche » de la vidéo originale. L'agent doit-il cependant rejouer tous les comportements multimodaux du personnage (mêmes gestes des

mains, mêmes mouvements de tête) ? Peut-on percevoir des variations d'expression corporelle entre plusieurs animations (effectuées avec des paramètres expressifs différents) d'un même extrait ? Peut-on retrouver, parmi plusieurs animations différentes, celle qui a été générée à partir des annotations de la vidéo originale ? Quel niveau de fidélité est nécessaire pour obtenir une perception similaire ?

1.3 *Approche développée durant la thèse*

Les objectifs décrits à la section précédente ont demandé une approche pluridisciplinaire couvrant des aspects théoriques, conceptuels, méthodologiques et logiciels. La méthodologie expérimentale utilisée s'appuie sur une approche empirique de collection et annotation de données pour constituer des corpus permettant de construire des systèmes de détection/synthèse. Nous en détaillons ici les différentes étapes.

1.3.1 Collection de vidéos émotionnelles spontanées d'interviews télévisées

La première difficulté consiste à trouver les sources vidéo contenant des comportements émotionnels spontanés et riches en comportements multimodaux. Quels sont les choix possibles ? Enregistrement en laboratoire ? extraits issus de la télévision ? Si oui, faut-il collecter des extraits provenant d'émissions de télé réalité ? De reportages ? D'interviews de journaux télévisés ? Les extraits doivent-ils contenir des monologues ou des interactions entre plusieurs personnes? Comment la caméra doit-elle cadrer le/les personnages ? Nous détaillerons plus loin l'ensemble de critères que nous avons définis pour sélectionner les vidéos pour notre étude.

1.3.2 Exploration de différents schémas d'annotation

L'approche que nous avons définie dans cette thèse a consisté à explorer plusieurs directions en définissant différents protocoles d'annotation de données d'interviews télévisées à la fois au niveau des émotions et de leurs expressions à travers différentes modalités.

Sept expériences d'annotation ont été réalisées, 4 portant sur l'annotation des émotions et 3 sur l'annotation des comportements multimodaux. Différentes itérations d'annotation / analyse / amélioration des schémas de codage ont été effectuées. Cette approche a permis de

définir expérimentalement des schémas de codage pour l'annotation des émotions spontanées et des comportements multimodaux expressifs. Pour ces expériences, des sujets ayant des niveaux d'expertise différents (dans le domaine des émotions et de la multimodalité) ont annoté des extraits vidéo provenant du corpus. Nous avons évalué l'expertise (naïf / intermédiaire / expert) selon plusieurs critères : la formation initiale (psychologie / sciences cognitives / autre), la profession (étudiant - ingénieur - chercheur dans le domaine / autre), la durée de travail dans le domaine (ponctuel / mois / années).

Les différences entre les schémas d'annotations concernent le choix: 1) des étiquettes verbales : texte libre vs. liste finie d'étiquettes (avec différentes manières de sélectionner ces étiquettes); 2) des dimensions abstraites : intensité/valence/contrôle/activation ; 3) de combiner ou non plusieurs étiquettes ; 4) d'affecter ou non une variation temporelle d'intensité à une émotion dans un même segment ; 5) des dimensions contextuelles ; 6) des indices multimodaux. Nous avons mis en place des protocoles de validation des annotations : mesures de cohérence entre les différentes représentations, Kappa multi label. Ces différentes expériences nous ont permis d'explorer la complexité de l'annotation et de la représentation de données émotionnelles spontanées.

1.3.3 Démarche d'analyse/synthèse avec un agent expressif

Nous avons proposé d'utiliser la génération de comportement émotionnel multimodal à l'aide d'un ACA pour tester la fiabilité des annotations et la validité de notre approche en demandant à des sujets d'annoter non plus les vidéos originales mais les animations spécifiées à partir des annotations. Les ACAs permettent en effet ces études perceptives contrôlées puisqu'on peut sélectionner individuellement les modalités à tester.

1.4 *Cadre national et international*

Ces travaux ont été effectués dans le département Communication Homme-Machine du LIMSI (Laboratoire d'Informatique et de Mécanique pour les Sciences de l'Ingénieur), laboratoire du CNRS. Cette thèse a été co-encadrée par Jean-Claude Martin du groupe AMI (Architectures et Modèles pour l'Interaction), et Laurence Devillers, du groupe TLP (Traitement du Langage Parlé).

Notre thème de recherche nécessitait une approche pluridisciplinaire, et a demandé de l'expertise dans les domaines de l'informatique (modéliser, représenter, analyser et générer les comportements émotionnels); de la psychologie cognitive (coopération avec Stéphanie Buisine du LIMSI/ENSAM pour les analyses statistiques de certains tests perceptifs) ; mais également du traitement du signal audio et vidéo (détection automatique d'indices audio et vidéo).

Cette thèse a été financée par le réseau d'excellence européen HUMAINE (Human-Machine Interaction Network on Emotion http://emotion-research.net), dans le cadre du 6ème programme cadre de recherche (2004-2008). HUMAINE a pour but l'étude des émotions dans les interactions homme machine, la création d'une nouvelle communauté de recherche interdisciplinaire, et l'avancement de l'état de l'art sur les émotions. Le projet regroupe 33 partenaires provenant de 14 pays, et est divisé en 8 domaines thématiques : 1) les théories et modèles d'émotion, 2) des signaux aux signes d'émotion (détection automatique), 3) bases de données émotionnelles, 4) émotion dans l'interaction (ACA), 5) l'émotion dans la cognition et l'action, 6) l'émotion en communication et persuasion, 7) l'évaluation de systèmes orientés émotions, et enfin 8) les questions d'éthique.

Nous avons collaboré plus étroitement avec trois partenaires du réseau HUMAINE : l'équipe de Catherine Pelachaud du laboratoire LINC de l'Université Paris 8 (agent expressif) ; l'Université de Belfast (QUB) qui a collecté un corpus d'interviews télévisés en anglais et avec qui nous avons coopéré, dans le cadre d'une de nos expériences, sur la définition intégrée d'un schéma de codage avec des dimensions continues et contextuelles ; et enfin l'équipe d'ICCS (Grèce) spécialisée sur le traitement automatique de vidéos émotionnelles, pour étudier les liens entre indices extraits automatiquement avec leur système et nos annotations manuelles.

1.5 *Plan de la thèse*

La structure du rapport se présente sous la forme suivante.
Le chapitre 2 présente les travaux existants dans les domaines des émotions et de la multimodalité.
Le chapitre 3 décrit EmoTV, le corpus d'interviews télévisées que nous avons collecté, avec entre autres les critères de sélection des extraits et une description quantitative du corpus.

Les phases d'annotation sont présentées dans la section 3.3 en ce qui concerne l'annotation des émotions, et dans la section 3.4 pour l'annotation des comportements multimodaux. Le traitement automatique d'image (effectué en collaboration avec le laboratoire ICCS en Grèce, un partenaire de réseau HUMAINE) a été appliqué à notre corpus pour comparer/valider les annotations manuelles, cette étude est décrite en 3.5.

La méthode de copie-synthèse que nous avons développé et l'étude perceptive que nous avons menée à l'aide de l'agent expressif Greta (Pelachaud, 2005), utilisé pour la génération de comportement émotionnel multimodal pour tester la fiabilité des annotations, est décrite dans la section 3.6.

Enfin, nous concluons nos travaux et décrivont les perspectives qu'ils ouvrent pour l'étude des émotions spontanées et de leur expression multimodales

Partie 2) Etude de l'existant

2.1 *Plan*

La structure de cette partie sur l'existant se présente sous la forme suivante :
La section 2.2 décrit les recherches dans le domaine des émotions : les études sur les théories des émotions (leur représentation) et les corpus émotionnels existants. La section 2.3 concerne la communication humaine multimodale et émotionnelle. Plus précisément, elle présente les travaux existants sur les différentes modalités de communication humaine, liés ou non aux émotions : gestes, expressions faciales, parole ; mais également sur l'expressivité des mouvements, et enfin les corpus multimodaux. La section 2.4 explore les travaux sur les agents conversationnels. Enfin, la section 2.5 résume les limites de cette étude de l'existant en terme d'expressions multimodales des émotions spontanées.

2.2 *Émotions*

2.2.1 Définition

Il existe plusieurs définitions pour le terme « émotion ». Scherer (2000) définit une émotion comme « un épisode de changements synchronisés dans plusieurs composants en réponse à un événement ayant une signification majeure pour l'organisme ». Ces composants de l'émotion sont : le sentiment subjectif, l'expression motrice, les tendances à l'action, les changements physiologiques et le traitement cognitif. L'émotion d'après Caffi & Janney (Caffi & Janney, 1994) est un « phénomène empirique, généralement transitoire et d'une certaine intensité qui se manifeste par des indices dans la voix, le visage, les gestes etc. et par des changements physiologiques ».
Une définition absolue de l'émotion ne peut donc être formulée. Toutefois, trois composantes sont généralement acceptées par tous comme constituants de la réaction émotionnelle : la réaction physiologique, l'expression émotionnelle et le sentiment subjectif « feeling ».
Dans cette thèse, le terme émotion est utilisé au sens large et inclut les états effectifs décrits par Scherer : émotion, humeur, attitude, etc. Nous utiliserons la définition de Caffi pour associer les comportements émotionnels identifiés grâce à des indices expressifs multimodaux à des descripteurs linguistiques des émotions.

2.2.2 Théories de la représentation des émotions

Il y a beaucoup de réponses à la question « Qu'est-ce qu'une émotion ? » et peu de consensus mais les recherches s'accordent toutefois sur l'existence d'un ensemble minimal d'émotions primaires (dites de base) dont le nombre varie selon les chercheurs. Les émotions sont généralement classifiées en émotions primaires et universelles (peur, colère, surprise, joie, tristesse, dégoût, etc.) (innée) et émotions secondaires ou sociales (honte, fierté, amour, etc.) (acquises).

Plusieurs types de représentation sont généralement utilisés en recherche sur les émotions et sont issues des trois théories suivantes: les catégories verbales, les dimensions abstraites, et les dimensions d' « appraisal ».

Les catégories verbales incluent les étiquettes « primaires » (colère, peur, joie, tristesse, etc.) (Ekman, 1999) et les étiquettes « secondaires » pour les émotions sociales (amour, soumission). (Plutchik, 1994) a également combiné des émotions primaires, produisant d'autres labels pour les émotions « intermédiaires ». Par exemple, l'amour est une combinaison de joie et d'acceptation. La plupart des études sur la modélisation d'émotions ont utilisé un ensemble minimum de labels pour être gérables statistiquement (Batliner et al., 2000 ; Devillers et Vasilescu, 2004, Devillers et al., 2005).

Au lieu d'utiliser ce nombre limité de catégories, certains chercheurs définissent les émotions avec des dimensions abstraites continues telles que les axes Activation-Evaluation (Douglas-Cowie, et al., 2003) ou les axes Intensité-Evaluation (Craggs et Wood, 2004). Mais ces dimensions n'apportent pas toujours une représentation suffisamment précise de l'émotion, par exemple il est impossible de distinguer entre peur et colère.

Enfin, la théorie de l'évaluation cognitive, également nommée théorie de « l'appraisal », considère que l'émotion est le résultat des évaluations cognitives qu'un individu effectue au sujet d'un évènement (Scherer, 1984). Ce modèle d'appraisal est un modèle de perception/production de l'émotion, utilisé en psychologie cognitive. Une hypothèse commune à de nombreux chercheurs (Scherer, 1984; Weiner et al. 1979 ; Smith et Ellsworth, 1985) est que les émotions sont déclenchées à partir de stimuli. Scherer a défini un ensemble de critères, qu'il a nommé SECs (pour Stimulus Evaluation Checks), pour évaluer un évènement. Cette évaluation permet de prédire le type et l'intensité d'une émotion déclenchée par un évènement. L'avancée majeure dans cette théorie est la spécification détaillée des

dimensions d'appraisals utilisées pour l'évaluation séquentielle des événements antécédents à l'émotion (Scherer, 2000) : nouveauté (soudaineté, familiarité, prédictibilité), agrément (intrinsèque, global), significations des buts/besoins (causalité : agent, causalité : motivation, degré de certitude dans la prédiction des conséquences, inadéquation avec les attentes, opportunité, urgence), potentiel de maîtrise (contrôlabilité de l'évènement direct, contrôlabilité de l'évènement conséquence, puissance, ajustement), compatibilité avec les standards (standards externes, standards internes). Un autre modèle fréquemment utilisé (notamment dans les agents animés) est le modèle OCC (Ortony et al. 1988) qui est définit les émotions d'une personne comme des réactions à des événements concernant cette personne, les actions des autres personnes ou des objets.

Le choix d'une unique théorie pour notre étude exploratoire des émotions complexes spontanées n'est pas envisageable. Nous explorerons ces théories et leur combinaison. L'utilisation de plusieurs approches d'annotation pourra servir à valider des annotations subjectives.

2.2.3 Corpus émotionnels

Différentes communautés s'intéressent depuis longtemps à la collecte et à l'annotation de corpus de texte, parole, dialogues (action ASILA[1] Interaction Langagière et Apprentissage, action CATCOD[2] pour le catalogage et le codage de corpus oraux, Centre de Ressources CRDO pour la Description de l'Oral[3]) et plus récemment aux corpus multimodaux (Martin et al. 2006).

A l'heure actuelle, il existe peu de corpus émotionnels de données réelles. La plupart des travaux effectués sur les émotions utilisent des enregistrements vidéo dans lesquels des acteurs expriment des comportements émotionnels basiques. Nous décrivons ici quelques exemples de différents types de données récoltées selon différents protocoles : actées ou induits en laboratoire, extraits de fiction (cinéma), réunions de travail, et interviews télévisées. Pour une vue récente des différents corpus émotionnels, le lecteur pourra se reporter aux actes du workshop LREC sur ce thème (Devillers et al. 2006). La plupart de ces corpus sont récent et sont en cours d'exploitation pour l'étude des relations entre les émotions et leurs expressions.

Corpus Audio/Vidéo

Corpus actés

Le corpus GEMEP (Geneva Multimodal Emotion Portrayals) contient des portraits d'expressions émotionnelles actées (Bänziger et al. 2006). L'équipe de l'Université de Genève a collecté ce corpus à partir de performances réalisées par des acteurs entraînés et dirigés par un véritable metteur en scène. La Figure 1 montre un aperçu du corpus collecté. Afin de neutraliser l'impact du contenu sémantique, les acteurs doivent prononcer une suite de phonèmes n'ayant pas de signification en jouant différentes émotions avec différentes intensités.

[1] http://www.loria.fr/projets/asila/corpus.html

[2] http://icar.ens-lsh.fr/wiki/index.php/Catalogage_et_codage_de_corpus_oraux

[3] http://crdo.up.univ-aix.fr/index.php?langue=fr

Figure 1: Corpus GEMEP (Bänziger et al. 2006)

Nous décrirons plus loin une autre étude faisant intervenir des enregistrements vidéo d'acteur et se focalisant sur l'expression et la perception des mouvements (Wallbott 1998) mais n'ayant pas exploité de technologies de corpus numériques.

Corpus de fiction

Le corpus SAFE (Situation Analysis in a Fictional and Emotional Database) contient 400 séquences vidéos (8 secondes à 5 minutes) extraites de 30 films de fiction récents sur support DVD, pour un total de 7h d'enregistrements (Clavel et al. 2006). L'anglais américain est le plus souvent présent dans les extraits (70% du corpus). Ce corpus a été utilisé pour étudier les problèmes d'annotation dans le contexte de la détection acoustique des émotions de type « peur », pour des applications de surveillance. Les annotations effectuées par 2 personnes (validées par une 3ème personne au niveau acoustique) ont révélé deux stratégies : la première orientée sur le contexte et les informations audio et vidéo pour la catégorisation de l'émotion ; et la deuxième uniquement basée sur l'audio. Les extraits ont été répartis dans 2 groupes : situation normale et situation anormale (tremblement de terre, incendie, inondation, ou encore agressions physiques : kidnapping, prise d'otage…).

Corpus de réunions de travail

Le corpus AMI (Heylen et al. 2006) contient des enregistrements audiovisuels de réunions (Figure 2). La majorité des réunions étant des regroupements de 4 personnes jouant chacune un rôle différent. La taille du corpus est de plus de 100 heures. Son objectif principal est le

développement de plusieurs types de « technologies de réunions » (comme par exemple la possibilité d'accéder aux enregistrements effectués.

Figure 2: Corpus AMI (Heylen et al. 2006)

Le corpus AMI a été utilisé par des chercheurs de l'Université de Twente (Reidsma et al. 2006) pour tester deux méthodes d'annotation des émotions et des états mentaux. La première méthode effectue une labellisation continue des dimensions émotionnelles. La deuxième méthode consiste en une labellisation discrète des segments à l'aide de labels catégoriques.

Le corpus de meeting ISL a été annoté par 3 personnes avec un schéma de codage défini de manière à recueillir les comportements pertinents au niveau émotionnel (Laskowski et Burger, 2006). Entre autre, la valence a été annotée et des accords inter annotateurs ont été calculés. Les relations entre valence et comportements ont été étudiées, mais également les relations inter locuteurs (corrélation entre valence du locuteur et comportement du locuteur précédent).

Le projet CHIL a également donné lieu à la collecte d'un corpus de réunions de travail et des annotations multimodales (extraction de mouvements par traitement d'images) et d'annotations manuelles du rôle des participants (Pianesi et al. 2006).

Corpus d'interviews télévisées

Belfast Naturalistic Database (BND) est un des rares corpus émotionnels « naturels » existant à l'heure actuelle (Douglas-Cowie et al. 2000). Il contient 298 extraits vidéo, avec 125 personnes différentes (31 hommes, 94 femmes). Chaque personne apparaît au moins dans 2 extraits, un dans lequel son état est jugé relativement émotionnel, et un deuxième dans un état relativement neutre. Les extraits durent entre 10 et 60 secondes et contiennent des émotions provenant d'interactions de tous les jours mais également des exemples d'émotions fortes

telles que la colère ou la peur. Les vidéos proviennent principalement d'émissions télévisées (Figure 3).

Figure 3: Corpus Belfast Naturalistic Database (Douglas-Cowie et al. 2000)

Ces mêmes chercheurs ont également collecté les 3 corpus suivants.

Le corpus Casteway Reality TV, contient un total de 5h d'enregistrements vidéo provenant d'une émission de télé réalité nord irlandaise. Les extraits vidéo peuvent être répertoriés dans 5 catégories : interview individuelle d'un(e) candidat(e) hors épreuve, interview à chaud après une épreuve physique (Figure 4), interview de groupe, interaction entre un groupe d'individus, enregistrement lors d'une épreuve physique.

Figure 4: Corpus Castaway Reality TV de l'Université QUB

Corpus d'émotions induites récoltées en laboratoire

Le corpus SAL (Sensitive Artificial Listener), d'une durée totale de 10 heures (4 locuteurs enregistrés 20 minutes chacun), contient des comportements émotionnels induits en parlant avec un expérimentateur (Figure 5). Le locuteur interagit (« discute ») avec l'expérimentateur (qui simule 4 personnalités différentes). Chacune des 4 personnalités a un ensemble de réponses caractéristiques visant à modifier l'état émotionnel du locuteur. L'annotation continue des émotions a été effectuée à l'aide de l'outil FeelTrace (Cowie et al. 2000).

Figure 5: Deux exemples provenant du corpus SAL

Enfin, un corpus collecté à partir d'enregistrements de situations dans lesquelles 4 femmes (se connaissant toutes) participent à une série d'épreuves physiques en extérieur. Leurs comportements ont été filmés à l'aide de deux caméras tenues à la main et d'une caméra fixée sur la tête des participantes (filmant leur visage). Une première série d'enregistrements de 10 minutes a été effectuée, chaque caméra filmant le visage ou le corps des participantes dans les 2 tâches (observation des autres participantes et action).

Des chercheurs de la division Recherche et Développement des Technologies à France Télécom ont collecté un corpus d'expressions multimodales émotionnelles spontanées (Le Chenadec et al. 2006). Ce corpus contient des interactions entre un acteur virtuel et des sujets apprenant un texte de théâtre. Ces travaux font partie d'une étude sur la fusion d'informations multimodales (verbal, facial, gestuel, postural, et physiologique) visant à long terme à contribuer à la détection et la caractérisation d'expressions émotionnelles dans les interactions homme-machine.

Le corpus expressif Sound Teacher d'E-Wiz (Aubergé et al. 2004) a été collecté à partir d'enregistrements de 17 sujets (11 femmes, 6 hommes). Les sujets sont filmés (Figure 6) dans un scénario de magicien d'Oz. L'expérience consiste à apprendre à prononcer des voyelles dans des langues du monde entier, en utilisant un logiciel d'apprentissage des langues (contrôlé en réalité par le magicien).

Figure 6: Corpus Sound Teacher d'E-Wiz (Aubergé et al. 2004)

La Figure 7 montre un exemple d'annotation d'une vidéo du corpus à l'aide de l'outil d'annotation multimodale ELAN (Brugman et Russell., 2004).

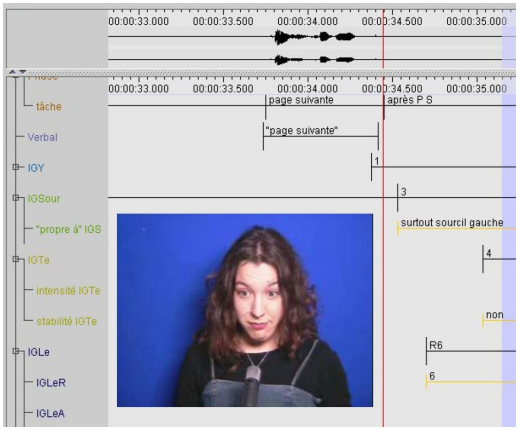


Figure 7: Exemple d'annotation d'un extrait du corpus Sound Teacher d'E-Wiz avec l'outil ELAN (Brugman et Russell., 2004)

Corpus Audio

GVEESS

Le corpus audio acté GVEESS (Geneva Vocal Emotion Expressions Stimulus Set), collecté à l'UNIGE (Université de Genève), contient 224 enregistrements (panel de 14 émotions). Les expressions sont en allemand, elles ont été enregistrées à Munich et digitalisées à Genève (Banse et Scherer., 1996).

Corpus Basque

Des chercheurs de l'Université du Pays Basque ont collecté un corpus audio émotionnel dans la langue Basque (Saratxaga et al. 2006). Ces données ont pour utilité l'étude de la synthèse audio basée sur le corpus ainsi que l'étude de modèles prosodiques pour les émotions. Le

corpus contient les mêmes textes enregistrés dans les 6 émotions de base plus la neutralité. Les enregistrements ont été effectués par 2 doubleurs professionnels (1 homme et 1 femme).

CREST

(Campbell, 2002) a collecté le corpus audio naturel CREST, contenant un large éventail d'états émotionnels et d'attitudes liées à des émotions. Ce corpus de données spontanées, enregistré dans 3 langues différentes (anglais, japonais, chinois) est collecté de la manière suivante : des volontaires enregistrent leurs interactions avec leur entourage par périodes tout au long d'une journée. L'objectif visé par Campbell, en terme de quantité, est d'atteindre 1000 heures d'enregistrement sur 5 ans.

AIBO (Erlangen Database)

(Batliner et al. 2004) ont enregistré des interactions entre des enfants et un AIBO (chien robot développé par Sony). 51 enfants allemands, âgés de 10 à 12 ans, ont participé à cette expérience. Ce corpus audio allemand contient près de 52 000 mots prononcés par les enfants pour diriger l'AIBO dans le cadre d'un parcours. La reconnaissance de la parole est simulée, et les réactions du robot (comme aller dans la direction opposée de celle ordonnée) ont pour objectif de générer des comportements émotionnels chez les enfants.

Centres d'appel

(Devillers et al. 2004, Devillers et al. 2006b) ont étudié les émotions dans plusieurs corpus audio provenant de centres d'appels. Le premier corpus provient d'un centre boursier (100 dialogues, 5000 tours de parole) ayant, d'après les résultats des annotations, 12% de tours de parole contenant des émotions (principalement la peur de perdre de l'argent). Le second corpus provient d'un centre bancaire (250 dialogues, 5000 tours de parole), dans lequel 10% des tours de parole contiennent des émotions (encore principalement la peur de manquer d'argent). Enfin, ces chercheurs ont étudié les émotions dans des enregistrements audio provenant d'un centre d'appels médical CEMO (Devillers et al., 2005, 2006c), contenant un plus large éventail de comportements émotionnels que les centres d'appels financiers. La peur dans ce cadre est liée à la peur de la maladie, ou même à la survie (agressions). Ces études ont mis en avant la présence d'émotions complexes mélangées dans des données naturelles. Des mélanges d'émotions positives et négatives ont été observés comme par exemple le soulagement et la peur pour l'appelant ou l'énervement et la compassion pour l'agent (Devillers et Vidrascu, 2006c), (Devillers, 2006).

ProGmatica

(Braga et al., 2006) ont collecté un corpus audio provenant d'émissions télévisées portugaises. Ce corpus de données spontanées se nomme ProGmatica. Son objectif est de combiner l'information prosodique dans un cadre pragmatique. Il est utilisé dans le but d'une part, d'analyser, de décrire et de prédire les patterns prosodiques impliqués dans les actes de dialogue et les discours ; et d'autre part de trouver des relations entre prosodie, pragmatique, émotion, style et attitude. L'approche utilisée consiste à sélectionner et extraire des actes de dialogue prototypiques à l'aide d'outils d'analyse de la parole. L'objectif à long terme de ces travaux est de fournir des dimensions pragmatiques et émotionnelles aux systèmes de synthèse vocale portugais.

NoTa

Le corpus audio norvégien NoTa (Johannessen et al., 2006) contient 2 millions de mots. Quelques objectifs lors de la création du corpus étaient d'avoir une représentation multimédia (transcription de la parole, audio et vidéo), d'inclure des situations variées de parole (en particulier de la parole contenant des émotions), et d'annoter le corpus de différentes manières (émotions, dialogue, grammaire, indices extralinguistiques).

Corpus d'Augsburg et de Bielefeld

Deux corpus audio, le premier contenant des enregistrements audio émotionnels joués par des acteurs, et le deuxième contenant des données spontanées provenant d'une expérience simulée par un magicien d'Oz, ont été utilisés par des chercheurs allemands des Universités d'Augsburg et de Bielefeld pour étudier l'impact du sexe du locuteur sur la reconnaissance des émotions dans la parole (Vogt et André., 2006). Ils ont observé que les systèmes de reconnaissance des émotions gérant les différences de sexe fonctionnent mieux que ceux qui ne font pas cette distinction. Les travaux effectués par ces chercheurs ont pour objectif de proposer un moyen d'améliorer encore la qualité discriminative des paramètres dépendant du sexe de la personne. Un système reconnaissant automatiquement si le locuteur est un homme ou une femme est utilisé avant d'exécuter le système de reconnaissance automatique des émotions. Les résultats ont montré un taux de reconnaissance de 90% pour la reconnaissance du sexe, et une amélioration de 2 à 4% de la reconnaissance des émotions en utilisant le système de reconnaissance du sexe au préalable.

Conclusion

L'étude des émotions soulève plusieurs questions théoriques mais aussi pratiques dans le cas des approches expérimentales à base de corpus (problèmes de collecte, d'éthique, de naturel d'une émotion, de dépendance contextuelle).

L'enregistrement de comportements joués par des acteurs a permis de fournir des réponses partielles à ces questions et montre l'importance de se tourner vers les corpus émotionnels plus spontanés et multimodaux.

Ces derniers nécessitent l'étude de la communication non-verbale que nous décrivons dans la section suivante afin de mieux comprendre comment elle est influencée par les émotions.

2.3 *Communication humaine multimodale et émotionnelle*

La communication non verbale joue un rôle central dans les comportements humains sociaux (Argyle, 2004 ; Knapp et Hall, 2006). Elle permet d'exprimer différentes fonctions communicatives (Poggi et al. 2000) : communiquer des attitudes interpersonnelles (par exemple établir et maintenir l'amitié à l'aide de signaux tels que la proximité, le ton de la voix, le toucher… ; accompagner et renforcer la parole ; ou encore exprimer des émotions et autres états mentaux (Baron-Cohen et al., 1996).

Il existe différents signaux non verbaux, (Argyle, 2004) : les expressions faciales (bouche, sourcils, rougeur ou blancheur de la peau, mouvements faciaux), le regard (la durée des regards, le niveau d'ouverture des yeux, l'expansion des pupilles), les mouvements (gestes et autres mouvements du corps), la posture (tendu ou relaxé, la droiture), le contact corporel, le comportement spatial, les habits (et autres aspects de l'apparence), les vocalisations non verbales (pitch, vitesse d'élocution, volume, rythme, perturbations), et enfin l'odeur. Ces signaux corporels peuvent être subtils et inconscients.

Nous nous focalisons ici sur les modalités pertinentes pour notre étude (notamment qui sont observables dans des interviews télévisées) : la parole, les gestes et mouvements expressifs, les expressions faciales.

2.3.1 Communication verbale et émotion

D'après les travaux de (Collier, 1985), bien que les indices vocaux associés à la parole semblent porter des émotions de la même manière que les indices non verbaux, le rôle du langage est assez différent. La grammaire et le contenu verbal représentent une méthode de plus haut niveau pour traduire les sentiments et pensées en mots. Les mots peuvent être choisis de manière directe, comme par exemple une personne disant à son interlocuteur qu'elle est en colère, ou alors plus indirecte, en utilisant des aspects plus subtils du langage. Lorsque le langage est utilisé pour exprimer une émotion, il représente une tentative pour décrire nos sentiments à nous même et aux autres. Les labels que nous utilisons pour décrire nos émotions ne sont pas les émotions elles mêmes. Et c'est parce que les descriptions verbales de nos propres émotions ne peuvent pas être prises en compte directement que les psychologues intéressés par les expressions émotionnelles se sont souvent focalisés sur des aspects plus subtils du langage, sur lequel le sujet a peu de contrôle conscient, par exemple la

répétition des thèmes de discussion, pouvant refléter ce en quoi le sujet est le plus intéressé, le plus préoccupé. Enfin les disfluences apparaissant dans la parole apportent des informations sur les sentiments réels du sujet. (Freud, 1901) a montré que les erreurs ont souvent lieu en raison de pensées en concurrence ou d'associations inconscientes.

Deux techniques d'analyse sont utilisées pour détecter des émotions dans la parole, l'analyse du contenu de la parole visant à identifier les thèmes récurrents, l'analyse en contingence permettant de déterminer la fréquence avec laquelle certains thèmes apparaissent ensemble.

2.3.2 Les vocalisations non verbales

L'état émotionnel a un impact direct à la fois sur la production de la parole et de la prosodie. Les comportements verbaux sont généralement composés de parole, mais parfois également d'écriture, ou de gestes reproduisant des lettres ou mots. La parole est accompagnée par un ensemble de signaux non verbaux, apportant des illustrations, feedback, et aidant la synchronisation. Certains de ces signaux font vraiment partie du message verbal, tels que les signaux prosodiques de rythme la fréquence fondamentale (F0), et l'emphase. D'autres signaux non verbaux sont indépendant du contenu de la parole, tels que les signaux paralinguistiques, par exemple des sauts de F0, des affects bursts (rire, souffle, etc.) ou des disfluences (Devillers et al., 2004) . Certains signaux non verbaux traduisent des émotions, des attitudes ou expériences qui ne sont pas facilement exprimables avec des mots. Nous savons que la voix peut exprimer une émotion, mais certains aspects de la prosodie sont déterminés par des contraintes linguistiques, telles que la montée du ton lors dans une phrase interrogative, ou l'emphase marquée par le stress (Feyereisen, 1991).

La fréquence fondamentale (ou pitch) est un indice de mesure globale de la voix, utile pour déterminer si une voix est aiguë ou grave. Elle est calculée sur les parties voisées du signal. Pour une femme, la F0 moyenne est de 250 Hz alors que pour un homme elle est estimée à 150 Hz. Elle peut être exprimée en plusieurs unités : en Hertz, échelle Mel et Bark. Les facteurs pouvant influencer la F0 sont à la fois physiologiques et sociaux. Ces facteurs sont la langue maternelle (origine géographique), l'âge, le sexe, le système hormonal (en période de puberté on non). Le contexte situationnel a également une très forte influence sur la F0 (discours, etc.). La F0 s'avère être un paramètre important pour la reconnaissance des émotions.

La communication des émotions à travers les vocalisations non verbales a été étudiée dans des expériences de simulation, dans lesquelles les sujets devaient par exemple lire un texte en se mettant dans un certain état émotionnel, ou encore à partir de scénarios joués, ou à l'aide de procédures d'induction pour produire les émotions recherchées. Mais en communication non verbale, les expressions jouées ne donnent pas les mêmes résultats qu'en situation réelle. Par exemple, d'après (Williams et Stevens 1972), la peur simulée augmente correctement la hauteur F0 maximum de la voix, mais n'augmente pas la hauteur F0 minimum contrairement à une peur réelle. Ce type de résultat est à nuancer car il existe beaucoup de peurs différentes et il est difficile d'isoler les facteurs exacts des différences constatées mais il est certain que des différences sont notables entre données réelles et artificielles. Pour un état de l'art plus détaillé, se reporter à (Devillers, 2006).

Dans le but de neutraliser les effets du contenu verbal de la parole, plusieurs méthodes sont employées, l'une d'entre elles consiste à lire des passages neutres émotionnellement. (Davitz et al, 1964) ont effectué plusieurs expériences, lors desquelles les sujets simulaient jusqu'à quatorze expressions émotionnelles en lisant des extraits neutres tels que « Je pars maintenant. Si quelqu'un appelle dis lui que je reviendrai bientôt ». Les résultats de ces expériences, ainsi que d'autres expériences similaires, montrent un pourcentage moyen de reconnaissance des émotions de l'ordre de 56% (ce qui est à peu près équivalent à la reconnaissance des expressions faciales). Les émotions négatives sont plus facilement reconnues que les émotions positives, dans l'ordre du mieux reconnu au moins bien : la colère, la tristesse, l'indifférence, et le mécontentement (Scherer, 1981). La joie et la haine sont plus faciles à reconnaître que la honte et l'amour. Des confusions apparaissent entre certaines émotions, en raison de leurs propriétés acoustiques similaires. (Scherer, 1986) a étudié l'encodage des émotions en terme de prosodie, les résultats pour certaines émotions donnent, pour la joie, une élévation des variables : pitch, pitch moyen dans une expression, variation du pitch, intensité, vitesse d'élocution ; pour le désespoir, une diminution des variables : pitch, pitch moyen, intonation, intensité, vitesse d'élocution.

(Kienast et Sendlmeier, 2000) ont étudié les modifications temporelles et spectrales dans les expressions vocales émotionnelles. Ils ont notamment mesuré la suppression de segments, qui se calcule à partir de la formule LMQ (Lautminderungsqutient) développé par (Hildebrandt, 1963) et permettant de comparer le nombre de segments effectivement produits dans la parole (expressions émotionnelles) avec le nombre de segments qui aurait du être produit dans le cas

d'une prononciation standard (parole neutre). La formule est la suivante :

$$LMQ = 10 - \frac{10 \cdot n_{emot}}{n_{neut}}.$$

Les valeurs positives représentent des suppressions de segments, tandis que les valeurs négatives indiquent des insertions de segments en comparaison avec la parole neutre. Les résultats de l'analyse acoustique montraient : des suppressions de segments plus fréquentes dans les expressions de peur et de tristesse, une articulation plus précise dans les expressions de colère en comparaison avec toutes les autres émotions, ainsi que la version neutre. Les voyelles sont prolongées dans les expressions de colère, ce qui explique la production plus précise des voyelles. Dans le cas de la tristesse et de l'ennui, les réductions de segments et formants sont bien visibles car la vitesse d'élocution est plus faible que pour les autres émotions ainsi que pour la version neutre.

(Devillers et al., 2004) ont montré la présence de plus de disfluences : hésitation ("euh") et silences pour des émotions de type peur que colère. De plus l'importance d'indices de type affect bursts (rire, souffle) est relatée dans (Devillers, 2006). (Campbell, 2004) a également montré l'importance du facteur de qualité vocale pour différencier les voix « grinçantes » et « respirées », pertinentes pour la détection des émotions.

(Mozziconacci, 2001) a souligné la présence de nombreuses variations en terme de prosodie dans les discours contenant des émotions, plus précisément au niveau du pitch, de la vitesse d'élocution, du rythme, de la qualité de la voix, du volume, et de l'articulation/prononciation. Les indices prosodiques apportent également des informations telles que le sexe du sujet, son âge, sa condition physique, mais aussi son point de vue, son émotion, et son attitude à propos du thème abordé, de son interlocuteur, ou de la situation.

2.3.3 Les gestes

Adam Kendon définit un geste comme « une action visible utilisée comme un énoncé ou une partie d'un énoncé » (Kendon 2004). Nous résumons ci-dessous quelques travaux concernant la structure temporelle des gestes (« phase ») et les catégories de gestes (« phrase ») (Kendon 2004, McNeill 2005).

Structure temporelle des gestes

(Efron, 1941) a introduit la notion de partition d'un geste en 3 **phases** ; la préparation, le stroke, et la retraction. (McNeill, 1992) a étendu et opérationnalisé cette structure. Basés sur

ce système, (Kita et al., 1998) ont développé des critères pour la segmentation des mouvements et l'identification des phases à l'aide de mesures objectives. Un geste commence lorsque les mains bougent à partir d'une position de repos, et se termine quand les mains reviennent en position de repos. La position de repos correspondant à la position des mains sur une partie du corps, un objet sur lesquels les mains peuvent être posées, ou encore les bras le long du corps. Pour la transcription de la structure d'un geste, il est nécessaire d'identifier les limites de chaque phase, puis de les classifier. La Figure 8 montre trois phases d'un geste (Kipp, 2004): la « préparation », partie allant de la position de repos jusqu'au déclenchement du stroke, puis le « stroke », qui est la partie la plus énergétique du geste (McNeill, 1992), et portant sa signification (Kendon, 1983), et enfin la « retraction », allant de la fin du stroke jusqu'à la position de repos. Les autres phases sont : le « hold » qui correspond au maintien d'une position au cours du geste (avant ou après le stroke), l' «independant hold » lorsqu'il n'y a pas de stroke, et la « retraction partielle», phase pendant laquelle les mains reviennent vers la position de repos mais retournent à une nouvelle phase de préparation avant de l'atteindre.

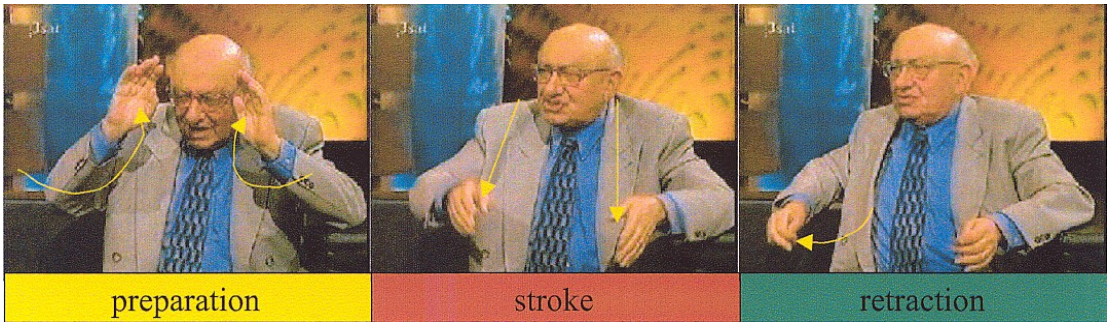


Figure 8: Exemple de transcription des phases d'un geste (Kipp, 2004)

Catégories de gestes

(McNeill, 2005) a utilisé des méthodes psycholinguistiques pour étudier les gestes communicatifs. Il a décrit les différentes catégories de gestes suivantes (appelées aussi fonctions des gestes ou types de gestes).

Les gestes **iconiques**, ayant une relation formelle avec le contenu sémantique de la parole : par exemple (Figure 9), une personne décrit une scène dans laquelle un personnage tire sur des branches pour se faire un passage dans la brousse ; le locuteur fait semblant d'attraper quelque chose et de le tirer sur le coté.

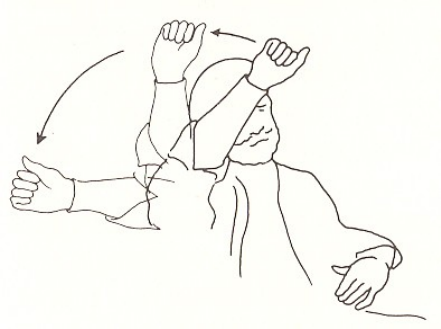


Figure 9: Geste iconique (McNeill, 2005)

Les gestes **métaphoriques**, étant imagés comme les iconiques, à la différence près que ce geste imagé représente une idée abstraite plutôt qu'un objet ou événement concret. Dans l'exemple de la Figure 10, le locuteur précise que ce dont il va parler est un dessin animé, et le geste qu'il effectue est métaphorique, dans le sens où il fait référence au concept de « dessin animé », en le présentant comme un container borné, porté par les mains.

Figure 10: Geste métaphorique (McNeill, 2005)

Les **beats**, également nommés « bâtons » (Efron, 1941 ; Ekman et Friesen, 1969). Le terme beat vient du fait que ce type de geste donne l'impression de battre le rythme musical. Les mains bougent en rythme avec la parole, sans être parfaitement synchrones (McClave, 1994). Contrairement aux iconiques et métaphoriques, les beats ont souvent la même forme de geste, peu importe le contenu de la parole (McNeill et Levy, 1982). Dans l'exemple de la Figure 11, l'interlocuteur dit « toutes les fois où elle le regarde », et effectue, en prononçant l'expression soulignée, le geste de lever légèrement la main gauche puis de la redescendre.

Figure 11: Beat (McNeill, 2005)

Les gestes **déictiques**, ayant la fonction d'indiquer non seulement des objets et événements dans le monde concret, mais également de pointer des objets et événement abstraits : par exemple, dans la Figure 12, le locuteur demande « d'où viens-tu ? » et pointe un endroit de l'espace abstrait, auquel le locuteur fait référence.

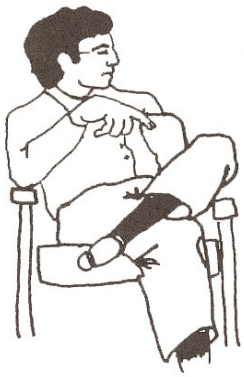


Figure 12: Déictique (McNeill, 2005)

Les « **unités gestuelles** » (gesture units) sont composées de plusieurs gestes qui s'enchaînent les uns aux autres sans retour à une position de repos (Figure 13).

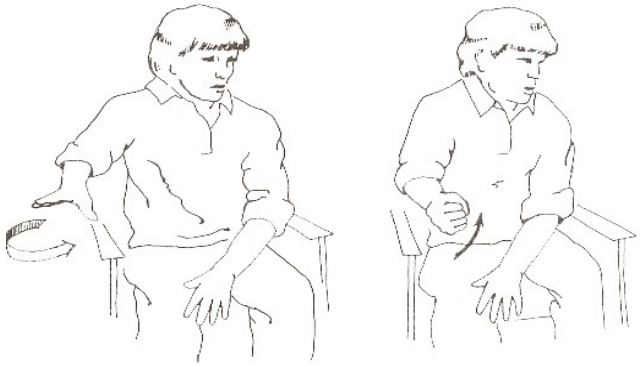


Figure 13: Unité gestuelle (McNeill, 2005)

La transcription des gestes, définie par McNeill, suit le schéma suivant : l'identification des gestes, puis l'identification des phases du geste, le codage du type du geste (Iconique, métaphorique, déictique, ou beat…), le codage de la forme du geste (si ce n'est pas un beat), le codage de la signification du geste.

Différentes classifications des gestes ont été proposées par plusieurs chercheurs. (Kipp, 2004) a utilisé un ensemble de 6 classes de gestes détaillés dans les travaux de (Efron, 1941), (Ekman et Friesen, 1969), et (McNeill, 1992). Ces classes sont : les adaptateurs, les emblèmes, les déictiques, les iconiques, les métaphoriques, et enfin les beats. La Figure 14 montre la classification des 6 classes utilisées par Michael Kipp.

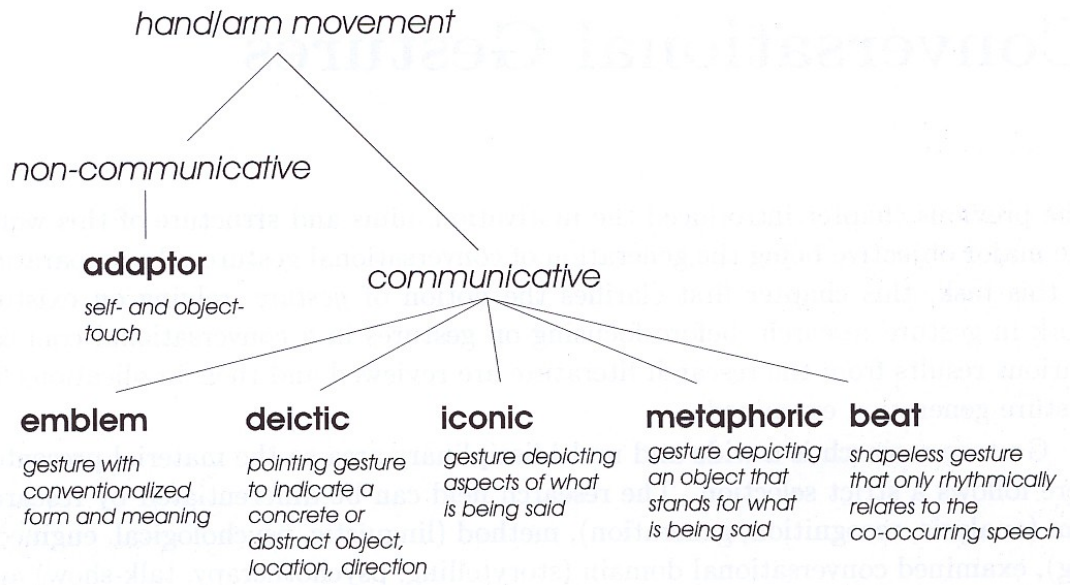


Figure 14: Les six classes de geste et leur définition (Kipp, 2004)

En plus des quatre classes : déictique, iconique, métaphorique, et beat, décrites par McNeill, Kipp a utilisé la classe « adaptateur » provenant des travaux de (Ekman et Friesen, 1969). Les adaptateurs sont des gestes non communicatifs, consistant à toucher un objet ou son propre corps (par exemple se gratter la tête). Plusieurs chercheurs ne considèrent pas les adaptateurs comme des gestes car n'étant pas conversationnels en raison de leur faible relation avec la parole (Kendon, 1983), McNeill (1992), Duncan (1983) et Webb (1997),. Les adaptateurs peuvent néanmoins nous informer sur l'état du locuteur. (Freedman, 1977) a par exemple précisé que ce type de geste signale un besoin de concentration. La Figure 15 montre 2 exemples d'adaptateurs.

Figure 15: Exemples d'adaptateurs (Kipp, 2004)

Enfin, la 6ème classe utilisée par Kipp est la classe emblème (Figure 16). Geste pouvant être utilisé indépendamment de la parole. La signification du geste peut être directement traduite en mots, ces derniers étant parfois prononcés en conjonction avec l'emblème. Etant donné qu'ils peuvent être utilisés en l'absence de parole, ils sont souvent employés lorsque le canal audio est restreint pour cause de bruit ou de distance par exemple. Les emblèmes ont été introduits par (Efron, 1941), puis repris par (Ekman et Friesen, 1969) et (Johnson et al., 1975). D'autres termes ont été utilisés par certains chercheurs pour décrire ces mêmes emblèmes, gestes autonomes (Kendon, 1983), gestes sémiotiques (Barakat, 1973), gestes symboliques (Bitti et Poggi, 1991).

Figure 16: Exemple d'emblème : «ok » former un anneau avec pouce et index (Kipp, 2004)

2.3.4 Expressions faciales et émotions

De nombreuses études ont été réalisées sur la mesure des mouvements faciaux, ces études ayant entre autre pour objectif de distinguer les mouvements signalant une émotion, ou de tester si les mouvements faciaux sont identiques dans des contextes sociaux similaires mais dans différentes cultures (Harrigan et al. 2005).

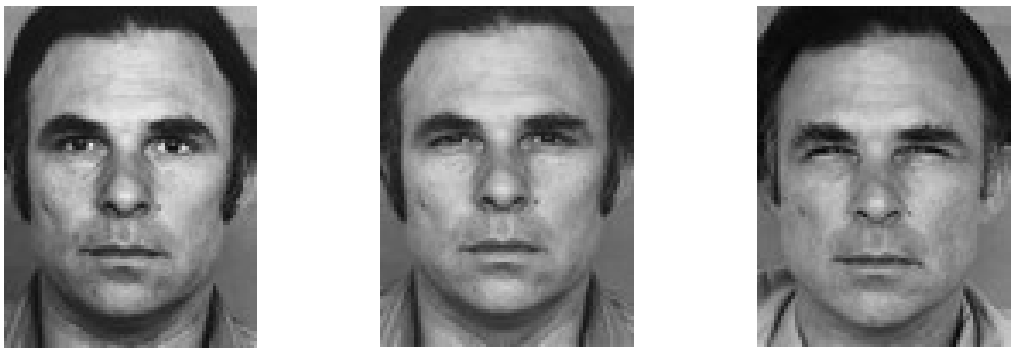
(Darwin, 1872/1965) a tenté de comprendre les expressions émotionnelles des humains en observant leur développement pendant l'enfance. Dans ses travaux, il a décrit les expressions faciales liées à la douleur, la colère, la tristesse, la détermination, le dégoût, la surprise, la peur, ainsi qu'une variété d'autres états émotionnels. Darwin a détaillé la morphologie de nombreuses expressions faciales, en indiquant la plupart du temps leur origine anatomique (musculaire). Il a présenté des hypothèses concernant la fonction adaptative de certaines composantes des expressions faciales, par exemple, lever la lèvre supérieure de dégoût pourrait servir à empêcher, exclure une odeur offensive. Pendant près d'un siècle, les résultats de Darwin sur les expressions faciales étaient largement ignorées ou rejetées. Mais depuis 1970 et le nouvel intérêt porté aux expressions émotionnelles, la recherche a avancé aussi bien en théorie qu'en méthodologie, avec le développement de systèmes basés sur l'anatomie pour annoter les expressions faciales et leur application pour l'étude des adultes, adolescents et enfants (Ekman & Friesen, 1978 ; Friesen & Ekman, 1984 ; Izard, 1979, Izard, Dougherty, & Hembree, 1983).

(Ekman et al., 1976, 1978, 2002) ont développé le système FACS (Facial Action Coding System) dans le but d'étudier les expressions faciales. Plus précisément, l'objectif principal était de développer un système « exhaustif » qui pourrait distinguer tous les mouvements faciaux possibles visuellement distinguables. FACS est le système le plus largement utilisé pour mesurer les comportements faciaux et les émotions. La première version du système est apparue dans les années 70, les développeurs de FACS avaient alors étudié de quelle façon la contraction de chaque muscle facial (seul ou en combinaison avec d'autres muscles) modifiait l'apparence du visage.

Le codage des actions faciales nécessite l'identification des muscles produisant les expressions faciales (Figure 17).

Figure 17: Les muscles faciaux, produisant les expressions faciales

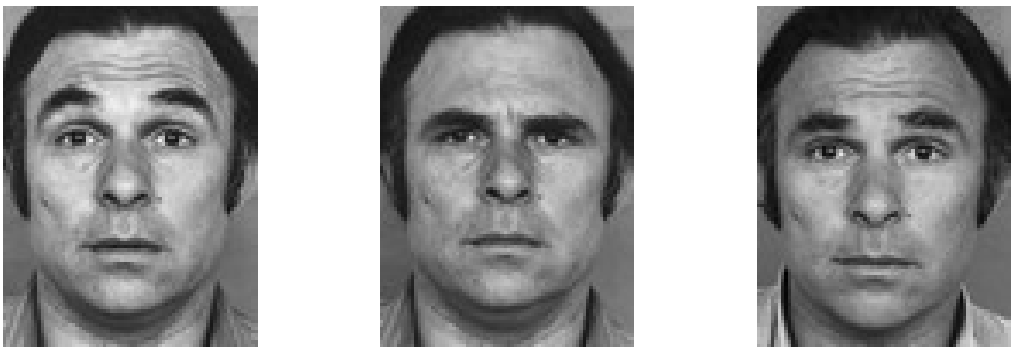
Les unités de mesure de FACS sont les Action Units (AUs), et pas les muscles. La contraction de certains muscles n'est pas toujours visible extérieurement. Les AUs combinent parfois plusieurs muscles faciaux. Le manuel FACS précise qu'il est nécessaire d'avoir une expression neutre d'une personne, pour avoir une meilleure estimation des expressions qu'elle réalisera ensuite, l'expression neutre servant de référence. L'intensité est également prise en compte dans l'estimation des AUs. Elle est indiquée à la suite du numéro de l'AU, par une lettre, échelonnée entre A (faible intensité) et G (forte intensité). La Figure 18 montre l'expression neutre, ainsi que 2 images, l'une montrant l'AU 7 (paupières resserrées) et l'autre montrant la même AU, avec une intensité très élevée E.



AU 0 AU 7 AU 7^{E}

Figure 18: Trois expressions faciales : l'expression neutre (AU 0), l'expression paupières resserrées, d'intensité normale (AU 7), et de forte intensité (AU 7^{E})

La Figure 19 montre deux exemples de combinaison d'AUs provenant du manuel FACS. La première image montre une première combinaison d'AUs (1+2) qui correspond aux sourcils élevés (parties interne et externe). La deuxième image montre l'AU 4 (sourcils froncés). Enfin, la troisième image montre la combinaison des AUs 1, 2 et 4.



AUs 1+2 AU 4 AUs 1+2+4

Figure 19: Trois expressions faciales : l'expression sourcils élevés (AU 1+2), l'expression sourcils froncés (AU 4), et la combinaison des deux (AU 1+2+4)

Un annotateur utilisant FACS disséquera une expression observée, la décomposant en différentes AUs produisant le mouvement. Le résultat d'annotation d'une expression faciale est la liste des AUs la produisant. La durée, l'intensité et l'asymétrie peuvent également être annotées. FACS ne fournit pas d'information concernant la signification du comportement, mais seulement des informations descriptives.

Dans le but de mesurer uniquement les actions faciales liées aux émotions, Ekman et Friesen ont développé le système EMFACS, une version alternative du système FACS, ne considérant que les expressions émotionnelles, et uniquement les AUs et les combinaisons d'AUs les plus souvent reconnues dans les recherches empiriques ou dans la théorie en tant que signaux émotionnels.

(Ekman, 2003) a étudié les expressions faciales associées à des états émotionnels. Il traduit une émotion en une superposition de processus automatiques. Il fait une distinction entre émotions, humeur et personnalité au niveau de la durée de chacun de ces états. Une émotion dure entre quelques secondes et quelques minutes, l'humeur dure entre plusieurs heures et plusieurs jours, tandis que la personnalité correspond a une section de la vie ou parfois toute la vie, associée à d'éventuels désordres mentaux. D'après Ekman, à chaque émotion est lié un trigger (élément déclanchant l'émotion), une fonction, des sensations corporelles, ainsi que des signes infimes. Il s'est limité à 7 émotions (tristesse, colère, peur, dégoût, mépris, joie), à chacune d'elle correspond une expression faciale universelle distincte. Chacune de ces émotions peut être catégorisée en tant que classe d'émotion, la classe colère par exemple contenant toutes les variétés de colère en terme de force (allant de l'irritation à la rage), et en terme de type (colère froide ou colère chaude). La variation d'intensité de ces émotions est clairement marquée sur le visage. La Figure 20 montre 6 photographies utilisées lors d'une expérience effectuée par Ekman sur les différences de perception cross culturelle. Un des

objectifs de cette étude était d'évaluer si les 6 émotions basiques (joie, surprise, peur, colère, dégoût, et tristesse) étaient bien reconnues par 21 pays différents. L'expérience a révélé un taux d'accord très élevé entre les 21 pays, qui ont tous majoritairement correctement perçu les 6 émotions de base présentées.

Ekman a également étudié l'expression faciale d'émotions complexes d'un point de vue théorique en combinant artificiellement par copier-coller des différentes zones du visage sur des photos d'acteurs exprimant les six émotions de base (Ekman 1975).

Figure 20: 6 photographies utilisées lors d'une expérience effectuée par Ekman sur les différences de perception cross culturelle des émotions

2.3.5 Expressivité des mouvements et émotions

Dès 1872, (Darwin, 1872) a proposé une liste de mouvements du corps liés à certaines émotions (Table 1).

Joie	Sauter, danser, applaudir, faire oui de la tête, garder le corps droit…
Tristesse	Pas de mouvements, passif…
Fierté	Tête et corps droits…
Honte	Tourner le corps, plus particulièrement le visage, mouvements nerveux…
Peur/terreur/ horreur	Pas de mouvements, ou mouvements convulsifs, mouvements larges des bras (au dessus de la tête), épaules relevées, bras ramenés contre soi...
Colère/rage	Tremblement du corps entier, intention de pousser ou frapper violement, gestes chaotiques, secouer le poing, tête droite, torse ouvert, pieds ancrés fermement dans le sol, coudes fermement relevés sur les cotés…
Dégoût	Gestes de répulsion ou de protection, bras serrés contre soi, épaules relevées…
Mépris	Tourner le corps…

Table 1: Mouvements corporels liés aux émotions (Darwin, 1872).

Les expressions émotionnelles décrites par Darwin sont actées et majoritairement basiques. Ces indices doivent donc être adaptés pour être utilisables avec des comportements émotionnels non actés. Nous verrons dans les sections décrivant nos travaux que nous avons cependant conservé certains de ces indices, tels que la position de la tête, la position des mains, certaines caractéristiques de la qualité du mouvement (expansion spatiale, force), ainsi que la symétrie pour certaines parties du corps.

Laban a définit une théorie sur l'art du mouvement, et sa pratique à travers plusieurs exercices (Newlove 1993). Trois dimensions du mouvement sont définies : la dimension haut-bas, la dimension gauche-droite, et la dimension avant-arrière. Les gestes et autres mouvements du corps sont interprétés, par exemple, l'envie de communiquer peut être exprimée par des mouvements vers l'extérieur, et l'envie de rester fermé par des mouvements vers soi-même. Ces travaux montrent le besoin de combiner les informations provenant de la forme et du mouvement de la tête, du torse, et des mains pour pouvoir exprimer et reconnaître les émotions. Plusieurs termes sont utilisés par Newlove dans ses exercices exploratoires de description des mouvements, tels que « torsion », « courbe », « symétrie », « simultanéité », et « alternance ».

Afin d'étudier le décodage des émotions à partir des gestes et mouvements du corps, l'expérience réalisée par (Montepare et al. 1999) a consisté à demander à de jeunes adultes et à de plus âgés, d'évaluer les paramètres suivants à partir de performances d'acteurs (uniquement les mouvements corporels, pas les expressions faciales ni la voix) : âge, sexe, et ethnie du sujet, position des mains, posture, variations dans la forme du mouvement, tempo, et direction ; mais également la qualité du mouvement. Les acteurs restaient silencieux et leur visage était électroniquement floué dans le but d'isoler les indices corporels. 82 sujets (un groupe de jeunes adultes, et un d'adultes plus âgés) ont participé à une étude en deux parties de décodage des émotions à partir de gestes et mouvements corporels. Dans la première partie, les sujets ont identifié les émotions apparaissant dans de courts enregistrements vidéo montrant des situations émotionnelles jouées par des acteurs. Dans la deuxième partie, les sujets ont évalué les enregistrements vidéo à l'aide de six caractéristiques fondamentales du mouvement (la qualité du mouvement) explorées dans une précédente étude sur les sources dynamiques de l'information émotionnelle (forme, tempo, force, et direction), échelonnés sur sept niveaux avec des descripteurs verbaux (très fluide-très saccadé, très raide-très relâché, très doux-très dur, très lent-très rapide, très étendu-très contracté, pas d'action-beaucoup d'action). Les estimations des deux groupes de sujets (jeunes et moins jeunes) ont un fort taux d'accord et apportent des informations fiables sur des indices corporels particuliers liés aux émotions. Les erreurs faites par les sujets âgés étaient liées aux indices corporels exagérés ou ambigus. Comme nous le verrons plus loin, pour notre corpus, nous n'avons conservé que 3 caractéristiques (sur une échelle à 3 niveaux) pour définir la qualité des mouvements de tête, des épaules, et du torse : fluidité (fluide, normal, saccadé), la vitesse (lent, normal, rapide), et la force (faible, normal, fort). Nous avons utilisé ces 3 caractéristiques plus une 4ème, l'extension spatiale, dans le cas des gestes.

Dans l'expérience de (Wallbott, 1998), douze étudiants en art dramatique ont eu à annoter les mouvements et postures effectués par des acteurs. Seules les catégories suivantes, ayant un taux d'accord inter-annotateurs supérieur à 75% ont été sélectionnées : position du haut du corps (distant de la caméra, penché), position de la tête (basse, arrière, tournée sur le coté, penchée sur le coté), position des bras (mouvements latéraux, mouvements étirés vers l'avant, mouvements étirés sur les cotés, bras croisés sur la poitrine, bras croisés au niveau du ventre, bras croisés en haut du ventre, mains posées sur les hanches), position des mains (poing, ouvert/fermé, arrière de la main sur le coté, manipulateur (soi-même), illustrateur, emblème, pointage), et qualité du mouvement (activité du mouvement, extension spatiale, dynamiques

du mouvement, énergie, et force). Mouvements corporels : saccadés ou actifs, posture : proximité, ouverture, symétrie, asymétrie des différents membres du corps.

Pour l'analyse des gestes, Wallbott a utilisé les classifications fonctionnelles (Table 2) développées par (Ekman et Friesen, 1969/1972).

Illustrateurs	Mouvements accompagnant, illustrant, ou accentuant le contenu verbal, comme la parole accompagnant les gestes.
Manipulateurs / Adaptateurs	Contact avec son propre corps: mouvements tels que se gratter la tête…
Emblèmes	Mouvements ayant une signification précise et définie culturellement, tel qu'un clin d'oeil...

Table 2: Classification fonctionnelle des gestes (Ekman et Friesen, 1969/1972).

(Boone et Cunningham 1996) ont étudié la précision des enfants à identifier les émotions dans des mouvements corporels expressifs effectués par des danseurs. 103 sujets (79 enfants, 24 adultes) ont participé à 2 tâches de décodage d'expressions non verbales. Pour cette expérience, les sujets ont du identifier, sur un écran de télévision, les émotions exprimées par deux danseurs (lequel des deux est content, triste, en colère, ou apeuré). Les indices spécifiques présentés dans la Table 3 ont été utilisés par les sujets adultes pour décrire les émotions.

Colère	Changements de tempo. Changements de direction au niveau du torse et du visage.
Joie	Fréquence des mouvements de bras vers le haut Durée des mouvements de bras loin du corps
Peur	Tension des muscles
Tristesse	Durée des mouvements de corps vers l'avant

Table 3: Mouvements corporels liés aux émotions (Boone et Cunningham 1996).

(Boone et Cunningham 1998) ont rapporté des moyens de décoder la joie, la tristesse, la colère, et la peur dans les mouvements corporels expressifs de danseurs, et la capacité à détecter les différences d'intensité de colère et de joie lorsque la quantité relative d'indices corporels spécifiques à chaque émotion sont modifiés. Les enfants (à partir de 5 ans) ont montré qu'ils se focalisaient sur les mouvements spécifiques aux émotions pour décoder l'intensité de la colère et de la joie.

(DeMeijer 1991) a étudié les effets de trois variables sur l'inférence d'émotions à partir de mouvements corporels : le sexe de l'acteur (danseur, encodeur), le sexe du sujet (décodeur), et

l'expressivité des mouvements effectués par l'acteur. 42 adultes, âgés entre 18 et 32 ans (21 hommes et 21 femmes), ont participé à une expérience d'analyse d'enregistrements vidéo comprenant 96 différents mouvements effectués par 1 homme et 2 femmes, tous les trois étudiants en danse expressive. Les sujets ont utilisé les 7 dimensions dichotomiques suivantes pour décrire les mouvements : mouvements de torse (étirer, pencher), mouvements des bras (ouverture, fermeture), direction verticale (haut, bas), direction sagittale (avant, arrière), force (fort, faible), vitesse (rapide, lent), forme du mouvement (mouvement droit ou en courbe).

Des séquences d'émotions provenant d'un show télévisé ont été analysées par (Atifi et Marcoccia 2001): ces séquences montrent les comportements d'une jeune étudiante exprimant son désespoir et sa colère en raison des difficultés d'insertion professionnelle des jeunes adultes en France. Les auteurs associent les comportements suivants aux émotions vécues par la jeune femme : mouvements latéraux de la tête, nombreux et continus, interprétés comme de la résignation. La combinaison de phénomènes continus, tels que les balayages du regard, et locaux, tels que lever fréquemment les sourcils avec des yeux grands ouverts, synchronisés avec des mots spécifiques, peut être interprétée dans cet exemple comme un regard affligé ou apeuré ; garder les yeux fermés peut être interprété comme du désespoir. Ces travaux montrent que l'analyse d'interactions à la télévision requiert la prise en compte de la dimension multimodale du média.

Nous avons résumé dans la Table 4 ces études sur les paramètres du mouvement.

	Paramètres	Acté	Annotés	Modalités	Type de Comportement	Sujets	Émotions
(Newlove 1993)	- Dimensions du mouvement (haut-bas, gauche-droite, avant-arrière) - Symétrie - Simultanéité	Oui	Non	Toutes	Danse	Danseurs	
(Montepare et al. 1999)	- Positions des mains - Posture - Fluidité / Force / Vitesse / Extension spatiale - Rigidité - Activité	Oui	Non	Toutes sauf les expressions faciales	Simulation de situations émotionnelles	82 adultes plus ou moins jeunes percevant des émotions dans de brefs extraits vidéo	- Joie - Tristesse - Colère - Neutre
(Wallbott 1998)	- Haut du corps - Epaules (haut, arrière, avant) - Tête (bas, arrière, tournée / penchée sur les cotés) - Bras - Mains - Qualité du mouvement (activité, extension spatiale, dynamique du mouvement, énergie, force) - Symétrie	Oui	Oui	Toutes	Mouvements et postures effectués par des acteurs pour chaque émotion	12 étudiants en art dramatique, et observateurs entraînés	- Joie - Tristesse - Fierté - Honte - Peur / terreur - Colère/rage - Dégoût - Mépris
(Boone and Cunningham 1996; 1998)	- Changements dans le tempo - Changements directionnels - Fréquence - Tension musculaire - Durée	Oui	Oui	Visage, torse, muscles faciaux, corps	Danse	103 sujets: 79 enfants (4 à 8 ans), 24 adultes	- Colère - Joie - Peur - Tristesse
	Paramètres	**Acté**	**Annotés**	**Modalités**	**Type de Comportement**	**Sujets**	**Émotions**

(DeMeijer 1991)	- Torse (étirer sur les cotés, pencher en avant / arrière) - Bras (ouverture, fermeture) - Direction verticale (haut, bas) - Direction sagittale (avant, arrière) - Force (fort, faible) - Vitesse (rapide, lent) - Droiture	Oui	Non	Torse, bras	Expressivité de l'agressivité et de la peine dans 16 expressions distinctes. 7 dimensions dichotomiques décrivent les mouvements	42 adultes (21 hommes et 21 femmes)	- Agressivité - Peine
(Magno Caldognetto et al. 2004)	- Type de geste (bâtonique, pantomimique, pictographique, symbolique, déictique) - Segmentation des gestes/mouvements (préparation, stroke, retraction) - Relations mains (miroir, asymétriques, indép.) - Mouvements de tête (direction, vitesse) - Regard (direction) - Yeux (ouverture, fermeture) - Sourcils (froncés, levés)	Oui et Non	Oui	Bras, mains, tête, expr. faciales	5 exemples: - Parole et gestes dans une interview - Parole et prosodie dans des journaux télévisés - Langage des signes italien - Parole et émotions - Parole et attitudes	5 sujets, 1 dans chaque exemple	- Colère

Table 4: Résumé des études sur les paramètres expressifs du mouvement.

2.3.6 Corpus de communication multimodale

Plusieurs études se sont intéressées aux comportements multimodaux, sans toutefois être directement reliées aux émotions. Nous nous limitons à trois exemples illustratifs de trois types d'approches pertinentes pour nos travaux.

Profils individuels de gestes conversationnels

Dans un but de génération de comportements gestuels individuels pour des personnages virtuels, Michael Kipp a collecté des extraits vidéo provenant d'un show télévisé allemand littéraire dans lequel 4 écrivains discutent à propos de leurs livres récemment publiés (Kipp, 2004). Il a étudié les gestes de 2 personnes à partir des extraits vidéo collectés. Son choix de collecter les extraits à partir d'un show télévisé s'est justifié pour trois raisons : 1) le matériel est facile à obtenir, 2) les orateurs sont expérimentés en terme de communication, 3) la motivation et les situations de conflits des personnes donne lieu à des comportements gestuels riches. L'objectif de ces travaux étant de générer des gestes conversationnels individuels dans des équipes d'agents de présentation animés dans une interface homme-machine, des critères de sélection ont été posés pour la sélection des vidéos, tels que : les personnes doivent faire face à la caméra, tout en interagissant les unes avec les autres ; les extraits doivent contenir de la parole, des gestes conversationnels et des interactions entre les locuteurs ; les gestes conversationnels doivent être propres à chaque locuteur, nombreux et de qualité (intéressants à regarder) ; la parole doit contenir à la fois des passages de monologues, des actes de dialogue, des actes de gestion des tours de parole ; enfin, les extraits vidéo doivent contenir des gestes visibles, et les interruptions dues aux changements de vues doivent être les plus courtes possible. Les 23 extraits de ce corpus ont des durées allant de 29s à 4m12s, pour un total de 46 minutes.

Multimodal Score

Dans le but de comprendre le fonctionnement de la communication multimodale humaine, plusieurs chercheurs construisent des lexiques de signaux non verbaux, consistant à trouver les correspondances entre les signaux comportementaux et leur signification (Poggi 2001). Pour cela, une des approches consiste à analyser des corpus multimodaux en utilisant des méthodes de segmentation, transcription et annotation des signaux dans les différentes modalités.

(Magno Caldognetto et al. 2004) ont ainsi développé le "Multimodal Score", une procédure permettant de transcrire et d'analyser les signaux multimodaux classifiés séparément ainsi que dans leur interaction mutuelle. Cette méthode permet de transcrire : la parole, la prosodie, les gestes, les expressions faciales (bouche, regard, yeux, sourcils), les mouvements de tête et la posture, en étiquetant les signaux sur cinq différents niveaux d'analyse : description, typologie descriptive, sens, typologie du sens, et fonction sémantique. Les auteurs ont utilisé les paramètres suivants : le type de geste (bâtonique, pantomimique, pictographique, symbolique, déictique), la segmentation des gestes/mouvements (préparation, stroke, retraction), les relations entre les mains (miroirs, asymétriques, indépendantes), les mouvements de la tête (direction, vitesse), le regard (direction), les yeux (ouverture, fermeture), et les sourcils (froncés, levés). Elle a été appliquée à différents types de vidéos (par exemple publicités).

MUMIN

Les travaux effectués par (Allwood et al. 2004) décrivent le schéma de codage MUMIN, développé pour l'annotation de 3 interviews télévisées. Il permet quelques annotations d'expressions faciales (sourcils, yeux, regard, bouche), de mouvements corporels (mains, tête, posture), et de parole (segmentale, suprasegmentale). Le schéma de codage se focalise sur la description des interactions (feedback, sequencing, turn-taking).

Conclusion

La revue des nombreux travaux existants en communication humaine multimodale et émotionnelle nous a incité à sélectionner et approfondir les points suivants dans nos travaux : l'étude des paramètres prosodiques de la voix et de la transcription de la parole, montrant l'importance du canal audio et les conflits qui peuvent exister entre communication non verbale et communication verbale ; l'étude de la structure des gestes, et notamment l'impact de la structure temporelle des gestes sur certaines caractéristiques de l'émotion (par exemple son intensité) ; l'étude des mouvements corporels et de leur expressivité. La Table 4, contenant le résumé des études sur les paramètres du mouvement, montre les éléments soulignés sur lesquels nous nous focalisons dans notre étude des comportements multimodaux.

2.4 *Agents conversationnels animés expressifs*

2.4.1 Agent conversationnel

Définition

D'après (Cassell, 2000), un agent conversationnel animé (ACA) est une interface machine représentée sous la forme d'un personnage virtuel. Il a la capacité de reconnaître et répondre à des entrées verbales et non verbales, de générer des sorties verbales et non verbales, mais également de gérer des fonctions conversationnelles telles que les tours de parole, la capacité de donner des signaux pour indiquer l'état de la conversation, ou de relancer la discussion. Le terme ACA correspond au terme anglo-saxon ECA pour Embodied Conversational Agent qui aurait pu être traduit par Agent Conversationnel Incarné (ACI), c'est-à-dire ayant une apparence.

Les ACAs peuvent jouer plusieurs rôles (Pesty et Sansonnet, 2006) : 1) celui d'assistant pour accueillir les utilisateurs et les aider à comprendre et à utiliser la structure et le fonctionnement d'applications et de services informatiques ; 2) celui de partenaire : composant logiciel capable d'interactions multimodales avec l'usager ; et 3) celui de tuteur : des apprenants dans les Environnements Interactifs d'Apprentissage Humain (EIAH).
Les 3 mots Agent Conversationnel Animé peuvent être définis de la manière suivante :

- Agent : composant logiciel capable de raisonnement sur les représentations qu'il a du système et aussi de l'usager pour assumer son rôle (assistant, partenaire, tuteur).
- Conversationnel : composant logiciel capable d'interactions multimodales avec l'usager
- Animé : composant logiciel doté d'une apparence effective face à l'usager.

Nous présentons dans cette section quelques technologies, modèles et prototypes d'ACA. Nous commençons par décrire une technologie (MPEG-4) utilisée par plusieurs prototypes d'ACA, notamment pour l'expression faciale d'émotions complexes. Nous décrivons ensuite quelques prototypes d'ACA en nous focalisant sur leurs capacités à exprimer des émotions.

Le standard MPEG-4

MPEG-4 (Ostermann, 1998) est le premier standard international pour la communication multimédia incluant l'audio naturelle et synthétique, la vidéo naturelle et synthétique, ainsi que les graphismes 3D. Ce standard intègre la possibilité de définir et d'animer des personnages virtuels composés de têtes et corps synthétiques. Pour la tête (visage et cheveux), plus de 70 paramètres d'animation indépendants du modèle permettent de définir les actions de bas niveau telles que déplacer le coin gauche de la lèvre vers le haut, et les actions de haut niveau telles que des expressions faciales complètes.

MPEG-4 spécifie un ensemble de paramètres d'animation du visage (FAPs pour Facial Animation Parameters), chacun correspondant à une action faciale particulière déformant un modèle du visage. La valeur associée à une FAP indique la magnitude de l'action correspondante, par exemple pour différencier un grand sourire d'un petit. Les FAPs sont basées sur l'étude des actions minimalement perceptibles et sont fortement liées aux actions musculaires. La Table 5 présente les 68 FAPs, catégorisés dans 10 groupes, et liés à différentes parties du visage.

Groupes	**Nombre de FAPs**
1 : visèmes et expressions	2
2 : mâchoire, menton, lèvre inférieure, coins des lèvres, centre des lèvres	16
3 : yeux, pupilles, paupières	12
4 : sourcils	8
5 : joues	4
6 : langue	5
7 : rotation de la tête	3
8 : positions de l'extérieur des lèvres	10
9 : nez	4
10 : oreilles	4

Table 5 : Présentation des 68 FAPs catégorisés en 10 groupes.

L'ensemble des FAPs contient les 2 paramètres de haut niveau : visèmes et expressions (groupe 1). Un visème est une représentation visuelle correspondant à un phonème. Seulement 14 visèmes statiques clairement distincts sont inclues. Dans le but de gérer la coarticulation de la parole et des mouvements de bouche, des transitions d'un visème au suivant sont définis en mélangeant 2 visèmes, en leur affectant des poids. Similairement, les paramètres des expressions définissent 6 expressions faciales de haut niveau telles que la joie et la tristesse.

En contraste avec les visèmes, les expressions faciales sont animées avec une valeur définissant l'amplitude de l'expression.

Plusieurs expressions faciales peuvent aussi être mélangées en affectant un poids à chacune permettant ainsi l'expression d'émotions complexes.

Le système FAPs de MPEG-4 est la technologie la plus répandue en animation faciale. Il distingue la base faciale (FB) et l'affichage facial (FD). Un FB est un mouvement facial basique tel que « sourcil droit levé », « lèvre supérieure levée ». Les FBs incluent également les mouvements des yeux et de la tête tels que « faire oui ou non de la tête ». Un FB peut être représenté comme un ensemble de FAPs. Et une FD se compose d'une ou de plusieurs FBs, exemple : FD= FB1 + FB2 + FB3 + … + FBn, surprise = sourcils_levés + paupières_levées + bouche_ouverte. (Raouzaiou et al., 2002) ont utilisé les FAPs pour modéliser des expressions émotionnelles basiques et ont utilisé une technique à base de règles pour gérer les expressions intermédiaires. Ils ont établi une relation entre les FAPs et le paramètre « activation » utilisé dans les études psychologiques classiques. Ils ont relié l'activation aux mouvements musculaires faciaux. Ils ont défini les FAPs spécifiques à chacune des émotions de base. La Table 6 liste quelques-unes des FAPs pouvant servir à représenter la tristesse.

Action	**FAP correspondante**
Fermer paupière haut gauche	F19
Fermer paupière haut droite	F20
Fermer paupière bas gauche	F21
Fermer paupière bas droite	F22
Lever partie intérieure du sourcil gauche	F31
Lever partie intérieure du sourcil droit	F32
Lever partie centrale du sourcil gauche	F33
Lever partie centrale du sourcil droit	F34
Lever partie extérieure du sourcil gauche	F35
Lever partie extérieure du sourcil droit	F36

Table 6 : Certains des FAPs définissant la tristesse, d'après les travaux de (Raouzaiou et al., 2002)

L'agent Steve

Steve (*Soar Training Expert for Virtual Environments*) est un Agent pédagogique, (Rickel et Johnson, 1999). Il habite un environnement 3D dans lequel des étudiants sont immergés (par l'intermédiaire d'un casque de réalité virtuelle). L'objectif de Steve est de leur apporter des connaissances procédurales, par exemple comment faire fonctionner ou comment réparer des équipements complexes (il a notamment été utilisé dans le cadre d'entraînements pour la marine militaire). Steve est capable de réaliser ces actions lui-même, de répondre aux questions des étudiants et de contrôler leur exécution de ces tâches (les étudiants peuvent agir grâce à une souris 3D et des *datagloves*). Cet Agent est composé de trois modules principaux : un module de perception, un module de cognition et un module de contrôle moteur. Steve peut aussi exprimer des réactions émotionnelles, qui sont liées aux buts qui lui ont été définis (Elliott et al., 1999). Par exemple, un des buts de Steve est de donner des explications sur le domaine considéré. Il manifeste donc un certain enthousiasme lorsqu'il a l'opportunité de discuter un détail intéressant avec un étudiant. Un autre de ses buts est de maintenir l'attention de l'étudiant : il est donc parfois amené à rappeler à l'ordre de manière ferme un étudiant qui ne regarde pas ce que Steve est en train de montrer. Enfin, un autre but de Steve est que les étudiants retiennent ses leçons : il manifeste donc de la joie s'il constate que l'étudiant a retenu un point important ; ou au contraire il apparaît préoccupé s'il constate l'inverse.

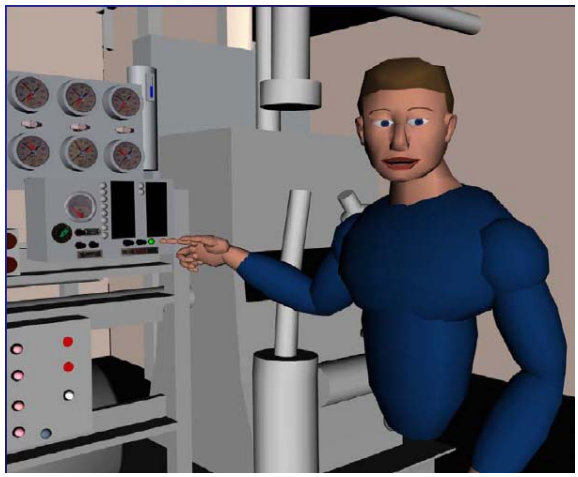


Figure 21 : Steve (Rickel et Johnson, 1999).

BEAT

L'outil d'animation BEAT (Behavior Expression Animation Toolkit) (Cassell et al. 2001) permet d'animer le corps d'un ACA en utilisant du texte en entrée. Trois étapes sont proposées pour la génération de comportements multimodaux : 1) suggestion de plusieurs comportements possibles à partir du texte, 2) sélection du comportement le plus adapté au contexte, et enfin 3) planification des signes correspondant à ce comportement dans les différentes modalités. Cet outil est utilisé dans l'agent MACK.

Max

Max (Kopp et al. 2003) utilise le langage de spécification MURML (Krandsted et al .2002), traitant les tâches d'animation et d'organisation dans la production de comportement multimodal. Cette étude est ciblée sur les interrelations entre les différentes modalités en conversation face à face (Figure 22) en fournissant aux agents des moyens d'interactions similaires à ceux employés par les humains, en particulier la réception et la génération d'information verbales ou non verbales simultanées et synchrones. Max est un médiateur intégré dans un environnement virtuel 3D et effectuant des tâches d'assemblage. Il a été également utilisé dans un musée (Kopp et al. 2005). Il est capable de générer incrémentalement des gestes en synchronisation avec la parole (Kopp et al. 2004). Il inclue également des capacités de simulation d'émotions (liées à une application donnée ou indépendante d'une application) et de leur interaction avec les processus cognitifs de l'agent (Becker et al. 2006).

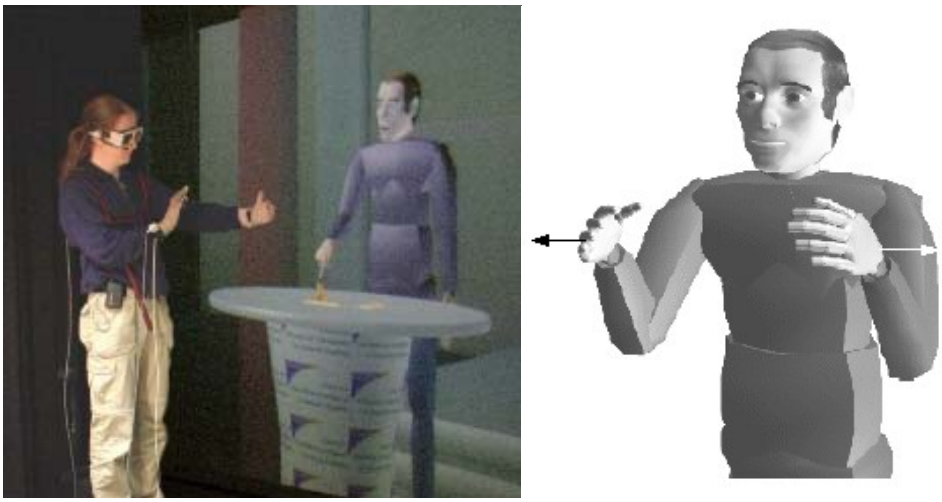


Figure 22 : Interaction multimodale avec Max (Kopp et al. 2003).

REA

Le système REA (*Real Estate Agent*) permet l'interaction en temps réel entre un Agent Conversationnel et un utilisateur humain (Cassell et al. 1999, Bickmore et Cassell, 2005). Le contexte d'interaction est un entretien entre l'agent immobilier REA et un client potentiel. L'agent a une apparence humanoïde (Figure 23) lui permettant d'avoir des comportements verbaux et non verbaux (regard, expressions faciales, posture et gestes des mains). Pour renforcer le réalisme de l'interaction, l'Agent et son environnement 3D sont projetés sur grand écran, mettant ainsi REA à taille humaine. Le domaine d'interaction de l'agent (les transactions immobilières) étant fortement lié à une notion de confiance entre Agent et utilisateur, (Cassell et Bickmore, 2001) se sont interrogés sur la manière d'instaurer un climat de confiance avec l'utilisateur. Ils ont ainsi donné à REA la capacité, avant d'aborder les questions principales relatives au domaine immobilier, de parler de la pluie et du beau temps (*small talks*). Cet agent a été développé de manière à combiner les différentes modalités aussi bien en entrée qu'en sortie. Le langage de spécification de REA fait intervenir des fonctions de haut niveau de spécification combinant plusieurs modalités, par exemple: la fonction «Give turn» (en génération), qui entraîne la relaxation des mains, le regard vers l'utilisateur et l'élévation des sourcils, ou encore la fonction «Open interaction» qui entraîne le regard vers l'utilisateur, un sourire et un mouvement de tête. (Cassell et al., 2001) ont apporté des données empiriques sur les relations entre les modifications de postures d'un agent et la structure de son discours, et un algorithme de génération de ces postures utilisant le système de dialogue Collagen (Lochbaum, 1998).

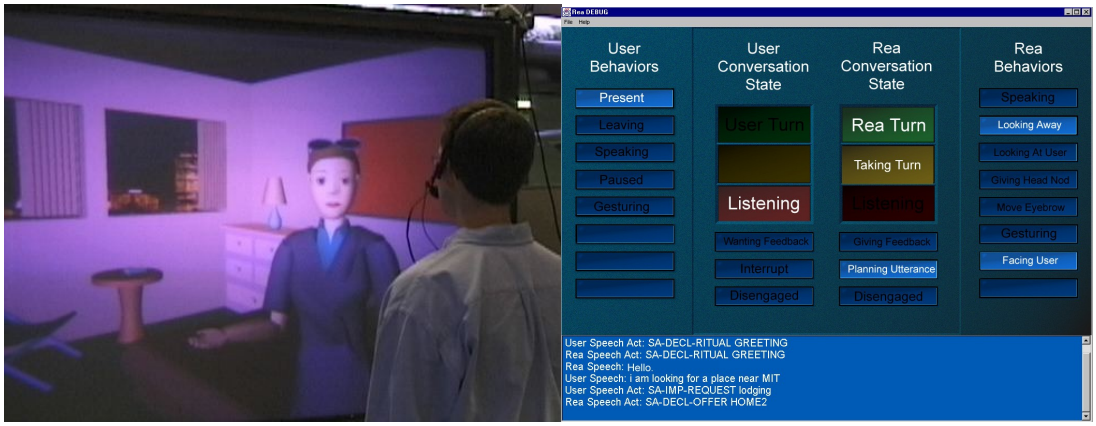


Figure 23 : L'Agent REA (Bickmore et Cassell, 2005)

NECA / RRL

Dans le cadre du projet NECA (pour Net Environment for Embodied Emotional Conversational Agents), système générant des interactions entre deux ou plus personnages animés, RRL (pour Rich Representation Language), développé par (Piwek et al. 2002) est utilisé pour représenter les échanges d'information entre les différents modules du système. L'objectif de ces travaux est de développer un langage formel permettant de représenter le comportement d'agents conversationnels. Le système comprend un module de raisonnement affectif utilisant le modèle OCC (Ortony et al. 1988). A l'exécution, l'émotion spécifiée par ce modèle permet la sélection de l'expression faciale de l'émotion primaire la plus proche ou d'une interpolation de plusieurs expressions faciales.

Story teller

(Cavazza et al., 2001) ont étudié les interactions entre agents dans des scénarios provenant d'une série télévisée américaine, dans un système de narration interactive (Figure 24). L'approche utilisée ici est centrée sur les personnages virtuels, l'histoire émerge des rôles que joue chacun d'eux, et plus particulièrement des interactions dynamiques entre ces rôles. Ces études ont montré entre autre la capacité des agents à pouvoir modifier l'évolution d'une histoire uniquement à partir de leur comportement. Les personnages principaux sont contrôlés par des plans.

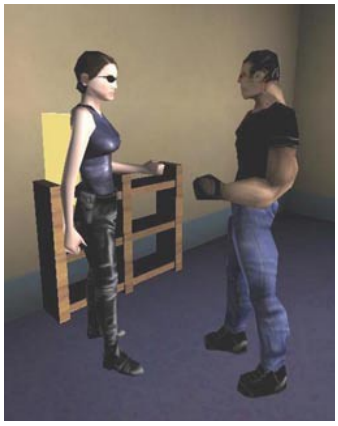


Figure 24: Deux des personnages du système Story teller (Cavazza et al., 2001)

La Figure 25 montre l'instanciation complète d'une histoire, appelé le scénario « jalousie ». Dans le but d'obtenir l'information qu'il cherche, Ross va chercher l'agenda de Rachel (a). Quand il arrive, il s'aperçoit que Rachel est en train d'écrire dans son agenda, il va donc voir Phoebe pour lui poser des questions sur Rachel (b). Au même moment, Rachel a fini d'écrire

et décide d'aller discuter avec Phoebe (c). Lorsqu'elle arrive près de Phoebe, elle voit que cette dernière discute joyeusement avec Ross (d). Elle devient jalouse et quitte la salle (e). Ross la voit partir et la suit (f). Chacun des personnages a son propre système de planification. Ce scénario montre que les interactions entre personnages peuvent considérablement contribuer à la variation de l'histoire.

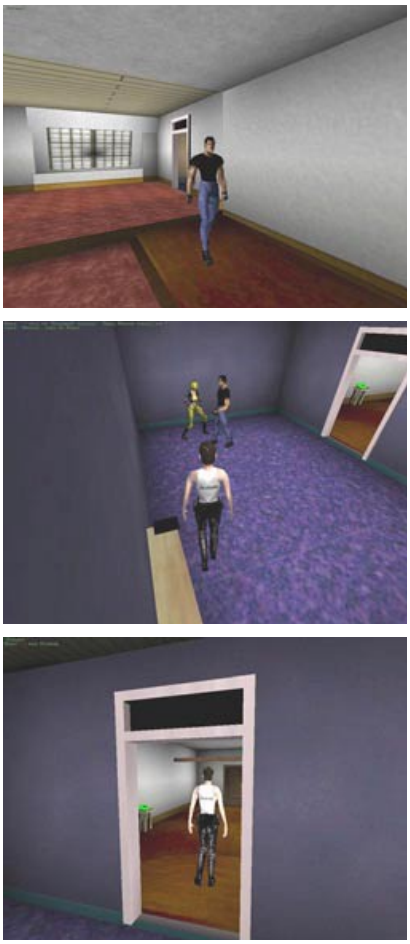


Figure 25 : Le scénario "jalousie" du système Story teller

Exemples d'agents Web

De nombreux ACAs se développent actuellement sur le Web. Leur comportement reste encore limité en terme d'expressivité et d'interactivité. L'agent Laura (Figure 26), développé par la société Cantoche (http://www.cantoche.com), donne des conseils sur l'économie d'énergie sur le site de EDF. L'interaction est très canalisée : l'agent propose des choix multiples et l'utilisateur répond en cliquant sur un des liens proposés. Les comportements sont pré-enregistrés.

Figure 26 : Laura (développé par la Cantoche)

Certains sites Web commercialisent des agents virtuels, la Figure 27 montre un exemple d'agent sur un site de renseignement médical.

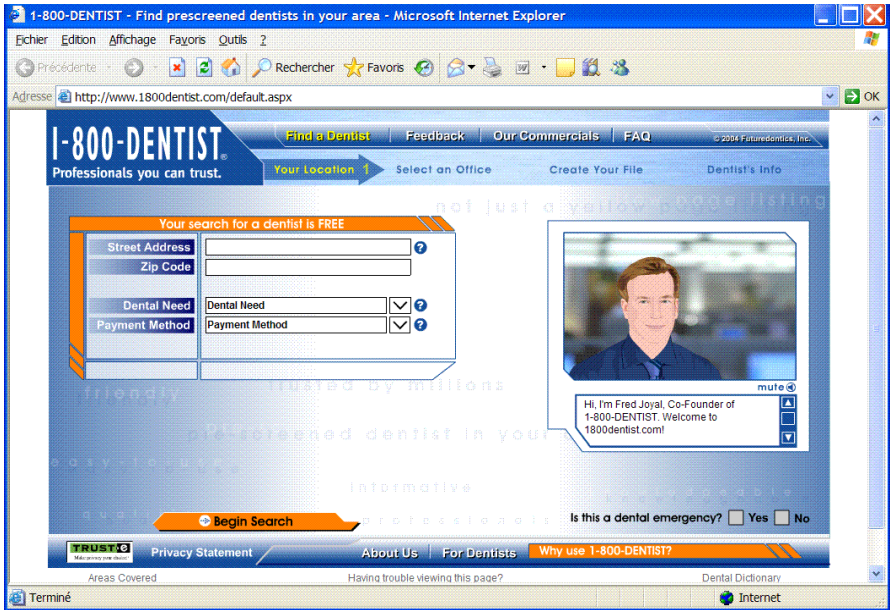


Figure 27 : Agent de type « head and shoulders » sur un site de renseignement médical

2.4.2 Agents conversationnels expressifs

L'agent Cosmo

(Lester et al., 1997/1999) ont développé une application pédagogique intégrant l'agent Cosmo à l'environnement 3D (Figure 28). L'agent donne des explications techniques sur la transmission de données numériques, propose des exercices aux utilisateurs, contrôle leurs actions et leur donne des conseils. Il est également capable de montrer des réactions « émotionnelles », du type « interrogation » quand l'utilisateur sollicite de l'aide, « concentration » quand Cosmo donne une explication, « félicitation » quand l'utilisateur fait un choix correct ou « déception » quand il fait une erreur (Towns et al., 1998). Ces réactions sont produites simultanément sur différentes parties du corps de l'Agent : expression faciale (en particulier par la forme de la bouche et des sourcils), position de la tête, gestes des mains, posture. Cosmo dispose également d'une modalité supplémentaire pour exprimer ses affects : la position de ses antennes (Lester et al., 2000).

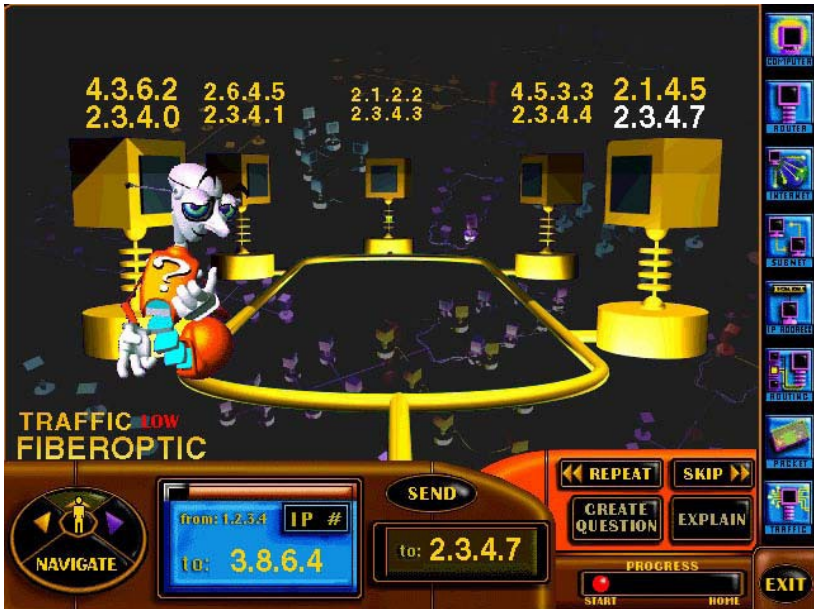


Figure 28: L'Agent Cosmo et son environnement pédagogique (Lester et al., 1997).

L'agent expressif Greta

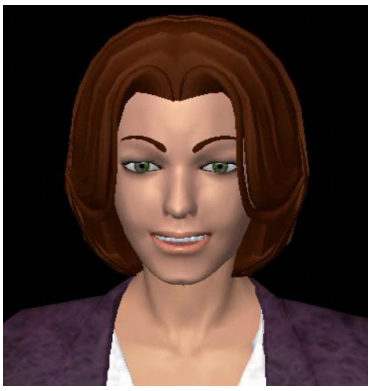
Greta est un agent conversationnel expressif pourvu de qualités communicatives, conversationnelles et émotionnelles (Pelachaud, 2005). Ce système se base sur une taxonomie de fonctions communicatives proposée par (Poggi et Pelachaud., 2000). Une fonction communicative y est définie par un couple (sens, signal) où le sens correspond à la valeur

communicative que l'agent souhaite communiquer, et le signal correspond au comportement utilisé pour porter ce sens. Dans la taxonomie, les fonctions communicatives sont différenciées en informations sur les « croyances » de l'utilisateur, ses intentions, son état affectif, et en informations métacognitives sur son état mental. Greta utilise un langage de représentation appelé APML pour Affective Presentation Markup Language, dans lequel les fonctions communicatives sont exprimées par des balises (De Carolis et al., 2004).
Afin de permettre la génération de comportements expressifs, une variable exprFactor a été ajoutée, spécifiant l'expressivité avec laquelle un acte communicatif doit être traité. Un exemple de texte en APML avec des fonctions communicatives est donné ci-dessous:

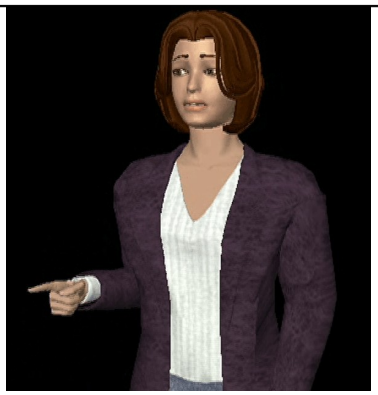
```
<performative type="criticize">
  <rheme affect="anger" exprFactor="1.3">
    Ils ont pris mon père <boundary type="HL"/>
    le jour où ils l'ont pris <boundary type="LH"/>
       ils m'ont pris <emphasis x-pitch-accent="Lstar"
deictic="self3"> moi</emphasis>
  </rheme>
</performative>
```

Ces spécifications sont utilisées par le système pilotant l'agent (Figure 29) a deux niveaux : 1) au niveau du texte balisé avec le langage APML que doit synthétiser vocalement l'agent, et 2) au niveau du profil comportemental de l'agent. Les balises APML, correspondant au sens d'une fonction communicative donnée, sont converties en signaux (faciaux, gestuels). Le système doit tout d'abord sélectionner les comportements les plus appropriés pour un acte communicatif particulier et un agent donné. Il y a deux phases de sélection (Maya et al., 2004) : sélection de la modalité et présélection des signaux. La première sélection détermine la modalité que l'agent utilise ; la seconde consiste à ordonner l'ensemble des comportements possibles ayant un sens similaire. Cet ordonnancement prend en compte l'expressivité de l'agent. Les sorties du système sont les fichiers d'animation et audio pilotant le modèle facial.
Le système prend plusieurs fichiers en entrée : la transcription étiquetée avec APML, les associations entre labels émotions et les expressions faciales, les paramètres expressifs des gestes. Le module gère la synchronisation entre parole, gestes et mouvements de tête. Le résultat final est une animation.

Si un geste annoté dans une vidéo n'est pas répertorié dans les gestes gérés par l'agent Greta, il est nécessaire de l'ajouter à l'aide de l'éditeur de configurations de gestes qui permet de spécifier la position et la configuration de la main. Ainsi plus le corpus traité s'agrandit, et plus la bibliothèque gestuelle de Greta s'enrichit de nouveaux gestes.



```
<apml>
 <performative type="agree">
  <rheme affect="joy">
   <emphasis x-pitchaccent="Hstar">
   oui
   </emphasis>
   on va y arriver
   <boundary type="LL"/>
  </rheme>
 </performative>
</apml>
```



```
<apml>
 <performative type="criticize">
  <theme affect="sadness">
   J'aurai préféré être à
   <emphasis x-pitchaccent="Hstar"
   deictic="there">
   sa place
   </emphasis>
  </theme>
 </performative>
</apml>
```

Figure 29: Deux exemples de génération de comportement émotionnel dans Greta à partir de spécifications avec le langage APML (Pelachaud, 2005).

Dans le domaine des agents expressifs, (Hartmann et al., 2002) ont proposé un ensemble d'attributs expressifs pour les expressions faciales (intensité, évolution temporelle), pour le regard (durée du regard mutuel), pour les gestes (force, tempo, dynamisme, amplitude), ainsi que des paramètres globaux (activation globale, extension spatiale, fluidité, force, répétitivité).

L'expressivité de l'agent est définie par un ensemble de six dimensions : l'extension spatiale, l'extension temporelle, la force, la fluidité, la répétition et l'activité globale. Le système Greta prend en entrée un texte balisé avec des fonctions communicatives et des valeurs pour ces dimensions d'expressivité qui caractérisent l'exécution des comportements de l'agent (Figure 30). Ce modèle a été récemment utilisé pour contrôler les mouvements de têtes combinés à

des expressions faciales complexes feintes, masquées et inhibées (Niewiadomski 2007, Bevacqua 2007).

Nous reviendrons sur l'agent Greta dans la description de nos travaux puisque nous l'avons utilisé pour rejouer des comportements annotés dans des vidéos que nous avons collecté.

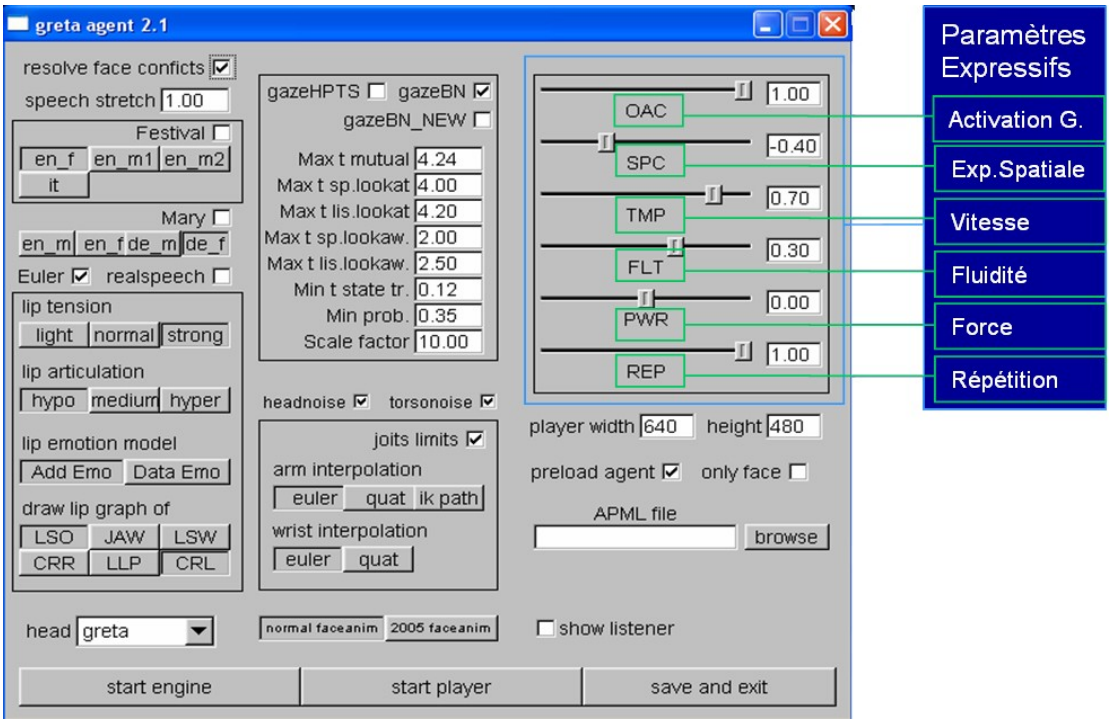


Figure 30: Interface et paramètres expressifs du système Greta (Pelachaud, 2005).

Petra / Cherry

(Lisetti et Marpaung, 2006) ont développé le robot Petra, un agent conversationnel autonome, faisant suite à un premier prototype Cherry (Brown et al, 2002). En plus de générer des comportements émotionnels à partir de modèles d'émotion, ce robot est équipé d'un sonar et d'une caméra pour la détection d'obstacles ainsi que la reconnaissance du visage. Petra a été modélisée de manière à pouvoir interagir avec des humains, accomplir un ensemble de tâches telles que faire une visite guidée de la faculté de science. Les expressions faciales de Petra changent pour refléter son état émotionnel. La modélisation de ces expressions est basée sur le système FACS décrit plus haut. La Figure 31 montre 5 expressions émotionnelles générées : joie, surprise, peur, tristesse, et colère.

Le système contient un module de représentation de la connaissance affective (AKR, pour Affective Knowledge Representation), produisant un schéma émotionnel pouvant être associé à un événement et à une émotion. L'AKR intègre le système SECs décrit plus haut (Scherer,

1984) déterminant les composants de l'émotion et les évaluations successives des différentes dimensions d'appraisal (Lisetti and Paleari accepté).

(a) (b) (c)

(d) (e)

Figure 31: Cinq émotions générées par Petra (Lisetti et Marpaung, 2006)
(a) Joie, (b) Surprise, (c) Peur, (d) Tristesse, (e) Colère

L'interface et le dispositif complet de Petra sont présentés dans la Figure 32. L'interface intègre plusieurs composants, tels que l'avatar, une carte, les barres de progression pour les changements d'émotion, une zone de texte pour la parole, deux vues : une pour la reconnaissance du visage, et l'autre pour le système de navigation, etc.

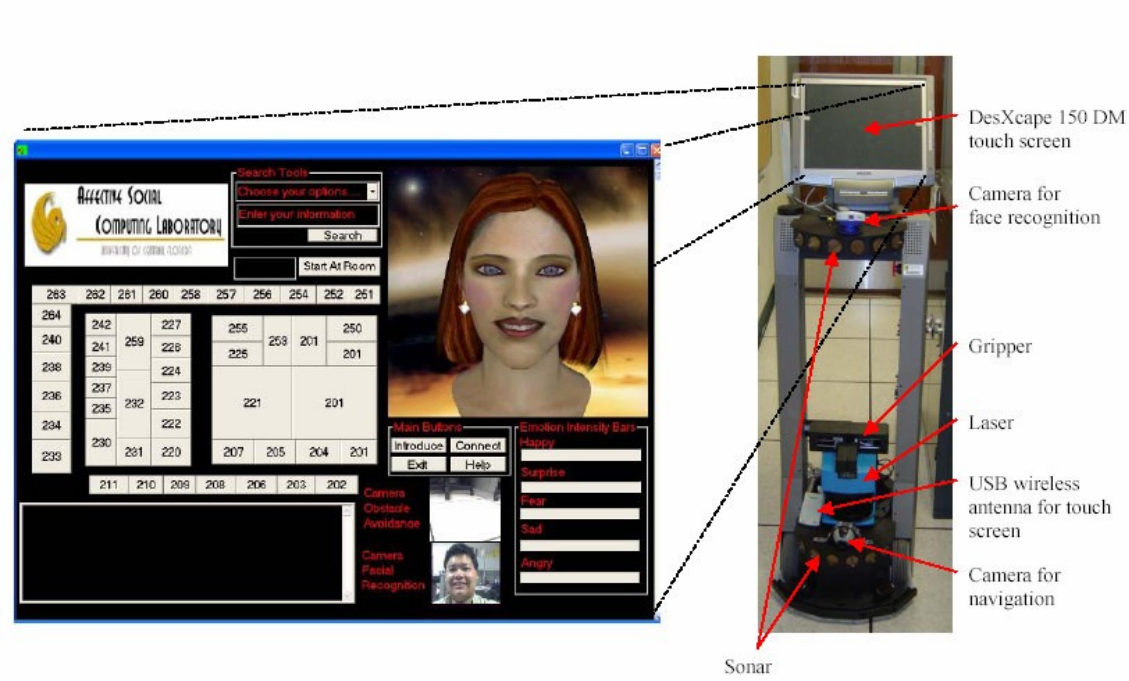


Figure 32 : Interface et dispositif complet du robot Petra (Lisetti et Marpaung, 2006)

Maestri

(Maestri, 1999) a spécifié une expression faciale unique pour 6 émotions de base, dans le but de générer des expressions émotionnelles dans des agents virtuels. Il décrit pour cela la position de chaque partie du visage. La Figure 33 montre les six expressions faciales réalisées avec un agent virtuel.

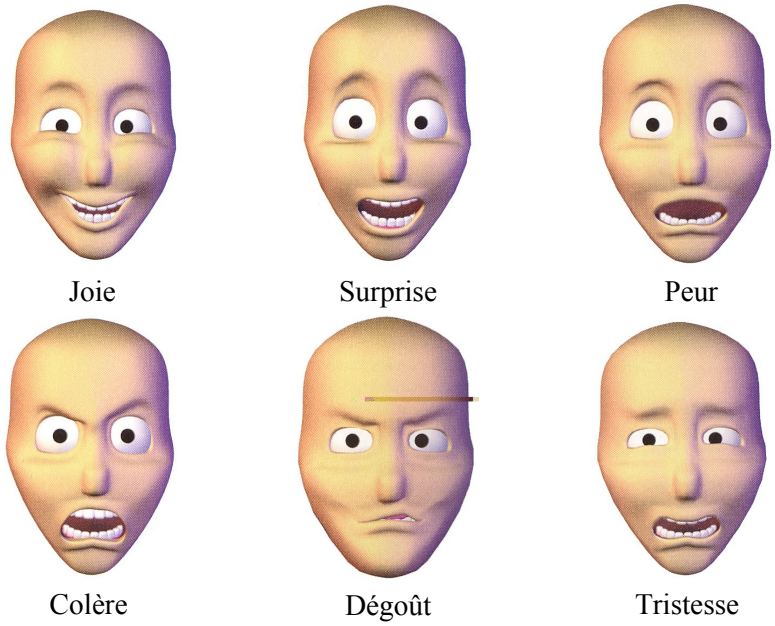


Figure 33 : Génération d'expressions émotionnelles (Maestri, 1999)

Le modèle PAR

Les agents conversationnels animés sont variés tant en apparence qu'au niveau de leurs possibilités. Ces possibilités n'étant pas uniquement les actions qu'ils peuvent faire, mais de quelle manière ils peuvent les faire. (Allbeck et Badler, 2002) proposent un modèle PAR (Parameterized Action Representation) permettant à un agent d'agir, de planifier et de raisonner sur ses actions et celles des autres (Figure 34). PAR est conçu pour construire les futurs comportements dans les agents conversationnels et contrôler les paramètres d'animation qui définissent l'humeur, l'affection, et la personnalité, à l'aide du modèle de personnalité OCEAN (Wiggins, 1996).

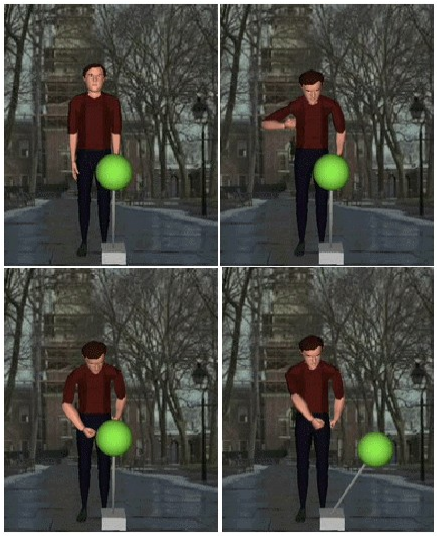


Figure 34 : Construction des comportements liés à l'action : frapper sur un ballon

Les systèmes immersifs VTE et MRE

(Gratch et al., 2004) ont utilisé l'agent VTE mobile dans un système de réalité virtuelle et augmentée, en contexte d'entraînement/action (Figure 35).

Figure 35 : L'agent VTE dans un système de réalité virtuelle et augmentée (Gratch et al., 2004)

(Traum et Rickel 2001) ont également utilisé ce même agent. Leurs travaux consistaient à gérer des interactions dans un groupe d'agents immergés dans un monde virtuel, et dont les spécifications sont sous la forme de tags conversationnels : make-contact, give-attention, start-topic, end-topic etc. (Figure 36).

Figure 36: Scénario de “secours militaire”, les personnages impliqués sont un sergent, une mère, son enfant blessé, et un médecin

Langages de représentations basés sur XML

Outre les langages que nous avons déjà mentionné, deux langages de représentation sont décrits dans les travaux de (Arafa et al. 2002). CML (Character Markup Language), définissant les attributs syntactiques, sémantiques et pragmatiques du personnage à l’aide de texte structuré (Figure 37), et AML (Avatar Markup Language), permettant aux utilisateurs de définir des séquences d’animation en sélectionnant et fusionnant des unités d’animation prédéfinies et ayant certaines propriétés. Ces langages sont tout deux basés sur les descriptions XML. AML a été développé dans le cadre du projet IST SoNG avec les partenaires IIS, Miralab et LIG, et dont l’objectif est de concevoir et développer une plate-forme multimédia MPEG 4 permettant entre autre de gérer des salons de discussion basés sur des avatars 3D et des personnages autonomes.

```
<cml>
 <character name="n1" personality="p1"
           role="r1" gender="M"
           disposition="d1"
           transition_Dstate="t1">
  <happy intensity="i1" decay="dc1"
           target="o1" priority="pr1">
    <move-to order="o1" priority="pr2"
            speed="s" object="obj1" begin="s1"
            end="s4"/>
    <point-to order="2" priority="pr3"
            object="obj1" begin="s2" end="s4"/>
    <utterance priority="pr2" begin="s2">
         UtteranceText
    </utterance>
   </happy>
   ……
 </character>
</cml>
```

Figure 37 : Exemple de script CML

Une initiative de standardisation a été récemment lancée pour favoriser l'échange de modules d'ACA. Elle s'intitule SAIBA et comporte un langage de représentation des fonctions communicatives (FML) et un langage de représentation des comportements (BML) (Kopp et al. 2006).

2.5 *Résumé de l'état de l'art*

Il existe très peu de corpus émotionnel naturel à l'heure actuelle. La plupart des études sur le comportement multimodal lors d'événements émotionnels sont basées sur des émotions basiques et actées, et elles n'utilisent pas toujours des outils et avancées récentes pour la gestion des corpus multimodaux.

L'utilisation d'outils informatiques pour l'annotation de corpus multimodaux a permis de mettre en place des schémas de codage informatisés implémentant certains schémas de représentation proposés par les Sciences Humaines. Des modèles informatiques construits à partir des annotations permettent ainsi l'étude de profils gestuels individuels. Cependant les dimensions expressives, qui semblent pertinentes pour l'étude des émotions, font l'objet de traitements automatiques sur des données récoltées en laboratoire mais n'ont pas été appliqués à l'annotation manuelle de vidéos émotionnelles spontanées. Il en est de même de la description des dimensions d'appraisal des événements émotionnels qui sont utilisés dans quelques systèmes d'ACAs mais qui n'ont pas fait l'objet d'annotation manuelle dans des corpus audiovisuels.

Du côté des ACAs, des agents expressifs existent mais ils ne sont pas toujours fondés sur des données expérimentales, ou quand ils le sont, ce sont sur des données actées qui permettent rarement l'expression d'émotions complexes sur plusieurs modalités en même temps comme les expressions faciales combinées avec les gestes.

Partie 3) Contributions

3.1 *Retour sur objectifs et approche choisie*

Nous avons vu que la plupart des recherches étudiant les comportements émotionnels étaient souvent limitées à l'étude d'émotions basiques non spontanées dans des modalités considérées de manière isolée.

L'objectif de cette thèse est d'apporter une contribution exploratoire et expérimentale à l'étude des émotions spontanées et de leur expression multimodale pouvant être utile pour la modélisation et la conception des futurs systèmes interactifs qui devront pouvoir détecter et afficher ces expressions multimodales spontanées. Nous nous sommes intéressé à la fois à l'étude des émotions, du multimodal, et du contexte, car une approche exploratoire doit selon nous se permettre d'avoir un spectre large, sachant que l'objectif de cette thèse n'est pas de collecter et annoter un grand corpus de données.

Cette thèse a un double but, d'une part de contribuer à l'étude des expressions multimodales des émotions humaines : étude des différentes théories de représentation des émotions et de leurs signes dans différentes modalités, et d'autre part d'informer la spécification d'ACA. Les travaux sur les émotions dans des contextes réels (ex : conversation dans des centres d'appels) montrent l'importance de la combinaison entre niveaux linguistiques et paralinguistiques dans la verbalisation des émotions (Devillers, 2006).

Cette 3ème partie présente les différentes étapes de notre travail de recherche. Des choix ont été effectués pour la collecte des sources vidéo contenant les événements émotionnels spontanés (section 3.2), pour leurs annotations à différents niveaux (émotion, contexte, multimodal), pour l'exploration de différents protocoles et schémas d'annotation sur nos données (4 expériences d'annotation des émotions et 3 d'annotation des comportements multimodaux).

Ces différentes expérimentations ont permis d'explorer conjointement les différentes directions d'études suivantes (ce qui était rare quand nous avons commencé cette thèse) : annotations manuelles des émotions, annotations manuelles des comportements multimodaux, traitement automatique (audio, image), copie-synthèse avec un agent animé, et tests perceptifs prenant en compte les différences individuelles.

A partir des dimensions étudiées par différents chercheurs, que ce soit au niveau émotionnel (Douglas-Cowie et al. 2000), contextuel (Scherer et al., 2001), ou mouvement expressif (Wallbott, 1998, Pelachaud 2005), nous avons adapté nos schémas au corpus collecté (dans le choix des étiquettes, dimensions abstraites, dimensions contextuelles, et indices multimodaux) et les avons affiné pour répondre à la complexité de la représentation de comportements émotionnels multimodaux spontanés. Nous avons adopté une approche exploratoire de trois approches théoriques des émotions (catégories, dimensions, appraisal). Nous avons cependant privilégié l'approche catégorielle car cela nous semblait plus adapté pour étudier la combinaison de plusieurs modalités lors d'émotions complexes et pour piloter un ACA.

Les phases d'annotation, présentées dans la section 3.3 en ce qui concerne l'annotation des émotions, et dans la section 3.4 pour l'annotation des comportements multimodaux, ont été effectuées à l'aide de plusieurs outils : l'outil d'annotation Anvil développé par (Kipp, 2001), et Praat pour la transcription manuelle de la parole et l'extraction automatique des paramètres audio « pitch » et « intensité » (importables dans Anvil). Les guides d'annotation mis à jour pour chaque phase d'annotation indiquaient de procéder piste par piste pour annoter de manière fiable. De plus, les annotateurs pouvaient jouer plusieurs fois le clip à annoter. Pour les dimensions d'appraisal, nous avons mené deux phases : une première pour lister les événements, et une seconde pour annoter les dimensions d'appraisal (ex : nouveauté) pour chaque événement.

Le traitement automatique d'image (effectué en collaboration avec le laboratoire ICCS en Grèce, un partenaire de réseau HUMAINE) a été appliqué à notre corpus pour comparer/valider les annotations manuelles, cette étude est décrite en 3.5. L'étude perceptive que nous avons effectuée à l'aide de l'agent expressif Greta (Pelachaud, 2005), utilisé pour la génération de comportement émotionnel multimodal pour tester la fiabilité des annotations, est décrite dans la section 3.6. Enfin, nous proposons en perspectives des mesures pertinentes entre émotions complexes et comportements multimodaux rendues possibles par nos annotations.

3.2 *Collecte du corpus EmoTV*

3.2.1 Buts

Nous avons collecté un corpus vidéo dans le but d'étudier les comportements émotionnels multimodaux humains, hors laboratoire. Deux choix ont été faits pour cette collecte, le premier concernant la taille du corpus à étudier, et le deuxième concernant le contenu du corpus. La taille d'un corpus et la granularité des annotations dépendent des objectifs de recherche visés. Pour la détection automatique, les approches statistiques requièrent de grands ensembles de données. Notre objectif étant l'étude exploratoire des émotions naturelles complexes et des comportements multimodaux, l'annotation fine d'un petit ensemble de vidéos peut être plus appropriée. Concernant les comportements multimodaux, l'objectif de ce corpus est de fournir des connaissances sur la coordination entre plusieurs modalités lors de comportements non actés émotionnellement riches. EmoTV peut être utilisé pour étudier la représentation de comportements individuels à plusieurs niveaux: émotions, comportements multimodaux, et contextes d'interaction. Il peut également être utilisé pour étudier l'extraction automatique de certains paramètres des comportements multimodaux spontanés mais de basse résolution (vidéo télévisée).

3.2.2 Critères de sélection

Plusieurs moyens peuvent être utilisés pour collecter des données vidéo, l'enregistrement en laboratoire, l'enregistrement « sur le terrain », ou la collecte d'extraits préenregistrés, provenant de la télévision par exemple. L'enregistrement en laboratoire permet de faciliter la contrôlabilité et l'exploitation du corpus (haute résolution, pas de problèmes concernant les droits de diffusion). Cependant il est nécessaire de trouver un contexte émotionnel contraint dans lequel la personne enregistrée fait abstraction de la caméra et ne simule pas ses comportements.

Notre choix s'est porté sur des extraits préenregistrés car nous avions pour objectif d'explorer les émotions spontanées exprimées dans divers contextes. Après enregistrement et visionnage de plusieurs types d'émissions (reportages, documentaires, journaux télévisés) susceptibles de contenir des comportements émotionnels et multimodaux, notre choix s'est porté sur les interviews provenant de journaux télévisés. Ces derniers sont en général plus spontanés que les talk show ou la télé réalité, car les interviews ont lieu « à chaud » dans l'environnement

des personnes, tandis que dans les cas de talk show ou de télé réalité, les personnes viennent en sachant qu'elles vont être filmées (même si elles peuvent plus tard faire abstraction de la caméra et agir spontanément). D'autres contraintes concernent le contexte d'interaction, dans lequel on trouve par exemple le type d'interaction (monologue ou interaction entre 2 ou plusieurs personnes), ou encore la position des personnes. Nous nous sommes ainsi fixé les critères suivants pour la récolte du corpus vidéo : les clips proviennent d'interviews télévisées en français (interaction entre 2 personnes, 1 seule visible), dans des situations réelles, contenant des émotions, et présentant des signes multimodaux : parole, expressions faciales, gestes, mouvements du corps.

3.2.3 Description quantitative

Notre corpus EmoTV est composé de 89 extraits vidéo contenant des comportements émotionnels d'individus interviewés dans des journaux télévisés (Abrilian et al. 2005a/2005b). La collecte du corpus EmoTV a été effectuée en deux parties. 1) La sélection et extraction de 50 interviews provenant de journaux télévisées français (d'une durée totale de 12 minutes, avec des extraits de 4 à 43 secondes, et un total de 2500 mots dont 800 distincts) à partir d'une douzaine de journaux télévisés provenant de chaînes différentes (TF1, France 2, TV5, LCI…), équivalent à environ 10h d'enregistrements). Une première étude de ce corpus (annotation des émotions et comportements multimodaux décrits dans les sections 3.3 et 3.4) a été effectuée. 2) Nous avons ensuite affiné les critères de sélection et complété le corpus pour arriver à un total de 89 extraits, présentant des émotions complexes uniquement, et des comportements riches en expressions multimodales.

3.2.4 Points forts et limites du corpus

Ce corpus présente plusieurs intérêts, tels que la présence de comportements émotionnels multimodaux spontanés dans une large variété de contextes : les personnes enregistrées sont de différents âges, équilibrées entre hommes et femmes, et les situations dans lesquelles elles interviennent sont variées. Il a également des désavantages, tels que le manque éventuel de visibilité des gestes et des expressions faciales (cachées par les lunettes, les cheveux ou encore la barbe), sachant que la caméra et les individus sont parfois en mouvement, en environnement publique quelque fois bruyant. De plus, le corpus EmoTV n'est pas diffusable.

3.2.5 Conclusion

Les sections suivantes présentent les différentes phases d'annotation effectuées sur le corpus EmoTV (manuelles et automatiques), ainsi que les analyses de ces annotations pour la représentation et la perception des émotions et comportements multimodaux, l'étude de leurs relations, et la copie-synthèse de comportements émotionnels multimodaux avec un ACA.

3.3 *Annotation manuelle des émotions*

Nous avons catégorisé les annotateurs ayant participé à nos expérimentations selon différents niveaux d'expertise dans le domaine des émotions (voir section 1.3.2). Ainsi, des sujets que nous avons évalué en tant qu'« experts » en émotion ont participé aux expériences décrites dans les sections 3.3.2, 3.3.3 et 3.3.5, et des sujets « naïfs », potentiels utilisateurs d'ACA même s'ils ne sont pas experts en émotion, ont participé à l'expérience décrite en 3.3.4.

3.3.1 Schéma d'annotation

Le but du corpus EmoTV est d'apporter des connaissances sur les relations entre modalités lors de comportements naturels émotionnellement riches. L'objectif à long terme de la série d'expériences réalisées pendant cette thèse était de modéliser les relations entre émotions complexes spontanées et comportements multimodaux dans des contextes variés. Quatre expériences d'annotation des émotions ont été réalisées au cours de la thèse. Elles ont permis d'explorer différentes directions d'annotation de ces émotions, et de proposer un système de représentation de ces émotions. L'outil d'annotation Anvil (Kipp, 2001) a été utilisé dans toutes les expériences d'annotation effectuées. L'observation de données naturelles nous a permis d'identifier les différents niveaux de codage à retenir pour l'étude des relations entre émotions réelles et comportements multimodaux. En effet, la difficulté principale dans la définition du schéma de codage permettant d'annoter et de représenter de tels comportements émotionnels réalistes, est de trouver les niveaux de description utiles en terme de granularité et de temporalité. Les spécificités du schéma de codage multi-niveaux utilisé pour EmoTV (dont nous détaillons les étapes dans cette section) sont : la possibilité d'annoter les étiquettes d'émotion ainsi que les dimensions abstraites dans le but d'étudier leur redondance et complémentarité ; la définition de combinaisons d'émotions non basiques ; l'utilisation de plusieurs étiquettes pour annoter un seul comportement émotionnel ; le contexte émotionnel incluant en autres les but communicatifs et des dimensions basées sur les théories d' « appraisal » ; une description temporelle de la variation d'intensité de l'émotion dans chaque segment ; une description globale des signes d'émotions perçus dans les différentes modalités ; et une description plus détaillée des comportements multimodaux de chaque segment (Abrilian et al., 2005).

3.3.2 Annotation des émotions avec un seul label et perception audio/vidéo

Nous avons demandé à 2 étudiants de maîtrise de psychologie d'annoter le corpus EmoTV à l'aide d'un premier schéma de codage des émotions. L'objectif de cette expérience était d'évaluer la complexité de l'annotation manuelle des émotions exprimées dans des données audiovisuelles spontanées, en particulier la difficulté de nommer une émotion ou les problèmes de segmentation des émotions. L'expérience a été réalisée dans 3 conditions, 1) audio seule, 2) vidéo seule, 3) audio et vidéo, de manière à évaluer l'impact de chaque modalité sur la perception de l'émotion. Chacune de ces conditions a été découpée en 2 phases. Dans la première phase, les annotateurs ont été amenés à détecter les évènements émotionnels perçus comme cohérents dans les extraits vidéo, et à les délimiter dans le temps par des segments. A la fin de cette phase, nous avons conservé une segmentation commune en fusionnant les deux segmentations individuelles (union des bornes temporelles choisies par les deux annotateurs). La deuxième phase a consisté à assigner un label à l'émotion perçue dans chaque segment, sous forme de texte libre. Par la suite, un schéma de codage des comportements multimodaux a également été utilisé par ces étudiants pour annoter les indices multimodaux. Cette autre expérience est décrite dans la section suivante.

Annotateurs

Deux étudiants en maîtrise de psychologie : Marianne et Fabien, qui ont travaillé pendant 3 mois sur ce sujet.

Vidéos sélectionnées

La première expérience a été réalisée sur les 50 premiers extraits vidéo d'EmoTV.

Schéma de codage

Le premier schéma était constitué d'une piste « émotion » contenant 3 attributs (Figure 38): un champ libre pour étiqueter l'émotion perçue (en utilisant un seul mot), et deux dimensions abstraites, l'intensité et la valence de l'émotion, échelonnées entre 1 (respectivement faible ou négatif) et 7 (respectivement forte ou positive).

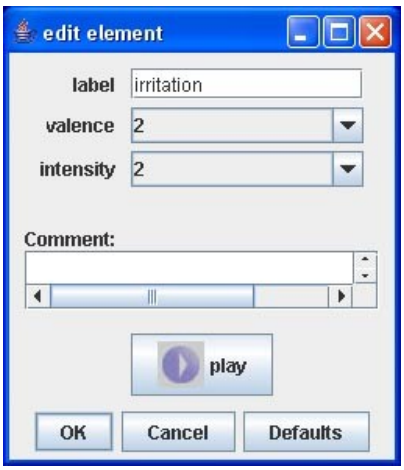


Figure 38: Fenêtre provenant de l'outil d'annotation Anvil (Kipp, 2001). Trois attributs sont définis pour annoter les émotions : un label (texte libre), la valence, et l'intensité.

Analyses des résultats

A l'issue de cette première phase, les deux ensembles de segments créés ont été regroupés en un ensemble de segments émotionnels consensuels par les deux annotateurs. Après normalisation, 176 labels différents ont été extraits. Nous avons ensuite demandé aux 2 annotateurs de regrouper les labels sémantiquement proches, ce qui a donné l'ensemble suivant, réduit à 14 labels : colère, désespoir, dégoût, doute, exaltation, peur, irritation, joie, neutre, douleur, tristesse, sérénité, surprise, et inquiétude.

Des mesures d'accord inter-annotateurs, Kappa (Carletta, 1996) pour les catégories d'émotion, et Alpha (Cronbach, 1951) pour les dimensions abstraites, ont été effectuées sur les annotations manuelles des émotions à partir des 14 labels. Ces trois mesures montrent que les accords inter-annotateurs étaient les plus bas pour la condition mixte « Audio et Vidéo », ensuite pour « Vidéo seule », et plus élevés pour « Audio seule » (Table 7). Le degré d'accord (Kappa ou Alpha) peut être considéré excellent si la valeur est supérieure à 0.8, bon entre 0.6 et 0.8, moyen entre 0.4 et 0.6, et faible en dessous.

Condition	Kappa sur les labels d'émotion	Alpha sur l'intensité des émotions	Alpha sur la valence des émotions
Audio seule	0.540	0.865	0.915
Vidéo seule	0.430	0.510	0.709
Audio et Vidéo	0.370	0.254	0.574

Table 7: Accords inter-annotateurs: Kappa sur les émotions, Alpha sur l'intensité et la valence

La Figure 39 montre la répartition des émotions annotées (en pourcentage) par les 2 annotateurs : joie, neutre, tristesse, et colère, sont les labels les plus fréquemment annotés, et totalisent à eux quatre 57.6% des annotations.

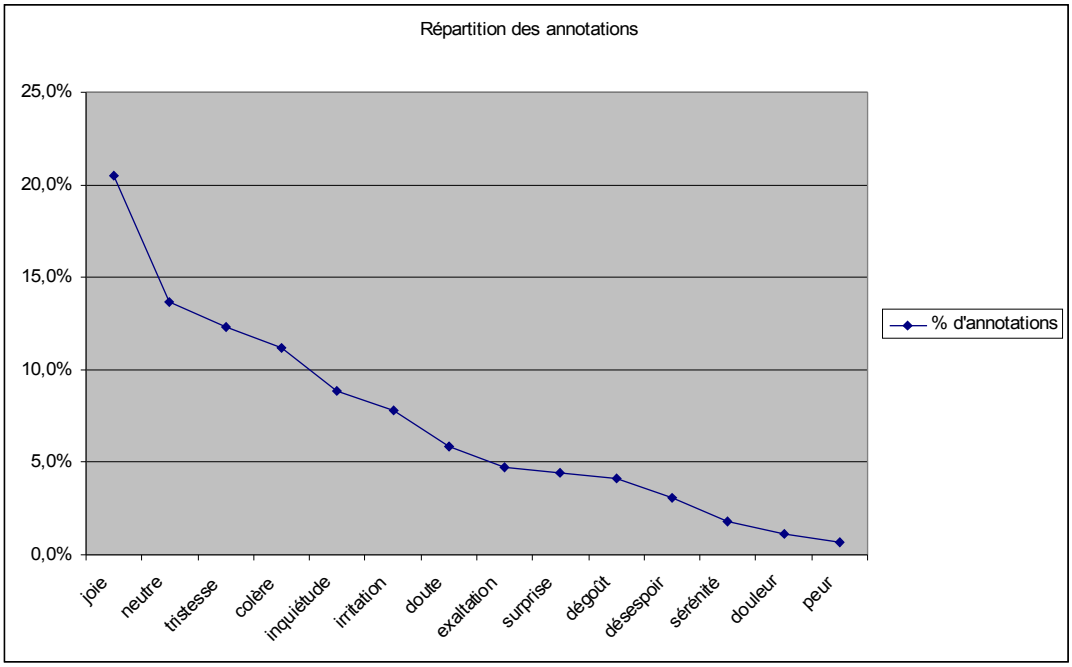


Figure 39: Répartition des annotations dans chaque émotion (en pourcentage, moyenne des 3 conditions).

Une étude individuelle de chaque cas de désaccord a montré que les faibles valeurs d'accord inter-annotateurs obtenues dans la condition « Audio et Vidéo » s'expliquent par le fait que les émotions naturelles contiennent des ambiguïtés et des indices conflictuels notamment entre le signal audio et les comportements multimodaux. Par exemple, les gens masquent parfois leurs émotions, ou essaient de les contrôler. Parmi les ambiguïtés retrouvées dans les analyses, certaines étaient même dues à des désaccords négatif/positif : 3% des désaccords en « Audio seul », 7% en « Vidéo seule », et 11% en « Audio et Vidéo ». Les ambiguïtés les plus fréquentes concernent cependant les classes négatives peur/colère/tristesse/dégoût, à différents niveaux d'intensité. En effet, dans plusieurs extraits, on peut percevoir de la tristesse et de la colère en même temps, soit en raison d'indices multimodaux conflictuels dans la parole et dans les expressions faciales, soit lors de transitions entre ces deux émotions. Les valeurs de valence sélectionnées par les deux annotateurs correspondaient bien à la valence des labels sérénité, joie, exaltation (c'est-à-dire les labels positifs) et du label neutre, mais ne correspondaient pas toujours à celle des labels négatifs. Cette ambiguïté peut s'expliquer par

le fait que, lors de comportements émotionnels naturels et complexes, deux types d'émotions peuvent être perçus en même temps, l'émotion de l'interviewé, basée sur son état émotionnel interne, et l'effet qu'il voudrait produire sur la personne qui l'écoute.

Conclusions

Cette première expérience a permis d'évaluer la difficulté de l'annotation des émotions dans des contextes naturels. L'étude des trois conditions audio, vidéo et audio/vidéo a permis de constater les conflits apparaissant parfois entre canal audio et vidéo dans la perception des émotions, et d'en tenir compte dans l'expérience suivante en ajoutant la possibilité de mettre 2 étiquettes verbales pour un même segment émotionnel. L'ensemble des 14 catégories d'émotion résultant de cette première expérience a été utilisé pour la phase d'annotation suivante.

3.3.3 Annotation des émotions avec deux labels et contexte d'enregistrement

Une partie du corpus EmoTV (11 vidéos) a été annotée par 3 sujets « experts » (Deux enseignant-chercheurs et un doctorant travaillant dans le domaine des émotions). Pour cette deuxième expérience, nous avons défini un nouveau schéma de codage tenant compte des résultats et difficultés rencontrées lors de la première expérience. Un des objectifs de cette 2ème expérience était l'évaluation de l'accord inter-annotateurs sur l'annotation des émotions et du contexte émotionnel. En effet, les émotions que l'on voit dans la réalité sont fortement dépendantes du contexte dans lequel elles apparaissent. Comment le représenter ? Nous avons donc ajouté une partie contextuelle au schéma de codage, dont l'objectif est d'apporter des informations qui pourraient expliquer l'état émotionnel de l'interviewé : les événements cause de l'émotion, la situation dans laquelle le personnage est interviewé, les conditions d'enregistrement…

Nous avons opté pour un nombre impair d'annotateurs, plus généralement utilisé pour calculer les accords inter-annotateurs, en permettant par exemple des votes majoritaires.

Pour l'annotation des labels d'émotion, les participants ont le choix entre les 14 labels provenant de la première expérience, ainsi que d'un label « autre » : si l'émotion qu'ils perçoivent ne se trouve pas dans la liste, ils peuvent ajouter un nouveau terme dans les commentaires.

Vidéos sélectionnées

11 vidéos contenant un total de 48 segments

Schéma de codage

Suite à notre première expérience, nous avons ajouté la possibilité d'annoter chaque segment émotionnel avec deux labels (Majeur et Mineur), et défini une typologie de combinaisons des émotions non basiques à partir des cas de désaccords observés dans la première expérience (Figure 40) :

- *mélangées* : deux émotions apparaissent en même temps
- *masquée-actée* : la personne filmée masque son émotion, comme par exemple en souriant mais avec une réelle déception
- *séquence* : deux émotions apparaissent peu de temps l'une après l'autre, dans un même segment émotionnel
- *conflit de cause à effet* : par exemple un conflit positif/négatif, pleurer de joie.
- *ambiguïté* : entre deux émotions de la même classe, par exemple colère et irritation

Cette typologie des combinaisons émotionnelles que nous proposons est basée sur nos données expérimentales. Elle est complémentaire à celle proposée par (Bevacqua et al. 2007) qui est inspirée des travaux théoriques d'Ekman et qui est composée de trois opérations appliquées sur des expressions d'émotions de base : modulation (diminuer ou augmenter l'intensité), falsification (par simulation, neutralisation ou masquage), et qualification (en ajoutant une expression feinte à l'expression d'une émotion ressentie).

La variation temporelle de l'intensité (croissant, décroissant, stable…) dans chaque segment a été ajoutée au schéma de codage. Cette information temporelle étant reconnue comme importante dans la perception dynamique des émotions et dans leur génération par des ACAs, nous voulions explorer la possibilité de l'annoter manuellement. En plus des deux dimensions abstraites présentes dans le schéma de codage de la première expérience, deux nouvelles dimensions ont été ajoutées : l'activation (échelle allant de 1-passif à 5-actif), et le contrôle (échelle allant de 1-incontrolé à 5-controlé) dans le but d'évaluer leur pertinence pour l'étude des émotions naturelles.

Cinq nouveaux ensembles d'attributs représentant le contexte ont été intégrés dans le schéma de codage. Ces cinq informations contextuelles nous sont apparues comme pertinentes durant la première expérience. Le premier ensemble se nomme « contexte émotionnel », il inclut des dimensions d'appraisal décrivant les événements causes des émotions : degré d'implication (faible à élevé), événement-cause (texte libre), relation personne-événement (sujet de société, histoire personnelle…), moment de l'évènement (passé, récent, présent, futur). Les autres ensembles sont le « contexte d'interview » : thème, place. ; la « personne filmée » : âge, genre, race ; le « but communicatif global » : message transmis (partager un sentiment, se plaindre, réclamer…), à qui (public, parent, collègue…) ; et enfin le « contexte d'enregistrement : caméra (statique, dynamique), personne filmée (statique, dynamique), qualité audio, qualité vidéo.

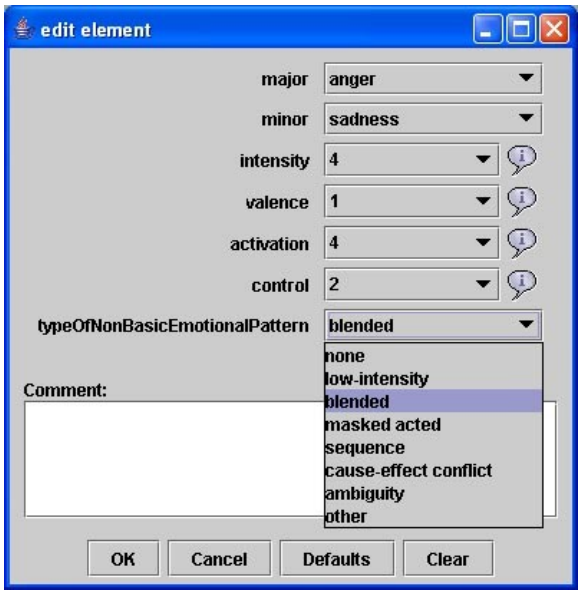


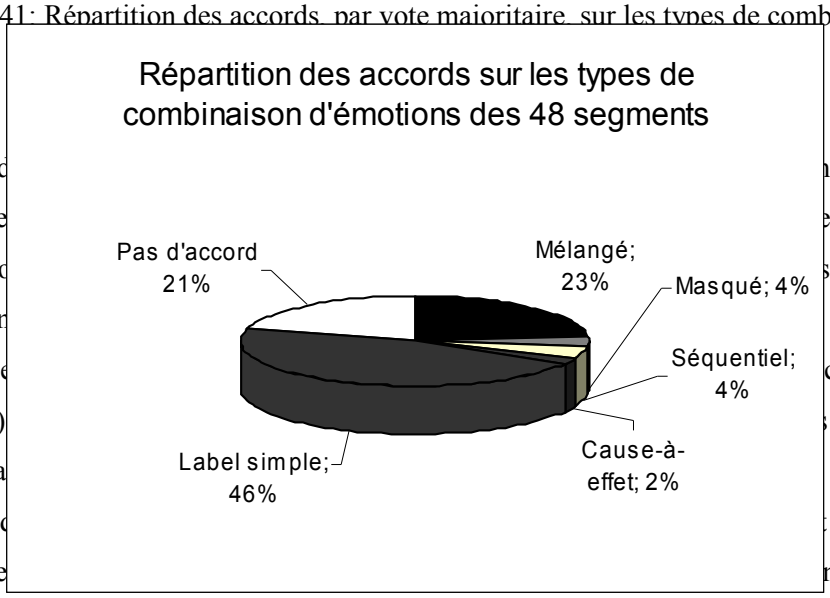
Figure 40: Pour la 2ème expérience, les annotateurs peuvent utiliser 2 étiquettes (majeur, mineur), 4 dimensions abstraites (intensité, valence, activation, contrôle), et un attribut de combinaison des émotions sont définis pour l'annotation.

Analyses des résultats

Les trois annotateurs experts ont annoté les 48 segments provenant de 11 extraits vidéo d'EmoTV. Un temps moyen de 15 minutes leur a été nécessaire pour l'annotation de chaque extrait. La Figure 41 résume les accords sur les types de combinaisons d'émotions annotés, en utilisant un vote majoritaire. Une description est acceptée si au moins 2 des 3 annotateurs l'ont utilisée. Sur ce critère, la proportion des segments sans accord sur la combinaison d'émotion est de 21%. Une information importante sortant de cette figure est que la différence

est relativement faible entre l'accord sur l'ensemble des combinaisons, 33% (23 + 4 + 4 + 2), et l'accord sur l'utilisation d'un label simple. Dans telle situation, il n'est pas surprenant d'avoir obtenu des désaccords entre les annotateurs lorsque nous avions demandé aux annotateurs lors de la première expérience d'assigner un unique label à l'émotion perçue.

Figure 41: Répartition des accords par vote majoritaire sur les types de combinaison



Le schéma c…nt (Majeur et Mineur), pe… moyen de représentatio… sidérant une combinaison… (poids de 2 pour le Maje… culés avec 3 annotateurs)… de fiabilité qu'en utilisa…

La fiabilité c… segmental). Le coefficie…ntinues. Ces résultats montrent un haut niveau de fiabilité (Table 8) pour l'intensité et la valence, et une plus haute fiabilité au niveau global qu'au niveau segmental. Dans cette expérience, l'activation (actif/passif) est la moins fiable des dimensions. L'intensité semble donc être plus fiable que l'activation.

Dimension	Alpha sur le clip global	Alpha sur les 48 segments
Intensité	0.896	0.750
Valence	0.739	0.761
Self-control	0.826	0.464
Activation	0.455	0.448

Table 8. Alpha de Cronbach pour chaque dimension annotée.

L'analyse des annotations du contexte montre qu'EmoTV contient des extraits prenant place dans des thèmes variés : politique, justice, sport, etc. Les buts communicatifs négatifs (critiquer, se plaindre, etc.), annotés dans le contexte émotionnel global, sont fortement liés aux valences négatives obtenues, après vote majoritaire. Mais lorsque le type de combinaison d'émotion est annoté avec la valeur « conflit de cause à effet » ou « acté-masqué », il n'y a pas d'accord inter-annotateur pour la valence, et de plus le but communicatif global n'explique pas la valence de la séquence (par exemple « expliquer », « partager », etc.). Ce

résultat est appuyé par le fait que la valence et les labels annotés pour chaque segment sont fortement liés, excepté en ce qui concerne les 2 types de combinaison d'émotion ci-dessus. La relation entre la valence et les labels verbaux est donc un moyen de valider nos annotations mais également de souligner la complexité de certains segments.

Les annotateurs ont également du identifier les indices multimodaux qu'ils ont jugés importants pour le choix de leur annotation. La Table 9 présente ces résultats. Le point principal est que les indices audio et faciaux étaient impliqués dans la grande majorité des segments. Dans tous les cas, au moins 2 modalités avaient été jugées pertinentes par les annotateurs pour la perception de l'émotion. Le nombre moyen d'indices multimodaux par segment est entre 3 et 4 (177/48). De plus, dans le sous-ensemble du corpus étudié, les gestes, les mouvements du torse et des épaules apparaissent dans des segments annotés avec des labels verbaux négatifs.

Une analyse des relations entre intensité et contrôle obtenus après vote majoritaire, et du nombre d'indices multimodaux par segment a été effectuée. 74% des segments contenant un nombre d'indices multimodaux supérieurs à la moyenne ont été annotés avec un haut niveau d'intensité. Quand le contrôle est élevé, le nombre d'annotations multimodales a tendance à diminuer.

Cette analyse permet de valider notre schéma de codage et d'étudier les relations entre la perception d'indices multimodaux annotés à un niveau global d'une vidéo et les représentations des émotions.

Indices Multimodaux	#segments (total de 48 segments)	% indices All segments	% indices Nég segments	% indices Neu/Pos segments
Parole	44	92%	96%	82%
Yeux	40	83%	84%	82%
Bouche	31	65%	64%	65%
Tête	29	60%	61%	59%
Sourcils	16	33%	29%	41%
Gestes	10	21%	29%	6%
Torse	4	8%	13%	0%
Epaules	3	6%	10%	0%

Table 9. Fréquence à laquelle différents indices multimodaux ont été annotés pertinents pour le jugement d'une émotion (un indice est accepté si au moins 2 des 3 codeurs l'annotent). Le

pourcentage de chaque indice est donné pour les segments Négatifs et Neutres/Positifs dans les 2 dernières colonnes. La proportion de segments (sur les 48 segments) annotée avec les labels négatifs (gagnant du vecteur) est de 65%.

Conclusions

Dans les comportements émotionnels collectés dans notre corpus TV, différents canaux sont actifs, et peuvent se superposer ou être en conflit. Nous avons identifié certains types d'émotions combinées, tels que les « conflits de cause à effet » et les émotions « actées/masquées », difficiles à juger et annoter, ne serait-ce que pour la dimension valence. Cette expérience a soulevé le problème des différences de perception des émotions entre différents annotateurs et nous a mené à l'expérience suivante sur la variabilité inter-annotateurs, pour étudier l'importance du nombre d'annotateurs, et les différences de perception entre sujets de sexes, d'âges, ou d'expertise (sur l'étude des émotions) différents.

3.3.4 Annotation par un grand nombre de sujets pour l'étude des différences individuelles

La troisième expérience était un test perceptif effectué sur 40 sujets. Elle avait pour buts de valider sur 1 vidéo les annotations précédentes de mélanges d'émotions, et d'étudier la variabilité d'annotation entre groupes d'annotateurs. Des différences entre hommes et femmes sont en effet rapportées dans la littérature en Sciences Sociales (Knapp et Hall 2006) mais finalement peu exploitées dans la communauté des chercheurs utilisant des corpus émotionnels multimodaux.

Ce test a impliqué un grand nombre d'annotateurs, en comparaison à des études expérimentales similaires dans la communauté des corpus.

Chaque sujet devait annoter les segments avec les informations suivantes : labels émotionnels, type de combinaison d'émotions, dimensions abstraites, indices multimodaux globaux. Chaque sujet devait également remplir un questionnaire, donnant des informations sur son âge, son expertise en annotation des émotions, ainsi que des commentaires sur la difficulté de l'expérience.

Annotateurs

La Table 10 montre les groupes d'annotateurs et leur répartition. Dans cette expérience, nous avons étudié les différences entre hommes et femmes. Les autres groupes se sont révélés trop déséquilibrés pour fournir des résultats exploitables. Les sujets étaient âgés de 21 à 55 ans (avec une moyenne de 28 ans). La classe Age était séparée en « Etudiant » / « Non étudiant ». Très peu d'experts en émotion ont été recensés, en comparaison des annotateurs « naïfs ».

Sexe	Homme (25)	Femme (15)
Age	Etudiant (30)	Non étudiant (10)
Expertise	Non expert (34)	Expert (6)

Table 10: Répartition des annotateurs dans 3 groupes.

Vidéos sélectionnées

L'expérience a été réalisée sur les 3 segments émotionnels de la vidéo #3 du corpus.

Schéma de codage

Le schéma utilisé était en grande partie le même que pour l'expérience 2. Afin de permettre une annotation plus fine et correspondant mieux aux données collectées, la liste des labels verbaux pour l'annotation des émotions est passée de 14 termes aux 18 termes suivants : colère, désespoir, déception, dégoût, doute, embarras, exaltation, peur, irritation, joie, neutre, content, fierté, tristesse, sérénité, honte, surprise, et inquiétude. Sept macros-classes (en gras dans la Table 11) ont été créées pour regrouper certains labels.

Labels négatifs	**Labels positifs**	**Autres labels**
Colère *Irritation*	***Joie*** *Exaltation* *Content* *Sérénité* *Fierté*	***Surprise***
Dégoût		
Peur *Doute* *Inquiétude*		
Embarras *Honte*		
Tristesse *Désespoir* *Déception*		

Table 11: Sept macro-classes de labels (sans *Neutre*).

Analyses des résultats

Plutôt que d'utiliser uniquement la variabilité inter annotateurs en tant que mesure de validation, notre stratégie de validation a considéré également la cohérence intra annotateur en vérifiant la compatibilité entre les dimensions et les labels verbaux sélectionnés par chaque annotateur. 11 des 40 sujets montrant des problèmes de cohérence de ce type ont été observés de cette manière. Mais ces erreurs n'apparaissaient que pour un attribut pour chaque codeur. Les annotations de ces 11 sujets ont donc été conservées pour l'analyse inter-annotateurs et la création des vecteurs d'émotions.

La Table 12 présente une comparaison des résultats d'annotations obtenus à partir de l'expérience 3 et de l'expérience 2 (en ne considérant que les annotations de la vidéo numéro 3). Comme nous l'avons vu lors de la section précédente, l'expérience 2 était réalisée avec 3 sujets (2 hommes/1 femme, 1 étudiant/2 non étudiants, 3 experts). 3 des segments annotés lors de l'expérience 2 sont les mêmes que les 3 segments annotés dans l'expérience 3. Les vecteurs de labels, de types de combinaisons d'émotions, ainsi que les dimensions : intensité, valence, contrôle, et activation ont été calculés à partir des annotations et sont listés dans cette table.

		Expérience 3.3.3 (3 sujets)	Expérience 3.3.4 (40 sujets)
Segment 1	Labels émergeants	1.00 **Colère**	0.282 **Colère**, 0.214 Irritation, 0.171 Inquiétude, 0.145 Désespoir
	Combinaison	1.00 **Simple**	0.32 Mélangé, 0.30 **Simple**, 0.17 Acté-Masqué, 0.15 Ambiguité
	Intensité moy.	**3.3**	**3.2**
	Valence moy.	2.0	2.3
	Control moy.	3.6	3.1
	Activation moy.	**3.3**	**3.1**
Segment 2	Labels émergeants	0.67 **Colère**, 0.11 **Désespoir**, 0.11 Déception, 0.11 **Tristesse**	0.317 **Colère** , 0.183 **Tristesse** , 0.158 Dégoût , 0.133 **Désespoir** , 0.100 Irritation
	Combinaison	0.67 **Mélangé** , 0.33 **Simple**	0.30 Séquentiel, 0.25 **Mélangé**, 0.22 Acté-Masqué, 0.20 **Simple**
	Intensité moy.	**4.0**	**3.7**
	Valence moy.	1.3	2.1
	Control moy.	2.6	2.8
	Activation moy.	**4.0**	**4.0**
Segment 3	Labels émergeants	0.56 **Désespoir**, 0.33 **Colère**, 0.11 **Tristesse**	0.383 **Désespoir**, 0.250 **Tristesse**, 0.208 **Colère**
	Combinaison	0.67 **Mélangé**, 0.33 **Séquentiel**	0.37 **Séquentiel**, 0.25 **Mélangé**, 0.15 Simple, 0.12 Ambiguité, 0.10 Acté-Masqué
	Intensité moy.	3.6	4.4
	Valence moy.	1.0	1.9
	Control moy.	1.3	2.1
	Activation moy.	**4.6**	**4.3**

Table 12: Comparaison de 2 expériences d'annotation, sur les trois mêmes segments d'une vidéo.

Ces résultats montrent que le label « gagnant » est le même dans les 2 expériences, pour chacun des 3 segments (colère pour les segments 1 et 2, désespoir pour le segment 3). De

plus, pour les segments 2 et 3, les 3 mêmes labels : désespoir, colère, et tristesse, sont annotés par les 2 groupes d'annotateurs.

Ce résultat permet de valider les annotations sur cette vidéo, de consolider l'utilisation de 3 annotateurs experts pour ce genre d'expérience, mais aussi de souligner la richesse de la perception fournie par un grand nombre d'annotateurs, qui fait apparaître des contributions plus subtiles pour certains labels.

Combinaisons d'émotions

Concernant les types de combinaisons d'émotions, 78% des 40 sujets (9% de plus pour les femmes que pour les hommes) ont sélectionné un type de combinaison différent de « simple » pour les 3 segments ; 69% pour le segment 1, 80% pour le segment 2, et 85% pour le dernier segment. Comme nous pouvons le voir dans la Figure 42 aucune femme n'a choisi le type : « simple », pour le segment 3.

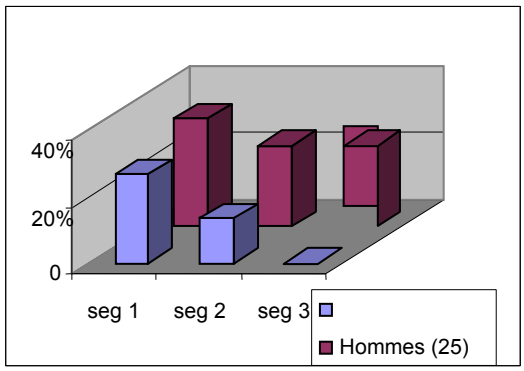


Figure 42: Proportion des types de combinaison égaux à "simple" pour les hommes et les femmes, et pour les 3 segments du clip.

La répartition des types de combinaison d'émotions des 40 annotateurs est donnée dans la Figure 43. Le premier segment a été perçu principalement avec des types de combinaison « mélangé » et « acté masqué », mais également avec un nombre important d' « ambiguïtés ». Les annotateurs ont perçu plus d'émotions « séquentielles » dans le 2ème segment, en comparaison au 1er. Dans le dernier segment, moins d'émotions « actées-masquées » sont perçues, ce qui correspond au fait que la femme apparaissant dans l'extrait vidéo ne contrôle plus ses émotions à la fin de l'extrait.

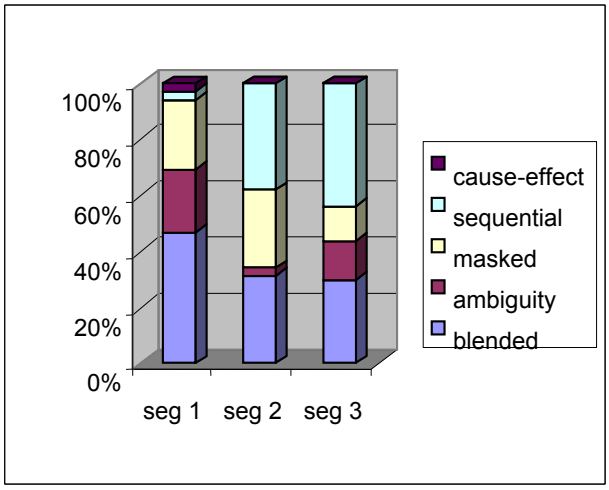


Figure 43: Répartition des 78% de types de combinaisons différents de "simple" pour les 3 segments (total de 69% pour le segment 1, de 80% pour le segment 2, et de 85% pour le segment 3).

Ces résultats montrent la complexité des combinaisons d'émotions. Ils confirment que la perception des émotions, et plus particulièrement celle des combinaisons d'émotions diffère entre hommes et femmes. Les femmes ayant participées à cette expérience ont perçu 25% plus d'émotions « ambiguës » et 15% plus d'émotions « masquées » que les hommes.
La Table 13 montre le label « gagnant » obtenu par vote majoritaire sur les 40 annotateurs.

Les différences sont uniquement visibles entre les groupes hommes et femmes.

	Seg 1	**Seg 2**	**Seg 3**
Tous les sujets	*Colère*	*Colère*	*Désespoir*
Hommes	*Colère*	*Colère*	*Tristesse*
Femmes	*Inquiétude*	*Tristesse*	*Désespoir*
Etudiants	*Colère*	*Colère*	*Désespoir*
Non étudiants	*Colère*	*Colère*	*Désespoir*
Experts	*Colère*	*Colère*	*Désespoir*
Non experts	*Colère*	*Colère*	*Désespoir*

Table 13: Le label "gagnant" obtenu par vote majoritaire sur chaque segment.

Dimensions

Concernant les dimensions, pour chaque label annoté, la moyenne et l'écart-type de l'intensité, de la valence, du contrôle, et de l'activation ont été calculés. Ces valeurs sont

résumées dans la Table 14. Les valeurs pour les 4 attributs sont échelonnées entre 1 et 5 ; 1 correspondant à « faible » ou « négatif », et 5 correspondant à « élevé » ou « positif ». Les variations sont similaires entre les segments, entre les 3 groupes (hommes, femmes, tous les sujets). Pour les 3 groupes, l'intensité et l'activation moyennes augmentent dans la séquence des 3 segments. La valence moyenne est négative et semble devenir encore plus négative à la fin de l'extrait. Le contrôle moyen décroît également à la fin de l'extrait.

Les femmes ont perçu le 3ème segment plus négatif et plus intense que les hommes. Ceci est cohérent avec le label Majeur choisi par les annotateurs ; désespoir pour les femmes et tristesse pour les hommes. Globalement, on retrouve un bon accord inter-annotateurs (écart-type souvent inférieur à 1). De plus, l'écart-type diminue dans le dernier segment, ce qui montre une meilleure fiabilité des annotations pour le dernier segment (ce qui pourrait être du à un apprentissage de la tâche d'annotation, à la disposition d'informations contextuelles qui sont plus riches pour le dernier segment (puisque le sujet a alors pris connaissance d'une bonne partie de la vidéo), ou au comportement exprimé dans le dernier segment, peut être moins complexe que celui des deux premiers segments).

Intensité	**Seg 1**	**Seg 2**	**Seg 3**
Tous les sujets	3.2 (1.0)	3.7 (0.8)	4.4 (0.7)
Hommes	3.2 (1.1)	3.8 (0.8)	4.3 (0.8)
Femmes	3.1 (0.7)	3.6 (0.7)	4.6 (0.5)
Valence	**Seg 1**	**Seg 2**	**Seg 3**
Tous les sujets	2.3 (0.6)	2.1 (1.0)	1.9 (1.3)
Hommes	2.4 (0.4)	2.2 (1.0)	2.0 (1.1)
Femmes	2.1 (0.6)	1.9 (1.0)	1.7 (1.4)
Activation	**Seg 1**	**Seg 2**	**Seg 3**
Tous les sujets	3.1 (1.0)	4.0 (0.8)	4.3 (0.7)
Hommes	3.1 (1.1)	4.2 (0.7)	4.2 (0.6)
Femmes	3.0 (0.8)	3.9 (0.8)	4.3 (0.8)
Contrôle	**Seg 1**	**Seg 2**	**Seg 3**
Tous les sujets	3.1 (1.1)	2.8 (1.0)	2.1 (1.1)
Hommes	3.2 (1.1)	2.6 (1.1)	2.0 (1.1)
Femmes	2.7 (1.0)	2.9 (0.8)	2.1 (1.1)

Table 14: Moyennes et écarts types des annotations dimensionnelles : Intensité, Valence, Activation, et Contrôle.

Dans la Table 15, les indices multimodaux sont égaux à 1 si au moins la moitié des annotateurs a sélectionné la modalité correspondante ; dans le cas contraire, la valeur est égale à 0. Nous observons quelques différences entre hommes et femmes ; les hommes ont

sélectionné plus de mouvements du torse et des sourcils que les femmes, ces dernières ayant sélectionné plus de mouvements des yeux.

	Seg 1			**Seg 2**			**Seg 3**		
Groupes de sujets	Tous	H	F	Tous	H	F	Tous	H	F
Parole	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
Torse	0	0	0	0	**1**	0	0	**1**	0
Tête	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
Epaules	0	0	0	0	0	0	0	0	0
Yeux	**1**	0	**1**	**1**	**1**	**1**	**1**	**1**	**1**
Sourcils	0	0	0	0	**1**	0	0	0	0
Bouche	0	0	0	0	0	0	0	0	**1**
Gestes	0	0	0	**1**	**1**	**1**	**1**	**1**	**1**

Table 15: Indices multimodaux par groupes: tous, hommes, et femmes.

Estimation de la difficulté de la tâche d'annotation

Chaque sujet a répondu à une question (dans le questionnaire) concernant la difficulté de la tâche. Nous avons mesuré la relation entre la difficulté réelle de la tâche (estimée par la durée du test perceptif), et le niveau de difficulté qu'ont noté les sujets dans le questionnaire (Table 16). La durée moyenne d'un test était de 20 minutes (19.7 pour les femmes, 20.8 pour les hommes). Nous avons observé une répartition équilibrée des jugements concernant la difficulté du test dans les différents groupes.

Difficile	**Facile**
23mns	18mns

Table 16: Durée moyenne du test en fonction de la difficulté notée dans le questionnaire (« facile » pour les sujets qui ont notés facile et très facile, et « difficile » regroupe les sujets qui ont noté difficile ou très difficile).

Représentation vectorielle de la perception collective

Nous avons calculé, à partir des annotations des 40 sujets, un vecteur d'émotion pour chacun des 3 segments. Ces vecteurs ont été utilisés comme références pour la comparaison avec les vecteurs calculés à partir de différents sous-ensembles d'annotateurs. Chaque vecteur a la même taille, correspondant au nombre de labels total proposé lors de l'annotation (18). La Table 17 montre les coefficients non nuls des éléments de chaque vecteur. En ne prenant en compte que les éléments ayant un coefficient supérieur à 1/18ème, on obtient 4, 5, et 3 labels différents respectivement pour les 3 segments, reflétant la complexité des émotions. Ces vecteurs montrent également l'évolution des émotions de segment à segment. On peut voir cette évolution dans la combinaison Colère-Irritation (somme des deux valeurs), diminuant du segment 1 (0.50) au segment 3 (0.23), et dans la combinaison Tristesse-Désespoir, augmentant du segment 1 (0.19) au segment 3 (0.63). Les 40 annotateurs ont utilisé un nombre total de 10 labels pour décrire l'émotion perçue dans le premier segment, 9 pour le deuxième segment, et enfin 8 pour le dernier segment. Ceci montre 1) la subtilité des émotions complexes observées ici (utilisation de plus de la moitié des labels proposés dans le schéma), et 2) l'évolution temporelle de la complexité de l'émotion (moins de labels annotés de segment à segment).

Seg1	(0.29 **Colère**, 0,21 **Irritation**, 0.17 **Inquiétude**, 0.14 **Désespoir**, 0.07 Déception, 0.05 Tristesse, 0.02 Dégoût, 0.02 Honte, 0.02 Peur, 0.01 Fierté)
Seg2	(0.32 **Colère**, 0.18 **Tristesse**, 0.16 **Dégoût**, 0.13 **Désespoir**, 0.1, **Irritation**, 0.05 Déception, 0.02 Honte, 0.02 Inquiétude, 0.02 Fierté)
Seg3	(0.38 **Désespoir**, 0.25 **Tristesse**, 0.21 **Colère**, 0.09 Dégoût, 0.02 Irritation, 0.02 Déception, 0.02 Exaltation, 0.01 Honte)

Table 17: Vecteurs de référence pour les labels, calculés à partir des 40 annotateurs.

Dans le but d'évaluer la relation entre le nombre d'annotateurs impliqués dans l'annotation et le vecteur référence de chaque segment, nous avons calculé la distance moyenne entre différents ensembles d'annotateurs (Figure 44). La courbe décroît jusqu'à 15 annotateurs puis semble se stabiliser.

Nous avons commencé par considérer les annotations effectuées par les sous-ensembles de 3 sujets. Nous avons calculé une distance Euclidienne entre les annotations produites par tous les groupes de 3 annotateurs et les 3 vecteurs de référence (1 pour chaque segment). Les distances minimum et moyennes sont listées dans la Table 18.

Les distances minimum et moyenne ont tendance à diminuer du segment 1 au segment 3, illustrant l'importance de la connaissance du contexte (et aussi de l'apprentissage de la tâche) pour la sélection des labels. Dans cette étude, les 3 sujets les plus représentatifs des annotations des 3 segments sont toutes des femmes (8/8, 1 est sélectionnée 2 fois), étudiants (8/8), et 1/8 est un expert.

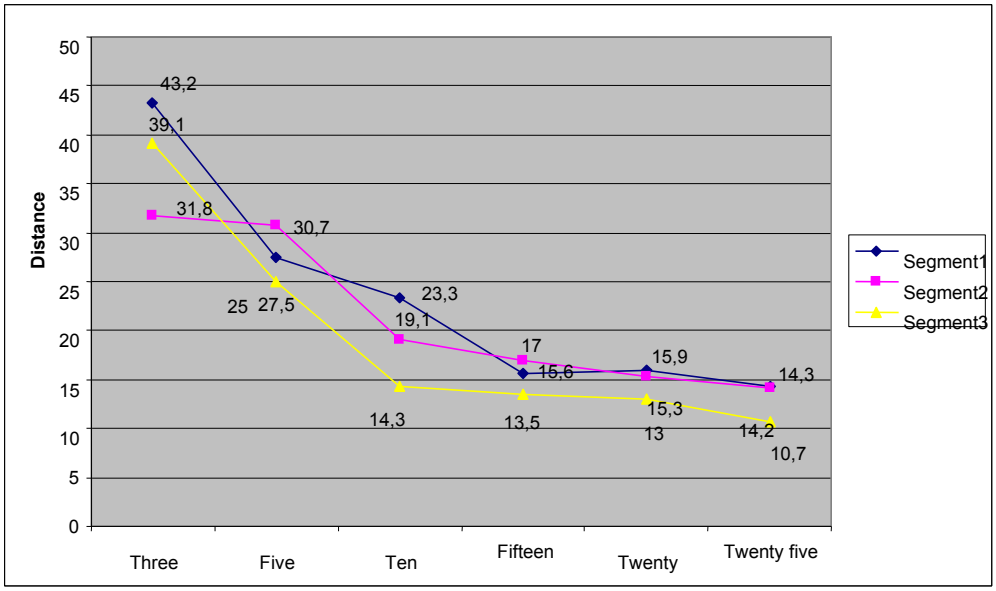


Figure 44: Distance moyenne entre le vecteur référence de chaque segment et le vecteur calculé à partir des sous-ensembles d'annotateurs (groupes de 3, 5, 10, 15, 20, 25 annotateurs) choisis (moyenne sur 10 tirages aléatoires).

3 annotateurs	Seg 1	Seg 2	Seg 3
Distance minimum	11.2	9.9	9.6
Distance moyenne	35.6	34.1	28.9

Table 18: Distances minimum et moyennes entre le vecteur calculé à partir d'un tirage aléatoire de 10 groupes de 3 sujets parmi les 40, et le vecteur référence, pour chaque segment.

La Table 19 montre les vecteurs calculés à partir des 7 macro-classes au lieu des 18 labels.

Seg1	(0.49 **Colère**, 0.26 **Tristesse**, 0.18 **Peur**, 0.03 Dégoût, 0.03 Embarras, 0.01 Joie)
Seg2	(0.41 **Colère**, 0.38 **Tristesse**, 0.15 **Dégoût**, 0.03 Peur, 0.03 Embarras, 0.01 Joie)
Seg3	(0.64 **Tristesse**, 0.24 **Colère**, 0.09 Dégoût, 0.02 Joie, 0.01 Embarras)

Table 19: Vecteurs de référence pour les macro-classes.

Ces vecteurs par macro-classes donnent une vue plus synthétique des émotions émergeant des annotations. Par exemple, dans le cas du segment 1, dans le vecteur de labels, colère, irritation et inquiétude sont les labels les plus élevés, mais les macro-classes permettent de voir ici que la tristesse est fortement présente, mais divisée dans les labels désespoir, tristesse et déception. La représentation avec des macro-classes amplifie les résultats observés avec les vecteurs de labels décrits précédemment ; colère et tristesse sont toujours présentes et varient de manière opposée de segment à segment. La Figure 45 présente l'évolution des distances moyennes entre vecteurs de macro-classes calculés à partir de tirages aléatoires de 3, 5, 10, 15, 20, 25 annotateurs, et le vecteur référence de macro-classes (des 40 sujets). En comparaison avec les courbes de convergence de la Figure 44, celles des macro-classes montrent des distances plus basses avec les vecteurs référence, et ce pour tous les sous-ensembles d'annotateurs. Nous remarquons dans les deux figures que les distances entre vecteurs référence et vecteurs de 3 sujets sont relativement élevées, puisque ces distances diminuent de manière significative pour les groupes de 5 et 10, et enfin la variation est moins visible pour les groupes de 15, 20 et 25 annotateurs.

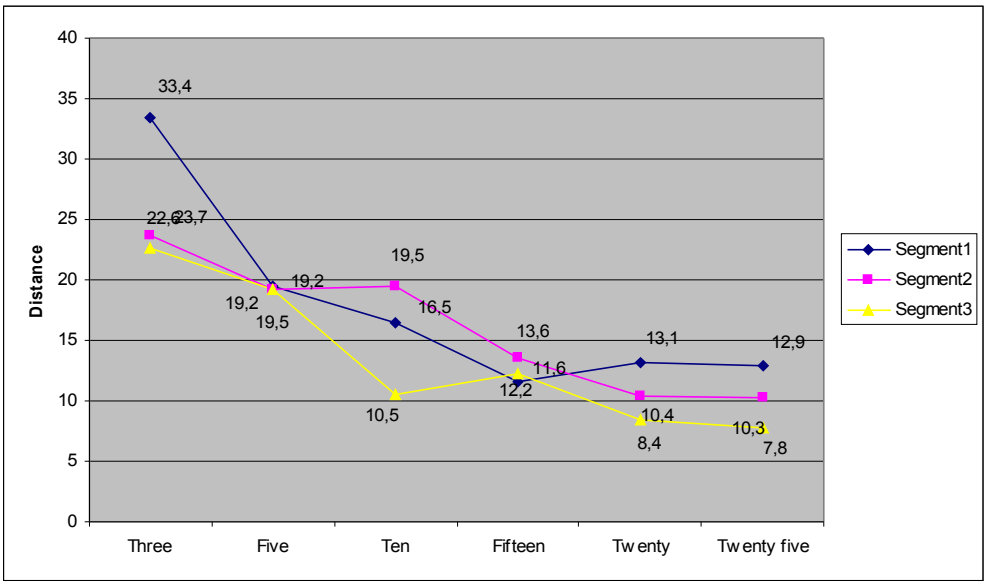


Figure 45: Distance moyenne entre les vecteurs références des macro-classes et les vecteurs macro-classes calculés à partir des annotations des groupes de 3, 5, 10, 15, 20, et 25 annotateurs.

Conclusions

Cette expérience, à laquelle a participé un nombre élevé de sujets, a permis d'évaluer la difficulté d'annotation des émotions complexes spontanées et de pointer les différences de perception entre hommes et femmes, bien connues dans la littérature. Des différences ont été montrées dans cette expérience, concernant d'une part l'affectation d'une étiquette verbale à une émotion, les hommes se focalisant dans notre exemple sur l'agressivité perçue (colère), et les femmes sur des sentiments plus profonds (désespoir, inquiétude) ; et d'autre part la perception des indices multimodaux, les hommes percevant plus les mouvements du torse et des sourcils que les femmes, qui ont détecté plus de mouvements des yeux.

Les similarités entre les vecteurs calculés dans les 2 expériences, celle effectuée sur 3 annotateurs experts, et celle-ci, montrent la cohérence des annotations, validant les annotations, mais également le schéma de codage défini pour ces 2 expériences. Il apparaît ainsi fondamental d'avoir des sujets des deux sexes pour enrichir les annotations des émotions.

Cette étude peut être complétée en ayant des groupes mieux équilibrés, notamment en ajoutant des experts, et des non étudiants. Concernant les autres groupes (âge, expertise), les résultats devraient être confirmés avec un éventail plus représentatif d'annotateurs.
Cette expérience montrant des différences individuelles nous a incité à effectuer des tests de personnalité pour étudier l'impact de la personnalité des annotateurs sur la perception des émotions dans les données naturelles (ces tests sont décrits plus loin).

3.3.5 Annotation des émotions et des dimensions contextuelles sur des vidéos françaises et britanniques

Une barrière principale dans le développement de modèles réalistes et précis d'émotions humaines est l'absence de base de données multi-langues de comportements naturels, et l'absence de protocole d'annotation pertinent et fédératif. Dans le but de remédier à ce manque, nous avons collaboré avec l'université de QUB (Queen's University of Belfast) pour définir un schéma de codage intégrant nos approches complémentaires (discrète vs. continue), ainsi que l'annotation du contexte et le typage des émotions. Nous avons appliqué ce schéma à des vidéos françaises et britanniques d'interviews télévisées.

Annotateurs

Quatre annotateurs (3 sujets français et 1 sujet britannique d'Irlande du Nord) ont annoté chaque clip avec l'outil Anvil, et cinq annotateurs (3 sujets français et 2 sujets britanniques d'Irlande du Nord) avec l'outil Trace. Tous ont utilisé l'outil Trace dans un premier temps, puis Anvil.

Vidéos sélectionnées

12 clips, 6 provenant d'EmoTV et 6 provenant de Belfast Naturalistic Database (Douglas-Cowie et al. 2000), ont été sélectionnés pour l'annotation.

Schéma de codage

Le schéma de codage multi-niveaux, défini à partir des 2 approches, combine les dimensions apparaissant comme utiles pour l'étude des émotions naturelles : les labels verbaux, les dimensions abstraites et les annotations du contexte (basées sur l'appraisal). La Table 20 présente la liste de labels définie conjointement par les deux équipes.

Emotions dans le sens strict	**Etats émotionnels**	
Colère (chaude)	Affection	Intérêt
Colère (froide)	Amusement	Irritation
Mépris	Anxiété	Content
Désespoir	Ennui	Soulagement
Dégoût	Courage	Relaxation
Joie	Déception	Satisfaction

Peur	Doute	Sérénité
Culpabilité	Embarras	Choc
Heureux	Empathie	Stress
Fierté	Excitation	Inquiétude
Tristesse	Amical	
Honte	Abandon	
Surprise	Espoir	

Table 20: Labels d'émotion sélectionnés pour l'étude.

La définition de cette liste a nécessité de faire un compromis entre les listes qui distinguent trop peu d'émotions pour être utilisables, et les listes qui distinguent trop d'émotions pour être gérables informatiquement. Les labels choisis doivent refléter les états émotionnels de la vie courante. Cette liste intègre les labels qui se sont avérés utiles dans les travaux des équipes du LIMSI et de QUB sur les corpus naturels.

A chaque émotion sont associés les dimensions suivantes :

- *Intensité* : force perçue de l'émotion (entre 1 et 5)
- *Valence* : mesure globale du sentiment positif/négatif associé à l'émotion (entre -2 et +2)
- *Activation* : degré d'activité (comportements corporels) (entre -2 passif et +2 actif)
- *Agressivité* : degré avec lequel la personne est en retrait (-2), ou en avant (+2)
- *Niveau acté* : degré avec lequel la personne acte/simule son émotion
- *Niveau masqué* : degré avec lequel la personne masque son émotion

Les expériences que nous avons décrites dans les sections précédentes ont montré que les interviews télévisées que nous avons collectées contiennent souvent des mélanges d'émotions, simultanés ou séquentiels. Pour refléter cela, nous avons développé les catégories suivantes décrivant les combinaisons temporelles d'émotion :

- *Simultané* : plusieurs émotions apparaissent au même moment
- *Séquentiel* : plusieurs émotions apparaissent l'une après l'autre dans un même clip
- *Simultané* + *séquentiel* : nécessaire pour l'annotation au niveau global d'une vidéo

D'après les théories des émotions (Sander et al. 2005), les émotions naissent en réaction à un événement; la réaction synchronise plusieurs systèmes et a un « sujet ». Nous avons défini les qualificateurs de temps, de contrôle, et de cible suivants :

- *Episodique* : conforme à la théorie, l'émotion apparaît en réaction à un événement extérieur, domine brièvement le mental.
- « *Simmering* » : l'émotion apparaît bien en réaction à un événement extérieur, mais ne domine pas encore le mental, elle est maîtrisée par la personne éprouvant l'émotion.
- « *Flitting* » : l'émotion varie d'un sujet à un autre (n'est pas ciblée sur un sujet unique).
- *Humeur* : implique un sentiment global, non ciblé sur un sujet spécifique.
- *Réactive* : l'émotion est ciblée et activée par des événements récents ou immédiats.
- *Etablie* : sentiments profondément installés proviennent d'événements passés.

Les « appraisals » sont les évaluations du ou des événements antécédents à l'émotion. A partir de l'ensemble des dimensions d'appraisals défini par (Scherer, 2000), nous avons défini l'ensemble suivant d'éléments regroupés dans 4 catégories (Table 21).

Classe d'appraisals	Appraisal	Valeur
Nouveauté	Soudaineté	entre 1 (faible) et 5 (élevé)
	Familiarité	entre 1 et 5
	Prédictibilité	entre 1 et 5
	Agrément intrinsèque	entre -2 (négatif) et +2 (positif)
	Agrément global	entre -2 et +2
Significations des buts/besoins	Causalité : agent	texte libre
	Causalité : motivation	entre -2 et +2
	Degré de certitude dans la prédiction des conséquences	entre -2 et +2
	Inadéquation avec les attentes	consonant ou dissonant
	Opportunité	aide ou bloque
	Urgence	entre 1 et 5
Potentiel de maîtrise	Contrôlabilité de l'événement direct	entre 1 et 5
	Contrôlabilité de l'événement conséquence	entre 1 et 5
	Puissance	entre 1 et 5
	Ajustement	entre 1 et 5
Compatibilité avec les standards	Standards externes	entre 1 et 5
	Standards internes	entre 1 et 5

Table 21: Liste des dimensions d'appraisal utilisée dans cette expérience.

Analyses des résultats

Cette étude a été évaluée à deux niveaux. Premièrement des mesures de fiabilité ont été effectuées sur les annotations manuelles : pour les catégories de labels annotés avec Anvil, le coefficient Kappa a été calculé ; pour les annotations avec l'outil Trace, des relations ont été calculées entre tableaux de résultats résumant les annotations. Deuxièmement, un questionnaire a été rempli par les 6 annotateurs, dans lequel on leur demandait de noter subjectivement les avantages, inconvénients, et difficultés de chaque élément du schéma de codage. L'objectif général étant de regrouper les descripteurs dans les catégories suivantes :

a) Eléments d'utilité confirmée
b) Eléments intermédiaires
c) Eléments semblant moins utiles

Labels négatifs	Labels positifs	Autres labels
Colère (Chaude)	**Joie** Heureux	**Surprise**

Colère (Froide) Irritation Mépris	Amusement Content Satisfaction	
Dégoût	**Excitation**	
Peur Anxiété Doute Choc Stress Inquiétude	**Affection** Amical	**Empathie** Intérêt
Embarras Honte Culpabilité	**Relaxation** Sérénité	
Tristesse Désespoir Déception Abandon	**Soulagement**	
Ennui	**Fierté**	
	Courage	

Table 22: Macro-classes de labels verbaux.

Trouver d'autres mesures d'accord inter-annotateurs n'est pas facile lorsque plusieurs labels peuvent être assignés à un même segment. Certaines mesures étendues du Kappa donnent des résultats bas sur des labels multiples, même s'ils sont en parfait accord (Rosenberg et Binkowski, 2005). Nous avons adapté l'approche de Rosenberg et Binkowski de manière à avoir un Kappa à 1 lorsque les annotateurs ont mis les mêmes labels multiples. La Table 23 présente ces Kappas, calculés pour les 35 labels et pour les 15 macro-classes de la Table 22.

Annotateurs	35 labels	15 labels
1 & 2	0.37	0.61
1 & 3	0.42	0.66
1 & 4	0.55	0.58
2 & 3	0.47	0.64
2 & 4	0.54	0.65
3 & 4	0.43	0.62
Kappa moyen	0.46	0.63

Table 23: Coefficients Kappa pour les labels d'émotion, pour la liste des 35 labels et pour celle des macro-classes.

Ces labels d'émotion sont fiables dans le cas de clips tels que ceux que nous utilisons. Cependant 9 des 35 labels n'ont jamais été utilisés : colère froide, mépris, intérêt, relaxation, amical, empathie, ennui, affection, et culpabilité. Ceci incite à élargir l'éventail des clips pour de futurs travaux.

Les cas où les annotateurs ont utilisé un unique label pour définir une émotion étaient rares (3 sur les 48 annotations). La moyenne était de 2.66 labels par clip. Le questionnaire subjectif post-expérimental indique que les descriptions des combinaisons d'émotion étaient bien formulées, faciles à annoter, et souvent très informatives. Néanmoins, ils n'étaient pas uniformément annotés, l'accord inter-annotateurs étant seulement de 16.6%. Cependant cet accord faible peut provenir de périodes dans le clip contenant des chevauchements ou transitions d'émotions.

Les dimensions liées au label ont été comparées dans le cas où les annotateurs ont assigné le même label au clip. Ces comparaisons donnent 71% d'accord dans le cas des qualificateurs de temps (réactif/établi), et 69% d'accord pour les qualificateurs de contrôle et de cible (épisodique, simmering, flitting, humeur). Les qualificateurs de temps étaient plus souvent annotés comme réactifs plutôt qu'établi (avec un facteur de 2.5). Les questionnaires utilisateurs indiquent que la distinction réactif/établi est évaluée comme très utile et informative. En ce qui concerne les qualificateurs de contrôle et de cible, ils ont été évalués comme bien formulés, mais deux des valeurs, « simmering » et « flitting », n'étaient pas toujours faciles à annoter, et occasionnellement informatives uniquement. Ceci reflète certainement la nature de l'ensemble des clips sélectionnés. Comme le montre la Figure 46, la plupart des clips ont été annotés comme épisodiques, laissant peu de cas où les autres catégories étaient annotées.

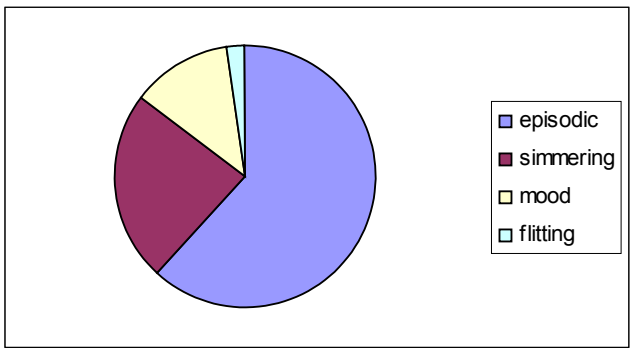


Figure 46: Répartition des annotations pour les qualificateurs de contrôle et de cible.

Des mesures de fiabilité ont été effectuées sur les dimensions des labels. Ces mesures étaient calculées en prenant le pourcentage de paires de sujets où les annotations étaient les mêmes où à 1 près (sachant que l'échelle des attributs allaient de 1 à 5 ou de -2 à +2). Ces résultats sont résumés dans la Table 24.

Acté	Masqué	Valence	Intensité	Activation	Agressivité
0.84	0.97	0.96	0.80	0.73	0.55

Table 24: Pourcentages d'accord sur les dimensions des labels globaux.

Les résultats montrent que, dans le cas des dimensions « classiques », la valence et l'intensité sont plus fiables que l'activation. Les dimensions « acté » et « masqué » sont moins classiquement utilisées dans les corpus émotionnels. Les accords sont élevés et les questionnaires utilisateurs indiquent que ces dimensions sont évaluées comme très utiles pour décrire les émotions complexes naturelles.

Le coefficient Kappa a été utilisé pour calculer l'accord inter-annotateurs sur les dimensions d'appraisal. La Table 25 résume les accords entre les 4 sujets qui ont effectué les annotations globales.

1) Soudaineté	0.1
2) Familiarité	0.19
3) Prédictibilité	0.0
4) Agrément intrinsèque	**0.53**
5) Agrément global	**0.4**
6) Causalité : motivation	0.13
7) Degré de certitude dans la prédiction des conséquences	0.01
8) Inadéquation avec les attentes	**0.27**
9) Opportunité	**0.7**
10) Urgence	0.22
11) Contrôlabilité de l'événement direct	0.19
12) Contrôlabilité de l'événement conséquence	0.3
13) Puissance	0.15
14) Ajustement	0.05
15) Standards externes	0.05
16) Standards internes	0.03

Table 25: Kappa pour les dimensions d'appraisal.

La Figure 47 étend ces mesures d'accord en montrant dans quelle proportion au moins 3 des 4 annotateurs étaient en accord. Des accords raisonnablement élevés ont été retrouvés pour: « Agrément intrinsèque », « inadéquation avec les attentes », « opportunité », « contrôlabilité de l'évènement conséquence ». Néanmoins, les autres accords restent faibles.

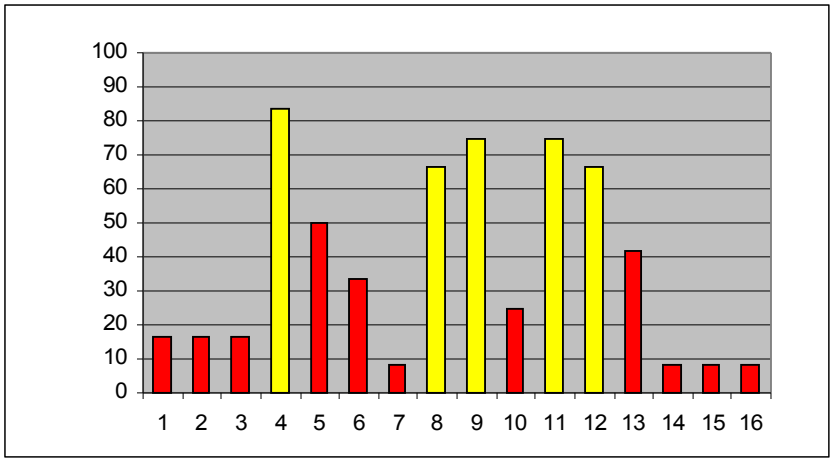


Figure 47: Pourcentages d'accords sur les 12 clips pour les 16 dimensions d'appraisal. Critère d'accord : au moins 3 des 4 annotateurs ont assigné la même valeur à l'attribut.

L'une des difficultés principales concernant les catégories d'appraisal est qu'elles sont liées à un évènement ou situation ; les clips d'émotions naturelles comportent souvent des

comportements émotionnels liés à plusieurs évènements (un traumatisme lointain, une maladie récente, l'interview présente). Parfois l'annotation des appraisals ne peut pas être correctement effectuée sans préciser l'évènement qui est considéré.
En perspectives, nous détaillerons notre proposition qui consiste à procéder en 2 étapes : 1) lister les évènements émotionnels, 2) annoter les appraisals pour chaque évènement. Ceci est intégré dans le schéma que nous fournissons en annexe. Ce schéma a été utilisé lors d'une première annotation avec trois annotateurs experts spécifiant les événements émotionnels et leurs dimensions d'appraisal.

Conclusions

L'une des innovations de cette expérience était la présence de dimensions d'appraisal dans le schéma de codage. Certains des éléments de cette liste d'appraisal ont eu un retour très positif, particulièrement les éléments de la catégorie « significations des buts/besoins » (« degré de certitude dans la prédiction des conséquences », « inadéquation avec les attentes », « opportunité », et « l'urgence » ont été évalués : bien formulés, faciles à annoter, et informatifs). Cette étude a aussi clarifié les difficultés en terme d'annotation des événements émotionnels et les difficultés qu'entraîne l'extension des annotations d'appraisal. Pour les éléments suivants, l'annotation séparée des différents événements est essentielle : « causalité : motivation », « degré de certitude dans la prédiction des conséquences », « contrôlabilité de l'événement direct », « contrôlabilité de l'événement conséquence », « ajustement ».
Des questions se posent sur la sélection des clips, qui doivent couvrir un panel approprié pour apporter une avancée aux techniques d'annotation. Plus de travaux sont nécessaires, sur plus de clips, pour confirmer les résultats de cette expérience. Les émotions épisodiques ne doivent pas être disproportionnellement dominantes.

De leur coté, l'équipe de QUB a fait annoter les mêmes extraits vidéos avec leur outil Trace et a analysé ces annotations. Suite à nos travaux en commun, nous avons écrit un schéma de codage Anvil combinant l'annotation discrète et continue des émotions. Une première phase informelle a permis d'établir une liste d'événements émotionnels qui sont ensuite déclarés dans le schéma de codage pour permettre une annotation distincte des dimensions d'appraisal pour chaque événement (cf. annexe). Nous n'avons fait ici qu'effleurer la dimension multi-culturelle des émotions. L'aspect cross-culturel est complexe, les différences d'expressions dans chacune des langues nous ont montré que l'établissement d'une terminologie et une mise en correspondance appropriées dans chacune des langues sont nécessaires.

3.3.6 Conclusion des expériences d'annotation des émotions

La Table 26 synthétise les quatre phases d'annotation manuelle des émotions.

Phases D'annotation	Nb vidéos annotées	Nb annotateurs	Conditions
Émotion 1	50	2	Audio Vidéo Audio/Vidéo
Émotion 2	11	3	Audio/Vidéo
Émotion 3	1	40	Audio/Vidéo
Émotion 4	12	5	Audio/Vidéo

Table 26: Synthèse des 4 expérimentations d'annotation manuelle des émotions.

Ces différentes phases d'annotation manuelle des émotions ont montré la difficulté et la subjectivité de la perception des émotions (par exemple les différences homme/femme) dans des interviews télévisées contenant des émotions spontanées. Les différences d'annotation nous ont amené à définir des vecteurs d'émotions de manière à regrouper les annotations des sujets, en partant du principe que toutes les annotations sont justes, si elles sont cohérentes individuellement. Elles ont également montré qu'un seul terme est rarement suffisant pour nommer la perception d'un état émotionnel spontané : deux ou plusieurs émotions apparaissent fréquemment en même temps, et il est également nécessaire de pouvoir annoter la manière dont ces émotions sont combinées. Enfin, nous avons intégré et expérimenté les dimensions contextuelles, au cours d'une collaboration avec l'Université de Belfast (Devillers et al, 2006a).

Nos travaux en terme de codage des émotions spontanées ont permis de définir itérativement un schéma adapté aux interviews télévisées.

3.4 *Annotation manuelle des comportements multimodaux*

3.4.1 Objectifs

L'annotation manuelle des comportements multimodaux et les modèles dérivés de ces annotations peuvent être utilisés comme source de connaissance pour la détection et la synthèse de comportement multimodal émotionnel spontané. L'un des objectifs est de comparer les données de la littérature sur les signes d'émotions avec ceux que l'on observe dans nos annotations. Comme nous le verrons, la définition itérative d'un schéma de codage et d'un protocole spécifique à ces données spontanées est nécessaire pour l'annotation des comportements multimodaux. Ce schéma de codage doit 1) permettre l'annotation (ou le calcul à partir de l'annotation) des attributs pertinents des comportements émotionnels spontanés observés dans nos vidéos de basse résolution TV ; 2) éviter des temps d'annotation trop longs ; et 3) permettre d'avoir des annotations fiables.

3.4.2 Annotation des comportements multimodaux par 2 sujets sur 35 vidéos

Un de nos objectifs est de contribuer à l'étude des relations entre émotions spontanées et comportement multimodaux. Les 2 premières expériences (la 1ère au niveau émotion, et la 1ère au niveau multimodal décrite dans cette section) avaient donc pour but l'annotation séparée, d'une part des émotions, et d'autre part des comportements multimodaux.

Vidéos sélectionnées

Les codeurs ayant participé à l'annotation des émotions sur les 50 premiers extraits collectés ont ensuite annoté les indices multimodaux. L'annotation multimodale étant longue, due au grand nombre d'attributs présents dans le schéma de codage et au niveau de segmentation fin qu'elle requiert, seulement 35 extraits ont été annotés par les 2 participants.

Schéma de codage

Les deux mêmes annotateurs ont utilisé le schéma de codage contenant les groupes de pistes d'annotations suivants. Le groupe « parole » contient une piste pour la transcription de la parole, incluant des marqueurs d'évènements non verbaux, et deux autres pistes affichant la prosodie et l'intensité, extraits automatiquement. Des groupes de pistes, contenant chacun

deux pistes : pose et mouvement, sont dédiés à chaque modalité : le torse, la tête, les épaules, les bras, les expressions faciales, les gestes. Le protocole consistait à annoter piste par piste. En ce qui concerne les mouvements des lèvres, seuls les mouvements exagérés perçus comme étant non directement reliés à la parole audiovisuelle ont été annotés. La Figure 48 montre un exemple d'annotation multimodale sur un extrait provenant du corpus.

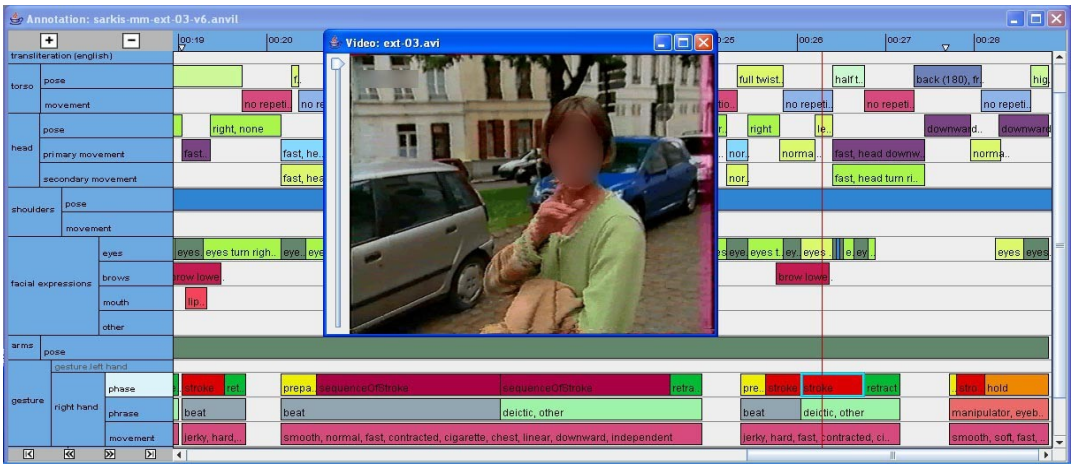


Figure 48: Annotation d'un du corpus avec Anvil (Kipp, 2001): le nom des pistes apparaît sur la gauche de la fenêtre. Le reste de la fenêtre contient des annotations pour chaque modalité.

Analyses des résultats

Les résultats ont révélé des différences dans l'annotation de la perception des indices multimodaux, et des difficultés intrinsèques dans l'annotation de certains attributs des modalités non-verbales. L'annotateur n°1 a annoté, pour toutes les vidéos, 1839 comportements multimodaux et l'annotateur n°2 seulement 914 au total. L'annotateur n°1 a effectué une annotation plus fine que l'annotateur n°2, pour toutes les modalités (les pourcentages d'annotation dans chaque modalité sont en effet similaires pour les deux annotateurs).

		Annotateur n°1	Annotateur n°2
Modalité	Expressions faciales	78.6%	80.4%
	Gestes	11.3%	11.9%
	Postures	10%	7.7%
Attribut	Direction du regard	26.8%	17%
	Mouvements de la tête	23.5%	21%
	Clignements des yeux	15.8%	17.6%
	Mouvements des Sourcils	10%	9.3%

Table 27: Comportements multimodaux les plus annotés (en pourcentage d'annotations effectuées).

La Table 27 montre les comportements multimodaux les plus fréquemment annotés pour chaque modalité ou attribut, et les similarités entre annotateurs. Dans l'ordre, les annotateurs ont le plus fréquemment annoté les expressions faciales, les gestes, puis les postures.

En ce qui concerne les attributs, l'annotateur n°1 a annoté une majorité d'attributs « direction du regard », et l'annotateur n°2 une majorité de « mouvements de la tête ». Les deux annotateurs étaient en accord sur le nombre d'annotations des attributs « preparation » et « stroke », utilisés classiquement dans l'annotation structurelle des gestes.

Les désaccords concernent les mouvements du corps, les types et énergie des gestes. L'annotateur n°2 a associé l'énergie des gestes avec leur vitesse, tandis que l'annotateur n°1 les a différenciés, annotant parfois des gestes avec une forte énergie et une faible vitesse d'exécution. De nombreux indices dans les annotations de l'annotateur n°1 sont partagés par plusieurs labels d'émotion (clignement des yeux, mouvements de la tête), mais les résultats montrent également que certaines émotions ont des indices multimodaux spécifiques, tels que baisser les mains (désespoir), ou la lenteur des mouvements corporels (sérénité). Une différence est observée entre les comportements multimodaux liés aux émotions fortes (colère, exaltation…) et les émotions plus faibles (irritation, sérénité…). Les attributs qui se révèlent les mieux appropriés pour distinguer les deux classes sont la vitesse et l'énergie des gestes, ainsi que la vitesse du corps en général. Par exemple, la sérénité n'implique pas de gestes, tandis que l'exaltation est souvent accompagnée de gestes rapides et énergétiques. De même, l'irritation est liée aux gestes lents et de faible intensité, tandis que la colère est liée aux gestes rapides et de forte intensité.

Conclusions

Ces premiers résultats sur l'annotation multimodale, incluant les désaccords observés entre annotateurs, sont compatibles avec ceux observés dans la littérature : les canaux expressifs peuvent être contradictoires ou complémentaires, les émotions peuvent entraîner des expressions très subtiles ou contrôlées (Collier, 1985 ; Ekman, 2003 ; Raouzaiou 2002). Notre corpus apporte cependant des annotations contextuellement plus riches et contient des comportements émotionnels complexes.

3.4.3 Perception globale des paramètres d'expressivité du mouvement

Nous avons mis en place une expérimentation de perception de l'expressivité des mouvements. Son objectif était d'étudier comment les gens perçoivent les comportements multimodaux dans des interviews télévisées. Nous avons considéré 6 paramètres expressifs utilisés en recherche sur l'expressivité du mouvement (Pelachaud, 2005; Wallbott, 1998) : l'activation globale, la répétition, l'extension spatiale, la vitesse, la force, et la fluidité.

Les objectifs de cette expérience étaient 1) d'évaluer la pertinence de ce modèle (élaboré à partir de données actées) à des émotions spontanées, 2) d'étudier les différences individuelles en terme de perception des émotions complexes, 3) de collecter des annotations globales de l'expressivité des mouvements sur un sous-ensemble de notre corpus EmoTV, 4) de comparer les annotations de l'expressivité des mouvements avec les annotations multimodales effectuées par 2 annotateurs experts sur les mêmes vidéos (décrites plus loin), 5) de comparer les annotations de l'expressivité des mouvements avec les annotations des émotions et des indices audio sur les mêmes vidéos.

Annotateurs

20 sujets

Protocole expérimental

Nous avons sélectionné 29 extraits vidéo riches en comportements multimodaux provenant de notre corpus EmoTV.

Les sujets devaient s'entraîner sur 2 extraits (l'un contenant beaucoup de mouvements et l'autre non) avant de commencer l'expérience sur les 27 extraits restants. 20 sujets (11 hommes, 9 femmes), âgés entre 19 et 54 ans (moyenne 32) ont dû annoter leur perception de l'expressivité des gestes et mouvements de tête et torse (Figure 49). Huit sujets étaient étudiants en informatique, huit étaient: chercheurs, professeurs ou ingénieurs, et enfin quatre travaillaient dans l'administration. Les sujets devaient visionner chaque extrait 3 fois avant de marquer sur une ligne (représentant les 2 valeurs extrêmes d'un paramètre expressif) le niveau des paramètres suivants : activation globale (passif/actif), répétition (1 seul geste / plusieurs répétitions), extension spatiale (contracté/étendu), vitesse (lent/rapide), force (faible/fort), et fluidité (saccadé/fluide). L'ordre de présentation des 27 extraits était aléatoire.

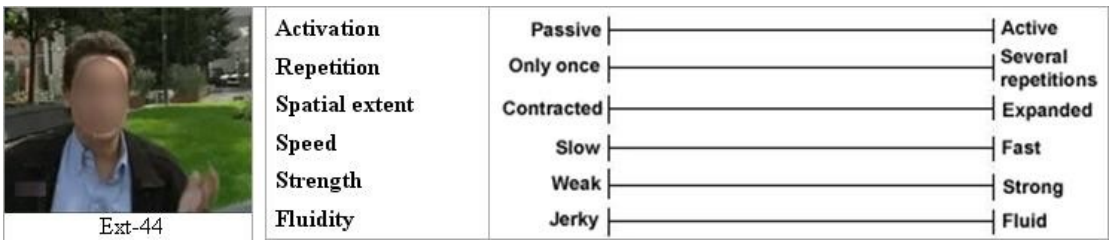


Figure 49 : Les sujets devaient évaluer manuellement leur perception de 6 paramètres d'expressivité (activation, répétition, extension spatiale, vitesse, force, et fluidité) pour chaque vidéo.

Résultats

Les corrélations entre annotations ont été calculées pour chaque paire de sujets. Les seuils utilisés étaient : pas de corrélation pour 0<r<0.3, faible corrélation pour 0.3<r<0.4, bonne corrélation pour 0.4<r<0.8, et forte corrélation pour 0.8<r<1. De bonnes corrélations ont été observées entre les sujets pour 4 des paramètres (Table 28): extension spatiale (indice de corrélation moyen r=0.58 entre sujets 2 à 2), activation (0.55), répétition (0.45), et force (0.44). Une corrélation faible a été observée pour le paramètre vitesse (0.35), et pas de corrélation pour la fluidité (0.13).

	Valeur de corrélation moyenne	Valeur de corrélation minimum	Valeur de corrélation maximum
Ext.Spat.	0,58	0,16	0,85
Activation	0,55	0,1	0,89
Répétition	0,45	-0,15	0,78
Force	0,44	-0,05	0,79
Vitesse	0,35	-0,21	0,72
Fluidité	0,13	-0,6	0,69

Table 28 : Corrélations des 6 paramètres expressifs, entre sujets 2 à 2.

Dans le but de positionner les extraits et les paramètres dans un unique nuage de points, nous avons effectué une analyse par composantes principales (ACP) sur les annotations des sujets. La Figure 50 montre l'espace à 2 dimensions créé par l'ACP. 88.4% de la variabilité du nuage de points obtenu est expliqué par cet espace (nous pouvons considérer que le modèle est acceptable à 70%). Si nous considérons l'axe horizontal, les 6 paramètres (en rouge dans la figure) sont tous positionnés dans la partie droite. Les extraits qui sont placés près des paramètres ont obtenu des scores élevés dans ces paramètres.

L'axe *horizontal* peut être interprété comme opposant les extraits ayant obtenu des scores élevés en moyenne pour les 6 paramètres (à droite) et les extraits ayant obtenu des scores bas (à gauche). Par exemple, les extraits 44 et 64 (situés sur la droite de l'axe horizontal) ont les valeurs les plus élevées pour la force et l'activation globale ; tandis que les extraits 78 et 15 (situés sur la gauche de l'axe horizontal) ont des valeurs très basses pour la force et l'activation globale. 7 extraits sont regroupés dans la partie la plus à gauche de l'axe horizontal. Les 5 extraits ne contenant pas de mouvements, intégrés dans l'expérience, font partie du groupe de 7 extraits. Les 2 autres extraits de ce groupe (le 78 et le 90) contiennent peu de mouvements, avec des extensions spatiales et des répétitions faibles.

L'axe *vertical* peut être interprété comme opposant la fluidité (en haut), et la répétition (en bas). Par exemple, le clip 78 (situé à gauche sur l'axe horizontal et légèrement au dessus de cet axe) contient plus de fluidité et moins de répétition que l'extrait 15 (situé à gauche sur l'axe horizontal et légèrement en dessous de cet axe). Les 4 clips contenant le plus de mouvements répétitifs sont les extraits 66, 02, 36, et 95, situés près et en dessous de la variable « repet ».

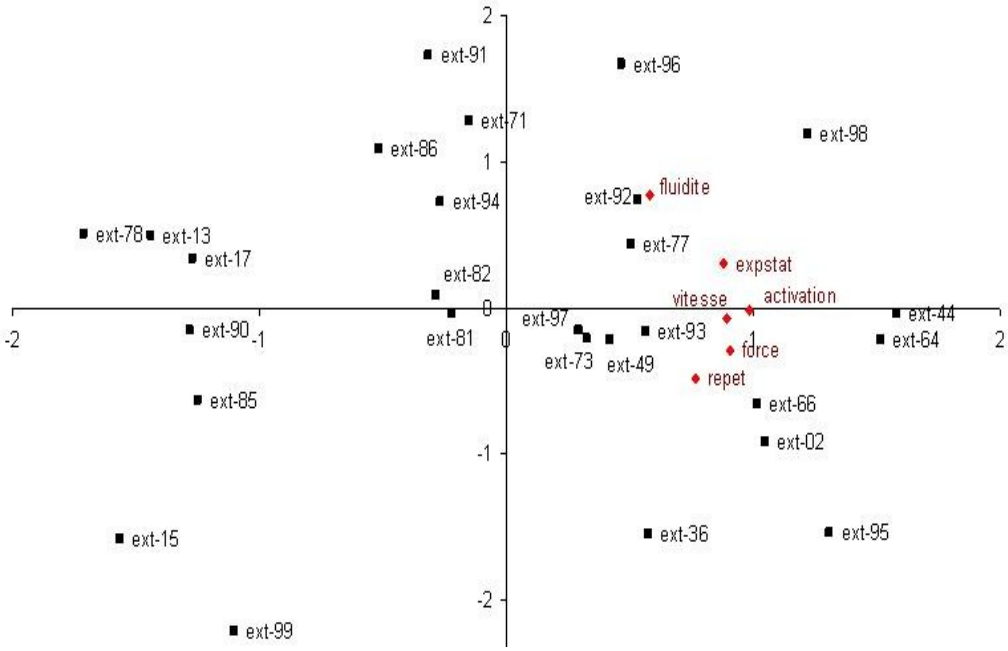


Figure 50. Analyse en Composantes Principales (ACP) effectuée en calculant les annotations moyennes de tous les sujets, pour positionner les extraits vidéo et les paramètres d'expressivité dans un unique nuage de points.

La Table 29 montre les valeurs moyennes calculées à partir des annotations des 20 sujets, pour chaque paramètre de quelques extraits. Les valeurs sont en pourcentage, chaque pourcentage correspondant à la portion de la ligne d'annotation comprise entre l'extrémité gauche et la marque mise par le sujet (exemple: si un sujet coche une marque à l'extrémité droite de la ligne activation globale, le pourcentage correspondant est de 100%). Cette table montre que l'activation la plus faible est observée dans l'extrait 78, et l'activation la plus élevée dans l'extrait 95. De plus, la répétition et la force sont les plus bas dans l'extrait 78, et les plus élevés dans l'extrait 95. Ceci illustre (pour ces extraits) la corrélation entre les 3 paramètres activation, répétition et force. De la même manière, la fluidité et l'extension sont les plus faibles pour le clip 99 ; la fluidité la plus élevée est observée dans l'extrait 98, et l'extension spatiale est l'une des plus élevées pour ce même extrait 98. Ceci montre que dans ces extraits les gestes étendus sont articulés de manière fluide alors que les gestes contractés sont plus saccadés.

Extrait	Activ.	Répét.	Ext.Spat	Vitesse	Force	Fluid.
13	20%	22%	18%	**24%**	30%	45%
64	73%	73%	60%	*76%*	65%	56%
78	**18%**	**21%**	12%	28%	**14%**	43%
92	54%	45%	*78%*	46%	53%	51%
95	*75%*	*83%*	52%	60%	*76%*	45%
98	73%	64%	61%	63%	50%	*67%*
99	29%	50%	**9%**	46%	32%	**23%**

Table 29. Valeurs moyennes des 6 paramètres, sur les 20 sujets. Seuls les clips contenant au moins 1 paramètre calculé comme le minimum (en rouge foncé gras) ou le maximum (en jaune clair italique) sont listés.

Suite à notre expérience sur les différences individuelles observées durant l'annotation des émotions (section 3.3.4), nous souhaitions étudier les différences individuelles en terme de perception de l'expressivité des mouvements spontanés vis-à-vis de l'âge, du sexe, de la personnalité et de la catégorie socio-professionnelle des sujets.

En terme de personnalité, des différences ont été observée entre personnes intraverties et extraverties dans la communication nonverbale (Knapp et Hall 2006), mais également en terme de perception des comportements multimodaux d'agents animés (Martin 2006 ; Buisine et Martin accepté). Nous avons donc demandé à chaque sujet de remplir un test de personnalité (l'Inventaire de Personnalité d'Eysenck (EPI)) permettant d'évaluer l'intraversion/extraversion.

Une analyse par composantes principales (ACP) a été effectuée pour mettre en évidence les différences interindividuelles. Etant donné que nous avons un tableau à trois entrées : Paramètres*Extraits vidéos*Individus, et qu'on ne peut analyser qu'un tableau à double entrée par ACP, nous avons cette fois-ci moyenné les extraits afin d'analyser Individus*Paramètres. La Figure 51 montre l'espace à deux dimensions résultant de l'analyse. Concernant le positionnement des paramètres entre eux, l'axe 1 fait apparaître un lien entre « répétition », « fluidité », « activation », « vitesse », et « extension spatiale », l'axe 2 montre un lien entre « répétition » et « force ». Nous avons groupé les sujets et regardé les positions de ces différents groupes :

- Sur la variable *age* : nous avons séparé l'échantillon en deux, les 19-30 ans (11 personnes) et les 34-54 ans (9 personnes). Les résultats montrent que cette variable n'est pas pertinente dans notre échantillon car les 2 groupes sont proches de l'origine et proches l'un de l'autre.

- Sur la variable *sexe* : idem

- Sur la variable *personnalité* : cette variable est pertinente, les triangles rouges montrent que les extravertis sont dans le négatif sur l'axe 1 et sur l'axe 2. Ce qui signifie qu'ils ont donné des notes plus basses que les introvertis sur tous les paramètres. Globalement nous pouvons interpréter cela par le fait que les extravertis ont minimisé l'expressivité des extraits vidéo qu'ils ont vus, ce qui témoigne d'un décalage de leur référentiel d'expressivité.

- Sur la variable *catégorie socioprofessionnelle* : la variable ne semble pas pertinente, il faudrait approfondir en menant une nouvelle expérimentation avec des catégories socioprofessionnelles (CSP) plus contrastées que dans notre échantillon.

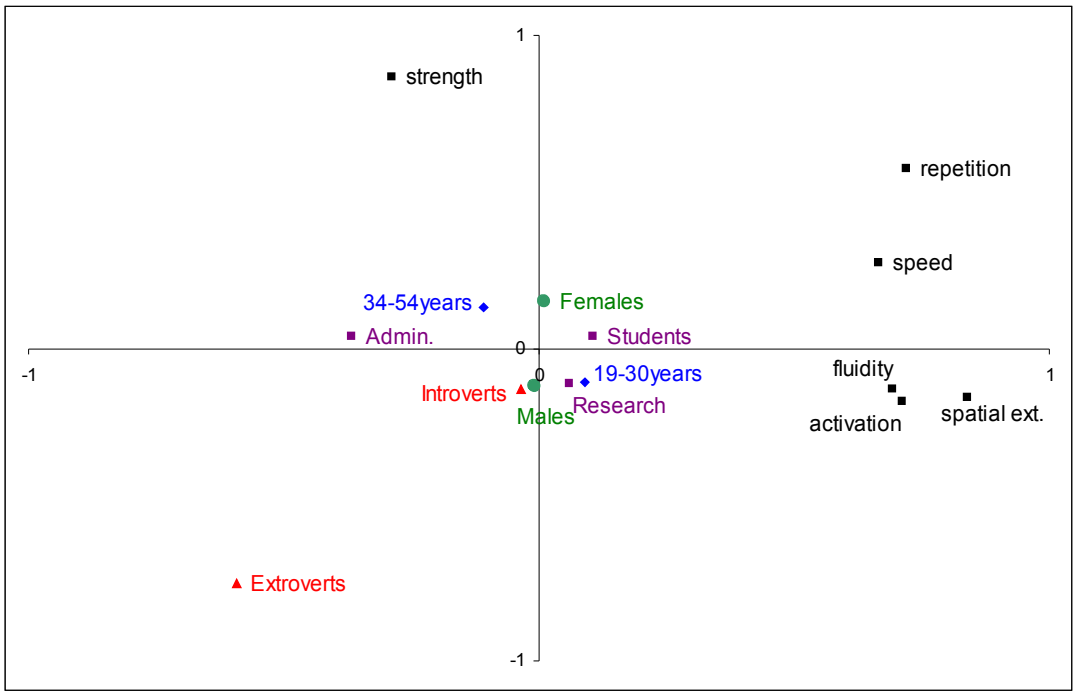


Figure 51. Analyse en Composantes Principales (ACP) effectuée en moyennant les extraits vidéo et en analysant les variables Paramètres*Individus.

Conclusion

Nous avons étudié la perception de l'expressivité des mouvements dans des interviews télévisées spontanées, alors que la plupart des études similaires sont effectuées sur des données actées enregistrées en laboratoire. Nous avons validé les annotations des sujets en vérifiant que les 5 extraits ne contenant pas ou peu de mouvements ont été correctement annotés, c'est-à-dire avec des valeurs faibles d'expressivité. Nous avons observé que, dans les extraits que nous avons sélectionné, les mouvements de grande extension spatiale sont également d'une grande fluidité ; et que la perception d'une grande activité est souvent liée à de nombreuses répétitions et à des mouvements d'une grande force. Finalement, nous avons observé que les 4 dimensions d'expressivité suivantes sont annotées de manière fiables puisque corrélées entre les sujets: l'extension spatiale, l'activation globale, la répétition, et la force. Une faible corrélation a été calculée pour le paramètre vitesse, et pas de corrélation pour le paramètre fluidité. Enfin, nous avons effectué plusieurs analyses en composantes principales (ACP) sur les annotations des sujets. Les résultats ont montré des différences d'annotation des paramètres expressifs entre les groupes de sujets « Extravertis » et les « Introvertis ». Les extravertis ont annoté tous les paramètres expressifs avec des valeurs plus basses que les introvertis. Les autres traits individuels (âge, sexe, catégorie socioprofessionnelle) n'ont pas donné de résultats.

3.4.4 Un schéma pour l'annotation de l'expression multimodale d'émotions spontanées

En raison des différences de segmentation importantes d'un annotateur à un autre, comme la première expérience l'a montrée, et en raison des temps d'annotation très coûteux pour les annotations multimodales (entre 1h et 2h pour annoter un extrait vidéo), nous avons opté pour une méthode d'annotation consistant à faire annoter chaque extrait par une unique personne « experte » (travaillant dans le domaine des émotions et des comportements multimodaux), puis à faire valider/corriger cette annotation par une deuxième personne experte. L'annotation a été effectuée, sur un schéma de codage raffiné itérativement avant d'arriver en fin de thèse à sa version finale. Un tiers du corpus (30 des 89 extraits vidéo d'EmoTV), contenant des comportements multimodaux riches et des émotions non basiques a été annoté de cette manière par un codeur expert et reste à valider par une deuxième personne.

Vidéos sélectionnées

28 vidéos du corpus ont été sélectionnées pour leur richesse en comportements multimodaux.

Schéma de codage

Le schéma a été défini itérativement durant la thèse. Les premières modifications ont été effectuées suite à la 1ère expérience, en prenant en compte les problèmes auxquels se sont confrontés les 2 annotateurs.

Expressions faciales

Le schéma précédent permettait d'annoter les expressions faciales à l'aide des Facial Animation Parameters (FAPs). Pour simplifier l'annotation et la rendre plus adaptée à des vidéos de basse résolution, nous les avons remplacées dans le nouveau schéma de codage par un sous-ensemble des Action Units de FACS, qui représentent un niveau supérieur aux FAPs (1 AU combine plusieurs FAPs). Des mouvements de tête combinés, comme faire oui de la tête tout en la tournant à droite, sont aussi possibles avec le nouveau schéma.

Nous avons défini dans un premier temps 27 AU (Action Unit) dans le schéma de codage : 12 pour les mouvements de bouche, 8 pour les yeux, 5 pour les sourcils, 1 pour le menton et 1 pour le nez. Cette définition semble trop précise, en ce qui concerne les yeux par exemple, les

AU permettent de donner la direction du regard ; pour la bouche, les AU permettent d'annoter des mouvements difficilement perceptibles sur de la qualité vidéo télévisée. Nous n'avons donc conservé qu'une partie des AU : les 5 des sourcils, celle du menton et du nez, mais retirer certaines AU pour les yeux et la bouche pour avoir au final 6 AU au maximum pour les yeux et 6 au maximum pour la bouche.

Phases des gestes

Nous avons détaillé les phases des gestes, en distinguant les 3 différents types de « hold » : prestroke hold / postroke hold / independent hold (quand il n'y a pas de stroke), ainsi que les 2 différents type de « retraction » : partial_retract / rectract. Ces annotations pourraient par exemple permettre de comparer l'impact de certaines caractéristiques de l'émotion (par exemple son intensité) sur la présence et la durée de chacune des phases.

Catégories des gestes

L'analyse des annotations précédentes a montré que les gestes déictiques sont correctement annotés, mais que les autres gestes, plus complexes, sont souvent ambigus (métaphoriques, iconiques, emblèmes et beats). Nous avons décidé de conserver toutes les catégories de geste, dans une perspective de schéma global même si finalement tous les types ne seront pas annotés en raison de la petite taille du corpus. Nous avons ajouté un attribut « confidence score » qui permet à l'annotateur de préciser à quel point il est sûr de son annotation, et l'annotateur a la possibilité de sélectionner plusieurs types pour un même geste, permettant d'avoir une approche dimensionnelle des catégories : par exemple un même geste peut être iconique et déictique (McNeill 2005). Similairement à la phase, ces annotations peuvent permettre d'étudier l'impact des émotions sur la fréquence de chacune des catégories.

Répétition

Dans le cas d'un geste répété plusieurs fois d'affilée, nous avons décidé de l'annoter par un unique segment « stroke » englobant la série de gestes répétés, et en ajoutant un attribut répétition pour annoter le nombre de répétitions du geste. Cela simplifie l'annotation et nous avons généralement observé que de toute façon les différents strokes avaient les mêmes paramètres expressifs.

Paramètres d'expressivité

Nous avons décidé de garder les 6 paramètres provenant des travaux de (Pelachaud, 2005), plutôt que de nous limiter aux 4 paramètres qui ont été validés par le test perceptif décrit plus haut. Les 2 paramètres pour lesquels peu ou pas de corrélation a été observée sont la vitesse et la fluidité. Néanmoins, ces 2 paramètres sont importants pour la spécification de comportement dans les ACA. Les faibles kappas montrent les différences interindividuelles de perception (par exemple en fonction de l'extraversion/introversion des individus), mais ne remettent pas en cause leur intérêt pour l'étude des comportements. Leur annotation pourra aussi servir ultérieurement via d'autres études perceptives impliquant des replays. Même si nous gardons les 6 paramètres dans notre schéma qui se veut fédérateur, les résultats de notre test perceptif suggère dans le cadre de futures recherches sur d'autres données, de vérifier au préalable à l'annotation proprement dite, via un test perceptif utilisant note protocole, la validité de ces 6 paramètres expressifs pour les données à annoter.

Accélérer l'annotation de l'expressivité

L'expressivité peut être annotée à trois niveaux : au niveau global de la vidéo, au niveau d'une unité gestuelle (gesture unit : série de gestes entre 2 positions de repos), et enfin au niveau d'un geste. Dans le but d'accélérer l'annotation, l'expressivité doit pouvoir être annotée à un seul niveau : le plus bas (au niveau du geste, ou au niveau global de la vidéo si aucune annotation n'a été effectuée au niveau local à un segment).

Cela permet d'envisager à terme, un programme Java analysant les annotations pourrait alors calculer les valeurs d'expressivité héritées des autres niveaux. Ainsi un niveau peut hériter des annotations d'un autre niveau (inférieur ou supérieur), sans avoir à l'annoter. Les outils d'annotations actuels ne permettent pas encore de gérer cette notion d'héritage des attributs.

Liens entre les 4 pistes des gestes

Dans le but de faciliter et fiabiliser la segmentation des 4 pistes des gestes : unité gestuelle / phase / phrase / expressivité, nous avons défini des liens entre ces 4 pistes dans le fichier XML constituant le schéma de codage. Ainsi, les phases des gestes doivent être annotées en premier, les phrases (liées aux bornes des segment de phase) sont annotées en deuxième, puis l'expressivité au niveau d'un geste (mêmes bornes que la piste phrase), et enfin les unités gestuelles sont annotées en dernier (liées aux bornes des segments de phrase). Cela rend l'annotation des gestes un peu plus contraignante mais en partie garantie l'alignement temporel strict des différentes pistes d'annotations concernant un même geste.

Utilisation de 3 groupes de pistes distincts pour l'annotation des gestes de la main droite / main gauche / deux mains

Trois groupes distincts sont définis pour l'annotation des gestes : 1 pour l'annotation des gestes effectués avec la main droite, un deuxième pour la main gauche, et un troisième pour les gestes effectués à deux mains. Les avantages d'utilisation de ces 3 pistes sont d'une part de pouvoir annoter 2 gestes différents effectués en même temps ou se chevauchant temporellement (par exemple un déictique effectué avec la main droite, et un emblème effectué avec la main gauche) ; et d'autre part d'utiliser la 3ème piste (deux mains) pour annoter des gestes synchrones et symétriques des deux mains (plutôt que de dupliquer l'annotation sur chacune des pistes main droite et main gauche).

Annotation des directions

L'annotation de la direction des mouvements n'est utile que pour la copie-synthèse avec un ACA, pour maximiser la ressemblance avec une vidéo originale. Elle n'est pas utile (en tout cas pas sur un corpus de petite taille) pour l'étude des relations entre émotions et comportements multimodaux (seule l'expressivité est utile dans ce cas). Nous avons quand même conservé ces attributs dans le schéma, même si nous ne les avons pas annotés, pour d'éventuels besoins en génération avec un agent conversationnel animé.

Modifications possibles

Les mouvements de la tête sont actuellement annotés relativement à la camera, ils pourraient également être annotés relativement au torse pour une spécification plus simple d'un ACA. Nous pourrions utiliser les mêmes termes d'angles du torse (full twist, high twist…) pour annoter les angles des mouvements de tête. Nous pourrions également définir deux Action Units pour chaque mouvement facial, pour avoir un schéma de codage plus clair et simplifier l'annotation.

L'annotation de l'évolution temporelle détaillée (à l'aide de courbes) n'est pas entièrement disponible dans notre schéma (une partie concernant la régularité peut être annotée), elle nécessite des connaissances supplémentaires pour la spécification d'ACA.
Ce schéma de codage permet de décrire les répétitions à plusieurs niveaux (répétition des annotations d'un geste apparaissant plusieurs fois dans la vidéo, répétition des attributs des mouvements individuels). Le calcul de métriques telles que: redondance / complémentarité / spécialisation / équivalence pourrait être ajouté, ainsi qu'un facteur de confiance, dans le but de prendre en compte les imprécisions dues à la qualité vidéo.

Besoins en terme de schéma de codage

Plusieurs questions sont apparues lors de la création du schéma de codage, telles que : Combien de valeurs doivent être utilisées pour évaluer les paramètres de qualité du mouvement? Après avoir essayé avec 5 valeurs (très bas, bas, moyen, élevé, très élevé), nous nous sommes finalement mis d'accord sur 3 valeurs (bas, moyen, élevé), suffisantes pour chaque paramètre de qualité du mouvement, et permettant une annotation et une analyse des annotations plus rapide ainsi qu'un meilleur accord. L'extension spatiale n'est pas appropriée pour le torse et la tête (du moins en ce qui concerne le corpus EmoTV).

3.4.5 Conclusion

Le schéma multimodal que nous avons incrémentalement défini et corrigé tout au long de la thèse permet d'annoter les attributs principaux de comportements multimodaux émotionnels observés dans nos vidéos d'interviews télévisées. Nous l'avons défini de manière à éviter des temps d'annotation trop longs et à avoir des annotations fiables.
Le corpus est exploratoire et ne permet pas pour l'instant des modélisations statistiques. Cependant, il permet d'illustrer la complexité des comportements multimodaux dans l'expression d'émotions naturelles.

Nous avons obtenu des résultats compatibles avec ceux observés dans la littérature, comme par exemple les contradictions ou complémentarités pouvant exister entre différents canaux expressifs. Nous avons utilisé un corpus contextuellement varié et contenant des comportements émotionnels complexes. Le corpus EmoTV nous a permis d'étudier la perception de l'expressivité des mouvements lors d'événements émotionnels spontanés, alors que la plupart des études existantes étaient effectuées sur des données actées enregistrées en laboratoire.

Les analyses en composantes principales (ACP) effectuées lors du test perceptif ont montré d'une part des corrélations entre quatre paramètres d'expressivité (extension spatiale, activation globale, répétition, et force), et d'autre part des différences d'annotations en fonction de la personnalité des sujets. Les extravertis ont annoté tous les paramètres expressifs avec des valeurs plus basses que les introvertis.

Nous proposons d'intégrer ces différentes connaissances sur les comportements multimodaux (annotations manuelles de l'expressivité, résultat du test perceptif) avec les représentations des émotions construites dans les expérimentations précédentes. L'ensemble représente un profil émotionnel expressif d'une vidéo que nous utiliserons pour la génération de comportements émotionnels expressifs (section 3.6).

3.5 *Annotation automatique des comportements multimodaux*

3.5.1 Introduction

Les annotations manuelles décrites dans les sections précédentes ont une dimension subjective et sont coûteuses en temps. Il parait donc intéressant d'étudier l'apport possible d'un traitement automatique sur la vidéo. Ces annotations manuelles pourraient par exemple être enrichies par le traitement d'image via la détection automatique de segments vidéo émotionnels. L'estimation de la quantité de mouvement par traitement automatique de l'image pourrait peut être valider les annotations manuelles des mouvements ponctuels ainsi que de l'activation émotionnelle au niveau de la vidéo globale. Enfin, l'annotation automatique pourrait faciliter le processus d'annotation manuelle en fournissant la segmentation des mouvements ainsi que des valeurs précises des paramètres expressifs calculées objectivement tels que la vitesse, l'expansion spatiale ou la fluidité d'un geste. Aujourd'hui l'annotation manuelle et le traitement d'image apportent des informations à des niveaux d'abstraction différents et leur intégration n'est pas claire. En outre, la plupart des travaux en traitement d'image de comportement émotionnel ont été effectués sur des enregistrements vidéo de haute qualité réalisés en laboratoire (Camurri et al. 2003) et impliquant souvent des émotions simples jouées par des acteurs (De Silva et al. 2005), où l'on peut s'attendre à ce que les émotions soient moins spontanées que lors d'interviews télévisées.

Le but de ces travaux d'intégration entre annotation manuelle et traitement d'image est double : 1) explorer l'applicabilité des techniques de traitement d'image sur des vidéos de basse résolution provenant de la télévision, et 2) étudier de quelle façon le traitement d'image peut être utilisé pour valider l'annotation manuelle de comportements émotionnels spontanés.

Pour cela, nous avons sélectionné 10 clips d'EmoTV montrant des mouvements dans différents contextes (intérieurs / extérieurs).

Le traitement d'image développé par notre partenaire ICCS est utilisé pour fournir des estimations sur les mouvements de tête et de mains en combinant 1) la localisation des zones de peau et 2) l'estimation du mouvement de ces zones. La tâche de localisation de la tête et des mains dans les séquences d'images est basée sur la détection des zones continues de couleur de peau. Pour notre étude, un modèle simplifié a été suffisant puisqu'il n'est ni

nécessaire ni envisageable de reconnaître la forme des mains. Le corpus traité est basé sur des situations naturelles et donc la posture initiale de la personne est arbitraire et non sujette à des contraintes spatiales telles que « main droite sur le coté droit de la tête lorsqu'elle a les bras croisés ». Certaines régions ressemblantes à des zones de peau peuvent aussi tromper le système de détection automatique et l'algorithme de tracking. Pour éviter ce genre de problèmes, une initialisation assistée par un utilisateur est nécessaire comme point de départ pour l'algorithme de tracking. Lors de ce processus, l'utilisateur confirme les régions proposées par le système telles que les mains et la tête de la personne ; après cela, sachant que les conditions d'éclairage et de couleur ne changent généralement pas dans un même extrait, la détection et le tracking sont effectués automatiquement.

Un autre problème fréquent en traitement d'image d'extraits télévisés est le fait que les mouvements de caméra ne sont pas contrôlés et peuvent conduire à des déplacements abrupts de régions de peau dans un extrait sans que la personne montre une activité particulière.

La mesure du mouvement entre images successives est calculée comme la somme des pixels en mouvement dans les masques de peau en mouvement, normalisés sur les régions de peau. La normalisation est effectuée dans le but de supprimer les facteurs de zoom de caméra, qui pourraient faire apparaître les régions de peau en déplacement plus grandes sans pour autant avoir plus d'activité. Les zones en mouvement possibles sont trouvées en plaçant un seuil sur la différence de pixels calculés entre 2 images consécutives. Le masque de mouvement utilisé ne contient pas d'information sur la direction ou l'amplitude du mouvement. Les masques de couleur et de mouvement contiennent tous les deux un grand nombre de petits objets dus à la présence de bruits et d'objets de couleur similaire à la peau. Pour éviter cela, des filtres morphologiques sont employés sur les deux masques pour retirer les petits objets.

3.5.2 Comparaison des annotations manuelles et des résultats du traitement automatique

Activation globale des comportements émotionnels dans chaque vidéo

L'activation globale est considérée comme la quantité de mouvement liée à l'émotion. Dans le traitement d'images, elle est calculée comme la somme des normes des vecteurs de mouvements.

Les valeurs obtenues pour (1) annotation manuelle de l'activation émotionnelle (moyenne des annotations par 3 annotateurs experts), (2) l'estimation de la quantité de mouvement au niveau global de la vidéo, et (3) le % de secondes pour chaque vidéo pour lequel il y a au moins une annotation manuelle du mouvement (torse, mains ou tête) sont données dans la Table 30.

Ces trois valeurs apportent différentes estimations de la quantité d'activité multimodale liée à l'émotion. L'analyse de ces valeurs montre que les mesures (1) et (2) sont significativement reliées ($r = 0.64$, $p < 0.05$). Ceci montre que le traitement automatique des 10 extraits vidéo valide l'annotation manuelle de l'activation au niveau global de chaque vidéo. Les mesures montrent également que (1) et (3) peuvent être corrélées ($r = 0.49$). Enfin, l'analyse des résultats suggère que (2) et (3) peuvent être corrélées ($r = 0.42$). Cependant, compte tenu de la faible taille de l'échantillon, ces deux mesures n'apportent pas de signification statistique. Plus de données sont nécessaires pour confirmer ces deux résultats.

Video #	(1) **Manuel** annotation de l'activation émotionnelle 1:faible 5: forte	(2) **Automatique** estimation de la quantité de mouvement	(3) **Manuel** % de sec. avec au moins 1 annotation manuelle de mouvement dans le fichier Anvil
Annotateurs	3 annotateurs experts	Système	1 annotateur expert + vérification par un 2nd annotateur
01	4	3398,50	81,2
02	3	269,64	72,9
03	4,33	1132,80	92,6
22	4,33	3282,80	81,1
36	4,66	2240,50	94,4
41	3	959,60	73,6
44	3,33	1771,30	92,3
49	4,33	1779,00	91,2
71	2,67	904,73	86,1
72	3,33	330,92	56,7

Table 30 : Mesures manuelles (1)(3), et mesures automatiques (2) de l'activation émotionnelle globale dans les 10 vidéos.

Estimation temporelle et annotation du mouvement

Au niveau local des événements temporels, nous avons voulu comparer les annotations manuelles (des mouvements de tête, de mains, et de torse) avec l'estimation automatique des mouvements. La Figure 52 montre un exemple de traitement d'image automatique pour l'estimation de la quantité de mouvement (en noir et blanc), et l'intégration dans l'outil Anvil du résultat du traitement automatique (courbe continue rouge).

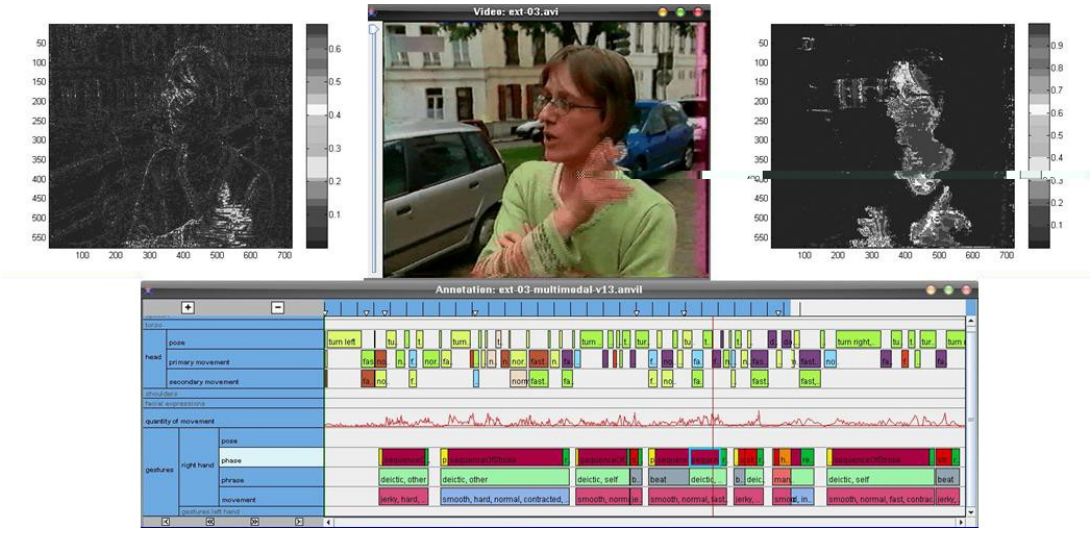


Figure 52 : Intégration dans l'outil Anvil (Kipp 2001) des annotations manuelles (blocs colorés) des mouvements de tête (pistes du haut) et gestes (pistes du bas). Les résultats du traitement d'image pour l'estimation de la quantité de mouvement sont également intégrés dans Anvil (courbe continue rouge au centre).

Le module de traitement d'image fournit une estimation du mouvement entre chaque image pour le corps entier. Il ne fournit pas d'estimation du mouvement séparée pour chaque partie du corps. Ainsi, nous avons comparé l'union des annotations manuelles des mouvements des modalités torse, mains et tête avec l'estimation automatique des mouvements. Lorsque le module de traitement d'image détectait un mouvement, nous estimions qu'il y avait un accord avec les annotations manuelles si un mouvement avait été annoté manuellement dans au moins une des trois parties du corps.

Pour être comparé avec les annotations manuelles (segment de mouvement ou pas de mouvement), un seuil doit être appliqué aux valeurs continues de l'estimation du mouvement fournies par le module de traitement d'image afin de fournir une valeur booléenne (mouvement détecté ou non),. Différents seuils entraînent des valeurs d'accord différentes entre annotations manuelles et détection automatique du mouvement. La valeur de ce seuil au dessus duquel le module de traitement d'image décide qu'un mouvement a été détecté doit donc être la valeur minimale à laquelle un mouvement a été perçu et annoté. Nous avons évalué l'accord entre l'union des annotations manuelles des mouvements et l'estimation du mouvement avec différentes valeurs de seuil. Les seuils testés étaient compris entre 0.1% et 40% de la valeur maximum de l'estimation de la quantité de mouvement. Nous avons utilisé un intervalle temporel de 0.04 secondes pour le calcul de l'accord puisque c'est, entre 2 images, l'intervalle utilisé par le module de traitement d'image. Les taux d'accord et de

désaccord sont présentés dans les tables 31 et 32. L'accord est le plus grand pour les vidéos 22 et 03 qui contiennent de nombreux mouvements (tête, mains), une zone de peau visible dans la partie supérieure du torse, et aucune personne ne se déplaçant en arrière plan. L'accord le plus faible est obtenu pour les vidéos 36 et 71 dans lesquelles des gens se déplacent dans l'arrière plan, les mouvements de ces gens n'ayant pas été annotés puisque nous nous sommes focalisés sur la personne interviewée. Une valeur intermédiaire a été obtenue pour la vidéo 41 qui ne montre que quelques légers mouvements de tête et de torse. Les interviews filmées en extérieur ont un accord plus élevé que celles filmées en intérieur, révélant l'impact de la qualité vidéo et de la luminosité. Il n'y a pas de relations systématique entre les désaccords : pour 6 vidéos, le nombre de désaccords « auto 0 – manuel 1 » est supérieur au nombre de désaccords « auto 1 – manuel 0 ».

Vidéo #	**Seuil**	**Accords**		
		Auto 0 - Manuel 0	**Auto 1 - Manuel 1**	**Total**
01	0,004	0,050	0,799	0,849
02	0,004	0,203	0,611	0,814
03	0,001	0,009	0,892	0,901
22	0,016	0,113	0,799	0,912
36	0,001	0,039	0,449	0,489
41	0,002	0,186	0,483	0,669
44	0,001	0,063	0,550	0,613
49	0,003	0,013	0,858	0,871
71	0,042	0,139	0,307	0,446
72	0,047	0,340	0,355	0,695
MOY	**0,012**	**0,115**	**0,610**	**0,726**

Table 31 : Pourcentages d'accord entre annotation manuelle du mouvement et estimation automatique de la quantité de mouvement (par exemple la colonne "Auto 0 – Manuel 0" décrit les accords "pas d'annotation manuelle du mouvement" / "pas de détection automatique du mouvement"). Le seuil est multiplié par la valeur maximum de l'estimation du mouvement.

Vidéo #	Seuil	Désaccords		
		Auto 0 Manuel 1	Auto 1 Manuel 0	Total
01	0,004	0,014	0,138	0,151
02	0,004	0,118	0,068	0,186
03	0,001	0,034	0,065	0,099
22	0,016	0,013	0,075	0,088
36	0,001	0,494	0,017	0,511
41	0,002	0,254	0,077	0,331
44	0,001	0,373	0,014	0,387
49	0,003	0,054	0,075	0,129
71	0,042	0,554	0,000	0,554
72	0,047	0,213	0,092	0,305
MOY	**0,012**	**0,212**	**0,062**	**0,274**

Table 32 : Pourcentages de désaccord entre annotation manuelle du mouvement et estimation automatique de la quantité de mouvement.

Ces 10 extraits vidéo provenant de EmoTV sont riches en annotations manuelles des mouvements, que ce soit des mouvements de tête, de mains ou de torse : le pourcentage d'images pour lesquelles il n'y a pas d'annotation manuelle de mouvement est de 26% pour la vidéo 41, 7% pour la vidéo 3, et 5% pour la vidéo 36.

Dans le but de pouvoir calculer des mesures statistiques de l'accord entre annotations manuelle et automatique, nous avons balancé le nombre d'images avec et sans annotations manuelles en 1) calculant le nombre d'images sans annotation manuelle du mouvement, et 2) par une sélection aléatoire du même nombre d'images contenant une annotation manuelle du mouvement. Les taux d'accord et désaccord résultants se trouvent dans les tables 33 et 34. Le nouvel accord moyen est supérieur (0.794) à celui obtenu dans la Table 31 sans le balancement du nombre d'images (0.726). La Table 34 montre que les désaccords ne sont plus balancés : le nombre d'images pour lesquelles il y avait une annotation manuelle du mouvement et pour lesquelles aucun mouvement n'était détecté par le traitement d'image est supérieur à l'inverse pour 8 des 10 vidéos.

Vidéo #	Seuil	Accords		
		Auto 0 Manuel 0	Auto 1 Manuel 1	Total
01	0,047	0,353	0,358	0,711
02	0,013	0,418	0,407	0,825
03	0,076	0,442	0,423	0,865
22	0,071	0,450	0,467	0,917
36	0,034	0,500	0,300	0,800
41	0,008	0,421	0,283	0,704
44	0,010	0,460	0,316	0,776
49	0,084	0,476	0,357	0,833
71	0,048	0,500	0,283	0,783
72	0,044	0,385	0,345	0,730
MOY	**0,043**	**0,440**	**0,354**	**0,794**

Table 33 : Pourcentages d'accord entre annotation manuelle du mouvement et estimation automatique de la quantité de mouvement pour un ensemble balance d'images avec ou sans annotation manuelle du mouvement.

Vidéo #	Seuil	Désaccords		
		Auto 0 Manuel 1	Auto 1 Manuel 0	Total
01	0,047	0,142	0,147	0,289
02	0,013	0,093	0,082	0,175
03	0,076	0,077	0,058	0,135
22	0,071	0,033	0,050	0,083
36	0,034	0,200	0,000	0,200
41	0,008	0,216	0,080	0,296
44	0,010	0,185	0,039	0,224
49	0,084	0,143	0,024	0,167
71	0,048	0,217	0,000	0,217
72	0,044	0,155	0,115	0,270
MOY	**0,043**	**0,146**	**0,059**	**0,206**

Table 34. Pourcentages de désaccord entre annotation manuelle du mouvement et estimation automatique de la quantité de mouvement pour un ensemble balance d'images avec ou sans annotation manuelle du mouvement.

Les valeurs maxima du Kappa et les seuils pour lesquels ils ont été obtenus sont listés dans la Table 35, colonne (1). Les valeurs du Kappas sont entre 0.422 et 0.833, dépendamment des vidéos. Ces valeurs peuvent être considérées plutôt bonnes pour la qualité de résolution de nos extraits vidéo. Dans les résultats décrits dans la Table 35 colonne (1), nous avons sélectionné pour seuils les valeurs donnant les Kappas maxima. Les différences entre les seuils obtenus

pour les différentes vidéos, probablement dues aux différences de qualité vidéo et de conditions d'enregistrement des différentes interviews, montrent que cette valeur de seuil doit être spécifique pour chaque vidéo.

Nous avons exploré l'utilisation des phases durant lesquelles aucun (ou très peu) mouvement n'était perceptivement visible. Nous avons calculé l'estimation moyenne du mouvement fourni par le module de traitement d'image pour chacune de ces phases. L'utilisation de ces valeurs moyennes comme seuil a entraîné des mesures du Kappa inférieurs, dans la Table 35, colonne (2). Des explorations expérimentales futures seront donc nécessaires pour étudier comment ce seuil doit être spécifié.

	(1) **Seuil correspondant au Kappa maximum**		**(2)** **Seuil sélectionné dans la phase de 2 sec sans mouvement**	
Vidéo #	**Kappa Max**	**Seuil**	**Kappa**	**Seuil**
01	0,422	0,047	0,275	0,056
02	0,649	0,013	0,547	0,016
03	0,731	0,076	0,57	0,059
22	0,833	0,071	0,633	0,079
36	0,600	0,034	0,6	0,037
41	0,407	0,008	0,342	0,039
44	0,553	0,01	0,19	0,066
49	0,667	0,084	0,428	0,089
71	0,565	0,048	0,304	0,008
72	0,459	0,044	0,327	0,034
MOY	**0,589**	**0,043**	**0,421**	**0,048**

Table 35 : Kappas obtenus pour le même nombre d'images impliquant une annotation manuelle du mouvement et sans annotation manuelle du mouvement: (1) Le seuil affiché correspond au Kappa maximum, (2) le seuil affiché correspond à la moyenne de l'estimation automatique de la quantité de mouvement sur un intervalle de 2 secondes dans lequel il n'y a pas ou peu de mouvement perçu dans la vidéo.

La Table 36 donne une description informelle des extraits vidéo traités automatiquement.

Video #	Description informelle
01	Rue ; mouvements de tête, expressions faciales ;
02	Rue; mouvements (torse, mains, tête) ; gens bougeant en arrière plan
03	Rue; mouvements (torse, mains, tête) ; expressions faciales ; zone de peau sur le torse
22	Plage ; mouvements (torse, mains, tête) ; expressions faciales ; nombreuses zones de peau (maillot de bain)
36	Intérieur sombre ; mouvement (mains, tête) ; expressions faciales ; gens bougeant en arrière plan
41	Intérieur; mouvements de tête, expressions faciales
44	Extérieur ; mouvements (mains, tête) ; expressions faciales
49	Extérieur; mouvements (mains, tête) ; expressions faciales ; gens bougeant en arrière plan
71	Intérieur ; mouvements (mains, tête) ; expressions faciales ; gens bougeant en arrière plan
72	Extérieur; mouvements (mains, tête)

Table 36 : Description informelle des 10 extraits vidéo utilisés pour l'étude.

3.5.3 Conclusion

Nous avons exploré la façon dont le traitement automatique de l'image peut valider les annotations manuelles de mouvements émotionnels dans des interviews télévisées, que ce soit au niveau global des extraits vidéo qu'au niveau de l'annotation individuelle du mouvement.

3.6 *Copie-Synthèse d'expressions multimodales d'émotions complexes*

3.6.1 Introduction

Les émotions complexes produisent souvent des expressions complexes, par exemple elles peuvent produire plusieurs expressions faciales simultanément (Ekman 1975). En fonction du type d'émotion complexe, les expressions faciales résultantes ne sont pas identiques. Une émotion masquée peut par exemple transparaître à travers l'émotion montrée, tandis que la superposition de deux émotions sera montrée par différentes parties du visage (une émotion étant montrée par le haut du visage et une autre par le bas du visage) (Niewiadomski 2007). La distinction de ces différents types d'émotions complexes dans les systèmes ACA est pertinente, sachant que des études perceptives ont montré que les gens sont capables de reconnaître les expressions faciales liées à des émotions ressenties, mais également à des émotions simulées, aussi bien chez les humains que dans les ACAs (Rehm & André 2005).

Nous avons mené une étude expérimentale dans le but d'évaluer si les gens détectent proprement les signes de différentes émotions dans plusieurs modalités (parole, expressions faciales, gestes), et ceci dans le cas d'émotions superposées ou masquées. Cette étude compare la perception de comportement émotionnel dans des extraits vidéo d'interviews télévisées, avec des comportements similaires reproduits par un agent expressif. La méthode de copie-synthèse que nous décrivons dans une première section a été appliquée à plusieurs extraits vidéos (videos 3, 36 et 41 issues de EmoTV ; vidéo 61 issue de Belfast Naturalistic Database (Douglas-Cowie et al. 2000)). Les vidéos 3 et 41 et les animations correspondantes ont été utilisées pour une étude perceptive que nous détaillons dans une deuxième section.

Les expressions faciales de l'agent sont définies suivant deux approches, l'une à partir d'un modèle d'émotions complexes, et l'autre à partir des annotations des expressions faciales d'extraits vidéo. Nous avons ainsi souhaité apporté une contribution expérimentale à l'étude des différences existantes entre perception visuelle uniquement et perception audio-visuelle, ainsi qu'aux différences entre perception des femmes et perception des hommes (Knapp et Hall 2006). L'un des objectifs était de tester si les femmes ont effectivement tendance à mieux reconnaître les expressions faciales liées aux émotions y compris pour des émotions complexes.

Le système ACA utilisé pour cette expérience était celui développé par l'Université Paris 8, Greta (Pelachaud 2005). Le système analyse le texte en entrée et sélectionne les comportements à générer. Les gestes et autres comportements non-verbaux (expressions faciales, regards) sont synchronisés avec la parole. Le système cherche les mots d'emphase. Il aligne les expressions faciales et la production d'un geste avec ces mots. Il calcule ensuite la phase de préparation du geste, ainsi que la co-articulation avec le geste suivant dans le cas d'un geste maintenu, si le temps entre les gestes consécutifs le permet, et sinon retourne à la position de repos.

Nous avons défini deux approches basées sur le corpus pour générer différentes animations de Greta à partir des annotations vidéo. La première approche, que l'on nommera « replay à partir des annotations », utilise l'annotation des émotions, et l'annotation bas niveau des comportements multimodaux (tels que l'expressivité du geste permettant d'assigner des valeurs aux paramètres d'expressivité de l'ACA, ou encore l'annotation manuelle des expressions faciales). La seconde approche, que l'on nommera « replay à l'aide d'un modèle d'émotions mélangées », est identique à la première approche excepté pour les expressions faciales : elle utilise un modèle pour générer les expressions faciales d'émotions mélangées (Niewiadomski 2007). Dans le but d'éviter le biais provenant de la qualité de la synthèse de la parole dans l'évaluation des similarités entre les animations de l'ACA et la vidéo originale, la voix réelle provenant de la vidéo originale a été intégrée aux animations de l'agent.

3.6.2 De l'annotation à la synthèse

Nous avons défini une méthode que nous avons appelé « copie-synthèse » permettant d'aller en plusieurs étapes d'une vidéo à l'animation rejouant les comportements observés dans la vidéo. Certaines étapes sont manuelles, d'autres sont automatiques (Figure 53). Notre but ici n'est pas de générer directement les animations à partir des annotations, mais plutôt 1) d'identifier les niveaux de représentations permettant de passer de la vidéo à l'animation, et 2) de raffiner itérativement nos annotations et l'agent grâce à l'évaluation de la copie-synthèse. Nous n'effectuons donc pas proprement dit de la « génération de comportement » dans des ACA : il n'y a pas ici de phase de suggestions, sélection et planification des comportements (Cassell et al. 2001).

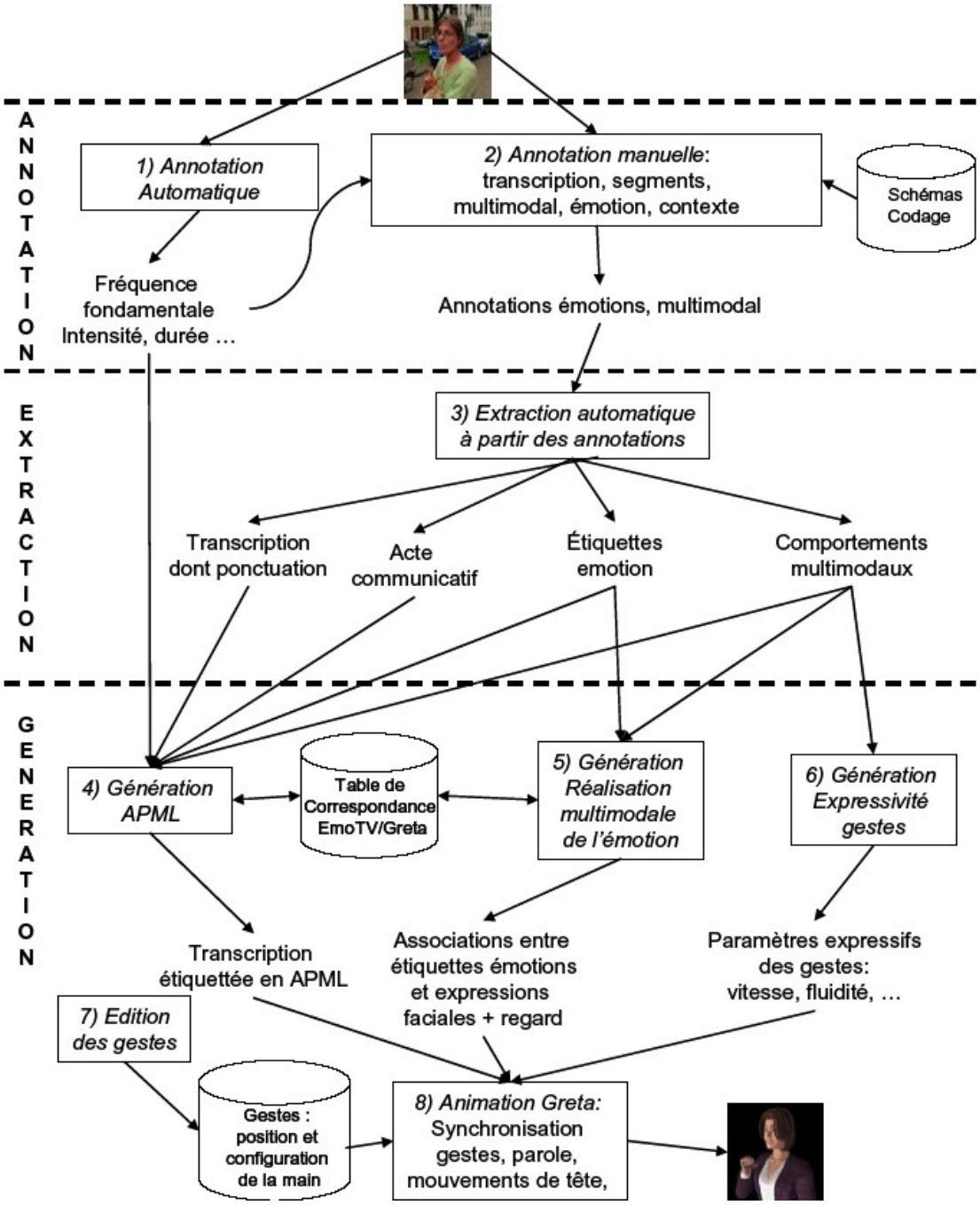


Figure 53: Méthode de copie-synthèse d'expressions multimodales d'émotions complexes spontanées. Chaque étape numérotée est détaillée dans le texte.

Annotation

Etape 1 : Annotation automatique

Cette étape vise à extraire de manière automatique, à partir de la vidéo d'origine, les informations utiles pour la spécification du comportement de l'agent ou pour l'annotation : fréquence fondamentale, intensité… Cette étape est réalisée actuellement avec l'outil Praat.

Etape 2 : Annotation manuelle

Cette étape manuelle effectue plusieurs niveaux d'annotation. La transcription mot à mot avec ponctuation est réalisée en suivant les normes LDC pour les hésitations, les respirations, les rires… La vidéo est ensuite annotée à plusieurs niveaux temporels (vidéo entière, segments de la vidéo, comportements observés à des instants précis) et à plusieurs niveaux d'abstraction (multimodalité, émotion, contexte). Le comportement global de la séquence est décrit par des annotations sur le contexte, les émotions et les types d'indices multimodaux. Les segments sont annotés avec les émotions, les modalités perçues, et intègrent aussi une variation temporelle de l'intensité dans le segment. Cette étape fait intervenir les outils Praat, Anvil et les schémas de codage décrits dans les sections précédentes.

Extraction

Etape 3 : Extraction automatique à partir des annotations

Un programme a été développé pour extraire, à partir des différentes annotations effectuées sur la vidéo, les informations qui ont été identifiées durant la première expérience comme nécessaires à la synthèse avec l'agent expressif Greta : la transcription, l'acte communicatif, les étiquettes et dimensions de l'émotion, les comportements multimodaux (et notamment le nombre d'occurrences et la durée de chaque comportement multimodal pendant chaque segment émotionnel).

Génération

Etape 4 : Génération du fichier APML

Cette étape consiste à générer le fichier APML utilisé par l'agent à partir des données extraites des annotations, et plus particulièrement des informations audio (transcription et fréquence fondamentale) et de l'acte communicatif. La transcription est directement utilisée dans le fichier APML, car elle correspond au texte que l'agent Greta devra synthétiser vocalement. Elle est complétée par différentes balises. La fréquence fondamentale permet de valider/corriger le passage de la ponctuation de la transcription à l'annotation de contours prosodiques adaptés du modèle ToBe utilisés par APML. Par exemple, la « terminaison » LL

correspond à un point de fin de phrase, et LH à une virgule. Nous avons défini une table de correspondance (voir en annexe les tables de correspondances EmoTV/Greta) associant l'acte communicatif utilisé pour l'annotation et le performatif le plus proche dans l'agent. Ainsi, le but communicatif « se plaindre » utilisé pour annoter l'extrait vidéo 3 du corpus est traduit en performatif « critiquer », correspondant à une spécification au niveau global de l'agent (critiquer = regarder_interlocuteur + froncer_sourcils + bouche_critiquer). Cette performative sera combinée aux valeurs d'expressivité des gestes et des expressions faciales de l'agent, pour obtenir l'animation finale.

Etape 5 : Spécification de la réalisation multimodale de l'émotion

Dans les vidéos étudiées, les comportements émotionnels sont complexes et sont généralement annotés avec plusieurs étiquettes différentes. Comme nous l'avons décrit précédemment, ces annotations effectuées par différents annotateurs sont ensuite regroupées en un vecteur d'émotion. Ainsi, le segment 3 de la vidéo 3 a été représenté par le vecteur suivant : 56% de désespoir, 33% de colère et 11% de tristesse. Les deux catégories les plus représentatives de ce vecteur (désespoir et colère) sont regroupées en une étiquette combinée « DespairAnger ». Les annotations des comportements multimodaux observés au niveau du visage durant ce segment sont ensuite utilisées pour associer à cette étiquette une combinaison de comportements des yeux et des sourcils.

Pour les mouvements des yeux, les annotations apportent des informations sur la direction du regard (droite, gauche, haut, bas). Les durées des regards extraits des annotations permettent de pondérer les regards à droite et à gauche, et de spécifier les durées maximales pendant lesquelles l'agent va regarder son interlocuteur ou regarder ailleurs. Dans le cas de notre segment 3, qui a une durée totale de 13 secondes, 41 segments ont été annotés pour les yeux, regard gauche (3 segments, 1.6 secondes (12% du temps)), regard droite (15sgts, 5.84s (45%)). Les résultats montrent que le sujet à également regardé en bas (2sgts, 0.32s (2%)), son interlocuteur (1sgt, 0.03s (0.2%)), et a cligné ou fermé les yeux (20sgts, 3.79s (29%)). Dans notre première expérience de génération d'animation à partir d'annotations, nous avions traité uniquement les annotations droite et gauche, et considéré que le reste du temps (ici 43% du temps), l'agent regarde son interlocuteur. L'utilisation automatique des annotations de bas niveaux permet ici un niveau plus fin et plus fidèle à la vidéo d'origine. Pour les sourcils, deux informations sont apportées par les annotations, leur mouvement (froncés ou relevés, avec la durée), et leur intensité. Dans le cas du segment 3, 4 segments ont été annotés pour les sourcils, tous avec la valeur « froncés » (4sgts, 3.4s (26% de 13 secondes)), et avec une intensité forte. Dans l'agent, ces annotations sont traduites en « high_frown », correspondant à

un froncement de sourcil de forte intensité pendant 26% (environ 1/4) du segment émotionnel. Au final, l'affect « DespairAnger» est défini de la manière suivante, sachant que les valeurs avant chaque parenthèse fermante correspondent aux probabilités d'exécution de chacun des comportements :

DespairAnger:=
// 12% regard gauche dont 1/4 de sourcils froncés (eyes_left + high_frown, 0.03),
(eyes_left, 0.09),
// 45% regard à droite dont 1/4 de sourcils froncés (eyes_right + high_frown, 0.11), (eyes_right, 0.34),
// 43% regard vers l'interlocuteur // dont 1/4 de sourcils froncés (look_at + high_frown, 0.10), (look_at, 0.33)

Etape 6 : Calcul des valeurs des paramètres expressifs des gestes

En ce qui concerne les paramètres d'expressivité des gestes, la correspondance est effectuée de la manière suivante : 5 segments ont été annotés, d'une durée totale de 11.68 secondes ; les attributs en rapport à la qualité du mouvement, c'est-à-dire la fluidité : fluide (3 segments, 9.16 secondes), la force : forte (2sgts, 2.5s), faible (1sgt, 1.1s), la vitesse : rapide (5sgts, 11.68s), et l'expansion spatiale : contractée (5sgts, 11.68s), déterminent les valeurs des paramètres expressifs correspondant de l'agent Greta. Les durées d'annotation de chaque valeur présente dans la fenêtre temporelle du segment émotionnel sont utilisées pour calculer les valeurs de chacun des paramètres d'expressivité. Par exemple, dans le cas du 3ème segment, la durée totale du segment étant de 13 secondes, les valeurs obtenues sont les suivantes : pour le paramètre « fluidité », 9.2 secondes ont été annotées avec la valeur « fluide », 2.4 secondes avec la valeur saccadée, et 1.4 secondes avec la valeur « normale »; sachant que le paramètre fluidité (FLT) de Greta est compris entre -1 (saccadé) et +1 (fluide), la valeur affectée est égale à 0.52 (9.2s = 70% fluide, 2.4s = 18% saccadé, 1.4s= 12% normal => 0.70 - 0.18 = 0.52).

Etape 7 : Création des gestes

Si un geste annoté dans une vidéo n'est pas répertorié dans les gestes gérés par l'agent Greta, il est nécessaire de l'ajouter à l'aide de l'éditeur de configurations de gestes qui permet de spécifier la position et la configuration de la main. Ainsi plus le corpus traité s'agrandit, et plus la bibliothèque gestuelle de Greta s'enrichit de nouveaux gestes.

Etape 8 : Animation Greta

L'agent Greta prend en entrée plusieurs fichiers : la transcription étiquetée avec APML, les associations entre labels émotions et les expressions faciales, les paramètres expressifs des gestes. Le module gère la synchronisation entre parole, gestes et mouvements de tête. Le résultat est une animation.

Conclusion

Nous avons défini une approche par copie/synthèse pour étudier les différents niveaux intervenant dans la réalisation multimodale des comportements émotionnels. Globalement deux niveaux d'indices sont extraits des annotations ; l'un pour simuler le comportement de l'utilisateur se servant des annotations sur les comportements émotionnels et des buts communicatifs, l'autre pour synthétiser ce comportement à l'aide d'indices multimodaux (combinaisons de gestes et d'expressions faciales ce qui a été peu fait dans la littérature).

Les annotations contextuelles contiennent d'autres informations (dont des dimensions d' « appraisal » comme l'implication, la nouveauté…) qui seraient à terme intéressantes de considérer dans le modèle de comportement de l'agent. De même, d'autres niveaux d'annotations multimodales annotés comme pertinents dans la vidéo (par exemple mouvements de tête) pourraient aussi être considérés dans la génération du comportement de l'agent afin d'envisager différents niveaux de fidélité du comportement de l'agent par rapport à la vidéo d'origine. Notre approche par copie/synthèse est particulièrement intéressante pour la création d'agents expressifs basés sur des émotions plus complexes que les émotions de base.

3.6.3 Test perceptif utilisant des vidéos et des animations d'ACA

3.6.3.1 Protocole expérimental

Les objectifs de cette expérience étaient de 1) tester si les sujets perçoivent une combinaison d'émotions dans les replays de l'agent expressif, comme dans la vidéo originale, et 2) comparer les 2 approches pour rejouer des émotions complexes. Nous avons sélectionné 2 extraits vidéo d'interviews télévisés pour cette étude. Le premier extrait est le 3ème segment de la vidéo #3 (une femme réagissant à une décision de justice récente condamnant son père et son frère à la prison). Les annotations manuelles de cette vidéo par 3 annotateurs experts ont montré que son comportement était perçu comme une superposition de colère et de désespoir, et ce dans la parole et dans plusieurs modalités visuelles (regard, mouvements de tête, mouvements du torse et gestes). Ces annotations ont été validées par celles faites par 40 sujets ayant différents niveaux d'expertise comme nous l'avons décrit à la section 3.3.4.

Le second clip (vidéo #41) montre une femme masquant sa déception après les résultats négatifs de son parti politique aux élections. Cette vidéo a été annotée comme une combinaison de labels négatifs (déception, tristesse, colère) et positifs (content, sérénité). L'annotation des comportements multimodaux a révélé que les lèvres dessinaient un sourire tendu (lèvres pincées). Les bras et le haut du corps sont visibles sur le premier extrait, tandis que le deuxième ne montre que la tête de la personne enregistrée.

Quarante sujets (23 hommes, 17 femmes), âgés entre 19 et 36 ans (moyenne 24 ans) ont comparé les vidéos originales et les différentes animations effectuées avec l'agent 3D Greta. Trente trois des sujets étaient étudiants en informatique, 7 étaient chercheurs, professeurs ou ingénieurs. L'expérience incluait 2 conditions : la première sans audio, puis la deuxième avec audio. Dans chaque condition, les sujets devaient visionner la vidéo originale et 4 animations différentes. Deux animations étaient spécifiées avec des données provenant de la littérature sur les émotions basiques dans les expressions faciales (Ekman 2003) et les mouvements corporels (Wallbott 1998, Pelachaud 2005). Les 2 autres animations étaient générées avec les approches mentionnées plus haut pour rejouer les comportements annotés. Ainsi, dans le cas de l'exemple de superposition d'émotion dans l'extrait #3, 4 animations ont été conçues : 1) Colère, 2) Désespoir, 3) Replay multiples niveaux issus de notre méthode de copie-synthèse décrite à la section précédente, et 4) Replay mélange facial (Niewiadomski 2007).

Pour l'animation 4), les valeurs assignées aux paramètres d'expressivité du geste étaient les mêmes que pour l'animation 3) (e.g. à partir de l'annotation manuelle de l'expressivité perçue dans les gestes). De la même manière, pour l'exemple d'émotion masquée présent dans l'extrait #41, les 4 animations étaient : 1) Joie, 2) Déception, 3) Replay multiples niveaux, et 4) Replay mélange facial.

Les sujets devaient assigner une valeur entre 1 (forte ressemblance) et 4 (faible ressemblance) à chaque animation (Figure 54). L'ordre de présentation des extraits #3 et #41, ainsi que la position des animations dans l'interface graphique étaient contrebalancés entre les sujets, pour les conditions audio et sans audio. Les sujets avaient la possibilité d'assigner la même valeur de ressemblance à plusieurs animations.

Après chaque condition, les sujets devaient remplir la partie du questionnaire correspondant à la tâche qu'ils venaient d'effectuer. Ils devaient noter le degré de certitude avec lequel ils ont assigné les valeurs de ressemblance, pour cela ils avaient le choix entre 5 scores de confiance :

1) J'ai clairement perçu des différences dans les 4 vidéos et j'ai pu facilement comparer avec la vidéo référence (4 points), 2) J'ai perçu quelques différences qui m'ont aidé(e) à faire mon classement (3 points), 3) J'ai perçu quelques différences mais j'ai eu du mal à comparer avec la vidéo référence (2 points), 4) J'ai perçu peu de différences entre les vidéos, j'ai eu beaucoup de mal à les classer (1 point), 5) Je n'ai perçu aucune différence entre les vidéos (0 point).

Dans le questionnaire, les sujets devaient également annoter les émotions qu'ils avaient perçu dans l'animation classée comme la plus ressemblante à la vidéo de référence. Ils avaient la possibilité de sélectionner un ou plusieurs labels dans la même liste de 18 labels émotionnels que celle utilisée pour l'annotation des vidéos dans les expériences décrites plus haut.

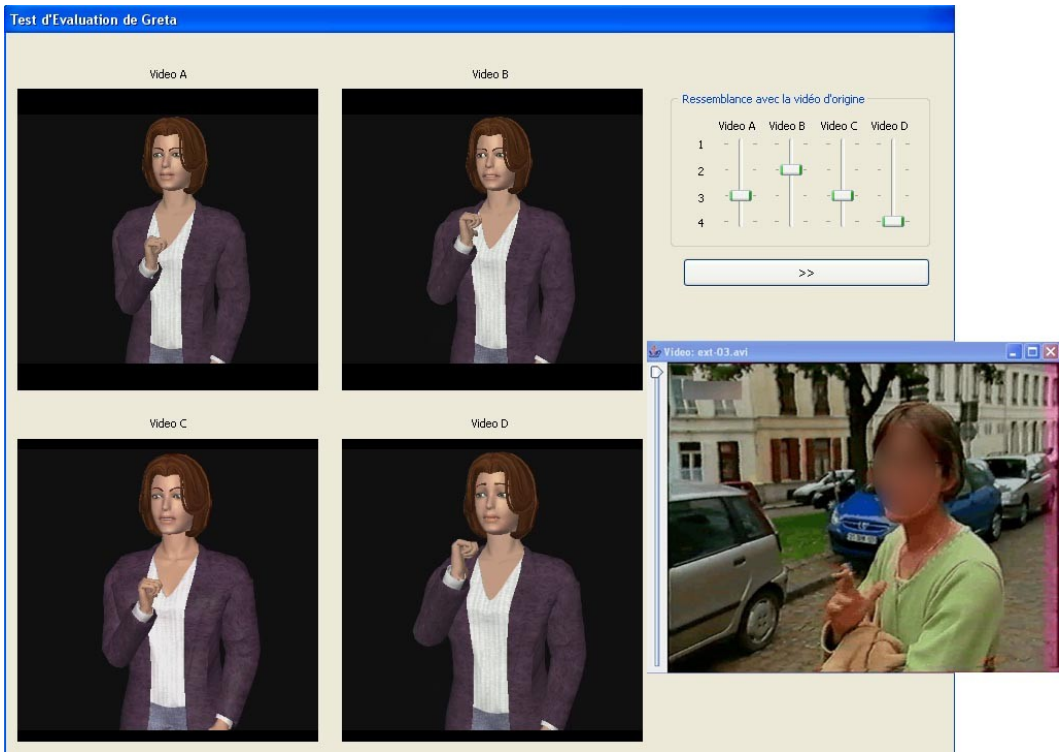


Figure 54 : Capture d'écran de l'exemple superposition; 4 animations différentes et 4 valeurs de ressemblance à affecter à chacune des animations ; la vidéo originale (non masquée lors de l'expérience) est affichée séparément ; la vidéo originale et les animations contiennent des expressions faciales et des gestes ; l'exemple avec l'émotion masquée est similaire mais est ciblée sur le visage.

3.6.3.2 Résultats

Superposition colère et désespoir

Nous avons calculé le nombre de fois où chaque animation a été classée comme la plus similaire à la vidéo. Dans la condition sans son, la colère est classée 1ère par 61% des sujets (niveaux multiples 20%, mélange facial 9%, désespoir 9%). Dans la condition avec son, la colère est classée 1ère par 33% des sujets (niveaux multiples 26%, mélange facial 24%, désespoir 17%). Nous avons mené une analyse de variance avec la Condition sonore (avec, sans son) et l'Animation (niveaux multiples, mélange facial, colère, désespoir) comme facteurs intra-sujets. Le classement des animations a été converti en scores de similarité (le 1er rang est devenu un score de similarité de 3 points, le 4ème rang un score de similarité de 0 point). L'effet principal de l'Animation s'est montré significatif ($F(1/114)=15.86$; $p<0.001$, cf. Fig. 55), ainsi que l'interaction entre Condition sonore et Animation ($F(1/114)=5.98$; $p=0.001$, cf. Fig. 56) : l'effet de l'Animation est hautement significatif dans la condition sans son ($F(3/114)=24.11$; $p<0.001$) alors qu'il est en tendance dans la condition avec son ($F(2/114)=2.42$; $p=0.087$).

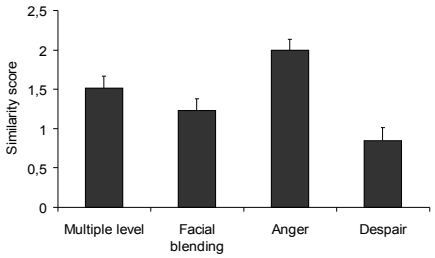


Figure 55. Scores de similarité des 4 animations avec l'exemple de superposition.

Figure 56. Scores de similarité des 4 animations en fonction de la condition sonore.

Enfin, la table 37 résume les résultats de la tâche d'annotation à l'aide de macro-catégories (colère, tristesse). Ces résultats montrent que, même si l'animation la mieux classée était l'animation « colère », les sujets l'ont perçue comme une combinaison de plusieurs émotions.

	Colère + Tristesse	Colère sans Tristesse	Tristesse sans Colère	Ni Colère ni Tristesse	Total
Sans son	52,5 %	40 %	0 %	7,5 %	100 %
Avec son	85 %	0 %	12,5 %	2,5 %	100 %

Table 37. Pourcentages des sujets ayant perçu chaque macro-catégorie d'émotion dans l'animation classée comme la plus similaire à la vidéo témoin.

Déception masquée par la joie

Dans la condition sans son, la joie a été classée 1ère par 40% des sujets (mélange facial 33%, niveaux multiples 20%, déception 7%). Dans la condition avec son, le mélange facial a été classé 1er par 38% des sujets (niveaux multiples 27%, joie 24%, déception 11%). Une analyse de variance avec la Condition sonore (avec, sans son) et l'Animation (niveaux multiples, mélange facial, joie, déception) comme facteurs intra-sujets, montre un effet principal de l'Animation ($F(1/114)=18.07$; $p<0.001$, cf. Figure 57). Par ailleurs, la table 38 résume les résultats de la tâche d'annotation.

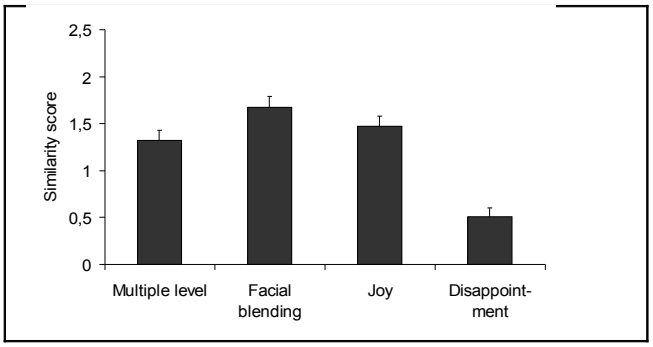


Figure 57. Scores de similarité des 4 animations avec la vidéo initiale dans l'exemple de masquage.

	Joie + Tristesse	Joie sans Tristesse	Tristesse sans Joie	Ni Joie ni Tristesse	Total
Sans son	7,5 %	10 %	57,5 %	25 %	100 %
Avec son	12,5 %	40 %	30 %	17,5 %	100 %

Table 38. Pourcentages des sujets ayant perçu chaque macro-catégorie d'émotion dans l'animation classée comme la plus similaire à la vidéo témoin.

Effets des modèles d'émotions complexes

Une analyse de variance a été menée sur les deux exemples (superposition et masquage), les deux conditions sonores (avec et sans son) et nos deux approches (niveaux multiples et mélange facial). Les résultats ne montrent aucun effet de l'approche ($F(1/38)=0.01$; NS), c'est-à-dire aucune différence globale significative entre les reproductions niveaux multiples et mélange facial.

3.6.3.3 Conclusion

Le 1er objectif de cette étude était de tester si les sujets perçoivent une combinaison d'émotions dans nos animations d'ACAs. Nos résultats montrent que les sujets one tendance à percevoir une émotion prédominante : dans l'exemple de la superposition, les sujets ont classé ce que nous avons appelé « colère basique » en premier ; dans l'exemple de déception masquée par la joie, la joie basique et le mélange facial ont été classés de manière équivalente. Cependant, ce résultat est partiellement contredit par l'analyse de la tâche d'annotation dans laquelle la plupart des sujets ont associé plusieurs labels aux animations classées les plus proches de la vidéo témoin.

Par ailleurs, notre protocole expérimental nous a permis de comparer deux conditions, avec et sans le son. Les 4 animations correspondant à l'exemple de superposition ont été mieux discriminées dans la condition sans son ; l'ajout d'indices verbaux a eu l'effet de diminuer les différences entre les animations. Dans les animations de superposition, les sujets ont mieux perçu la dimension tristesse / désespoir dans la condition avec le son. Nous pouvons en conclure que, dans cette séquence particulière, la colère a été principalement exprimée par des comportements non verbaux alors que le désespoir a été exprimé par le canal verbal.
Dans l'exemple de la déception masquée par la joie, les sujets ont mieux perçu les émotions négatives dans la condition avec le son. Ceci suggère que la personne dans cette séquence vidéo contrôlait mieux ses comportements non verbaux (émotions positives perçues dans la condition sans son) que ses comportements verbaux (indices négatifs perçus dans la condition avec le son). Cet effet important du comportement verbal peut être dû au fait que nous avons utilisé la voix originale et non une synthèse vocale. Ceci est compatible avec de précédentes études suggérant que les indices acoustiques expriment bien la colère et la douleur (Abrilian et al. 2005).

Le second objectif de cette étude était d'évaluer nos deux approches pour reproduire des émotions mélangées (reproduction niveaux multiples et mélange facial). Notre analyse globale montre qu'aucune des deux approches n'a été préférée de manière univoque. Il faut préciser que les deux approches de reproduction partagent des traits communs, ce qui peut partiellement expliquer pourquoi il n'était pas si simple de les discriminer : elles ont été élaborées sur la base des mêmes annotations et incluaient des comportements générés automatiquement par le système Greta. Cependant, nous pouvons remarquer que la reproduction du mélange facial a été significativement mieux notée que la reproduction niveaux multiples dans l'exemple du masquage. A l'inverse dans l'exemple de superposition, les deux types de reproduction ne différaient pas significativement. Ces résultats suggèrent que la reproduction du mélange facial était plus appropriée à l'exemple de masquage et la reproduction niveaux multiples plus appropriée à l'exemple de superposition. De nouvelles données sont nécessaires pour comprendre cette interaction.

Nous avons cependant observé une différence entre les hommes et les femmes. Dans le cas de la superposition, les femmes ont mieux noté l'animation issue de notre méthode de copie-synthèse. Elles se sont également déclarées plus confiantes que les hommes dans leurs sélections. Ceci est compatible avec l'expérience menée par (Hall et Matsumoto 2004) qui ont demandé à des sujets d'annoter avec plusieurs labels des images statiques censées représentée des émotions simples. Ils ont observé dans une condition de temps limité que les femmes se rapprochaient plus d'un vecteur idéal représentant la perception collective calculé dans une condition sans contrainte de temps.

Un inconvénient de notre approche par corpus basée sur des comportements spontanés pendant des interviews télévisées est qu'il est difficile de collecter suffisamment de données pour entraîner des modèles statistiques. C'est la raison pour laquelle nous avons utilisé une approche exploratoire avec des cas illustratifs d'émotions complexes (superposition et masquage) pour valider nos représentations. Nous pensons que de telles études expérimentales vont permettre d'identifier les facteurs comportementaux les plus critiques à la perception et à l'expression d'émotions multimodales réalistes.

3.7 *Pistes de recherche*

Nous décrivons ici quelques pistes d'utilisation de nos contributions pour des futures recherches.

3.7.1 Exploitation et extension des schémas de codage

Les schémas de codage ont déjà été étendus dans trois directions.

Les schémas et expériences d'annotation des différentes dimensions de l'émotion ont récemment contribué à la définition des besoins en terme d'annotation des émotions au sein du groupe de travail W3C Emotion Incubator Group visant à proposer des éléments de **standardisation** des émotions (http://www.w3.org/2005/Incubator/emotion/).

Une des expériences que nous avons menée sur l'annotation du **contexte émotionnel** a montré que les émotions spontanées de la vie de tous les jours sont souvent complexes car déclenchées par plusieurs évènements de différente nature qui peuvent être liés temporellement et donnant lieu à différentes émotions. Ce constat nous a poussé à explorer une annotation précise des évènements émotionnels et de leur localisation dans le temps. Nous avons ainsi défini une première méthode d'annotation manuelle des événements émotionnels et de leurs dimensions d' « appraisal » en deux étapes. La première étape produit une liste des événements émotionnels pertinents pour chaque vidéo et est validée par accord consensuel par les annotateurs. Les caractéristiques temporelles de chaque événement sont annotées en terme de couverture temporelle (choix multiple parmi passé lointain (> 1 semaine), passé proche (moins d'1 semaine), présent (le jour même), futur proche moins d'1 semaine), futur lointain (> 1 semaine)). La durée de l'évènement est également annotée (mois et plus, semaines, jours, heure, minutes). Ce qui permet de déduire l'enchaînement temporel des évènements. L'interview est un évènement commun à toutes les vidéos (couverture: présent, durée: minutes) et n'est pas annotée. Cette liste d'événements est alors insérée dans le fichier d'annotation Anvil. La deuxième étape consiste à annoter les dimensions d' « appraisal » pour chaque événement pris indépendamment. Cette méthode a été appliquée par 3 codeurs experts sur 28 clips d'EmoTV. Les annotations des dimensions d' « appraisal » sur ces événements qui sont maintenant décrits individuellement restent à effectuer ainsi que les mesures d'accord correspondantes et de corrélation avec les vecteurs d'émotions. Cette phase d'annotation précise sur des données réelles est très onéreuse, de plus la perception du contexte est

extrêmement sollicitée et donc cette annotation est également subjective. Les premiers résultats de ces annotations montrent qu'il y a dans la majorité des cas plus 2 évènements (sans inclure l'évènement interview qui peut stresser, décontenancer ou le locuteur) pour chaque segment émotionnel du sous corpus étudié. Ce chiffre est corrélé avec la présence d'émotions mixtes dans ces vidéos. Des travaux récents proposent des modèles d'émotions ambivalentes dans des agents virtuels dues à la présence simultanée de plusieurs événements émotionnels conflictuels (Lee et al. 2006).

La difficulté d'obtenir des mesures automatiques fiables sur le signal audio (due aux niveaux disparates de qualité d'enregistrement et aux bruits ambiants présents dans l'ensemble des vidéos) nous a poussé à décrire un **schéma d'annotation sur les paramètres prosodiques**. Nous avons élargi ce schéma avec des annotations sur la qualité vocale qui sont difficilement extraites automatiquement même sur des données audio de bonne qualité. De plus, l'expérience encourageant sur l'annotation perceptive des mouvements nous a poussé à envisager ce test perceptif. Le schéma suivant a été défini à partir de indices perceptifs décrits dans la littérature principalement dans des travaux de Klaus Scherer et d'Ellen Douglas-Cowie, travaux repris dans les schémas du réseau HUMAINE. Ce test sera appliqué à un sous-ensemble d'EmoTV :

- Rythme (Allongement début, Allongement milieu, Allongement final, Rapide, Lent, Irrégulier, Autre rythme)
- Qualité (Tremolo, Grinçante, Criée, Chuchotée, Nasiarde, Hyperarticulée, Dure, Timbrée, Détimbrée, Soufflée),
- Plaisante
    - Etendue : relaxation du conduit oral, plus d'énergie dans les basses fréquences
    - Etroite : tension du conduit vocal, plus d'énergie en hautes fréquences
    - Intermédiaire
    - Relaxation
    - Détendue : relaxation de l'appareil vocal, F0 bas, peu d'amplitude
    - Tendue : tension de l'appareil vocal, F0 et amplitudes élevées
    - Intermédiaire
- Contrôle
    - Tendue : tension de l'appareil vocal, augmentation de F0 et d'amplitude

    - Relâchée : hypotension de la musculature, F0 bas et peu d'amplitude
    - Intermédiaire
- Puissance
    - Pleine : registre de poitrine, F0 bas et beaucoup d'énergie
    - Fine : registre de tête, F0 haut et peu d'énergie
    - Intermédiaire
- Energie : Forte, Faible, Irrégulière, Normale
- Intonation : Haute, Basse, Grand saut, Irrégulière, Grande variabilité, Peu de variabilité
- Valence : Négative, Positive, Neutre
- Mot accentué
- « Affect burst » (Devillers, 2006) (Rire, Toux, Respiration, Cris, Larmes, Souffles, Reniflements, Interjection, Hésitations)
- Silence (Interruption, Silences très longs, Silences trop fréquents, Silences courts, Débit lent, Débit rapide)

Ces deux extensions pourront donner lieu à des tests perceptifs visant à étudier les différences individuelles déjà observées dans certaines expérimentations effectuées durant cette thèse.

3.7.2 Mesures des relations entre niveaux d'annotation

Cette thèse a permis d'identifier différents niveaux de représentation pertinents pour les émotions spontanées et leurs expressions multimodales. Les annotations à ces différents niveaux permettront de mesurer des relations entre les différents niveaux. Nous listons ici quelques mesures qu'il nous parait pertinent d'effectuer et qui sont suggérées par la littérature sur des données différentes (corpus conversationnels non émotionnels ou données émotionnelles actées) :

Relations entre modalités

- Redondance des mouvements de tête, de torse, et des gestes
- Relation temporelle segment parole / segment de mouvement :
    - Relation entre mot et « stroke » du geste
    - Corrélation entre répétition des mouvements et de la parole
- Corrélation entre annotation manuelle de l'activation au niveau global et estimation de la quantité de mouvement

Relations entre annotation manuelle de l'activation et indices audio extraits automatiquement

- Expressions faciales et expressivité des gestes
- Pitch et intensité de la parole liés à la vitesse, expansion, force des mouvements
- Type de geste et mot prononcé

Relations entre émotions et modalités

- Expressivité des gestes et catégorie d'émotion
- Expressions faciales / mot prononcé / émotion
- Mouvements liés aux « affects burst/disfluences »
- Mouvements de tête et classe d'émotion (les mouvements de tête « non » sont-ils liés à des émotions négatives ? Et les « oui » à des émotions positives ?)
- Relations entre les mouvements des 2 bras et les émotions (utilisation d'un seul bras lié à une émotion moins forte que l'utilisation de 2 bras, quand les 2 bras sont disponibles)
- Type de geste et émotion

Nous illustrons ci-dessous deux de ces mesures : 1) la redondance des mouvements de tête, de torse, et des gestes, et 2) les relations entre annotation manuelle de l'activation et les indices audio extraits automatiquement.

Redondance des mouvements de tête, de torse, et des gestes

Un sous-ensemble du corpus EmoTV, constitué de 28 extraits vidéo (sélectionnés pour leur richesse en mouvements (gestes, tête, torse) et contenu audio, a été annoté par 1 sujet expert avec un sous-ensemble du schéma de codage multimodal limité aux paramètres expressifs des mouvements et gestes des mains. Les annotations ont été analysées par programme Java, et les résultats de cette analyse sont présentés dans les tables ci-dessous, pour un des extraits du sous-ensemble annoté (vidéo 71 annotée aussi par un deuxième codeur expert). Durée totale de la vidéo = 6.64 secondes (166 images, avec un taux de 25 images par secondes). Le nombre et la durée des mouvements annotés par les 2 sujets sont donnés dans la Table 39.

Modalité		Mesure	Coder1	Coder2
Torse		Nombre	0	0
		Durée totale (sec)	0	0
		Durée moyenne (sur les segments)	0	0
Tête		Nombre	6	8
		Durée totale	4.48	5.88
		Durée moyenne	0.75	0.74
Gestes	Main droite	Nombre	1	1
		Durée totale	0.8	0.96
		Durée moyenne	0.8	0.96
	Main gauche	Nombre	0	0
		Durée totale	0	0
		Durée moyenne	0	0
	Deux mains	Nombre	1	1
		Durée totale	1.24	0.88
		Durée moyenne	1.24	0.88

Table 39 : Nombre et durée des mouvements annotés par les 2 sujets, pour la vidéo 71.

La Table 40 montre les mesures de redondance des mouvements de tête, torse, et des gestes pour chacun des 2 sujets, pour la vidéo 71. Le programme vérifie si une annotation est présente pour chaque image (toutes les 0.04 secondes) et pour chaque modalité. Par exemple, les résultats montrent que le Coder1 n'a annoté aucun mouvement pour 30 % de la vidéo

(correspondant à 50 images sur le total de 166), a annoté 42 % de la vidéo avec uniquement 1 modalité (39 % de mouvements de tête et 3 % de geste avec la main droite), et 27 % avec 2 modalités au mêmes instants (18 % pour les modalités tête/deux mains, et 9% pour les modalités tête/main droite).

<table>
<tr><td>Nb modalités</td><td colspan="2">Coder1</td><td colspan="2">Coder2</td></tr>
<tr><td>0</td><td colspan="2">30 % (50 images)</td><td colspan="2">12 % (20 images)</td></tr>
<tr><td rowspan="2">1</td><td rowspan="2">42 %
(70 images)</td><td>Tête : 39 %</td><td rowspan="2">54 %
(90 images)</td><td rowspan="2">Tête : 54 %</td></tr>
<tr><td>Main droite : 3 %</td></tr>
<tr><td rowspan="3">2</td><td rowspan="3">27 %
(46 images)</td><td>Tête/Deux mains : 18%</td><td rowspan="3">33 %
(56 images)</td><td>Tête/Deux mains : 13 %</td></tr>
<tr><td rowspan="2">Tête/Main droite : 9%</td><td>Tête/Main droite : 14%</td></tr>
<tr><td>Tête/Main gauche : 6 %</td></tr>
<tr><td>3</td><td colspan="2">0 %</td><td colspan="2">0 %</td></tr>
<tr><td>4</td><td colspan="2">0 %</td><td colspan="2">0 %</td></tr>
</table>

Table 40 : Mesures de redondance des mouvements de tête, torse, et des gestes pour chacun des 2 sujets, pour la vidéo 71.

La présence de 4 modalités est possible s'il y a recouvrement des annotations de main droite et main gauche (mouvements de torse + tête + main droite + main gauche).

Relations entre annotation manuelle de l'activation et indices audio extraits automatiquement

14 vidéos ont été traitées automatiquement pour l'extraction d'indices audio (en moyenne 3 segments par vidéo). Ces vidéos ont été sélectionnées car des mesures automatiques sur l'audio pouvaient être menées de façon fiable. Ce n'est pas le cas de la majorité du corpus ; les conditions d'enregistrement sont souvent mauvaises et les bruits ambiants très élevés. Les segments ont été annotés par 3 personnes sur le niveau d'activation, une valeur moyenne a été calculée. 70% des segments contenaient l'activation annotée manuellement la plus élevée, le Max F0, et le F0 moyen le plus élevé. Concernant l'intensité, aucune différence n'a été remarquée entre les segments.

La Table 41 montre les relations entre le F0 max et la phase des gestes pour la vidéo 36.

Transcription	**F0 max local**	**Phase du geste**	**Mvt de tête**
2ème tour	474	Retraction	Mvt
Pourri	582	Retraction	Mvt

Table 41 : Relations entre F0 max et phase des gestes.

La Table 42 montre les relations temporelles entre le F0 max et le « stroke » des gestes pour la vidéo 44.

Transcription	Max F0 local	Geste	Mvt de tête
extérieur	152	Pendant le stroke	Mvt
police	187	Fin du geste 1 seconde avant de prononcer le mot "police"	Aucun
puis	191	Fin du mot contigu avec le début de la préparation	Aucun
eux	176	Mot contigu avec les gestes précédent et suivant	Mvt
point	171	Pendant le stroke	Mvt

Table 42: Relations temporelles entre F0 max et stroke des gestes.

L'emphase des gestes avec la parole a lieu sur les bornes syntaxiques ou au niveau des « lexical affiliates ».

Enfin, la Table 43 montre les relations entre le F0 max, F0 moyen, et expressivité, pour la vidéo 36.

Segments	Vitesse des mvts de tête	Fluidité des mvts de tête	F0 max	F0 moyen	Max Int
S1	rapide	saccadé	484,865	480,865	77,346
S2	lent	fluide	194,063	191,063	78,232
S3	rapide	saccadé	484,884	480,884	76,687

Table 43 : Relations entre F0 max, F0 moyen, et expressivité, pour la vidéo 36.

Ces exemples illustrent le type de mesures qui pourront être effectuées à l'aide des schémas développés dans cette thèse et leurs extensions futures.

Les différents niveaux de représentation identifiés pourront être intégrés dans un modèle fédérant des connaissances hétérogènes multi niveaux en terme de profil expressif émotionnel : intégration annotation continue / discrète, manuelles / automatiques, résultats de tests perceptifs sur vidéo et animations.

Conclusion générale

Les prototypes actuels d' « affective computing » sont généralement limités à la détection et à la génération de quelques émotions simples et se fondent sur des données audio ou vidéo jouées par des acteurs et récoltées en laboratoire. L'objectif de cette thèse était d'étudier les émotions spontanées et les signes multimodaux associés pour contribuer à long terme à la conception des futurs systèmes affectifs interactifs. Notre étude portait tout d'abord sur le choix de la représentation. Que peut-on conclure sur ce choix?

Nous avons privilégié une approche catégorielle des émotions lors de nos travaux, mais également exploré deux autres théories, de manière complémentaire : « Component Process Model of Emotion » de Scherer pour la représentation des dimensions « d'appraisal » des événements émotionnels ; et l'approche dimensionnelle (Osgood et al., 1975) : représentation de l'intensité, valence, activation et autres dimensions continues. Pour représenter des émotions complexes, il était plus aisé de combiner des catégories d'émotions que des dimensions. Les émotions complexes ont été étudiées principalement de manière théorique par (Ekman 1975/1992, Izard 1994, Plutchik 1980, Scherer 1984) et observée expérimentalement par (Scherer 1998). Notre étude nous a permis de montrer expérimentalement l'existence d'émotions mélangées de valences différentes dans des interactions émotionnelles de la vie de tous les jours, ici des interviews télévisées. En parallèle de notre étude expérimentale, une étude « Grid Study» (Roesch et al. 2006) a été menée par UNIGE sur les relations entre les trois théories de représentation impliquant 24 labels et 15 nationalités.

L'approche exploratoire que nous avons choisie dans cette thèse était nécessaire afin de pouvoir représenter les émotions spontanées et leurs expressions dans différentes modalités. Cette approche a consisté à collecter et annoter un **corpus vidéo d'interviews télévisées**. Ce type de corpus comporte des émotions plus complexes que les 6 émotions de base (colère, peur, joie, tristesse, surprise, dégoût). Nous y avons en effet observé dans les comportements émotionnels spontanés des superpositions, des masquages, des conflits entre émotions positives et négatives. Ce corpus illustre ainsi des situations et événements émotionnels variés, et est de taille appropriée pour l'étude exploratoire de ces différents niveaux de représentations, et de leurs relations.

Nous avons rapporté plusieurs expérimentations ayant permis d'apporter des éléments de réponses aux questions suivantes :

Comment annoter de manière fiable et représenter des émotions complexes spontanées ?
Nos différentes expériences sur **l'annotation des émotions** nous ont montré la difficulté d'annoter des émotions dans des contextes naturels, la difficulté de nommer une émotion, et les conflits pouvant apparaître entres différentes modalités d'expression, comme la voix et les gestes : nos résultats sont compatibles avec ceux observés dans la littérature, les canaux expressifs peuvent être contradictoires ou complémentaires, les émotions peuvent entraîner des expressions très subtiles ou contrôlées. Les expériences ont montré par exemple que l'accord inter-annotateur était plus haut pour les données audio, puis vidéo et enfin pour les données audiovisuelles. L'utilisation des deux canaux audio et vidéo a rajouté de la complexité contrairement à notre première intuition. Nos expériences ont également montré que les émotions naturelles sont fréquemment des combinaisons de plusieurs émotions, et nous avons identifié plusieurs types de combinaisons d'émotions (masque, conflit, mixage…). Le choix de nos données est en ce sens provocatif, c'est-à-dire qu'il correspond à une collection de données audio-visuelles sélectionnées pour leur richesse expressive et leur complexité émotionnelle. Le suivi du guide d'annotation, des instructions et définitions qu'il contient est important. L'approche séquentielle utilisée pour annoter doit être respectée, en se focalisant à chaque fois sur la dimension que l'on souhaite annoter. La représentation des émotions complexes spontanées nécessite la définition de schémas de codage adaptés au corpus utilisé, dans le choix des étiquettes, des dimensions abstraites, des dimensions contextuelles, et des indices multimodaux.
Une des premières expériences menées avec des annotations libres a permis d'évaluer la difficulté de l'annotation des émotions dans des contextes naturels. La perception des émotions est subjective, dépendante du contexte et des différences individuelles (notamment les différences de perception entre hommes et femmes), l'analyse des annotations a montré des différences de perception entre groupes de sujets : variations entre gens de sexes, d'âges, ou d'expertises différents. Nous avons pu ainsi montrer des différences de perception entre hommes et femmes ce qui est en accord avec les résultats de la littérature.
Les différences notées entre la perception des émotions des hommes et des femmes sontpar exemple le choix du label : les hommes ont donné plus d'importance à la colère et les femmes au désespoir ; et les indices multimodaux :les hommes ont détecté plus de froncement de sourcils et les femmes ont perçu les modifications dans la direction du regard.

A quels niveaux d'abstraction représenter une émotion?

Nos expériences ont permis la **définition de plusieurs niveaux de représentation** des émotions et des paramètres comportementaux multimodaux apportant des informations pertinentes pour la perception de ces émotions complexes spontanées.

Trois niveaux d'abstraction ont été explorés dans cette thèse : l'émotion elle-même, le comportement multimodal associé, et le contexte émotionnel. Aux deux niveaux, multimodal et émotion, principalement étudiés et pour lesquels nous avons défini des guides et schémas d'annotation distincts, nous avons ajouté le niveau contextuel.

Comment représenter le contexte d'une émotion ?

Pour représenter le contexte, nous avons intégré la notion d'événement émotionnel au niveau global du schéma de codage émotion. Nous avons défini ces événements avec les dimensions « d'appraisal » proposées par (Scherer, 2000).

A quel niveau temporel effectuer l'annotation?

Le niveau temporel d'annotation choisi dépend de l'objectif visé, et du niveau de précision voulu, sachant que plus le niveau d'annotation sera bas, plus le temps nécessaire à l'annotation d'une vidéo sera élevé. Pour une génération fidèle de comportement par un ACA, un niveau précis (comportements observés à des instants précis) sera choisi. Un niveau temporel global à la vidéo permettrait par exemple de catégoriser plus rapidement des vidéos (e.g. regrouper les extraits contenant beaucoup de gestes, contenant l'émotion : tristesse, etc.).

Quels sont les paramètres des comportements multimodaux pertinents pour la perception des émotions?

Nous avons défini un schéma **d'annotation des expressions multimodales** des émotions adapté à nos données. Nous avons étudié la perception de l'expressivité des mouvements dans des données naturelles, alors que la plupart des études similaires sont effectuées sur des données actées enregistrées en laboratoire. Nous avons observé des corrélations entre quatre des six dimensions d'expressivité annotées par nos sujets: l'extension spatiale, l'activation globale, la répétition, et la force. Nous avons collecté des informations sur la personnalité des sujets à l'aide du questionnaire EPI (Eysenck Personality Inventory). Plusieurs analyses en composantes principales (ACP) sur les annotations des sujets ont montré des différences d'annotation des paramètres expressifs entre les groupes de sujets « extravertis » et les « introvertis », les extravertis ayant annoté les paramètres expressifs avec des valeurs plus basses que les introvertis. Le guide d'annotation des comportements multimodaux présenté en annexe contient la liste finale des paramètres conservés après la série d'expériences effectuée

durant nos travaux. Les expressions faciales ont un fort impact sur la perception des émotions mais sont également difficiles à annoter, particulièrement dans des vidéos de résolution télé contenant des comportements spontanés. La possibilité d'annoter les mouvements de sourcils, des yeux, du nez et de la bouche a été ajoutée au schéma de codage, et nous avons simplifié l'annotation de ces paramètres en limitant le nombre de positions possibles pour chaque modalité. Notre schéma final contient également les paramètres qui pourraient avoir un impact sur certaines caractéristiques de l'émotion (ex : intensité), tel que la phase des gestes.

Quelles sont les relations entre les émotions et les comportements multimodaux associés ?
L'analyse de ces relations nécessite des données conséquentes. L'approche exploratoire que nous avons choisie nous a permis de collecter des annotations à plusieurs niveaux (multimodal, émotions) sur un corpus varié mais de petite taille, mais n'est pas suffisante pour une étude statistique des relations entre émotions et comportements multimodaux. Nous avons défini la notion de profil émotionnel pour chaque vidéo, incluant les vecteurs d'émotions et les paramètres d'expressivité du mouvement. Des expériences différentes effectuées sur les mêmes vidéos ont révélé des similarités entre vecteurs d'émotions et ont montré l'uniformité des annotations, validant les annotations et le schéma de codage. Nos différentes expériences ont permis de lister les mesures pertinentes suivantes entre ces deux niveaux pour de futurs travaux avec plus d'annotations : expressivité des gestes et catégorie d'émotion, expressions faciales et émotion, mouvements liés aux « affects burst/disfluences » présents dans le signal vocal, mouvements de tête et classe d'émotion, relation entre mouvements de bras et émotions, type de geste et émotion.

Comment appliquer ce type de connaissance à la spécification d'agents animés expressifs montrant des émotions complexes ?
Nous avons mis en place une **méthode de copie-synthèse** pour étudier les différents niveaux intervenant dans la réalisation multimodale des comportements émotionnels, en utilisant l'annotation des émotions et l'annotation de bas niveau des comportements multimodaux pour permettre l'expression d'émotions complexes dans un ACA. Lors de cette étude, nous avons spécifié des comportements émotionnels complexes dans le système Greta (Pelachaud, 2005) en effectuant un alignement temporel des annotations émotion et multimodale et en définissant les paramètres d'expressivité de Greta à partir des annotations des dimensions abstraites. (Il serait intéressant de considérer les informations provenant des annotations contextuelles dans le modèle de comportement de l'agent. Globalement deux niveaux d'indices sont extraits des annotations ; l'un pour simuler le comportement de l'utilisateur se servant des

annotations sur les comportements émotionnels et des buts communicatifs, l'autre pour synthétiser ce comportement à l'aide d'indices multimodaux).

Comment ces expressions multimodales d'émotions complexes issues des vidéos d'origines ou générées dans des animations d'agent sont-elles perçues par différents utilisateurs ?
Le test perceptif décrit en 3.6.3 apporte des éléments de réponse à cette question : les combinaisons d'émotions sont bien perçues par les sujets dans les animations d'ACAs ; les sujets ont tendance à percevoir une émotion prédominante. Notre approche par analyse/synthèse, particulièrement intéressante à long-terme pour la création d'agents expressifs basés sur des émotions plus complexes que les émotions de base, a permis lors d'une expérience de comparer la perception des comportements émotionnels annotés dans un corpus d'interviews télévisées à celle rejouée par un agent expressif à différents niveaux d'abstraction. Les résultats ont montré que les deux approches de replay peuvent être intéressantes pour différents types de mélanges d'émotions. Des différences homme / femme ont été observée lors d'un test perceptif sur la perception multimodale d'émotions complexes faisant intervenir vidéos et agent animé (l'analyse des formulaires montre par exemple que dans leurs sélections, les femmes se sont déclarées plus confiantes que les hommes).

Nous avons également exploré, en comparant les annotations manuelles des mouvements des modalités torse, mains et tête avec l'estimation automatique des mouvements (mesures d'accord), la façon dont le **traitement automatique de l'image** peut valider les annotations manuelles de mouvements émotionnels dans des interviews télévisées, que ce soit au niveau global des extraits vidéo qu'au niveau de l'annotation individuelle du mouvement.

En résumé, ces expériences apportent trois contributions originales à l'étude des émotions et à la conception des futurs systèmes affectifs. Une première originalité de nos travaux est que nous avons étudié la complexité des émotions présentes dans les interactions naturelles, contrairement aux travaux existant sur des émotions basiques et actées. Une deuxième originalité est d'avoir étudié l'expression des émotions dans plusieurs modalités (gestes, parole, expressions faciales). Enfin, nous avons intégré dans un même cadre d'étude différentes sources de connaissances issues des annotations manuelles, des annotations automatiques, de la génération avec un ACA expressif et des tests perceptifs montrant certaines différences individuelles.

Perspectives

En perspectives, les outils développés durant cette thèse (schémas d'annotation, programmes de mesures, méthode de copie-synthèse, protocoles d'annotation) pourront être utilisés à long terme pour concevoir des modèles utilisables par des systèmes interactifs affectifs capables gérer des expressions multimodales d'émotions spontanées. Pour la détection, il est nécessaire de disposer de grands corpus de données annotés pour pouvoir construire des systèmes de détection à fort pouvoir de généralisation. Ces schémas sont aujourd'hui trop complexes pour être utilisés sur de grands corpus. Par contre, ce type d'analyse menée sur des corpus de taille réduite avec des schémas en profondeur est indispensable si on veut avancer sur la connaissance des indices et contextes caractérisant les émotions et l'importance relative de ces différents paramètres. Pour la synthèse, ces schémas ont prouvé leur utilité. Des études sont en cours sur des schémas dérivés de ceux obtenus dans cette thèse. De plus, cette thèse a un impact important sur des travaux en cours ou sur le point de démarrer : coopération avec les psychologues d'UNIGE pour comparer les données spontanées d'EmoTV et les donnée actées GEMEP ; méthode d'annotation manuelle de gestes expressifs utilisée sur le corpus EmoTABOO d'interactions dyadiques (Zara et al. 2007) ; approche copy-synthèse allant vers une approche plus automatique ; étude des différences individuelles étendue à d'autres traits de personnalité (limitées à l'extra-intraversion dans EmoTV).

Les schémas de codage et les protocoles pourront être utilisés pour annoter d'autres vidéos émotionnelles spontanées. Il pourrait également être intéressant de tester leur pouvoir de généralisation à d'autres données y compris des données actées. En effet, (Hall et Matsumoto 2004) ont observé que des sujets percevaient plusieurs émotions lorsqu'on leur présentait un stimulus censé représenter une émotion primaire actée. Leur application à des données multilingues et multi-culturelles pourra être également envisagé.

D'autres manières de comparer les annotations manuelles et automatiques peuvent être investiguées. Les futurs travaux incluront l'estimation séparée des quantités de mouvement pour les différentes parties du corps (incluant le « tracking » de ces zones) dans le but de gérer les gens se déplaçant dans l'arrière plan, l'extraction automatique des paramètres expressifs des mouvements (tels que l'expansion spatiale), la validation des annotations manuelles de l'activation au niveau des segments émotionnels des vidéos, les relations entre estimation de la quantité de mouvement et phases de geste (préparation, « stroke », « retraction »), l'utilisation de filtres temporels pour améliorer la détection automatique des mouvements, et

enfin l'inclusion de l'annotation du torse dans l'union des mouvements annotés uniquement si une zone de peau est visible au niveau du torse.

Enfin, d'autres niveaux d'annotations multimodales annotés comme pertinents dans les vidéos (par exemple les mouvements de tête) peuvent aussi être considérés dans la génération du comportement de l'agent afin d'envisager différents niveaux de fidélité du comportement de l'agent par rapport à la vidéo d'origine. Nous avons l'intention d'améliorer le rendu des animations des émotions basiques et la méthode pour les comparer avec les approches de « replay ». Certaines étapes de la méthode de copie-synthèse qui sont actuellement manuelles pourront être automatisées afin de pouvoir passer plus rapidement d'une vidéo à l'animation de l'agent expressif lors d'étapes de validation des schémas d'annotation et des spécifications de l'agent. La fiabilité des annotations manuelles de la dynamique temporelle que nous avons effectuées et leur utilisation pour la génération d'animation pourra également être évaluée. Enfin la méthode de copie-synthèse pourra être adaptée à d'autres comportements que les émotions.

De façon général, ce champ de recherche est en plein évolution. Nos travaux ont permis de montrer l'existence d'émotions complexes en interaction naturelle, et nous avons apporté des premières pistes pour les représenter. De nombreuses recherches sont cependant encore nécessaires pour doter les systèmes de demain d'émotions complexes, que ce soit en génération ou en détection.

Références bibliographiques

Abrilian, S., Devillers, L. Buisine, S., Martin, J-C. (2005a). EmoTV1: Annotation of Real-life Emotions for the Specification of Multimodal Affective Interfaces. HCI International 2005, Las Vegas, USA.

Abrilian, S., Martin, J-C., Devillers, L. (2005b). A Corpus-Based Approach for the Modeling of Multimodal Emotional Behaviors for the Specification of Embodied Agents. HCI International 2005, Las Vegas, USA.

Allbeck, J. and Badler, N. (2002). Toward Representing Agent Behaviors Modified by Personality and Emotion. Proceedings of the workshop on Embodied conversational agents – let's specify and evaluate them! In conjunction with The First International Joint Conference on Autonomous Agents & Multi-Agent Systems. July 16, Bologna

Allwood, J., Cerrato, L., Dybkær, L., & Paggio, P. (2004). The MUMIN multimodal coding scheme. Proceedings of Workshop on Multimodal Corpora and Annotation, 21-22 June.

Arafa, Y., Kamyab, K., Mamdani, E., Kshirsagar, S., Magnenat-Thalsmann, N., Guye-Vuillème, A., Thalmann, D. (2002). Two approaches to Scripting Character Animation. Proceedings of the workshop on Embodied conversational agents – let's specify and evaluate them! In conjunction with The First International Joint Conference on Autonomous Agents & Multi-Agent Systems. July 16, Bologna Italy.

Argyle, M. (2004). Bodily communication. Second edition. London and New York, Routledge. Taylor & Francis.

Atifi, H., & Marcoccia, M. (2001). L'expression et la mise en scène des émotions à la télévision : l'articulation des émotions vécues, racontées et attribuées. Proceedings of Oralité et gestualité (ORAGE 2001) : Interactions et comportements multimodaux dans la communication, 18-22 juin, pp. 179-182.

Aubergé V., Audibert N., Rilliard A., "E-Wiz: A Trapper Protocol for Hunting the Expressive Speech Corpora in Lab", 4th LREC, 2004, p. 179-182.

Banse, R. and Scherer, K. (1996). Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology, 70(3):614–636.

Bänziger, T., Pirker, H., & Scherer, K. (2006). Gemep - geneva multimodal emotion portrayals: a corpus for the study of multimodal emotional expressions. In L. Deviller et al. (Ed.), Proceedings of LREC'06 Workshop on Corpora for Research on Emotion and Affect (pp. 15-19). Genoa. Italy.

Barakat, R.A. (1973). Arabic Gestures. Journal of Popular Culture 6(4): 749-892.

Baron-Cohen, S., Riviere, A., Fukushima, M., French, D., Hadwin, J., Cross, P., Bryant, C. and Sotillo, M. (1996). "Reading the Mind in the Face: A Cross-cultural and Developmental Study". Visual Cognition 3: 39–59.

Batliner, A & al. Desperately seeking emotions or: actor, wizards, and human beings, ISCA ITRW Speech and Emotion, 2000.

Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Arcy, S., Russel, M., & Wong, M. (2004). "you stupid tin box" - children interacting with the aibo robot: a cross-linguistic

emotional speech corpus. Proceedings of the 4th International Conference of Language Resources and Evaluation (LREC 2004) (pp. 171-174).

Becker, C., Leßmann, N., Kopp, S. and Wachsmuth, I. (2006). Connecting feelings and thoughts - modeling the interaction of emotion and cognition in embodied agents. Seventh International Conference on Cognitive Modeling (ICCM-06).

Bevacqua, E., Mancini, M., Niewiadomski, R. and Pelachaud, C. (2007). An expressive ECA showing complex emotions. AISB'07, Newcastle, UK.

Bickmore, T. et Cassell, J. (2005). Social Dialogue with Embodied Conversational Agents. Advances in Natural, Multimodal Dialogue Systems. New York, Kluwer Academic.

Bitti, P.E.R., and Poggi, I. (1991). Symbolic Nonverbal Behavior: Talking Through Gestures. In Feldman, R.S., and Rimé, B., eds., Fundamentals of Nonverbal Behavior. New York: Cambridge University Press. 433-457.

Boersma, P. and Weenink, D. "Praat: doing phonetics by computer." http://www.fon.hum.uva.nl/praat/.

Boone, R. T., & Cunningham, J. G. (1996). The Attribution of Emotion to Expressive Body Movements: A Structural Cue Analysis.

Boone, R. T., Cunningham, J. G. (1998). Children's decoding of emotion in expressive body move-ment: The development of cue attunement. Developmental Psychology 34 5.

Braga, D., Coelho, L., Teixeira, J. P.; Freitas, D. R. ProGmatica: a Prosodic and Pragmatic Database for European Portuguese. LREC 2006.

Brown, S. Lisetti, C. and Marpaung, A. (2002). Cherry, the Little Red Robot with a Mission and a Personality. In *Working Notes of the AAAI Fall Symposium Series on Human-Robot Interaction*, Menlo Park, CA: AAAI Press. Cape Cod, MA, November.

Brugman, H. et Russell, A. (2004). Annotating Multimedia/Multi-modal resources with ELAN. In: Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation.

Buisine, S. and Martin, J. C. (accepted). "The effects of speech-gesture cooperation in animated agents' behavior in multimedia presentations." International Journal "Interacting with Computers: The interdisciplinary journal of Human-Computer Interaction". Elsevier.

Caffi C. et Janney R. W. (1994). Toward a pragmatics of emotive communication. Journal of Pragmatics 22. Pp. 325-373.

Campbell, N. (2002). Recording techniques for capturing natural everyday speech. LREC, Las Palmas, 2002.

Campbell, N. (2004). "What do people hear? - a study of the perception of non-verbal affective information in conversational speech", Journal of the Phonetic Society of Japan, pp.9-28, V.8,N.1, 2004.

Cañamero L. and Blanchard A. and Nadel J. (2006). Attachment Bonds for Human-Like Robots International Journal of Humanoïd Robotics. Special issue of the Journal of Humanoid Robotics on "Achieving Human-like Qualities in Interactive Virtual and Physical Humanoids". Eds: C. Pelachaud, L. Canamero. 3(3).

Camurri, A., Lagerlöf, I. and Volpe, G. (2003). "Recognizing Emotion from Dance Movement: Comparison of Spectator Recognition and Automated Techniques." International Journal of Human-Computer Studies 59(1-2): 213-225.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics, 22(2), 249–254.

Cassell, J., Bickmore, J., Billinghurst, M., Campbell, L., Chang, K., Vilhjalmsson, H., and Yan, H. (1999). Embodiment in conversational interfaces: Rea, CHI'99, Pittsburgh, PA, pp. 520-527.

Cassell, J., et al.: Embodied Conversational Agents. 2000: MIT Press.

Cassell, J., Vilhjálmsson, H., Bickmore, T. (2001). BEAT: the Behavior Expression Animation Toolkit. Proceedings of SIGGRAPH '01, pp. 477-486. August 12-17, Los Angeles, CA.

Cassell, J., Bickmore, T. (2001). "A Relational Agent: A Model and Implementation of Building User Trust." Proceedings of the CHI'01 Conference, pp. 396-403. March 31-April 5, Seattle, Washington.

Cavazza, M., Charles, F., and Mead, S.J., 2001. Agents' Interaction in Virtual Storytelling. Intelligent Virtual Agents, Madrid, Spain.

Clavel, C., Vasilescu, I., Devillers, L., Ehrette, T., Richard, G. (2006). “Fear-type emotions of the SAFE Corpus : annotation issues”, LREC 2006.

Collier, G. (1985). Emotional expression, Lawrence Erlbaum Associates. 0-89859-505-3 http://faculty.uccb.ns.ca/~gcollier/

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M. and Schröder, M. (2000). 'FEELTRACE’: An Instrument for Recording Perceived Emotion in Real Time. ISCA Workshop on Speech & Emotion. Northern Ireland: 19-24.

Craggs, R. and Wood, M. (2004). A 2 dimensional annotation scheme for emotion in dialogue. In Proceedings of AAAI Spring Symposium on Exploring Attitude and A ect in Text: Theories and Applications, Stanford University.

Cronbach, L.J., Coefficient alpha and the internal structure of tests. In Psychometrika (pp. 16, 297-334), 1951.

Darwin, C. (1872). The Expression of Emotion in Man and Animals. Murray, London (reprinted by University of Chicago Press, 1975).

Davitz, J.R., and al. (1964). The Communication of Emotional Meaning. New York, McGraw-Hill, Ed.: 2nd, Pages: vii-214 pp

De Carolis, B., Pelachaud, C., Poggi, I., and Steedman, M. (2004). APML, a Mark-up Language for Believable Behavior Generation . In H. Prendinger, ed., "Life- like Characters. Tools, Affective Functions and Applications", 65- - 85. Berlin: Springer.

De Silva, P. R., Kleinsmith, A., & Bianchi-Berthouze, N. (2005). Towards unsupervised detection of affective body posture nuances. 1st International Conference on Affective Computing and Intelligent Interaction (ACII'2005), Beijing, China, 22-24 october.

DeMeijer, M. (1991). The attribution of agression and grief to body movements : the effect of sex-stereotypes. European Journal of Social Psychology 21.

Devillers, L. (2006) Les émotions dans les interactions Homme-Machine : perception, détection et génération. Habilitation à Diriger des Recherches en Informatique de l’Université Paris XI – Orsay, soutenue le 4 décembre 2006.

Devillers, L., Vasilescu, I., Vidrascu, L. (2004). Anger versus Fear detection in recorded conversations', Speech Prosody, Japon, mars.

Devillers, L., Vasilescu, I. (2004). Reliability of Lexical and Prosodic cues in two real-life spoken dialog corpora, LREC, Lisbonne, mai.

Devillers, L. Vidrascu, L. Lamel, L. (2005). “Emotion detection in real-life spoken dialogs recorded in call center”, Journal of Neural Networks, numéro spécial “Emotion and Brain”, volume 18, numéro 4, 407-422, mai 2005.

Devillers, L., Martin, J.-C., Cowie, R., Douglas-Cowie, E. and Batliner, A. (2006). Workshop "Corpora for research on emotion and affect". 5th International Conference on Language Resources and Evaluation (LREC'2006), Genova, Italy. http://www.limsi.fr/Individu/martin/tmp/LREC2006/WS-Emotion/LREC06-WSemotion-proceeding-12.pdf

Douglas-Cowie, E., Cowie, R., Schröeder, M. A New Emotion Database: Considerations, Sources and Scope. Proceedings of the ISCA Workshop on Speech and Emotion, Belfast 2000.

Douglas-Cowie, E., Campbell, N., Cowie, R., and Roach, P. (2003). Emotional speech: Towards a new generation of databases. Speech Communication Special Issue Speech and Emotion, 40(1–2):33–60.

Duncan, S. (1983). Studies of Structure and Individual Differences. In Wiemann, J.M., and Harrison, R.P., eds., Nonverbal Interaction. London : Sage. 149-177.

Efron, D. (1941). Gesture, race and culture, Mouton, Paris.

Ekman, P. and Friesen, W. (1969). The Repertoire of Nonverbal Behavior: Categories, Origins, and Coding. Semiotica 1: 49-98.

Ekman, P. and Friesen W. V. (1972). Hand movements. Journal of Communication, 22, 353-374.

Ekman, P. (1975). The universal smile: face muscles talk in every language, Psychology Today, September 34: 35–9.

Ekman, P., Friesen, W. V., & Scherer, K. (1976). Body movement and voice pitch in deceptive interaction. B Semiotica.16(1), 23-27.

Ekman, P. and Friesen, W. V. (1978). Facial action coding system: A technique for the measurement of facial movement. Palo Alto, Calif.: Consulting Psychologists Press.

Ekman P. (1979). Human ethology: Claims and limits of a new discipline: contributions to the Colloquium, In: M. von Cranach, K. Foppa, W. Lepenies and D. Ploog (Eds.): About brows: Emotional and conversational signals, pp. 169-248. Cambridge University Press, Cambridge, England; New-York.

Ekman, P. (1999). Basic emotions. In Dalgleish, T. and Power, M. J. eds., Handbook of Cognition & Emotion. John Wiley, New York, pp. 301–320.

Ekman, P., Friesen, W.V., & Hager, J.C. (2002). THE FACIAL ACTION CODING SYSTEM Second edition. Salt Lake City: Research Nexus eBook. London: Weidenfeld & Nicolson (world).

Ekman, P. (2003). Emotions revealed. Understanding faces and feelings. Weidenfeld.

Elliott, C., Rickel, J., and Lester, J. (1999). Lifelike pedagogical agents and affective computing: An exploratory synthesis. In Wooldridge, M. and Veloso, M., editors, Artificial Intelligence Today, pages 195-212. Springer-Verlag, Berlin.

Engle, R. A. (1998). Not channels but composite signals: Speech, gesture, diagrams and object demonstrations are integrated in multimodal explanations. In M. A.

Gernsbacher & S. J. Derry (Eds.), Proceedings of the Twentieth Annual Conference of the Cognitive Science Society (pp. 321–326). Mahwah, NJ: Erlbaum.

Feyereisen, P. (1991). La compréhension d'expressions émotionnelles après lésions hémisphériques droites et gauches, Revue Internationale de Psychopathologie, 4, 417-433.

Freedman, N. (1977). Hands, Words and Mind: On the Structuralization of Body Movements During Discourse and the Capacity for Verbal Representation. In Freedman, N., and Grand, S., eds., Communicative Structures. A Psychoanalytic Interpretation of Communication. New York: Plenum Press. 219-235.

Freud, S. (1901). The psychopathology of everyday life. In The Standard Edition of the Complete Works of Sigmund Freud Volume 6, Hogarth, London.

Friesen, W., and Ekman, P. (1984). EMFACS-7: Emotional Facial Action Coding System, Version 7. Unpublished.

Gratch, J., Marsella, S., Egges, A., Eliëns, A., Isbister, K., Paiva, A., Rist, T., & ten Hagen, P. (2004). Design criteria, techniques and case studies for creating and evaluating interactive experiences for virtual humans. Working group on ECA's design parameters and aspects, Dagstuhl seminar on Evaluating Embodied Conversational Agents (organized by Z. Ruttkay, E. André, K. Höök, W. Lewis Johnson, C. Pelachaud).

Gunes, H., & Piccardi, M. (2005). Fusing Face and Body Display for Bi-modal Emotion Recognition: Single Frame Analysis and Multi-Frame Post Integration. 1st International Conference on Affective Computing and Intelligent Interaction (ACII'2005), Beijing, China, 22-24 october.

Hall, J. A. and Matsumoto, D. (2004). "Gender differences in judgments of multiple emotions from facial expressions." Emotion 4(2): 201-206.

Harrigan, J. A., Rosenthal, R. and Scherer, K. (2005). The new handbook of methods in nonverbal behavior research, Oxford University Press.

Hartmann, B., Mancini, M., Pelachaud, C. (2002). Formational Parameters and Adaptive Prototype Instantiation for MPEG-4 Compliant Gesture Synthesis. Computer Animation 11.

Heylen, D.K., Reidsma, D., & Ordelman, R.J. (2006). Annotating state of mind in meeting data. In L. Devillers and J.C. Martin and R. Cowie and A. Batliner (Ed.), Proc. of the LREC2006 Workshop on Corpora for Research on Emotion and Affect (pp. 84–170). Genoa, Italy.

Hildebrandt, B. (1963): Die arithmetische Bestimmung der durativen Funktion. Eine neue Methode der Lautdauerbewertung. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* 14, p. 328-3368. Lindblom,

Izard, C. E. (1979). The maximally discriminative facial movement coding system (MAX). Unpublished manuscript. Available from Instructional Resource Center, University of Delaware, Newark, Delaware.

Izard, C. E., Dougherty, L. M, & Hembree, E. A. (1983). A system for identifying affect expressions by holistic judgements (Affex). Newark: University of Delaware, Instructional Resources Center.

Johannessen, J., Bondi, J., Hagen, K., Priestley, J., Nygaard, L. A speech corpus with emotions. LREC 2006, Workshop for corpora of emotions and affect; 23.05.2006 - 26.05.2006.

Johnson, H.G., Ekman, P., and Friesen, W.V. (1975). Communicative Body Movements: American Emblems. Semiotica 15:335-353.

Kapur, A., Virji-Babul, N., Tzanetakis, G., & Driessen, P. F. (2005). Gesture-Based Affective Computing on Motion Capture Data. 1st International Conference on Affective Computing and Intelligent Interaction (ACII'2005), Beijing, China, 22-24 october.

Kendon, A. (1983). Gesture and Speech. How They Interact. In Wiemann, J.M., and Harrison, R.P., eds., Nonverbal Interaction. London: Sage. 13-45.

Kendon, A. (1988). How Gestures Can Become Like Words. In F. Poyatos (ed.), Cross-Cultural Perspectives in Nonverbal Communication, pp. 131-141. Toronto: Hogrefe.

Kendon, Adam. (2004). Gesture: Visible action as utterance. Cambridge: Cambridge University Press.

Kienast, M., Sendlmeier, W.F. (2000). Acoustical analysis of spectral and temporal changes in emotional speech. In: Proc. ISCA ITRW on Speech and Emotion, Newcastle, 5-7 September 2000, Belfast, Textflow, pp. 92-97

Kita, S., Van Gijn, I., and Van Der Hulst, H. (1998). Movement Phases in Signs and Co-Speech Gestures, and their Transcription by Human Coders. In Wachsmuth, I., and Fröhlich, M., eds., Gesture and Sign Language in Human-Computer Interaction, 23-35. Berlin: Springer.

Kipp, M. (2001), "Anvil - A Generic Annotation Tool for Multimodal Dialogue". In Eurospeech'2001.

Kipp, M. (2003). "Gesture Generation by Imitation - From Human Behavior to Computer Character Animation". PhD Thesis, Saarland University, December 2003.

Kipp, M. (2004), "Gesture Generation by Imitation - From Human Behavior to Computer Character Animation", Boca Raton, Florida: Dissertation.com.

Knapp, M.L. and Hall, J.A. (2006). Nonverbal Communication in Human Interaction. 6th Edition. Thomson/Wadsworth.

Kopp, S., Jung, B., Lessmann, N., Wachsmuth, I. (2003). Max - A Multimodal Assistant in Virtual Reality Construction. KI-Küstliche Intelligenz 4/03, pp 11-17, Bremen: arenDTap Verlag.

Kopp, S. and Wachsmuth, I. (2004). "Synthesizing Multimodal Utterances for Conversational Agents." Journal of Computer Animation and Virtual Worlds 15(1): 39-52.

Kopp, S., Gesellensetter, L., Krämer, N. and Wachsmuth, I. (2005). A conversational agent as museum guide -- design and evaluation of a real-world application. Intelligent Virtual Agents, Berlin: Springer-Verlag.

Kopp, S., Krenn, B., Marsella, S., Marshall, A., Pelachaud, C., Pirker, H., Thórisson, K. and Vilhjalmsson, H. (2006). Towards a Common Framework for Multimodal Generation: The Behavior Markup Language. 6th International Conference on Intelligent Virtual Agents (IVA'06), Marina del Rey, CA, Springer.

Laskowski, K. and Burger, S. (2006). "Annotation and Analysis of Emotionally Relevant Behavior in the ISL Meeting Corpus". Fifth International Conference on Language Resources and Evaluation (LREC2006), Genoa, Italy, 24-26 May.

Lechenadec, G., Maffiolo, V., Chateau, N. and Colletta, J. (2006). Creation of a corpus of multimodal spontaneous expressions of emotions in Human-Machine Interaction. 5th International Conference on Language Resources and Evaluation (LREC'2006), Genova, Italy, 24-26 May.38-42.

Lee, B., Kao, E. and Soo, V. (2006). Feeling Ambivalent: A Model of Mixed Emotions for Virtual Agents. 6th International Conference on Intelligent Virtual Agents (IVA'06), Marina del Rey, CA, Springer.

Lester, J., Voerman, J., Towns, S., & Callaway, C. (1997). Cosmo: A life-like Animated Pedagogical Agent with deictic believability. Proceedings of IJCAI '97 Workshop on Animated Interface Agents: Making Them Intelligent, pp. 61-69.

Lester, J., Zettlemoyer, L., Grégoire, J., & Bares, W. (1999). Explanatory lifelike avatars: Performing user-centered tasks in 3D learning environments. Proceedings of International Conference on Autonomous Agents, pp. 24-31.

Lester, J., Towns, S., Callaway, C., Voerman, J., and FitzGerald, P. Deictic and Emotive Communication in Animated Pedagogical Agents, In Embodied Conversational Agents, J. Cassell, J. Sullivan, et al, Eds, Cambridge: MIT Press, 2000.

Lisetti, C. and Marpaung, A. (2006). BDI+E Framework: an affective cognitive modeling for autonomous agents based on Scherer's emotion theory. 1rst Workshop on Emotion and Computing at KI'2006, 29th Annual Conference on Artificial Intelligence, June, 14-19, Bremen, Germany.

Lisetti, C. L. and Paleari, M. (accepté). "Facial Expression Generation for Embodied Conversational Agents based on Psychological Component Process Theory (CPT): Lessons Learned." International Journal on Multimodal Interfaces. Springer.

Lochbaum, K. (1998). A Collaborative Planning Model of Intentional Structure, Computational Linguistics, vol. 24, pp. 525-572.

Maestri, G. (1999). Digital Character Animation 2. Volume 1 – Essential Techniques. New Riders publishing.

Magno Caldognetto, E., Poggi, I., Cosi, P., Cavicchio, F., Merola, G. (2004). Multimodal Score: an Anvil Based Annotation Scheme for Multimodal Audio-Video Analysis. Workshop "Multimodal Corpora: Models Of Human Behaviour For The Specification And Evaluation Of Multimodal Input And Output Interfaces" In Association with the 4th International Con-ference On Language Resources And Evaluation (LREC'2004) Lisbon, Portugal.

Martin, J. C. (2006). Multimodal Human-Computer Interfaces and Individual Differences. Annotation, perception, representation and generation of situated multimodal behaviors. Habilitation à diriger des recherches en Informatique. Université Paris XI. 6th december 2006.

Martin, J.-C., Kuhnlein, P., Paggio, P., Stiefelhagen, R. and Pianesi, F. (2006). Workshop "Multimodal Corpora: from Multimodal Behaviour Theories to Usable Models ". In Association with the 5th International Conference on Language Resources and Evaluation (LREC2006), Genoa, Italy. http://www.limsi.fr/Individu/martin/tmp/LREC2006/WS-MM/final/proceedings-WS-MultimodalCorpora-v3.pdf

Maya, V., Lamolle, M., Pelachaud, C. (2004). Influences on Embodied Conversational Agent's Expressivity: Toward an Individualization of the ECAs. AISB 2004

convention. Symposium on Language, Speech and Gesture for Expressive Characters 75-85.

McClave, E. (1994). Gestural Beats: The Rhythm Hypothesis. Journal of Psycholinguistic Research 23(1): 45-65.

McNeill, D. and Levy, E. (1982). Conceptual Representations in Language Activity and Gesture. In R.J. Jarvella & W. Klein (eds.), Speech, Place, and Action, pp 271-296. Chichester, Eng: John Wiley & Sons.

McNeill, D. (1992). Hand and mind - what gestures reveal about thoughts: University of Chicago Press.

McNeill, D. (2005). Gesture and thought. Chicago: University of Chicago Press.

Montepare, J., Koff, E., Zaitchik, D., Albert, M. (1999). The use of body movements and gestures as cues to emotions in younger and older adults. Journal of Nonverbal Behavior 23 2.

Mozzinconacci, S. Speech Variability and Emotion, Production and Perception. CIP-Data Library Technische Universiteit Eindhoven (1998).

Newlove, J. (1993). Laban for actors and dancers. Routledge New York.

Niewiadomski, R. (2007). A model of complex facial expressions in interpersonal relations for animated agents, PhD. dissertation, University of Perugia.

Ortony, A., Clore, G., Collins, A. (1988). The cognitive structure of emotions. Cambridge University Press, Cambridge, England.

Osgood, C.E.; May, W.H. & Miron, M.S. (1975) Cross-Cultural Universals of affective Meaning. Urbana: University of Illinois Press.

Ostermann, J. (1998). "Animation of synthetic faces in MPEG-4", in Computer Animation'98, Philadelphia, USA, June, 49–51.

Pelachaud, C.: Multimodal expressive embodied conversational agent. ACM Multimedia, Brave New Topics session (2005) Singapore 683 – 689

Pesty, S., et Sansonnet, J-P. (2006). GT ACA : Groupe de Travail sur les Agents Conversationnels Animés. 15ème Congrès Francophone AFRIF-AFIA, Reconnaissance des Formes et Intelligence Artificielle (RFIA 2006), 25-27 janvier.

Pianesi, F., Zancanaro, M. and Leonardi, C. (2006). Multimodal Annotated Corpora of Consensus Decision Making Meetings Workshop "Multimodal Corpora: from Multimodal Behaviour Theories to Usable Models " in Association with the 5th International Conference on Language Resources and Evaluation (LREC2006), Genoa, Italy. http://www.limsi.fr/Individu/martin/tmp/LREC2006/WS-MM/final/proceedings-WS-MultimodalCorpora-v3.pdf

Piwek, P., Krenn, B., Schröder, M., Grice, M., Baumann, S., Pirker, H. (2002). RRL : A Rich Representation Language for the Description of Agent Behaviour in NECA. Proceedings of the workshop on Embodied conversational agents – let's specify and evaluate them! In conjunction with The First International Joint Conference on Autonomous Agents & Multi-Agent Systems. July 16, Bologna Italy.

Plutchik, R. The psychology and Biology of Emotion, HarperCollins College, New York, 1994.

Poggi, I. and Pelachaud, C. (2000). Emotional meaning and expression in animated faces. In: A. Paiva (Ed): Affective Interactions. Springer, New York.

Poggi I., Pelachaud C., DeRosis F. (2000). "Eye communication in a conversational 3D synthetic agent", AI Communications. Special Issue on Behavior Planning for Life-Like Characters and Avatars., vol. 13, n° 3, p. 169-181.

Poggi, I. (2001). Mind markers. In C.Mueller and R.Posner, editors, The Semantics and Pragmatics of Everyday Gestures. Berlin Verlag Arno Spitz, Berlin, 2001 (in press).

Raouzaiou, A., Tsapatsoulis, N., Karpouzis K. and Kollias S. (2002). Parameterized facial expression synthesis based on MPEG-4, EURASIP Journal on Applied Signal Processing, 10:1021-1038.

Rehm, M. and André, E. (2005). Catch Me If You Can - Exploring Lying Agents in Social Settings. Int. Conf. Autonomous Agents and Multiagent Systems (AAMAS'2005), Utrecht, the Netherlands.

Reidsma, D., Heylen, D.K., & Ordelman, R.J. (2006). Annotating emotions in meetings. Proc. of the fifth international conference on Language Resources and Evaluation, LREC'2006 (pp. 1117–2238). Genoa, Italy.

Rickel, J., and Johnson, W.L. (1999). Virtual Humans for Team Training in Virtual Reality, in Proceedings of the Ninth International Conference on AI in Education, pp. 578-585, July, IOS Press.

Roesch, E.B., Fontaine, J.R., Scherer, K.R. (2006). The world of emotions is two-dimensional .. or is it?. Presentation at the 3rd HUMAINE Summer School, Genova (September 21-28, 2006). pdf – http://emotion-research.net/ws/summerschool3/

Rosenberg and Binkowski (2005). "Augmenting the kappa statistic to determine interannotator reliability for multiple labelled data points", HLT-NAACL.

Sander, D., Grandjean, D. and Scherer, K. (2005). "A systems approach to appraisal mechanisms in emotion." Neural Networks 18: 317-352.

Saratxaga I, Navas E., Hernaez I., Luengo I. (2006). Designing and Recording an Emotional Speech Database for Corpus Based Synthesis in Basque. Proceedings of the LREC 2006, pp 2126-2129.

Scherer, K. R. (1981). Speech and emotional states. In J. Darby (Ed.), Speech evaluation in psychiatry (pp. 189-220). New York: Grune & Stratton.

Scherer, K. R. and Ekman, P., eds. (1984). Approaches to Emotion, Erlbaum, Hillsdale, NJ.

Scherer, K. R. (1986). Voice, stress, and emotion. In M. H. Appley, & R. Trumbull, (Eds.), Dynamics of stress (pp. 159- 181). New York: Plenum.

Scherer, K. R. (2000). Emotion. In M.H.W. Stroebe (Ed.), Introduction to Social Psychology: A European perspective, (pp. 151-191): Oxford: Blackwell.

Scherer, K. R., Schorr, A. and Johnstone, T., eds. (2001). Appraisal Processes in Emotion. New York: Oxford University Press.

Scherer, K. R.and Ellgring, H.(2007)."Multimodal Expression of Emotion:Affect Programs or Componential Appraisal Patterns?" Emotion 7 (1), http://content.apa.org/journals/emo

Smith, C. A., Ellsworth, P.C. (1985). "Patterns of cognitive appraisal in emotion." Journal of Personality and Social Psychology. 48. pp. 813 – 838.

Towns, S., FitzGerald, P., & Lester, J. (1998). Visual emotive communication in lifelike pedagogical agents. Proceedings of ITS'98, pp. 474-483.

Vogt, T., Andre, E. (2006), "Improving Automatic Emotion Recognition from Speech via Gender Differentiation". LREC 2006, Genoa, Italy.

Wallbott, H.G. (1998). Bodily expression of emotion. Europ. Journal of Social Psychology 28.

Webb, R. (1997). Linguistic Properties of Metaphoric Gestures. New York: UMI.

Weiner, B., Russell, D., Lerman, D. (1979). "The cognition-emotion process in achievement-related contexts. Journal of Personality and Social Psychology, 37, pp. 1211 – 1220.

Wiggins, J.S. (1996). The Five-Factor Model of Personality: Theoretical Perspectives. New York: The Guilford Press.

Williams, U. and Stevens, K.N. (1972). Emotions and speech: some acoustical correlates. JASA:52,1238–1250.

Zara, A., Maffiolo, V., Martin, J. C. and Devillers, L. (2007). Collection and Annotation of a Corpus of Human-Human Multimodal Interactions. 2nd International Conference on Affective Computing and Intelligent Interaction (ACII 2007), Lisbon, Portugal.

Annexes

Annexe 1 : Aperçu du Corpus EmoTV

Annexe 2 : Guide d'Annotation des Émotions

Annexe 3 : Guide d'Annotation des Comportements Multimodaux

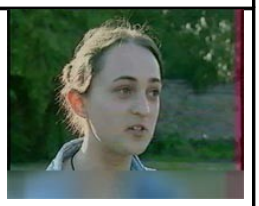
Annexe 4 : Schéma de Codage Émotions/Contexte/Parole

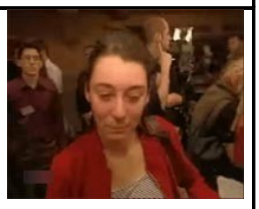
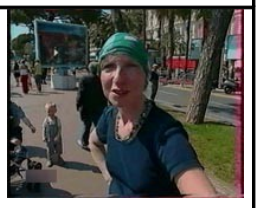
Annexe 5 : Schéma de Codage Multimodal

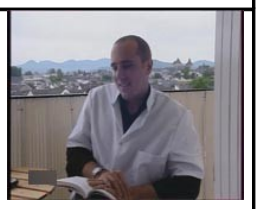
Annexe 6 : Tables de correspondances EmoTV/Greta

Annexe 7 : Publications effectuées durant la thèse

Annexe 1 : Aperçu du Corpus EmoTV

			
ext-01	ext-02	ext-03	ext-04
			
ext-05	ext-06	ext-07	ext-09
			
ext-10	ext-12	ext-13	ext-14
			
ext-15	ext-16	ext-17	ext-18
			
ext-19	ext-20	ext-21	ext-22
			
ext-23	ext-24	ext-25	ext-26

			
ext-27	ext-28	ext-29	ext-30
			
ext-31	ext-35	ext-36	ext-37
			
ext-38	ext-39	ext-40	ext-41
			
ext-42	ext-44	ext-45	ext-46
			
ext-47	ext-48	ext-49	ext-50
			
ext-51	ext-52	ext-53	ext-54
			
ext-55	ext-60	ext-61	ext-62

			
ext-63	ext-64	ext-65	ext-66
			
ext-67	ext-68	ext-69	ext-70
			
ext-71	ext-72	ext-73	ext-74
			
ext-75	ext-76	ext-77	ext-78
			
ext-79	ext-80	ext-81	ext-82
			
ext-83	ext-84	ext-85	ext-86
			
ext-87	ext-88	ext-89	ext-90

ext-91

ext-92

ext-93

ext-94

ext-95

ext-96

ext-97

ext-98

ext-99

Annexe 2 : Guide d'Annotation des Émotions

Les sujets ayant effectué nos expériences d'annotation des émotions avaient à leur disposition un guide. Ce guide, complété lors des différentes expériences d'annotation, apparaît ci-dessous dans sa version finale en anglais (pour sa mise à disposition aux partenaires européens du projet HUMAINE). Il présente les instructions pour l'annotation, mais apporte également des détails et définitions pour les variables à annoter.

1. Annotation instructions

This guide is for expert labellers.
Read this document before starting to annotate.
At the beginning of the annotation, launch a timer. At the end of the annotation, record the time it took to annotate this video in a file.
The annotation protocol consists of several sequential steps:

- *Clip level annotation: global emotion, global multimodal cues and context annotations*
- *Segment level annotation: full-time frontier segmentation*

You can replay the video and come back at any time.
Save frequently your annotation.

2. Global annotation

Watch entirely the video sample at least 3 times.
Annotate the following dimensions giving information on the context. Add comments on emotion labelling if necessary in the specified field.
For annotating each dimension, in the Anvil main frame, go in "View", then "Sets", and then click on corresponding dimension. A window will open, click on the "add" button. Another window will open with corresponding attributes.

2.1. Annotation of perceived emotional event(s)

The scheme permits to annotate up to 3 emotional events, each of them contains following attribute:

Appraisal is the process by which an organism assesses its overall relationship with its environment including not only its current condition but past events that led to this state as well as future prospects. Cognitive appraisal theory argues that an organism may possess many distributed processes for interpreting this relationship (e.g., planning, explanation, perception, etc.) but that appraisal maps characteristics of these disparate processes into a common set of intermediate terms (intermediate between stimuli and response, between organism and environment) called appraisal variables (Marsella, 2003).

Scherer appraisal dimensions (Sherer 2001) have been studied in emotion recall experiments and have been used to predict major modal emotions.

We use these dimensions in emotion perception investigation for describing complex blended emotional behaviour of the video-taped person. Here, the appraisal variables also depend on the real context which we are able to detect in the interview video-taped.

Temporality

- *Past (> 1 week)*
- *Close past (< 1 week)*
- *Present (today)*
- *Close future (< 1 week)*
- *Future (> 1 week)*

Duration

- *Monthes and more*
- *Weeks*
- *Days*
- *Hours*
- *Minutes*

Appraisal variables

Each variable can be annotated with following values, in addition to the values specified for each of them:

- *Normal*
- *Not relevant*
- *Impossible to annotate*

Relevance

Is the person's emotional state determined by his/her impressions of these aspects of some critical event?

- *Suddenness of the event (sensory-motor level)*
    - *Sudden*
    - *Unsudden*
- *Familiarity with the event (schematic level)*
    - *Familiar*
    - *Unfamiliar*
- *Predictability of the event (conceptual level)*
    - *Predictable*
    - *Unpredictable*
- *Pleasantness*
    - *Pleasant*
    - *Unpleasant*

Implication/Consequences

Is the person's emotional state determined by his/her impressions about these aspects of some critical event?

- *Cause agent*
    - *Self*
    - *Other character*
    - *Nature*
    - *Organisation*
    - *Other*
    - *Don't know*
- *Cause agent intention*
    - *Intentional*
    - *Negligence*
- *Outcome probability*
    - *Probable*
    - *Unprobable*
- *Relation to expectation*
    - *Consonant*
    - *Dissonant*
    - *No expectation*
- *Conduciveness to goals*
    - *Conducive*
    - *Obstructive*
- *Urgency of a response*
    - *Urgent*

- *Not urgent*

Coping potential

These items relate to how well the video-taped person feels he/she can cope with the event or events that are central to his/her emotional state, including their ability to reverse negative or maintain positive circumstances. Is the person's emotional state determined by his/her impressions about these aspects of the critical event?

- *Controllability of direct event*
  - *Controlable*
  - *Uncontrolable*
- *Controllability of consequence event*
  - *Controlable*
  - *Uncontrolable*
- *Power of person*
  - *High*
  - *Low*
- *Possible adjustment to person's own goals*
  - *Adjustable*
  - *Not adjustable*

Compatibility with standards

Significance of the event with respect to video-taped person self concept and social norms and values. Is the person's emotional state determined by his/her impressions about these aspects of the critical event?

- *External (norms or demands of a reference group)*
    - *Compatible*
    - *Uncompatible*
- *Internal (self ideal or internalized moral code)*
    - *Compatible*
    - *Uncompatible*

Consequence event of annotated event

- *Consequence event*
    - *Event1 (if annotated event is not event1)*
    - *Event2 (if annotated event is not event2)*
    - *Event3 (if annotated event is not event3)*

2.2. Global annotation of perceived emotion(s)

The scheme permits to annotate up to 5 global emotions.

Label(s) selection

Select up to 5 labels among the following list that we defined:

- *Anger*
- *Despair*
- *Disappointment*
- *Disgust*

- *Doubt*
- *Embarassment*
- *Exaltation*
- *Fear*
- *Irritation*
- *Joy*
- *Neutral*
- *Pleased*
- *Pride*
- *Sadness*
- *Serenity*
- *Shame*
- *Surprise*
- *Worry*

New label

If the label that you want to annotate is not the list above, write it here.

Information for each label and for the whole video

- *Acted-level*

From 1 (spontaneous/genuine) to 5 (acted)

- Masked-level

From 1 (spontaneous/genuine) to 5 (hidden)

– Intensity dimension

Estimate the intensity of perceived emotion.

1 (low), 2, 3 (normal), 4, 5 (strong)

– Valence dimension

The valence dimension is a global measure of the positive or negative feeling associated with the emotion. Estimate the valence of perceived emotion.

-2 (negative), -1, 0 (neutral), +1, +2 (positive)

– Activation dimension

The activation dimension measures the degree to which the emotion inspires action in humans. Estimate the activation of perceived emotion.

-2 (passive), -1, 0 (normal), +1, +2 (active)

– Control dimension

The control dimension measures the degree to which the character controls his emotion Estimate the control on perceived emotion.

1 (low), 2, 3 (normal), 4, 5 (high)

– Approach/avoid dimension

The dimension concerns the person behaviour with the emotional event/situation

-2 (withdrawn), -1, 0 (normal), +1, +2 (engaged)

Modalities involved in your perception

Select among the list below, the modality(ies) showing perceived emotion.

- *Speech, Torso, Head, Shoulders, Gaze, Brows, Mouth, Gestures, Other*

Missing modalities

Select in the same list, the modalities not visible in the clip, due to camera position or external elements (character's face blurred, body parts hidden by objects or people)

3. Segment level annotation

Segmentation rules :

- *Full-time frontiers*

- *Coherence between emotion and perception*

3.1. Speech Annotation

For the speech transcription, the coders have to annotate verbal sounds but also nonverbal sounds, as hesitations, breaths, etc..., listed in the Linguistic Data Consortium (LDC). Most annotated tags are [laugh], [cries], [b] (breath), [pff]. We suggest to do the transcription using Praat[4]. The transcription can be done either at the level of the utterance or at the level of the words. It depends on how much a fine-grained analysis of the synchrony between speech and other modalities is required.

Transliteration

Words and nonverbal sounds. Special marks signified by squared brackets for the mark-up of nonverbal sounds ([cries], [breath], [pff] etc...).

Sentence

This track contains the sentence, and includes translation in english if needed

Speech turn

This track contains speech turns, each speaker is identified by a number

Pitch

Curve exported from Praat

[4] http://www.fon.hum.uva.nl/praat/

Intensity

Curve exported from Praat

Rhythm

- *Allongement début, Allongement milieu, Allongement final, Rapide, Lent, Irrégulier, Autre rythme*

Quality

- *Tremolo*
- *Grinçante*
- *Criée*
- *Chuchotée*
- *Nasiarde*
- *Hyperarticulée*
- *Dure*
- *Timbrée*
- *Détimbrée*
- *Soufflée*
- *Pleasantness*
    - *Etendue : relaxation du conduit oral, plus d'énergie dans les basses fréquences*
    - *Etroite : tension du conduit vocal, plus d'énergie en hautes fréquences*
    - *Intermédiaire*
- *Relaxation*
    - *Détendue : relaxation de l'appareil vocal, F0 bas, peu d'amplitude*
    - *Tendue : tension de l'appareil vocal, F0 et amplitudes élevées*
    - *Intermédiaire*
- *Control*

- *Tendue : tension de l'appareil vocal, augmentation de F0 et d'amplitude*
- *Relâchée : hypotension de la musculature, F0 bas et peu d'amplitude*
- *Intermédiaire*

– *Power*

- *Pleine : registre de poitrine, F0 bas et beaucoup d'énergie*
- *Fine : registre de tête, F0 haut et peu d'énergie*
- *Intermédiaire*

Energy

– *Forte, Faible, Irrégulière, Normale*

Intonation

– *Haute, Basse, Grand saut, Irrégulière, Grande variabilité, Peu de variabilité*

Lexical/semantic valence

– *Negative, Positive, Neutral*

Emphasis word

Affect burst

– *Rire, Toux, Respiration, Cris, Larmes, Soufflés, Interjection, Hésitations, Reniflements*

Timing

– *Disruptive pausing, Too long pauses, Too frequent pauses, Short pauses, Slow rate, Fast rate*

3.2. Emotion Annotation

Specification of one track for each emotion label (there are at most 5 such tracks). For all the segments of the "emotion1" track: right click on the segment, "play element" at least 3 times. Then right click on that segment and "edit" it. A window will open, fill following attributes. The definitions of attributes are nearly the same as for global annotation of emotion. The segmentation will enable to compute the temporal relations (sequentiality / simultaneity of two segments). If you perceive several emotions in a segment, fill corresponding segments below emotion1 track (emotion2… emotion5 track). If the segment is annotated as "not-

visible", meaning that the subject is not visible on that part of the video, don't fill the other attributes of this segment.

- *Label (s) selection*
- *Seven dimensions (abstract, with scales) for each label of the combination and for the combination:*
    - *Acted*
    - *Masked*
    - *Intensity*
    - *Valence*
    - *Activation*
    - *Control*
    - *Approach/avoid*
- *"Missing modalities"*

4. Final validation

Replay the video.
Check the global annotation.
Check each segment annotation.
Store the time it took to annotate this video in a file.
In case of bug, save your work, close the annotation and reopen it.
Be prepared to explain your annotations with other annotators.

Annexe 3 : Guide d'Annotation des Comportements Multimodaux

Comme pour l'annotation manuelle des émotions, nous avons écrit un guide d'annotation des comportements multimodaux expliquant comment annoter les comportements multimodaux apparaissant dans le corpus EmoTV. Ci-dessous la version finale du guide, en anglais, comprenant les instructions d'annotation ainsi que les définitions des dimensions.

1. Annotation instructions

This document describes how to annotate multimodal behaviours occurring during non basic emotional behaviour in the EmoTV Corpus. The corpus and the goal of emotional and multimodal annotation are described in (Abrilian et al. 2005a), (Abrilian et al. 2005b).
Read this document before starting to annotate. When starting to annotate a video, launch a timer. At the end of the annotation, record the time it took you to annotate this video in an Excel file "annotation timing.xls".

- *Play the video once.*
- *Annotate each track one by one (hiding the other tracks in Anvil makes things clearer), focusing at corresponding body part in the video.*
- *Always annotate a movement direction relatively to the previous position of the moving body part.*
- *Mute the sound before annotating the behaviours others than the mouth and gesture.*
- *Annotate speed of movements relatively to this video.*
- *Annotate open mouth when the subject does not speak.*
- *Replay and check each track once.*

You can replay the video and modify your annotation at any time.
Save your annotation frequently.
You should be confident enough in your annotation so that you can explain and justify your annotations during validation with other coders.

2. Global description of the coding scheme

The coding scheme contains the following tracks or groups of tracks:

- *Torso: contains a description of the pose and movement.*
- *Head: contains pose, primary and secondary movement.*
- *Shoulders: contains pose track.*
- *Facial expressions: contains eyes, brows, mouth, chin, and nose.*
- *Gestures: contains 3 groups of tracks (left hand, right hand, both hands). Each group contains 4 track (gesture unit, phase, phrase, and expressivity).*

Torso, head and gestures contain a description of the pose, and of the movement. Pose and movement annotations alternate.

We need to be able to compute for each movement: its duration, its starting position, its ending position and the movement direction and quality. Starting and ending position can be found in the pose track. The direction of movement, its type (rotation vs. translation) and the angles can be computed from the annotations in the pose track. Thus the remaining required attributes are those related to movement quality. Duration of movement can be computed by the start and end of each segment in the movement quality track.

When the video starts by a movement, a short pose annotation is inserted at the beginning.

The annotator has to annotate multimodal behaviours that occur during emotional segments. Each modality and track should be annotated one after the other (e.g. the annotator starts by annotating the 1st track for the whole video and then proceed to the next track). Each attribute list of value contains an “other” value and a comment text field.

Currently, the hand shape is not annotated.

The fluidity annotation is necessary for isolated gestures, more especially as some of them are repetitive.

Most of the tracks contain an **asymmetry attribute** (shoulders, arms, eyes…) and a “none” or “other” value.

3. Torso

Torso movements are not so complex that head movements. Thus the direction of torso movements will be computed from the poses. For instance, if a pose has the value “Full twist right (-90)” and the next pose has the value “Front (0)”, we will compute that the movement was directed to the left.

3.1. Pose

Annotate the pose of the torso relatively to the camera.

Three dimensions / attributes: twist, side-side, bend.

Annotating absolute value of angles would be too long to annotate. A small set of approximating values is proposed (-90, -67, -45, -22, 0, +22, +45, +67, +90) along with labels.

Twist

Front (0)		Back (180)	
Full twist right (-90)		Full twist left (+90)	
High twist right (-67)		High twist left (+67)	
Half twist right (-45)		Half twist left (+45)	
Low twist right (-22)		Low twist left (+22)	

Side-side

Side right (+20)	
Front (0)	
Side left (-20)	

Bend

Bend front (+20)	
Front (0)	
Bend back (-20)	

4. Head

Head movements are numerous and sometimes combined (e.g. head nod and head turn right at the same time). That is the reason why the information about movement direction cannot be computed from pose as it is the case for torso.

4.1. Pose

Annotate the position and movements of the head relatively to the torso.

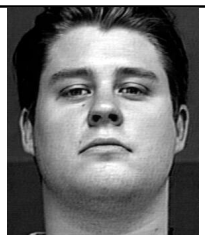
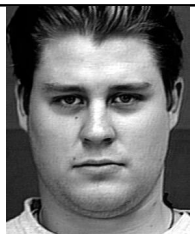
Primary position

Most of the following values and images come from FACS.

http://face-and-emotion.com/dataface/facs/description.jsp

Secondary position

In case of combination of positions (e.g. head to the right and down). Same table as the primary position attribute.

	FACS	Poser
Toward interlocutor		
Front		
Tilt left		
Tilt right		
Forward		
Backward		
Upward		
Downward		

4.2. Movement

Multiple complex movements can be grouped into a single annotation if needed.

Primary movement

Movement observed between the start and the end pose. Same table as the primary position attribute, except that the "front" value is not present in the movement list.

Secondary movement

Additional movement that enables the combination of several movements (e.g. head nod while turning the head). It has the same values as the primary movement. There is no priority between primary and secondary movements; if a combination of 2 head movements occurs, they will both be as relevant for the generation of emotional behaviour in a conversational agent.

5. Shoulders

5.1. Pose

Position

- *Normal / Upward / Downward / Forward / Backward*

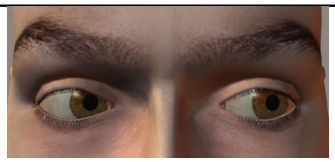
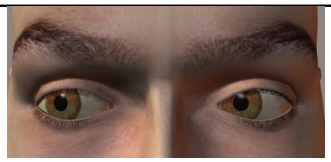
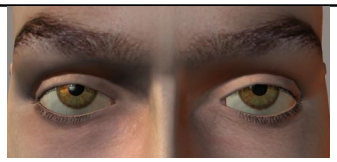
6. Facial Expressions

6.1. Eyes

Annotate the gaze direction relatively to the head.

The gaze direction is annotated but not the transitions which are very fast.

Movement

	FACS	Poser
Toward interlocutor		
Front		
Turn left (AU61)		
Turn right (AU62)		
Up (AU63)		
Down (AU64)		
Closed (AU43)		
Upper lid raiser (AU5)		
Squint (AU44)		

Needed to annotate combination of movements, as Turn left + Upper lid raiser.

6.2. Brows

Movement

Inner brow raiser (AU1)	
Outer brow raiser (AU2)	
Brow lowerer (AU4)	
Brow raiser (AU1 + AU2)	
Brow raiser and lowerer (AU1 + AU2 + AU4)	
Other	
None	

Intensity

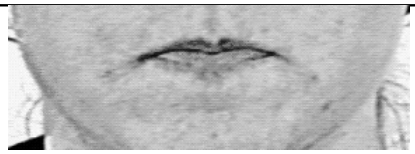
- *Low / Normal / High*

Mouth

Only exaggerated mouth movements are annotated in this track.

Primary movement

Upper lip raiser (AU10)	
Lip corner puller (AU12)	
Lip corner depressor (AU15)	
Lower lip depressor (AU16)	
Lip stretcher (AU20)	
Lip funneler (AU22)	
Jaw drop (AU26)	
Mouth stretch (AU27)	

Lips part	
Lip pressor	
Cheek puffer	
Lip tightener	

Secondary movement

Needed to annotate combination of mouth movements, as Mouth stretch + Lip corner puller.

6.3. Chin

Movement

Chin raiser (AU17)	

6.4. Nose

Movement

Nose wrinkler (AU9)	

Gestures

Three groups of tracks: “right hand”, “left hand”, and “both hands”. Each group contains four tracks (Gesture unit / Phase / Phrase / Movement) for each hand.

6.5. Gesture unit

Do the segmentation. A segment can be a single gesture or a succession of several gestures, the end of a gesture-unit is characterized by the hands going back to the rest position.
Annotate expressivity for these segments (see “Expressivity” track below for the description of parameters).

6.6. Phase

Type

Preparation	Bringing arm and hand into stroke position, note that changing hand shape before/after moving the arm belongs to the preparation
Stroke	The most energetic part of the gesture
SequenceOfStroke	A number of successive strokes; all strokes should be covered by this phase
Hold	A phase of stillness just before or just after the stroke usually used to defer the stroke so that it coincides with a certain word
Retract	Movement back to rest position; in sitting position this is usually the arm rest, the lap or folded arms

6.7. Phrase

Category

The gestures that are expected to be most frequent are listed first:

Manipulator	Contact with self or an object
Beat	Formless gesture shape
Deictic	Arm/hand is used to point at an existing or imaginary object
Illustrator	Movement which accompany, illustrate, or accentuate the verbal content
Emblem	Movement with a precise, culturally defined meaning, like the eye-wink

6.8. Expressivity

Repetition

- *Value between 1 and 10*

Fluidity

- *Smooth / Normal / Jerky*

Strength

- *Soft / Normal / Hard*

Speed

- *Slow / Normal / Fast*

Spatial expansion

- *Contracted / Normal / Expanded*

Annexe 4 : Schéma de Codage Émotions/Contexte/Parole

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-- ************************************************************** -->
<!-- ********** Scheme for coding emotional behaviors ************** -->
<!-- **** Sarkis Abrilian, Laurence Devillers, Jean-Claude Martin *** -->
<!-- ********************** LIMSI-CNRS ************************** -->
<!-- ************************************************************** -->
<!-- ******************** Anvil tool (Kipp 2004) ******************** -->
<!-- ************************************************************** -->

<!-- **************last modifications: 13 March 2007 *************** -->
<annotation-spec>

 <!-- ****************************** DEFINITIONS *************************** -->
 <head>
  <valuetype-def>

        <!-- ********************** durationType ********************** -->
        <valueset name="durationType">
          <value-el>monthes and more</value-el>
          <value-el>weeks</value-el>
          <value-el>days</value-el>
          <value-el>hours</value-el>
          <value-el>minutes</value-el>
        </valueset>

        <!-- ********************** suddennessType ********************** -->
        <valueset name="suddennessType">
                <value-el>sudden</value-el>
                <value-el>unsudden</value-el>
                <value-el>normal</value-el>
                <value-el>not relevant</value-el>
                <value-el>impossible</value-el>
        </valueset>

        <!-- ********************** familiarityType ********************** -->
        <valueset name="familiarityType">
                <value-el>familiar</value-el>
                <value-el>unfamiliar</value-el>
                <value-el>normal</value-el>
                <value-el>not relevant</value-el>
                <value-el>impossible</value-el>
        </valueset>

        <!-- ********************** predictabilityType ********************** -->
        <valueset name="predictabilityType">
                <value-el>predictable</value-el>
                <value-el>unpredictable</value-el>
                <value-el>normal</value-el>
                <value-el>not relevant</value-el>
                <value-el>impossible</value-el>
        </valueset>
```

```
<!-- ********************** pleasantnessType ********************** -->
<valueset name="pleasantnessType">
        <value-el>pleasant</value-el>
        <value-el>unpleasant</value-el>
        <value-el>normal</value-el>
        <value-el>not relevant</value-el>
        <value-el>impossible</value-el>
</valueset>

<!-- ********************** causeAgentType ********************** -->
<valueset name="causeAgentType">
        <value-el>self</value-el>
        <value-el>other character</value-el>
        <value-el>nature</value-el>
        <value-el>organisation</value-el>
        <value-el>other</value-el>
        <value-el>don't know</value-el>
        <value-el>not relevant</value-el>
        <value-el>impossible</value-el>
</valueset>

<!-- ********************** causeAgentIntentionType ********************** -->
<valueset name="causeAgentIntentionType">
        <value-el>intentional</value-el>
        <value-el>negligence</value-el>
        <value-el>normal</value-el>
        <value-el>not relevant</value-el>
        <value-el>impossible</value-el>
</valueset>

<!-- ********************** outcomeProbabilityType ********************** -->
<valueset name="outcomeProbabilityType">
        <value-el>probable</value-el>
        <value-el>unprobable</value-el>
        <value-el>normal</value-el>
        <value-el>not relevant</value-el>
        <value-el>impossible</value-el>
</valueset>

<!-- ********************** relationToExpectationType ********************** -->
<valueset name="relationToExpectationType">
<value-el>consonant</value-el>
        <value-el>dissonant</value-el>
   <value-el>no expectation</value-el>
        <value-el>normal</value-el>
        <value-el>not relevant</value-el>
        <value-el>impossible</value-el>
</valueset>

<!-- ********************** conducivenessType ********************** -->
<valueset name="conducivenessType">
        <value-el>conducive</value-el>
        <value-el>obstructive</value-el>
        <value-el>normal</value-el>
```

```
        <value-el>not relevant</value-el>
        <value-el>impossible</value-el>
</valueset>

<!-- ********************** urgencyType ********************** -->
<valueset name="urgencyType">
        <value-el>urgent</value-el>
        <value-el>not urgent</value-el>
        <value-el>normal</value-el>
        <value-el>not relevant</value-el>
        <value-el>impossible</value-el>
</valueset>

<!-- ********************** controlabilityOfDirectEventType ********************** -->
<valueset name="controlabilityOfDirectEventType">
        <value-el>controlable</value-el>
        <value-el>uncontrolable</value-el>
        <value-el>normal</value-el>
        <value-el>not relevant</value-el>
        <value-el>impossible</value-el>
</valueset>

<!-- ********************** controlabilityOfConsequenceEventType ********************** -->
<valueset name="controlabilityOfConsequenceEventType">
        <value-el>controlable</value-el>
        <value-el>uncontrolable</value-el>
        <value-el>normal</value-el>
        <value-el>not relevant</value-el>
        <value-el>impossible</value-el>
</valueset>

<!-- ********************** powerOfPersonType ********************** -->
<valueset name="powerOfPersonType">
        <value-el>high</value-el>
        <value-el>low</value-el>
        <value-el>normal</value-el>
        <value-el>not relevant</value-el>
        <value-el>impossible</value-el>
</valueset>

<!-- ********************** possibleAdjustmentType ********************** -->
<valueset name="possibleAdjustmentType">
        <value-el>adjustable</value-el>
        <value-el>not adjustable</value-el>
        <value-el>normal</value-el>
        <value-el>not relevant</value-el>
        <value-el>impossible</value-el>
</valueset>

<!-- ********************** compatibilityExternalType ********************** -->
<valueset name="compatibilityExternalType">
        <value-el>compatible</value-el>
        <value-el>uncompatible</value-el>
        <value-el>normal</value-el>
        <value-el>not relevant</value-el>
```

```
            <value-el>impossible</value-el>
        </valueset>

        <!-- ********************** compatibilityInternalType ********************** -->
        <valueset name="compatibilityInternalType">
            <value-el>compatible</value-el>
            <value-el>uncompatible</value-el>
            <value-el>normal</value-el>
            <value-el>not relevant</value-el>
            <value-el>impossible</value-el>
        </valueset>

        <!-- ********** emotionType ************** -->
        <!-- 18 following emotions selected for this corpus (Abrilian et al., LREC 2006) -->
        <valueset name="emotionType">
          <value-el>anger</value-el>
          <value-el>despair</value-el>
          <value-el>disappointment</value-el>
          <value-el>disgust</value-el>
          <value-el>doubt</value-el>
          <value-el>embarassment</value-el>
          <value-el>exaltation</value-el>
          <value-el>fear</value-el>
          <value-el>irritation</value-el>
          <value-el>joy</value-el>
          <value-el>neutral</value-el>
          <value-el>pleased</value-el>
          <value-el>pride</value-el>
          <value-el>sadness</value-el>
          <value-el>serenity</value-el>
          <value-el>shame</value-el>
          <value-el>surprise</value-el>
          <value-el>worry</value-el>
        </valueset>

        <!-- ********************** 1to5Type ********************** -->
        <valueset name="1to5Type">
          <value-el>1</value-el>
          <value-el>2</value-el>
          <value-el>3</value-el>
          <value-el>4</value-el>
          <value-el>5</value-el>
        </valueset>

<!-- ********************** -2to+2Type ********************** -->
        <valueset name="-2to+2Type">
          <value-el>-2</value-el>
          <value-el>-1</value-el>
          <value-el>0</value-el>
          <value-el>+1</value-el>
          <value-el>+2</value-el>
        </valueset>

  </valuetype-def>
 </head>
```

```
<!-- ************************ BODY ******************** -->
<body>

  <!-- ********************** emotional events ************************ -->
    <set-spec name="event1">
        <attribute name="event" valuetype="String" display="true" />

        <attribute name="past" valuetype="Boolean" display="true" />
        <attribute name="close past" valuetype="Boolean" display="true" />
        <attribute name="present" valuetype="Boolean" display="true" />
        <attribute name="close future" valuetype="Boolean" display="true" />
        <attribute name="future" valuetype="Boolean" display="true" />

        <attribute name="duration" valuetype="durationType" display="true" />

        <attribute name="r.suddenness" valuetype="suddennessType" display="true" />
        <attribute name="r.familiarity" valuetype="familiarityType" display="true" />
        <attribute name="r.predictability" valuetype="predictabilityType" display="true" />
        <attribute name="r.pleasantness" valuetype="pleasantnessType" display="true" />
        <attribute name="ic.cause.agent" valuetype="causeAgentType" display="true" />

        <attribute name="ic.cause.agent.intention" valuetype="causeAgentIntentionType"
                display="true" />
        <attribute name="ic.outcome probability" valuetype="outcomeProbabilityType"
                display="true" />
        <attribute name="ic.relation to expectation"
                valuetype="relationToExpectationType" display="true" />
        <attribute name="ic.conduciveness to goals"
                valuetype="conducivenessType" display="true" />
        <attribute name="ic.urgency of a response" valuetype="urgencyType"
                display="true" />
        <attribute name="cp.controllability of direct event"
                valuetype="controlabilityOfDirectEventType" display="true" />
        <attribute name="cp.controllability of consequence event"
                valuetype="controlabilityOfConsequenceEventType" display="true" />
        <attribute name="cp.power of person"
                valuetype="powerOfPersonType" display="true" />
        <attribute name="cp.possible adjustment to person's own goals"
                valuetype="possibleAdjustmentType" display="true" />
        <attribute name="cws.external"
                valuetype="compatibilityExternalType" display="true" />
        <attribute name="cws.internal"
                valuetype="compatibilityInternalType" display="true" />

        <attribute name="consequence" display="true">
                <value-el>event2</value-el>
                <value-el>event3</value-el>
        </attribute>
   </set-spec>

  <set-spec name="event2">
        <attribute name="event" valuetype="String" display="true" />
        <attribute name="past" valuetype="Boolean" display="true" />
        <attribute name="close past" valuetype="Boolean" display="true" />
```

```
    <attribute name="present" valuetype="Boolean" display="true" />
    <attribute name="close future" valuetype="Boolean" display="true" />
    <attribute name="future" valuetype="Boolean" display="true" />

    <attribute name="duration" valuetype="durationType" display="true" />

    <attribute name="r.suddenness" valuetype="suddennessType" display="true" />
    <attribute name="r.familiarity" valuetype="familiarityType" display="true" />
    <attribute name="r.predictability" valuetype="predictabilityType" display="true" />
    <attribute name="r.pleasantness" valuetype="pleasantnessType" display="true" />
    <attribute name="ic.cause.agent" valuetype="causeAgentType" display="true" />
    <attribute name="ic.cause.agent.intention" valuetype="causeAgentIntentionType"
            display="true" />
    <attribute name="ic.outcome probability"
            valuetype="outcomeProbabilityType" display="true" />
    <attribute name="ic.relation to expectation"
            valuetype="relationToExpectationType" display="true" />
    <attribute name="ic.conduciveness to goals"
            valuetype="conducivenessType" display="true" />
    <attribute name="ic.urgency of a response"
            valuetype="urgencyType" display="true" />
    <attribute name="cp.controllability of direct event"
            valuetype="controlabilityOfDirectEventType" display="true" />
    <attribute name="cp.controllability of consequence event"
             valuetype="controlabilityOfConsequenceEventType" display="true" />
    <attribute name="cp.power of person"
            valuetype="powerOfPersonType" display="true" />
    <attribute name="cp.possible adjustment to person's own goals"
            valuetype="possibleAdjustmentType" display="true" />
    <attribute name="cws.external"
            valuetype="compatibilityExternalType" display="true" />
    <attribute name="cws.internal"
            valuetype="compatibilityInternalType" display="true" />

    <attribute name="consequence" display="true">
            <value-el>event1</value-el>
            <value-el>event3</value-el>
    </attribute>
</set-spec>

<set-spec name="event3">
    <attribute name="event" valuetype="String" display="true" />

    <attribute name="past" valuetype="Boolean" display="true" />
    <attribute name="close past" valuetype="Boolean" display="true" />
    <attribute name="present" valuetype="Boolean" display="true" />
    <attribute name="close future" valuetype="Boolean" display="true" />
    <attribute name="future" valuetype="Boolean" display="true" />

    <attribute name="duration" valuetype="durationType" display="true" />

    <attribute name="r.suddenness" valuetype="suddennessType" display="true" />
    <attribute name="r.familiarity" valuetype="familiarityType" display="true" />
    <attribute name="r.predictability" valuetype="predictabilityType" display="true" />
    <attribute name="r.pleasantness" valuetype="pleasantnessType" display="true" />
    <attribute name="ic.cause.agent" valuetype="causeAgentType" display="true" />
```

```
    <attribute name="ic.cause.agent.intention"valuetype="causeAgentIntentionType"
            display="true" />
    <attribute name="ic.outcome probability" valuetype="outcomeProbabilityType"
            display="true" />
    <attribute name="ic.relation to expectation"
            valuetype="relationToExpectationType" display="true" />
    <attribute name="ic.conduciveness to goals"
            valuetype="conducivenessType" display="true" />
    <attribute name="ic.urgency of a response"
            valuetype="urgencyType" display="true" />
    <attribute name="cp.controllability of direct event"
            valuetype="controlabilityOfDirectEventType" display="true" />
    <attribute name="cp.controllability of consequence event"
           valuetype="controlabilityOfConsequenceEventType" display="true" />
    <attribute name="cp.power of person"
            valuetype="powerOfPersonType" display="true" />
    <attribute name="cp.possible adjustment to person's own goals"
            valuetype="possibleAdjustmentType" display="true" />
    <attribute name="cws.external"
            valuetype="compatibilityExternalType" display="true" />
    <attribute name="cws.internal"
            valuetype="compatibilityInternalType" display="true" />

    <attribute name="consequence" display="true">
            <value-el>event1</value-el>
            <value-el>event2</value-el>
    </attribute>
</set-spec>

    <!-- ********************** global annotation of emotions ********************** -->
    <set-spec name="globalEmotion1">
            <attribute name="label" valuetype="emotionType" display="true" />
            <attribute name="new label" valuetype="String" display="true" />
            <attribute name="acted-level" valuetype="1to5Type" display="true" />
            <attribute name="masked-level" valuetype="1to5Type" display="false" />
            <attribute name="intensity" valuetype="1to5Type" display="true" />
            <attribute name="valence" valuetype="-2to+2Type" display="true" />
            <attribute name="activation" valuetype="1to5Type" display="true" />
            <attribute name="control" valuetype="1to5Type" display="true" />
           <attribute name="approach-avoid" valuetype="-2to+2Type" display="true" />

            <attribute name="speech" valuetype="Boolean" display="true" />
            <attribute name="torso" valuetype="Boolean" display="true" />
            <attribute name="head" valuetype="Boolean" display="true" />
            <attribute name="shoulders" valuetype="Boolean" display="true" />
            <attribute name="gaze" valuetype="Boolean" display="true" />
            <attribute name="brows" valuetype="Boolean" display="true" />
            <attribute name="mouth" valuetype="Boolean" display="true" />
            <attribute name="gestures" valuetype="Boolean" display="true" />
            <attribute name="other" valuetype="Boolean" display="true" />
    </set-spec>

    <set-spec name="globalEmotion2">
            <attribute name="label" valuetype="emotionType" display="true" />
            <attribute name="new label" valuetype="String" display="true" />
```

```
        <attribute name="acted-level" valuetype="1to5Type" display="true" />
        <attribute name="masked-level" valuetype="1to5Type" display="false" />
        <attribute name="intensity" valuetype="1to5Type" display="true" />
        <attribute name="valence" valuetype="-2to+2Type" display="true" />
        <attribute name="activation" valuetype="1to5Type" display="true" />
        <attribute name="control" valuetype="1to5Type" display="true" />
      <attribute name="approach-avoid" valuetype="-2to+2Type" display="true" />

        <attribute name="speech" valuetype="Boolean" display="true" />
        <attribute name="torso" valuetype="Boolean" display="true" />
        <attribute name="head" valuetype="Boolean" display="true" />
        <attribute name="shoulders" valuetype="Boolean" display="true" />
        <attribute name="gaze" valuetype="Boolean" display="true" />
        <attribute name="brows" valuetype="Boolean" display="true" />
        <attribute name="mouth" valuetype="Boolean" display="true" />
        <attribute name="gestures" valuetype="Boolean" display="true" />
        <attribute name="other" valuetype="Boolean" display="true" />
</set-spec>

<set-spec name="globalEmotion3">
        <attribute name="label" valuetype="emotionType" display="true" />
        <attribute name="new label" valuetype="String" display="true" />
        <attribute name="acted-level" valuetype="1to5Type" display="true" />
        <attribute name="masked-level" valuetype="1to5Type" display="false" />
        <attribute name="intensity" valuetype="1to5Type" display="true" />
        <attribute name="valence" valuetype="-2to+2Type" display="true" />
        <attribute name="activation" valuetype="1to5Type" display="true" />
        <attribute name="control" valuetype="1to5Type" display="true" />
      <attribute name="approach-avoid" valuetype="-2to+2Type" display="true" />

        <attribute name="speech" valuetype="Boolean" display="true" />
        <attribute name="torso" valuetype="Boolean" display="true" />
        <attribute name="head" valuetype="Boolean" display="true" />
        <attribute name="shoulders" valuetype="Boolean" display="true" />
        <attribute name="gaze" valuetype="Boolean" display="true" />
        <attribute name="brows" valuetype="Boolean" display="true" />
        <attribute name="mouth" valuetype="Boolean" display="true" />
        <attribute name="gestures" valuetype="Boolean" display="true" />
        <attribute name="other" valuetype="Boolean" display="true" />
</set-spec>

<set-spec name="globalEmotion4">
        <attribute name="label" valuetype="emotionType" display="true" />
        <attribute name="new label" valuetype="String" display="true" />
        <attribute name="acted-level" valuetype="1to5Type" display="true" />
        <attribute name="masked-level" valuetype="1to5Type" display="false" />
        <attribute name="intensity" valuetype="1to5Type" display="true" />
        <attribute name="valence" valuetype="-2to+2Type" display="true" />
        <attribute name="activation" valuetype="1to5Type" display="true" />
        <attribute name="control" valuetype="1to5Type" display="true" />
      <attribute name="approach-avoid" valuetype="-2to+2Type" display="true" />

        <attribute name="speech" valuetype="Boolean" display="true" />
        <attribute name="torso" valuetype="Boolean" display="true" />
        <attribute name="head" valuetype="Boolean" display="true" />
```

```
                <attribute name="shoulders" valuetype="Boolean" display="true" />
                <attribute name="gaze" valuetype="Boolean" display="true" />
                <attribute name="brows" valuetype="Boolean" display="true" />
                <attribute name="mouth" valuetype="Boolean" display="true" />
                <attribute name="gestures" valuetype="Boolean" display="true" />
                <attribute name="other" valuetype="Boolean" display="true" />
        </set-spec>

        <set-spec name="globalEmotion5">
                <attribute name="label" valuetype="emotionType" display="true" />
                <attribute name="new label" valuetype="String" display="true" />
                <attribute name="acted-level" valuetype="1to5Type" display="true" />
                <attribute name="masked-level" valuetype="1to5Type" display="false" />
                <attribute name="intensity" valuetype="1to5Type" display="true" />
                <attribute name="valence" valuetype="-2to+2Type" display="true" />
                <attribute name="activation" valuetype="1to5Type" display="true" />
                <attribute name="control" valuetype="1to5Type" display="true" />
               <attribute name="approach-avoid" valuetype="-2to+2Type" display="true" />

                <attribute name="speech" valuetype="Boolean" display="true" />
                <attribute name="torso" valuetype="Boolean" display="true" />
                <attribute name="head" valuetype="Boolean" display="true" />
                <attribute name="shoulders" valuetype="Boolean" display="true" />
                <attribute name="gaze" valuetype="Boolean" display="true" />
                <attribute name="brows" valuetype="Boolean" display="true" />
                <attribute name="mouth" valuetype="Boolean" display="true" />
                <attribute name="gestures" valuetype="Boolean" display="true" />
                <attribute name="other" valuetype="Boolean" display="true" />
        </set-spec>

        <set-spec name="missing modalities in whole video">
                <attribute name="speech" valuetype="Boolean" display="true" />
                <attribute name="torso" valuetype="Boolean" display="true" />
                <attribute name="head" valuetype="Boolean" display="true" />
                <attribute name="shoulders" valuetype="Boolean" display="true" />
                <attribute name="gaze" valuetype="Boolean" display="true" />
                <attribute name="brows" valuetype="Boolean" display="true" />
                <attribute name="mouth" valuetype="Boolean" display="true" />
                <attribute name="gestures" valuetype="Boolean" display="true" />
                <attribute name="other" valuetype="Boolean" display="true" />
        </set-spec>

<group name="speech">
        <!-- ********************** audio waveform visualisation ********************** -->
        <track-spec name="audio" type="waveform" height="1.5"/>

        <!-- ********************** audio transliteration ********************** -->
        <track-spec name="transliteration" type="primary">
                <doc>
                This track contains the transliteration of the spoken
                discourse. The <b>token</b> attribute contains the word or sound
                uttered.
                </doc>
                <attribute name="token">
                        <doc>
```

```
                Words and nonverbal sounds. Special marks signified by
                squared brackets for the mark-up of nonverbal sounds ([cries],
                [breath], [pff] etc...)
                </doc>
        </attribute>
</track-spec>

<!-- ********************** sentence ********************** -->
<track-spec name="sentence" type="primary">
        <doc>
        This track contains the sentence, and includes translation in english
        </doc>
        <attribute name="token" />
</track-spec>

<!-- ********************** speech turn ********************** -->
<track-spec name="speech turn" type="primary">
        <doc>
        This track contains the speech turns
        </doc>
        <attribute name="speakerId" valuetype="1to5Type"/>
</track-spec>

<!-- ********************** pitch ********************** -->
<track-spec name="pitch" type="speech analysis" />

<!-- ********************** intensity ********************** -->
<track-spec name="intensity" type="speech analysis" />

<!-- ********************** rythm ********************** -->
<track-spec name="rythm" type="primary">
        <attribute name="allongement debut" valuetype="Boolean"/>
        <attribute name="allongement milieu" valuetype="Boolean"/>
        <attribute name="allongement final" valuetype="Boolean"/>
        <attribute name="rapide" valuetype="Boolean"/>
        <attribute name="lent" valuetype="Boolean"/>
        <attribute name="irregulier" valuetype="Boolean"/>
        <attribute name="autre rythme" valuetype="String"/>
</track-spec>

<!-- ********************** quality ********************** -->
<track-spec name="quality" type="primary">
        <attribute name="tremolo" valuetype="Boolean"/>
        <attribute name="grinçante" valuetype="Boolean"/>
        <attribute name="criée" valuetype="Boolean"/>
        <attribute name="chuchotee" valuetype="Boolean"/>
        <attribute name="nasiarde" valuetype="Boolean"/>
        <attribute name="hyperarticulée" valuetype="Boolean"/>
        <attribute name="dure" valuetype="Boolean"/>
        <attribute name="timbrée" valuetype="Boolean"/>
        <attribute name="détimbrée" valuetype="Boolean"/>
        <attribute name="soufflée" valuetype="Boolean"/>

        <attribute name="pleasantness" display="true">
                <value-el>étendue</value-el>
```

```
                <value-el>etroite</value-el>
                <value-el>intermédiaire</value-el>
        </attribute>

        <attribute name="relaxation" display="true">
                <value-el>détendue</value-el>
                <value-el>tendue</value-el>
                <value-el>intermédiaire</value-el>
        </attribute>

        <attribute name="control" display="true">
                <value-el>tendue</value-el>
                <value-el>relâchée</value-el>
                <value-el>intermédiaire</value-el>
        </attribute>

        <attribute name="power" display="true">
                <value-el>pleine</value-el>
                <value-el>fine</value-el>
                <value-el>intermédiaire</value-el>
        </attribute>

</track-spec>

<!-- ********************** energy ********************** -->
<track-spec name="energy" type="primary">
        <attribute name="forte" valuetype="Boolean"/>
        <attribute name="faible" valuetype="Boolean"/>
        <attribute name="irreguliere" valuetype="Boolean"/>
        <attribute name="normale" valuetype="Boolean"/>
</track-spec>

<!-- ********************** intonation ********************** -->
<track-spec name="intonation" type="primary">
        <attribute name="haute" valuetype="Boolean"/>
        <attribute name="basse" valuetype="Boolean"/>
        <attribute name="grand saut" valuetype="Boolean"/>
        <attribute name="irreguliere" valuetype="Boolean"/>
        <attribute name="grande variabilité" valuetype="Boolean"/>
        <attribute name="peu de variabilité" valuetype="Boolean"/>
</track-spec>

<!-- ********************** lexical/semantic valence ********************** -->
<track-spec name="lexical/semantic valence" type="primary">
        <attribute name="negative" valuetype="Boolean"/>
        <attribute name="positive" valuetype="Boolean"/>
        <attribute name="neutral" valuetype="Boolean"/>
</track-spec>

<!-- ********************** emphasis word ********************** -->
<track-spec name="emphasis word" type="primary">
        <attribute name="token" valuetype="String"/>
</track-spec>

<!-- ********************** affect burst ********************** -->
```

```
        <track-spec name="affect burst" type="primary">
                <attribute name="rire" valuetype="Boolean"/>
                <attribute name="toux" valuetype="Boolean"/>
                <attribute name="respiration" valuetype="Boolean"/>
                <attribute name="cris" valuetype="Boolean"/>
                <attribute name="larmes" valuetype="Boolean"/>
                <attribute name="souffles" valuetype="Boolean"/>
                <attribute name="interjection" valuetype="Boolean"/>
                <attribute name="hésitations" valuetype="Boolean"/>
                <attribute name="reniflements" valuetype="Boolean"/>
        </track-spec>

        <!-- ********************** timing ********************** -->
        <track-spec name="timing" type="primary">
                <attribute name="disruptive pausing" valuetype="Boolean"/>
                <attribute name="too long pauses" valuetype="Boolean"/>
                <attribute name="too frequent pauses" valuetype="Boolean"/>
                <attribute name="short pauses" valuetype="Boolean"/>
                <attribute name="slow rate" valuetype="Boolean"/>
                <attribute name="fast rate" valuetype="Boolean"/>
        </track-spec>

</group>

<group name="emotion">
        <!-- ********************** combination type and temporality ********************** -->
        <track-spec name="combination type and temporality" type="primary">
                <attribute name="blend" valuetype="Boolean" display="true" />
                <attribute name="mask" valuetype="Boolean" display="true" />
                <attribute name="conflict" valuetype="Boolean" display="true" />
                <attribute name="ambiguity" valuetype="Boolean" display="true" />

                <attribute name="sequential" valuetype="Boolean" display="true" />
                <attribute name="simultaneity" valuetype="Boolean" display="true" />
        </track-spec>

        <track-spec name="emotion1" type="primary">
                <attribute name="label" valuetype="emotionType" display="true" />
                <attribute name="new label" valuetype="String" display="true" />
                <attribute name="acted-level" valuetype="1to5Type" display="true" />
                <attribute name="masked-level" valuetype="1to5Type" display="false" />
                <attribute name="intensity" valuetype="1to5Type" display="true" />
                <attribute name="valence" valuetype="-2to+2Type" display="true" />
                <attribute name="activation" valuetype="1to5Type" display="true" />
                <attribute name="control" valuetype="1to5Type" display="true" />
               <attribute name="approach-avoid" valuetype="-2to+2Type" display="true" />

                <attribute name="speech" valuetype="Boolean" display="true" />
                <attribute name="torso" valuetype="Boolean" display="true" />
                <attribute name="head" valuetype="Boolean" display="true" />
                <attribute name="shoulders" valuetype="Boolean" display="true" />
                <attribute name="gaze" valuetype="Boolean" display="true" />
                <attribute name="brows" valuetype="Boolean" display="true" />
                <attribute name="mouth" valuetype="Boolean" display="true" />
                <attribute name="gestures" valuetype="Boolean" display="true" />
```

```
        <attribute name="other" valuetype="Boolean" display="true" />
</track-spec>

<track-spec name="emotion2" type="primary">
        <attribute name="label" valuetype="emotionType" display="true" />
        <attribute name="new label" valuetype="String" display="true" />
        <attribute name="acted-level" valuetype="1to5Type" display="true" />
        <attribute name="masked-level" valuetype="1to5Type" display="false" />
        <attribute name="intensity" valuetype="1to5Type" display="true" />
        <attribute name="valence" valuetype="-2to+2Type" display="true" />
        <attribute name="activation" valuetype="1to5Type" display="true" />
        <attribute name="control" valuetype="1to5Type" display="true" />
       <attribute name="approach-avoid" valuetype="-2to+2Type" display="true" />

        <attribute name="speech" valuetype="Boolean" display="true" />
        <attribute name="torso" valuetype="Boolean" display="true" />
        <attribute name="head" valuetype="Boolean" display="true" />
        <attribute name="shoulders" valuetype="Boolean" display="true" />
        <attribute name="gaze" valuetype="Boolean" display="true" />
        <attribute name="brows" valuetype="Boolean" display="true" />
        <attribute name="mouth" valuetype="Boolean" display="true" />
        <attribute name="gestures" valuetype="Boolean" display="true" />
        <attribute name="other" valuetype="Boolean" display="true" />
</track-spec>

<track-spec name="emotion3" type="primary">
        <attribute name="label" valuetype="emotionType" display="true" />
        <attribute name="new label" valuetype="String" display="true" />
        <attribute name="acted-level" valuetype="1to5Type" display="true" />
        <attribute name="masked-level" valuetype="1to5Type" display="false" />
        <attribute name="intensity" valuetype="1to5Type" display="true" />
        <attribute name="valence" valuetype="-2to+2Type" display="true" />
        <attribute name="activation" valuetype="1to5Type" display="true" />
        <attribute name="control" valuetype="1to5Type" display="true" />
       <attribute name="approach-avoid" valuetype="-2to+2Type" display="true" />

        <attribute name="speech" valuetype="Boolean" display="true" />
        <attribute name="torso" valuetype="Boolean" display="true" />
        <attribute name="head" valuetype="Boolean" display="true" />
        <attribute name="shoulders" valuetype="Boolean" display="true" />
        <attribute name="gaze" valuetype="Boolean" display="true" />
        <attribute name="brows" valuetype="Boolean" display="true" />
        <attribute name="mouth" valuetype="Boolean" display="true" />
        <attribute name="gestures" valuetype="Boolean" display="true" />
        <attribute name="other" valuetype="Boolean" display="true" />
</track-spec>

<track-spec name="emotion4" type="primary">
        <attribute name="label" valuetype="emotionType" display="true" />
        <attribute name="new label" valuetype="String" display="true" />
        <attribute name="acted-level" valuetype="1to5Type" display="true" />
        <attribute name="masked-level" valuetype="1to5Type" display="false" />
        <attribute name="intensity" valuetype="1to5Type" display="true" />
        <attribute name="valence" valuetype="-2to+2Type" display="true" />
        <attribute name="activation" valuetype="1to5Type" display="true" />
```

```
                <attribute name="control" valuetype="1to5Type" display="true" />
               <attribute name="approach-avoid" valuetype="-2to+2Type" display="true" />

                <attribute name="speech" valuetype="Boolean" display="true" />
                <attribute name="torso" valuetype="Boolean" display="true" />
                <attribute name="head" valuetype="Boolean" display="true" />
                <attribute name="shoulders" valuetype="Boolean" display="true" />
                <attribute name="gaze" valuetype="Boolean" display="true" />
                <attribute name="brows" valuetype="Boolean" display="true" />
                <attribute name="mouth" valuetype="Boolean" display="true" />
                <attribute name="gestures" valuetype="Boolean" display="true" />
                <attribute name="other" valuetype="Boolean" display="true" />
        </track-spec>

        <track-spec name="emotion5" type="primary">
                <attribute name="label" valuetype="emotionType" display="true" />
                <attribute name="new label" valuetype="String" display="true" />
                <attribute name="acted-level" valuetype="1to5Type" display="true" />
                <attribute name="masked-level" valuetype="1to5Type" display="false" />
                <attribute name="intensity" valuetype="1to5Type" display="true" />
                <attribute name="valence" valuetype="-2to+2Type" display="true" />
                <attribute name="activation" valuetype="1to5Type" display="true" />
                <attribute name="control" valuetype="1to5Type" display="true" />
               <attribute name="approach-avoid" valuetype="-2to+2Type" display="true" />

                <attribute name="speech" valuetype="Boolean" display="true" />
                <attribute name="torso" valuetype="Boolean" display="true" />
                <attribute name="head" valuetype="Boolean" display="true" />
                <attribute name="shoulders" valuetype="Boolean" display="true" />
                <attribute name="gaze" valuetype="Boolean" display="true" />
                <attribute name="brows" valuetype="Boolean" display="true" />
                <attribute name="mouth" valuetype="Boolean" display="true" />
                <attribute name="gestures" valuetype="Boolean" display="true" />
                <attribute name="other" valuetype="Boolean" display="true" />
        </track-spec>
</group>

        <track-spec name="missing modalities" type="primary">
                <attribute name="speech" valuetype="Boolean" display="true" />
                <attribute name="torso" valuetype="Boolean" display="true" />
                <attribute name="head" valuetype="Boolean" display="true" />
                <attribute name="shoulders" valuetype="Boolean" display="true" />
                <attribute name="gaze" valuetype="Boolean" display="true" />
                <attribute name="brows" valuetype="Boolean" display="true" />
                <attribute name="mouth" valuetype="Boolean" display="true" />
                <attribute name="gestures" valuetype="Boolean" display="true" />
                <attribute name="other" valuetype="Boolean" display="true" />
        </track-spec>
</body>
</annotation-spec>
```

Annexe 5 : Schéma de Codage Multimodal

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-- ************************************************************** -->
<!-- ******* Scheme for coding multimodal emotional behavior ******** -->
<!-- **** Sarkis Abrilian, Jean-Claude Martin, Laurence Devillers *** -->
<!-- ********************** LIMSI-CNRS ************************** -->
<!-- ************************************************************** -->
<!-- ******************** Anvil tool (Kipp 2004) ******************** -->
<!-- ************************************************************** -->

<!-- **************last modifications: 1 february 2007 *************** -->
<annotation-spec>

 <!-- ****************************** DEFINITIONS ************************** -->
 <head>
  <valuetype-def>
        <valueset name="fluidityType">
                <value-el color="light yellow">smooth</value-el>
                <value-el color="light green">normal</value-el>
                <value-el color="light red">jerky</value-el>
        </valueset>

        <valueset name="strengthType">
                <value-el color="light yellow">soft</value-el>
                <value-el color="light green">normal</value-el>
                <value-el color="light red">hard</value-el>
        </valueset>

        <valueset name="speedType">
                <value-el color="light yellow">slow</value-el>
                <value-el color="light green">normal</value-el>
                <value-el color="light red">fast</value-el>
        </valueset>

        <valueset name="spatExpType">
                <value-el color="light yellow">contracted</value-el>
                <value-el color="light green">normal</value-el>
                <value-el color="light red">expanded</value-el>
        </valueset>

        <valueset name="1to5Type">
          <value-el>1</value-el>
          <value-el>2</value-el>
          <value-el>3</value-el>
          <value-el>4</value-el>
          <value-el>5</value-el>
        </valueset>

        <valueset name="phaseType">
           <value-el color="light green">
                preparation
                <doc>
```

```
                bringing arm and hand into stroke position, note that changing hand shape
                before/after moving the arm belongs to the preparation, also code position info
                </doc>
            </value-el>
            <value-el color="light red">
                        stroke
                        <doc>
                        the most energetic part of the gesture.
                        </doc>
            </value-el>
            <value-el color="light blue">
                        prestroke hold
                        <doc>
                           a phase of stillness just before the stroke,
                           usually used to defer the stroke so that it coincides with a
                           certain word.
                        </doc>
                 </value-el>
            <value-el color="light blue">
                        poststroke hold
                        <doc>
                           a phase of stillness just after the stroke,
                           usually used to defer the stroke so that it coincides with a
                           certain word.
                        </doc>
                 </value-el>
            <value-el color="light blue">
                        independent hold
                        <doc>
                        a phase of stillness when there is no stroke
                        </doc>
                 </value-el>
                 <value-el color="light yellow">
                        retract
                        <doc>
                        movement back to rest position; in sitting position this is
                        usually the arm rest, the lap or folded     arms.
                        </doc>
            </value-el>
                 <value-el color="light yellow">
                        partial-retract
                        <doc>
                        partial retraction between 2 co-articulated gestures
                        </doc>
            </value-el>
         </valueset>

 </valuetype-def>
</head>

<body>
         <!-- ********************** GLOBAL ANNOTATION ********************** -->
         <set-spec name="global annotation of speech">
                 <attribute name="emotion in content" valuetype="Boolean" display="true" />
                 <attribute name="emotion in prosody" valuetype="Boolean" display="true" />
```

```
        </set-spec>
        <set-spec name="global annotation of torso movements">
                <attribute name="many torso movements" valuetype="Boolean"
                        display="true" />
        </set-spec>
        <set-spec name="global annotation of head movements">
                <attribute name="continuous head movements" valuetype="Boolean"
                        display="true" />
        </set-spec>
        <set-spec name="global annotation of shoulders movements">
                <attribute name="many shoulders movements" valuetype="Boolean"
                        display="true" />
        </set-spec>
        <set-spec name="global annotation of facial expressions">
                <attribute name="many brow movements" valuetype="Boolean"
                        display="true" />
                <attribute name="many mouth movements" valuetype="Boolean"
                        display="true" />
                <attribute name="many gaze movements" valuetype="Boolean"
                        display="true" />
        </set-spec>
        <set-spec name="global annotation of gestures">
                <attribute name="repeated gestures" valuetype="Boolean" display="true" />
                <attribute name="fast gestures" valuetype="Boolean" display="true" />
                <attribute name="slow gestures" valuetype="Boolean" display="true" />
                <attribute name="jerky gestures" valuetype="Boolean" display="true" />
                <attribute name="smooth gestures" valuetype="Boolean" display="true" />
                <attribute name="hard gestures" valuetype="Boolean" display="true" />
                <attribute name="soft gestures" valuetype="Boolean" display="true" />
                <attribute name="contracted gestures" valuetype="Boolean" display="true" />
                <attribute name="expanded gestures" valuetype="Boolean" display="true" />
        </set-spec>

<!-- ************************* SEGMENTAL ANNOTATION ************************** -->

<!-- ************************* SPEECH ************************** -->
<group name="speech">

         <!-- ***************** audio waveform visualisation ***************** -->
         <track-spec name="audio" type="waveform" height="1.5"/>

         <!-- ***************** transliteration ***************** -->
         <track-spec name="transliteration" type="primary">
          <doc>
           contains the transliteration of the spoken discourse; the token
           attribute contains the word or sound uttered.
          </doc>

          <attribute name="token">
           <doc>
            words and nonverbal sounds; special marks signified by squared
            brackets for the mark-up of nonverbal sounds
            ([cries], [b] (breath), [pff] etc...)
           </doc>
          </attribute>
```

```
            <attribute name="backlink" link-color="green"
                  valuetype="ReciprocalLink(gestures.left hand)"/>
           </track-spec>

        <!-- ********************** speech turn ********************** -->
        <track-spec name="speech turn" type="primary">
                  <attribute name="speakerId" valuetype="1to5Type"/>
        </track-spec>

        <!-- *************** audio transliteration in english *************** -->
        <track-spec name="transliteration (english)" type="primary">
           <doc>
            contains the english translation of the transliteration track.
           </doc>
          <attribute name="token" />
         </track-spec>

        <track-spec name="pitch" type="speech analysis" />

        <track-spec name="intensity" type="speech analysis" />
</group>


<!-- ************************ TORSO ************************** -->
<track-spec name="torso" type="primary">
        <attribute name="repetition" emptyvalue="false" defaultvalue="1"
                  display="true" valuetype="Number(1,10)" />

        <attribute name="twist" emptyvalue="false" defaultvalue="front"
                  display="true">
                  <value-el color="light red">full twist right</value-el>
                  <value-el color="light yellow">half twist right</value-el>
                  <value-el color="light green">front</value-el>
                  <value-el color="light yellow">half twist left</value-el>
                  <value-el color="light red">full twist left</value-el>
                  <value-el color="light blue">back</value-el>
        </attribute>

        <attribute name="side-side" emptyvalue="false" defaultvalue="front"
                  display="true">
        <value-el color="light red">side right</value-el>
                  <value-el color="light green">front</value-el>
                  <value-el color="light red">side left</value-el>
        </attribute>

        <attribute name="bend" emptyvalue="false" defaultvalue="front"
                  display="true">
                  <value-el color="light red">bend front</value-el>
                  <value-el color="light green">front</value-el>
                  <value-el color="light red">bend back</value-el>
        </attribute>
</track-spec>

<!-- ************************ HEAD ************************** -->
<track-spec name="head" type="primary">
```

```
        <attribute name="repetition" emptyvalue="false" defaultvalue="1"
                display="false" valuetype="Number(1,10)" />

        <attribute name="toward interlocutor" valuetype="Boolean" display="true" />
        <attribute name="front" valuetype="Boolean" display="true" />
        <attribute name="turn left" valuetype="Boolean" display="true" />
        <attribute name="turn right" valuetype="Boolean" display="true" />
        <attribute name="tilt left" valuetype="Boolean" display="true" />
        <attribute name="tilt right" valuetype="Boolean" display="true" />
        <attribute name="upward" valuetype="Boolean" display="true" />
        <attribute name="downward" valuetype="Boolean" display="true" />
        <attribute name="forward" valuetype="Boolean" display="true" />
        <attribute name="backward" valuetype="Boolean" display="true" />
 </track-spec>

<!-- ************************* SHOULDERS ************************** -->
 <track-spec name="shoulders" type="primary">
        <attribute name="repetition" emptyvalue="false" defaultvalue="1" display="false"
                valuetype="Number(1,10)" />

        <attribute name="normal" valuetype="Boolean" display="true" />
        <attribute name="upward" valuetype="Boolean" display="true" />
        <attribute name="downward" valuetype="Boolean" display="true" />
        <attribute name="forward" valuetype="Boolean" display="true" />
        <attribute name="backward" valuetype="Boolean" display="true" />

        <attribute name="asymmetry" display="true">
        <value-el color="light red">left</value-el>
        <value-el color="light red">right</value-el>
        </attribute>
 </track-spec>

<!-- ********************* FACIAL EXPRESSIONS ********************* -->
<!-- ************** adapted from the FACS ACTION UNIT list ********** -->
 <group name="facial expressions">
          <track-spec name="eyes" type="primary">
                <attribute name="toward interlocutor" valuetype="Boolean" display="true" />
                <attribute name="front" valuetype="Boolean" display="true" />
                <attribute name="turn left (AU61)" valuetype="Boolean" display="true" />
                <attribute name="turn right (AU62)" valuetype="Boolean" display="true" />
                <attribute name="up (AU63)" valuetype="Boolean" display="true" />
                <attribute name="down (AU64)" valuetype="Boolean" display="true" />
                <attribute name="closed (AU43)" valuetype="Boolean" display="true" />
               <attribute name="upper lid raiser (AU5)" valuetype="Boolean" display="true" />
                <attribute name="squint (AU44)" valuetype="Boolean" display="true" />

                <attribute name="asymmetry" display="true">
                <value-el color="light red">left</value-el>
                <value-el color="light red">right</value-el>
                </attribute>
        </track-spec>

        <track-spec name="brows" type="primary">
                <attribute name="inner brow raiser (AU1)" valuetype="Boolean"
display="true" />
```

```
                <attribute name="outer brow raiser (AU2)" valuetype="Boolean"
                        display="true" />
                <attribute name="brow lowerer (AU4)" valuetype="Boolean" display="true" />

                <attribute name="intensity" emptyvalue="false" defaultvalue="normal"
                        display="true">
                        <value-el color="light yellow">low</value-el>
                        <value-el color="light green">normal</value-el>
                        <value-el color="light red">high</value-el>
                </attribute>

                <attribute name="asymmetry" display="true">
                        <value-el color="light red">left</value-el>
                        <value-el color="light red">right</value-el>
                </attribute>
          </track-spec>

        <track-spec name="mouth" type="primary">
        <attribute name="upper lip raiser (AU10)" valuetype="Boolean" display="true" />
        <attribute name="lip corner puller (AU12)" valuetype="Boolean" display="true" />
        <attribute name="lip corner depressor (AU15)" valuetype="Boolean" display="true" />
        <attribute name="lower lip depressor (AU16)" valuetype="Boolean" display="true" />
        <attribute name="lip stretcher (AU20)" valuetype="Boolean" display="true" />
        <attribute name="lip funneler (AU22)" valuetype="Boolean" display="true" />
        <attribute name="jaw drop (AU26)" valuetype="Boolean" display="true" />
        <attribute name="mouth stretch (AU27)" valuetype="Boolean" display="true" />
        <attribute name="lips part" valuetype="Boolean" display="true" />
        <attribute name="lips pressor" valuetype="Boolean" display="true" />
        <attribute name="cheek puffer" valuetype="Boolean" display="true" />
        <attribute name="lip tightener" valuetype="Boolean" display="true" />

         <attribute name="asymmetry" display="true">
                <value-el color="light red">left</value-el>
                <value-el color="light red">right</value-el>
        </attribute>
        </track-spec>

          <track-spec name="other AUs" type="primary">
                <attribute name="chin raiser (AU17)" valuetype="Boolean" display="true" />
                <attribute name="nose wrinkler (AU9)" valuetype="Boolean" display="true" />

                <attribute name="asymmetry" display="true">
                        <value-el color="light red">left</value-el>
                         <value-el color="light red">right</value-el>
                </attribute>
          </track-spec>
</group>

 <!-- ************************** GESTURES ************************** -->
 <group name="gestures">
  <doc>
    arm and hand gestures; the members of this group give a functional
    view on gesture. overlap is only possible with beats that can
    occur in any type of gesture. look at description of phase
    values to see how that is coded. (Kipp 1999)
```

```
  </doc>

<!-- ************************* right hand ************************* -->
<group name="right hand">
        <track-spec name="phase" type="primary" color-attr="type">
                <attribute name="type" display="true" valuetype="phaseType" />
        </track-spec>

        <track-spec name="phrase" type="span" ref="gestures.right hand.phase"
                color-attr="category">
                <attribute name="category lemme" display="true" valuetype="String">
                <doc>
                  Enter manually your interpretation of the gesture category by using
                  one of the 6 category types defined below + "." + a word or expression
                  for naming the gesture. (Example: deictic.self, metaphoric.cup, iconic.box).
                  See the annotation guide for the definifions of the gestures categories.
                </doc>

                </attribute>
                <attribute name="category annotation confidence" display="true"
                        valuetype="Number(1,5)" />
                <attribute name="lexical affiliate" link-color="red"
                        valuetype="ReciprocalLink(speech.transliteration)" />
        </track-spec>

        <track-spec name="expressivity" type="span" ref="gestures.right hand.phase"
                color-attr="speed">
                <attribute name="fluidity" emptyvalue="false" defaultvalue="normal"
                        display="true" valuetype="fluidityType" />
                <attribute name="strength" emptyvalue="false" defaultvalue="normal"
                        display="true" valuetype="strengthType" />
                <attribute name="speed" emptyvalue="false" defaultvalue="normal"
                        display="true" valuetype="speedType" />
                <attribute name="spatial expansion" emptyvalue="false"
                        defaultvalue="normal" display="true" valuetype="spatExpType" />
                <attribute name="repetition" valuetype="Number(1,10)" defaultvalue="1"
                        display="true" />
          </track-spec>

        <!-- ************************* gesture-unit ************************* -->
        <track-spec name="gesture-unit" type="span" ref="gestures.right hand.phrase"
                color-attr="speed">
        <track-spec name="gesture-unit" type="primary" color-attr="speed">
                <attribute name="repetition" valuetype="Number(1,10)" defaultvalue="1"
                        display="true" />
                <attribute name="fluidity" emptyvalue="false" defaultvalue="normal"
                        display="true" valuetype="fluidityType" />
                <attribute name="strength" emptyvalue="false" defaultvalue="normal"
                        display="true" valuetype="strengthType" />
                <attribute name="speed" emptyvalue="false" defaultvalue="normal"
                        display="true" valuetype="speedType" />
                <attribute name="spatialExpansion" emptyvalue="false"
                        defaultvalue="normal" display="true" valuetype="spatExpType" />
                <attribute name="lexical affiliate" link-color="red"
                        valuetype="ReciprocalLink(speech.transliteration)" />
```

```
        </track-spec>
</group>

<!-- ************************* left hand ************************ -->
<group name="left hand">
        <track-spec name="phase" type="primary" color-attr="type">
                <attribute name="type" display="true" valuetype="phaseType" />
        </track-spec>

        <track-spec name="phrase" type="span" ref="gestures.left hand.phase"
                color-attr="category">
                <attribute name="category" display="true" valuetype="String" />
                <attribute name="category annotation confidence" display="true"
                        valuetype="Number(1,5)" />
                <attribute name="lexical affiliate" link-color="red"
                        valuetype="ReciprocalLink(speech.transliteration)" />
        </track-spec>

        <track-spec name="expressivity" type="span" ref="gestures.left hand.phase"
                color-attr="speed">
                <attribute name="fluidity" emptyvalue="false" defaultvalue="normal"
                        display="true" valuetype="fluidityType" />
                <attribute name="strength" emptyvalue="false" defaultvalue="normal"
                        display="true" valuetype="strengthType" />
                <attribute name="speed" emptyvalue="false" defaultvalue="normal"
                        display="true" valuetype="speedType" />
                <attribute name="spatial expansion" emptyvalue="false"
                        defaultvalue="normal" display="true" valuetype="spatExpType" />
                <attribute name="repetition" valuetype="Number(1,10)" defaultvalue="1"
                        display="true" />
        </track-spec>

        <!-- ************************* gesture-unit ************************* -->
        <track-spec name="gesture-unit" type="span" ref="gestures.left hand.phrase"
                color-attr="speed">
        <track-spec name="gesture-unit" type="primary" color-attr="speed">
                <attribute name="repetition" valuetype="Number(1,10)" defaultvalue="1"
                        display="true" />
                  <attribute name="fluidity" emptyvalue="false" defaultvalue="normal"
                        display="true" valuetype="fluidityType" />
                <attribute name="strength" emptyvalue="false" defaultvalue="normal"
                        display="true" valuetype="strengthType" />
                <attribute name="speed" emptyvalue="false" defaultvalue="normal"
                        display="true" valuetype="speedType" />
                <attribute name="spatialExpansion" emptyvalue="false"
                        defaultvalue="normal" display="true" valuetype="spatExpType" />
                <attribute name="lexical affiliate" link-color="red"
                        valuetype="ReciprocalLink(speech.transliteration)" />
        </track-spec>
</group>

<!-- ************************* both hands ************************ -->
<group name="both hands">
        <track-spec name="phase" type="primary" color-attr="type">
                <attribute name="type" display="true" valuetype="phaseType" />
```

```
        </track-spec>

        <track-spec name="phrase" type="span" ref="gestures.both hands.phase"
                color-attr="category">
                <attribute name="category" display="true" valuetype="String" />
                <attribute name="category annotation confidence" display="true"
                        valuetype="Number(1,5)" />
                <attribute name="lexical affiliate" link-color="red"
                        valuetype="ReciprocalLink(speech.transliteration)" />
        </track-spec>

        <track-spec name="expressivity" type="span" ref="gestures.both hands.phase"
                color-attr="speed">
                <attribute name="fluidity" emptyvalue="false" defaultvalue="normal"
                        display="true" valuetype="fluidityType" />
                <attribute name="strength" emptyvalue="false" defaultvalue="normal"
                        display="true" valuetype="strengthType" />
                <attribute name="speed" emptyvalue="false" defaultvalue="normal"
                        display="true" valuetype="speedType" />
                <attribute name="spatial expansion" emptyvalue="false"
                        defaultvalue="normal" display="true" valuetype="spatExpType" />

                <attribute name="repetition" valuetype="Number(1,10)" defaultvalue="1"
                        display="true" />
        </track-spec>

        <!-- ************************* gesture-unit ************************* -->
        <track-spec name="gesture-unit" type="span" ref="gestures.both hands.phrase"
                color-attr="speed">
        <track-spec name="gesture-unit" type="primary" color-attr="speed">
                <attribute name="repetition" valuetype="Number(1,10)" defaultvalue="1"
                        display="true" />
                <attribute name="fluidity" emptyvalue="false" defaultvalue="normal"
                        display="true" valuetype="fluidityType" />
                <attribute name="strength" emptyvalue="false" defaultvalue="normal"
                        display="true" valuetype="strengthType" />
                <attribute name="speed" emptyvalue="false" defaultvalue="normal"
                        display="true" valuetype="speedType" />
                <attribute name="spatialExpansion" emptyvalue="false"
                        defaultvalue="normal" display="true" valuetype="spatExpType" />
                <attribute name="lexical affiliate" link-color="red"
                        valuetype="ReciprocalLink(speech.transliteration)" />
        </track-spec>
  </group>
 </group>

 </body>
</annotation-spec>
```

Annexe 6 : Tables de correspondances EmoTV/Greta

1. Emotions / Affects

Emotion list from EmoTV and affects from Greta can be linked as following.

EmoTV table	Greta table
anger	anger
despair	distress / sadness
disappointment	disappointment / disliking
disgust	disgust
doubt	worried
embarrassment	embarrassment
exaltation	gloating / joy
fear	fear / fear_confirmed
irritation	disliking
joy	joy / small_joy
neutral	no affect
pleased	satisfaction / liking / haddy-for
pride	pride / gratification
sadness	sadness / sorry-for
serenity	satisfaction
shame	shame
surprise	surprise
worry	worried

2. Communicative goals / Performatives

EmoTV table	Greta table
to claim	request
to complain	criticize
to convince	suggest / propose / advice
to criticize	criticize
to demonstrate	no correspondance
to describe	inform
to divert	no correspondance
to encourage	approve / praise
to exhibit	announce
to explain	inform
to joke	no correspondance
to justify	no correspondance
to reassure	agree / confirm / praise
to revolt	criticize / disagree / warn
to sell	no correspondance
to share a story	inform
to share a feeling	inform
to teach	inform
to testify	propose

3. Multimodal annotations / Greta specification

We have defined new items in Greta specification files, from EmoTV relations generated with the AnnotSoft program. Each modality has been treated in this way, adding progressively new definitions in Greta files, as shown in the table below.

EmoTV attribute	Greta definition
fluidity (smooth, normal, jerky)	jerky (-1) FLD smooth (1)
strength (soft, normal, hard)	soft (-1) PWR hard (1)
speed (slow, normal, fast)	fast (-1) TMP slow (1)
spatial expansion (contracted, normal, expanded)	contracted (-1) SPT expanded (1)
torso	not implemented
head : primary position, secondary position, primary movement, secondary movement (turn left, turn right, tilt left, tilt right, forward, backward, upward, downward, front)	automatic generation
shoulders	not implemented
eyes : toward interlocutor, turn left (AU61), turn right (AU62), up (AU63), down (AU64), closed (AU43), upper lid raiser (AU5), squint (AU44)	gaze simulation parameters: max t sp.lookat = max time that the agent looks at the camera max t sp.lookaw = max time that the agent looks away
brows : frown (AU4), inner and outer brow raiser (AU1+ AU2), inner brow raiser (AU1), outer brow raiser (AU2) + intensity (low, normal, high)	small_frown, medium_frown, high_frown

Annexe 7 : Publications effectuées durant la thèse

2007

Martin, J.-C., **Abrilian, S.,** Buisine, S., Devillers, L. (2007) Individual differences in the perception of spontaneous gesture expressivity. Third conference of the ISGS "Integrating Gestures", June 18-21, Evanston, Illinois, USA. Poster.

Martin, J.-C., Caridakis, G., Devillers, L., Karpouzis, K. and **Abrilian, S.** (2007) Manual Annotation and Automatic Image Processing of Multimodal Emotional Behaviours: Validating the Annotation of TV Interviews. Special issue of the Journal on Personal and Ubiquitous Computing on 'Emerging Multimodal Interfaces' following the special session of the AIAI 2006 Conference. Springer.

2006

Abrilian, S., J.-C. Martin, L. Devillers, S. Buisine. (2006). Perception of Movement Expressivity in Emotional TV Interviews. Ecole d'été, Gênes, Italie, 25 sept. 2006.

Martin, J.-C., **Abrilian, S.**, Devillers, L., Lamolle, M., Mancini, M. and Pelachaud, C. (2006). Du corpus vidéo à l'agent expressif. Utilisation des différents niveaux de représentation multimodale et émotionnelle. Revue d'Intelligence Artificielle. Numéro spécial sur les interactions émotionnelles.

Buisine, S., **Abrilian, S.**, Niewiadomski, R., Martin, J.-C., Devillers, L., Pelachaud, C (2006). Perception of Blended Emotions: from Video Corpus to Expressive Agent. 6th Int. Conference on Intelligent Virtual Agents (IVA'2006), Marina del Rey, USA. Best Paper Award

Martin, J.-C., Caridakis, G., Devillers, L., Karpouzis, K., **Abrilian, S.** (2006) Manual Annotation and Automatic Image Processing of Multimodal Emotional Behaviours: Validating the Annotation of TV Interviews. Fifth international conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy.

Devillers, L., Cowie, R., Martin, J.-C., Douglas-Cowie, E., **Abrilian, S.**, McRorie, M. (2006) Real life emotions in French and English TV video clips: an integrated annotation protocol combining continuous and discrete approaches. 5th international conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy, 24-27 may.

Abrilian, S., Devillers, L., Martin, J.-C. (2006). Annotation of Emotions in Real-Life Video Interviews: Variability between Coders. 5th Int. Conf. on Language Resources and Evaluation (LREC 2006), Genoa, Italy, 24-27 may.

2005

Martin, J.-C., **Abrilian, S.** and L., D. (2005). Annotating Multimodal Behaviors Occurring during Non Basic Emotions. 1st International Conference on Affective Computing & Intelligent Interaction (ACII'2005) Beijing, China, October 22-24.

Devillers, L., **Abrilian, S.** and Martin, J.-C. (2005)Representing real life emotions in audiovisual data with non basic emotional patterns and context features. 1st International Conference on Affective Computing & Intelligent Interaction (ACII'2005) Beijing, China, October 22-24.

Martin, J.-C., **Abrilian, S.**, Devillers, L., Lamolle, M., Mancini, M. and Pelachaud, C. (2005). Levels of Representation in the Annotation of Emotion for the Specification of Expressivity in ECAs. 5th International Working Conference On Intelligent Virtual Agents (IVA'2005) Kos, Greece, September 12-14.

Douglas-Cowie, E., Devillers, L., Martin, J.-C., Cowie, R., Savvidou, S., **Abrilian, S.** and Cox, C. (2005). Multimodal Databases of Everyday Emotion: Facing up to Complexity. 9th European Conference on Speech Communication and Technology (Interspeech'2005) Lisbon, Portugal, September 4-8.

Lamolle, M., Mancini, M., Pelachaud, C., **Abrilian, S.**, Martin, J.-C. and Devillers, L. (2005). Contextual Factors and Adaptative Multimodal Human-Computer Interaction: Multi-Level Specification of Emotion and Expressivity in Embodied Conversational Agents. (Context'05) Paris, 5-8 July.

Abrilian, S., Devillers, L. Buisine, S., Martin, J-C. (2005a). EmoTV1: Annotation of Real-life Emotions for the Specification of Multimodal Affective Interfaces. HCI International 2005, Las Vegas, USA.

Abrilian, S., Martin, J-C., Devillers, L. (2005b). A Corpus-Based Approach for the Modeling of Multimodal Emotional Behaviors for the Specification of Embodied Agents. HCI International 2005, Las Vegas, USA.

Abrilian, S. (2005) Annotation de Corpus d'Interviews Télévisées pour la Modélisation de Relations entre Comportements Multimodaux et Emotions Naturelles. 6ème Colloque des Jeunes Chercheurs en Sciences Cognitives (CJCSC'2005), Bordeaux, France.