La branche émotion, un modèle conceptuel pour l'intégration de la reconnaissance multimodale d'émotions dans des applications interactives : application au mouvement et à la danse augmentée

Alexis Clay

N° d'ordre : 3935

THÈSE

PRÉSENTÉE À

L'UNIVERSITÉ BORDEAUX I

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET D'INFORMATIQUE

Par **Alexis CLAY**

POUR OBTENIR LE GRADE DE

DOCTEUR

SPÉCIALITÉ : INFORMATIQUE

La branche émotion, un modèle conceptuel pour l'intégration de la reconnaissance multimodale d'émotions dans des applications interactives : application au mouvement et à la danse augmentée

---

Soutenue le : 7 décembre 2009

Après avis des rapporteurs :

Alice Caplier	Professeur
Jean-Claude Martin	Professeur

Devant la commission d'examen composée de :

Alice Caplier	Professeur	Rapporteur
Jean-Claude Martin	Professeur	Rapporteur
Nadine Couture	Enseignant-chercheur	Directrice
Maylis Delest	Professeur	Directrice
Guy Melançon	Professeur	Président
Laurence Nigay	Professeur	Directrice
Jean-Baptiste de la Rivière	Docteur	Examinateur
Wilfrid Lefer	Professeur	Examinateur

– 2009 –

Résumé

La reconnaissance d'émotions est un domaine jeune mais dont la maturité grandissante implique de nouveaux besoins en termes de modélisation et d'intégration dans des modèles existants. Ce travail de thèse expose un modèle conceptuel pour la conception d'applications interactives sensibles aux émotions de l'utilisateur. Notre approche se fonde sur les résultats conceptuels issus de l'interaction multimodale : nous redéfinissons les concepts de modalité et de multimodalité dans le cadre de la reconnaissance passive d'émotions. Nous décrivons ensuite un modèle conceptuel à base de composants logiciels s'appuyant sur cette redéfinition : la branche émotion, facilitant la conception, le développement et le maintien d'applications reconnaissant l'émotion de l'utilisateur. Une application multimodale de reconnaissance d'émotions par la gestuelle a été développée selon le modèle de la branche émotion et intégrée dans un système d'augmentation de spectacle de ballet sensible aux émotions d'un danseur.

Mots-clés. Interaction homme-machine, émotion, reconnaissance, modèle conceptuel, composants logiciels, multimodalité, gestuelle, mouvement, réalité augmentée, danse.

Computer-based emotion recognition is a growing field which develops new needs in terms of software modeling and integration of existing models. This thesis describes a conceptual framework for designing emotionally-aware interactive software. Our approach is based upon conceptual results from the field of multimodal interaction : we redefine the concepts of modality and multimodality within the frame of passive emotion recognition. We then describe a component-based conceptual model relying on this redefinition. The emotion branch facilitates the design, development and maintenance of emotionally-aware systems. A multimodal, interactive, gesture-based emotion recognition software based on the emotion branch was developed. This system was integrated within an augmented reality system to augment a ballet dance show according to the dancer's expressed emotions.

Keywords. Man-machine interaction, emotion, recognition, conceptual model, software components, multimodality, gesture, movement, augmented reality, dance.

Remerciements

Il est amusant pour moi de me souvenir que cette thèse ne fut en aucun cas due à une quelconque vocation mais à une suite de circonstances inattendues. Si le DRH avait lu le CV et la lettre de motivation que j'avais envoyé pour un poste en VIE à Taiwan (il m'a été confirmé plus tard que le CV n'avait pas été lu, ou ma candidature aurait été acceptée) ; si en attendant cette réponse je n'étais pas allé à la fac pour m'inscrire, dans le but d'avoir la mutuelle étudiante ; si quelqu'un avait été présent pour m'inscrire aux Beaux-Arts de Pau (mon premier choix), puis en filière Langues Etrangères Appliquées (mon second choix) ; si, errant dans le bâtiment Informatique de l'UFR de Sciences de la faculté de Pau, je n'étais tombé sur la mère d'un excellent ami, secrétaire du doyen des Sciences, qui me permit de m'inscrire en DEA ; si je n'avais pas assisté aux cours, préférant une dernière année estudiantine à l'effroi de la vie active ; si enfin, je n'étais pas tombé, totalement par hasard et au gré de mes lectures bibliographiques de mon sujet de stage, sur une idée qui tout de suite enchanta Nadine Couture, mon encadrante... alors je n'aurais pas fait de thèse, de cela j'en suis certain.

Mes premiers remerciements lui sont donc tout naturellement adressés, ainsi qu'à mes autres encadrantes de thèse, Laurence Nigay et Maylis Delest. Merci Nadine pour avoir, à partir d'une simple idée jetée en l'air, mis en place cette thèse, défini un sujet de recherche, trouvé des financements, tout ceci pendant que je parcourais Taïwan (à la recherche de contacts !) ; merci à toutes les trois d'avoir accepté de co-diriger cette thèse, de m'accompagner et de me faire profiter de vos diverses expériences dans des domaines variés tout au long de cette aventure. Merci enfin de m'avoir donné ce goût pour la recherche ; Vous avez réussi ce sur quoi je n'aurais pas parié un sou quelques mois auparavant.

La réalisation de ce travail a bien évidemment nécessité des moyens et un financement. Je tiens à remercier l'ESTIA et son directeur Jean-Roch Guiresse de m'avoir accueilli au sein de leur équipe de recherche. Je remercie bien sûr également Didier Borotra et la Communauté d'Agglomération Bayonne-Anglet-Biarritz, qui a financé ce travail de thèse.

J'adresse mes sincères remerciements à l'ensemble du jury : son président Guy Mélançon, Jean-Baptiste de la Rivière, et Wilfrid Lefer. Je remercie particulièrement les rapporteurs de ce mémoire, Alice Caplier et Jean-Claude Martin, pour leur regard critique et leurs remarques avisées.

Je tiens à exprimer mes profonds remerciements à Thierry Malandain et au Centre Chorégraphique National Malandain Ballet Biarritz pour s'être joint à nous dans cette aventure. La

collaboration que nous avons mis en place m'a permis d'aborder un monde dont j'avais une totale ignorance, et a permis d'associer pleinement à nos travaux Gaël Domenger, danseur et chorégraphe au CCN d'Aquitaine.

Un paragraphe spécifique était d'ailleurs nécessaire, me semble-t-il, pour pleinement remercier cet énergumène extra-terrestre dont l'imagination et la créativité se le disputent à la rigueur et à la raison. Gaël, tu es un geek, sache-le ; si tu n'avais été danseur, tu aurais du être scientifique fou. Il n'en fallait pas moins pour que nous nous entendions aussi bien et que tu appréhendes avec une telle facilité un monde qui n'était pas le tien.

Un grand merci bien sûr à tout le personnel de l'ESTIA et aux membres du laboratoire, permanents comme doctorants. Comment ne pas citer les compagnons d'infortunes, galériens harassés par le triste labeur du doctorant ?

- La bande des hispanos tout d'abord : David et Alvaro qui m'ont permis de partager les frais (élevés) de logement sur la côte basque, Ric, Carmen, Keny, pour toutes les soirées, la bonne humeur, l'esprit festif des paisas, et tout ce qui va bien... sans oublier ce voyage à Medellin, mémorable ! Et Livier, bien évidement, pour ses concours d'insultes Shakespeariennes, nos débats Pratchettiens, et pour m'avoir prouvé avec Keny que l'ingestion d'une demi-bouteille de tabasco était une prouesse humainement réalisable.
- Guillaume R., motard dans le plus pur esprit Joe BAR Team, pour ses conseils, ainsi que pour sa double capacité à pouvoir s'endormir en soirée et ensuite travailler plus de 50 heures d'affilée.
- Olivier P., compagnon de rédaction, et dont les multiples théories ont lancé des débats qui ont égayé nos pauses café-clope. La thèse, un tour du monde en solitaire ? C'est pas si pire, mais c'est pas non plus déconnant.
- à Elric D., pour avoir pris sur lui du travail supplémentaire afin de me laisser rédiger sans distractions.
- à Emilie C., Olivier Z., Ionel V., et tous les autres pour tous ces super moments que nous avons pu passer ensemble !

Nearly last but not least, je remercie bien évidemment ma famille - mon père et mes frère et soeurs, mes amis. Je remercie Gaëlle de m'avoir supporté pendant ces années de thèse et en particuliers les mois de rédaction, ces mois difficiles où une épaule compréhensive sur laquelle pleurer l'infortune du doctorant en rédaction est inestimable.

Enfin, je dédie ce mémoire à ma mère, décédée le 31 mars 2009, à 13h31. Une seule phrase de dédicace me vient à l'esprit aujourd'hui pour lui rendre dignement hommage.

A ta santé.

Table des matières

Table des figures

Introduction

Sujet et constats

De nombreux domaines s'intéressent à l'émotion. La philosophie, l'histoire, l'ethnologie, l'éthologie, la neuroscience, la psychologie, ou encore les arts, pour lesquels l'émotion est une matière première. L'informatique s'est penchée sur l'émotion dans les années quatre-vingt dix. Le livre de Rosalind Picard *Affective Computing* [114] a marqué les débuts du domaine. Le domaine de l'informatique affective cherche à donner aux machines (ordinateurs ou robots) la capacité de prendre en compte l'émotion.

Cette prise en compte se fait selon deux axes, correspondant à deux des fonctions de l'émotion : une fonction de survie tout d'abord, permettant à un organisme de mieux réagir à un évènement ; une fonction de communication ensuite, que l'on peut diviser en ses deux sens : la reconnaissance et l'expression.

L'émotion agit comme mécanisme d'adaptation d'un organisme à un stimulus. Certaines émotions - les plus primaires, comme la peur ou la colère - provoquent des réactions motrices et physiologiques extrêmement rapides, servant à adopter une réaction appropriée à un évènement. Ainsi la peur provoque une rétractation des vaisseaux sanguins pour mieux irriguer les muscles et le cerveau et ainsi préparer à la fuite. Certains travaux en informatique et robotique cherchent à modéliser ce "câblage" d'un stimulus à une réaction pour adapter ces émotions à la machine : la "peur" peut ainsi être "ressentie" par un robot lorsqu'il perçoit une menace à son intégrité physique ou logicielle [107].

L'émotion intègre des réactions ayant pour but sa communication. Cette communication émotionnelle influence grandement la communication entre deux individus. L'homme est une créature profondément émotionnelle, chez qui les émotions et leur communication jouent un rôle prépondérant dans la communication d'humain à humain. Un axe de recherche sur la communication émotionnelle entre un humain et une machine est donc né, cherchant à composer reconnaissance des émotions de l'utilisateur et expression d'émotions par la machine (généralement au travers d'un avatar 3D animé [111]), afin de proposer une communication plus naturelle pour l'utilisateur. Dans ce cas, il ne s'agit pas de doter la machine d'émotions, mais de capacités expressives permettant d'offrir un retour émotionnel à l'utilisateur et ainsi faciliter la communication homme-machine.

Dans nos travaux, nous focalisons sur l'aspect reconnaissance de cette fonction de communication de l'émotion. La reconnaissance d'émotions a pour but de permettre à la machine de s'adapter à l'état émotionnel de l'utilisateur. Cette adaptation peut être traduite par une modification du fonctionnement du système interactif : par exemple, une application d'apprentissage peut détecter la frustration de l'utilisateur pour proposer une série d'exercices moins difficiles. L'adaptation peut également se faire au niveau des interactions que le système propose. Ainsi un cockpit doté d'un système pouvant détecter le stress du pilote peut, en cas de situation difficile, réduire la quantité d'information affichée afin de focaliser son attention sur les informations absolument nécessaires à la résolution de la situation. Si dans la même situation, le pilote n'exprime aucun stress, le système pourrait alors également afficher d'autres informations moins critiques pour la situation.

Objectifs et motivations

Le domaine de la reconnaissance d'émotions est donc un domaine très actif, marqué par un intérêt croissant depuis sa création. La première conférence internationale rassemblant les acteurs du domaine, à la fois en informatique et en psychologie, s'est déroulée en 2005[1] et a été reconduite en 2007 et 2009. Les systèmes de reconnaissance ont au fil des années considéré les différents vecteurs de la communication émotionnelle, en commençant par le visage et la voix, pour ensuite s'intéresser à la gestuelle et enfin aux réactions physiologiques de l'émotion.

Depuis les débuts du domaine, de nombreux systèmes de reconnaissance d'émotions ont été implémentés et testés [152]. Ces systèmes diffèrent grandement sur plusieurs points :

- Tout d'abord, sur les expressions émotionnelles reconnues : expressions faciales, voix, mouvements ou expressions physiologiques.
- Ensuite sur les méthodes utilisées pour interpréter les signaux considérés et les traduire en des états émotionnels.

Le dynamisme de cet axe de recherche se traduit par des avancées brutales aux facettes multiples. Les systèmes de reconnaissance sont répliqués, les techniques améliorées, les résultats plus robustes grâce à des systèmes plus complexes. Face à ce foisonnement des systèmes de reconnaissance proposés, notre objectif de recherche est de proposer **un cadre architectural de référence**, **cadre d'étude unificateur des systèmes de reconnaissance existants**, afin de pouvoir capitaliser et comparer les expériences. Nos travaux s'inscrivent donc clairement dans l'ingénierie des systèmes interactifs capables de reconnaître les émotions des utilisateurs.

Nous constatons que les recherches s'orientent vers des systèmes multicanaux de reconnaissance des émotions [153] [64] [26], c'est-à-dire capables d'analyser plusieurs canaux de communication émotionnelle (visage, voix, mouvement, réactions physiologiques). Ces systèmes plus complexes font naître le besoin d'un support à la conception logicielle de systèmes de reconnaissance d'émotions modifiables et extensibles. Ce constat a aussi motivé nos travaux sur l'architecture logicielle : fournir **un cadre architectural de référence** pour faire face à la

[1] International Conference on Affective Computing and Intelligent Interaction - ACII

complexité croissante des systèmes de reconnaissance et pour **fournir une forte modularité en vue de la modifiabilité et de l'extensibilité** des systèmes de reconnaissance.

Pour notre objectif de mise en place d'un cadre architectural de référence, notre approche de travail a été de s'appuyer sur les résultats en Interaction Homme-Machine (IHM) et en particulier en interaction multimodale. En effet le domaine de l'IHM dispose de plusieurs cadres d'étude établis fournissant donc une base et une inspiration pour la conception d'un cadre architectural pour des systèmes sensibles aux émotions. Dès les années 1980 la modifiabilité du code a été le centre d'intérêt de nombreux travaux en architecture logicielle des systèmes interactifs. La modifiabilité du code était alors la réponse incontournable adoptée dans le cadre d'un processus de conception itérative centrée sur l'utilisateur. Aussi de nombreux modèles de référence en IHM répondent à cette exigence de modifiabilité avec le principe de séparation fonctionnelle.

Contributions

Répondant à notre objectif de cadre d'étude de référence, nos contributions concernent les facettes conceptuelles et pratiques de la conception logicielle. Pour comprendre, comparer et unifier les expériences en reconnaissance d'émotions ainsi que cerner la complexité et l'extensibilité des systèmes de reconnaissance d'émotions, nous proposons un cadre conceptuel architectural de référence basé sur le principe de séparation fonctionnelle. Notre contribution est aussi pratique par la réalisation logicielle d'un système de reconnaissance d'émotions par les mouvements selon notre modèle architectural de référence. Cette réalisation logicielle constitue un canevas logiciel générique et extensible pour le cas de la reconnaissance d'émotions par les mouvements. Nous illustrons son application au cas d'un spectacle de danse.

Pour notre domaine d'application, nous considérons la danse dans son concept fondamental : le corps et son expression. Cette expression est distribuée selon trois dimensions : l'espace, le temps et l'autre. La réalité augmentée permet de peupler l'espace par des éléments virtuels sur la scène, donnant aussi des références spatiales au danseur, comme le ferait un décor ou d'autres danseurs. De plus, l'expression dans la danse se focalise sur l'expression émotionnelle et sur le déclenchement d'expériences émotionnelles chez le public. La reconnaissance de l'émotion du danseur permet de prendre en compte cette expression pour moduler et influencer les éléments virtuels, permettant ainsi de leur donner une dimension temporelle dépendante du danseur.

Structure du mémoire

La structure du mémoire reflète les deux facettes, conceptuelle et pratique, de nos contributions centrées sur la conception des systèmes multimodaux de reconnaissance d'émotions (figure 1). Après une première partie qui recense les travaux en reconnaissance d'émotions (Chapitres 1, 2 et 3), la thèse est organisée en deux parties : architecture conceptuelle (Chapitres 4 et 5) et canevas logiciel (Chapitres 6 et 7) appliqué à notre domaine d'application.

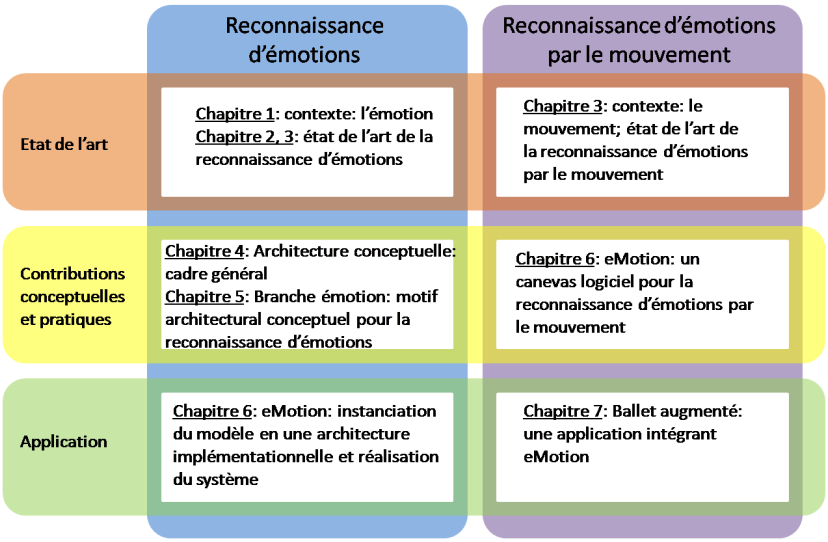


FIG. 1: Organisation du mémoire sur les facettes "reconnaissance d'émotions" et "reconnaissance d'émotions par la gestuelle"

Ainsi la première partie définit notre cadre de recherche en détaillant le concept d'émotion et d'expression émotionnelle, puis en analysant les travaux afin de recenser les divers choix conceptuels et implémentationnels des systèmes de reconnaissance existants. Ce recensement nous permet d'identifier les requis du modèle conceptuel que nous exposons en deuxième partie. Nous y traitons l'ingénierie logicielle de la reconnaissance d'émotions par une approche s'inspirant des travaux existants en ingénierie logicielle de l'interaction multimodale et nous y détaillons notre modèle conceptuel, la branche émotion. Enfin la troisième partie décrit l'implémentation de notre modèle conceptuel au cas de la reconnaissance d'émotions par les mouvements en un canevas logiciel générique de par son architecture implémentationnelle. Notre cas d'application, l'utilisation de l'émotion pour augmenter un spectacle de danse, nous permet alors de valider ce canevas logiciel générique.

Le premier chapitre **"Les émotions"** explore le concept d'émotion, indispensable à nos travaux. Nous y retenons une définition sur laquelle nous nous basons tout au long de ce mémoire. Nous y décrivons ensuite des théories et modèles de représentation de l'émotion utilisés dans le domaine de la reconnaissance d'émotions. Enfin, nous introduisons les expressions émotionnelles, qu'analysent les systèmes de reconnaissance pour en inférer une émotion.

Le chapitre deux **"Systèmes de reconnaissance d'émotions"** décrit plusieurs systèmes de reconnaissance des émotions existants. Nous examinons les différentes problématiques et choix conceptuels ou implémentationnels effectués à chaque niveau du processus de reconnaissance. Nous traitons ensuite de l'évaluation de tels systèmes.

Le chapitre trois **"Reconnaissance d'émotions par la gestuelle"** reprend l'analyse effectuée au chapitre deux en se focalisant sur la gestuelle, afin de cerner en détail le contexte

de nos contributions pratiques. Nous définissons le geste et étudions le mouvement du point de vue de la psychologie avant de décrire les enjeux liés à la reconnaissance informatique des émotions à partir des mouvements.

Le chapitre quatre **"Modèle d'architecture : un point de vue interaction homme-machine de la reconnaissance d'émotions"** décrit la première partie de notre contribution. Après une description de l'interaction multimodale, nous introduisons la branche émotion sous la forme d'un motif architectural de conception en trois niveaux fonctionnels et mettons en avant l'intégration de cette branche dans des modèles existants en ingénierie logicielle de l'interaction multimodale.

Le chapitre cinq **"La branche émotion"** décrit en détail notre contribution : la branche émotion, un modèle conceptuel pour les systèmes sensibles aux émotions basé sur le motif en trois niveaux présenté au chapitre quatre. La branche émotion est un modèle basé sur une approche à composants. Nous définissons dans ce chapitre les différents composants mis en jeu, leurs spécifications, et proposons un mécanisme situé en arrière-plan permettant la communication entre les composants. Première forme d'évaluation de notre modèle, nous modélisons des systèmes de reconnaissance existants selon notre modèle.

Le chapitre six **"eMotion, un canevas logiciel de reconnaissance d'émotions basée sur la gestuelle"** traite alors d'une autre forme de validation de notre modèle par la réalisation logicielle d'un canevas générique basé sur notre modèle conceptuel et dédié à la reconnaissance d'émotions par les mouvements.

Le chapitre sept **"eMotion appliqué à la danse : Reconnaissance d'émotions et réalité augmentée"** présente enfin une application de ce canevas logiciel au cas d'un ballet augmenté. L'émotion reconnue par les mouvements du danseur est utilisée par le noyau fonctionnel d'un système de réalité augmentée afin de moduler les éléments virtuels projetés sur scène.

Première partie

Etat de l'art

Chapitre 1

Les Émotions

Le terme “émotion” est un terme couramment utilisé mais qu'il est difficile de définir avec précision. Intuitivement, nous sentons que l'émotion est un phénomène à la fois physique et physiologique. Mais que caractérise une émotion ? La colère est sans conteste une émotion. Qu'en est-il de l'amour, du stress, de la nervosité ou de l'hostilité ?

Dans ce chapitre nous posons le cadre de travail de notre recherche sur la reconnaissance d'émotions en informatique. Nous présentons tout d'abord plusieurs définitions de l'émotion, avant d'en retenir une permettant d'isoler les émotions d'autres états affectifs, comme les humeurs ou les attitudes. Nous retenons deux théories intéressantes dans le cadre de ce mémoire, issues des grandes questions qui ont agitées la recherche sur les émotions : la théorie de l'évaluation cognitive et la théorie évolutionniste. Nous présentons ensuite trois modèles de représentation de l'émotion utilisés dans le domaine informatique : le modèle discret, le modèle continu et le modèle à composants. Enfin, nous développons dans la dernière section des notions concernant l'expression émotionnelle. Définitions, théories et modèles sont généralement liés car développés conjointement ; pour mieux appréhender le domaine, nous avons cependant décidé de les présenter séparément.

1.1 Définitions de l'émotion

Dans son livre de 1997 *Affective Computing* [114], Rosalind W. Picard constate qu'il existe près d'une centaine de définitions de l'émotion. Le terme “émotion' est en effet difficile à définir. Une première définition nous est donnée dans le dictionnaire Larousse en ligne[1] :

> **émotion :** nom féminin (de émouvoir, d'après l'ancien français motion, mouvement). Réaction affective transitoire d'assez grande intensité, habituellement provoquée par une stimulation venue de l'environnement.

En recherchant le terme “affectif” dans ce même dictionnaire, on obtient :

> **affectif :** adjectif (latin affectivus). Qui concerne les sentiments, les émotions, la sensibilité. Qui relève du sentiment, non de la raison ; sentimental.

[1] http ://www.larousse.fr

Enfin, en recherchant le terme “sentiment” :

> **sentiment :** nom masculin (de sentir). État affectif complexe et durable lié à certaines émotions ou représentations.

Ces trois définitions ont été tronqués afin de ne garder que le sens qui nous concerne. Cependant, nous constatons que ces définitions se renvoient à elles-mêmes. Il en est de même dans le dictionnaire AskOxford[2], où le terme “sentiment” (“feeling”) est utilisé pour définir le terme “émotion” (“emotion’) et vice-versa.

Afin d’obtenir une définition plus précise de l’émotion, il est nécessaire de se tourner vers le domaine de la psychologie. Le grand dictionnaire de la psychologie de Bloch *et al.* [9] nous donne comme définition de l’émotion :

> **émotion :** Constellation de réponses de forte intensité qui comportent des manifestations expressives, physiologiques et subjectives typiques. Elles s’accompagnent généralement de tendances d’action caractéristiques et s’inscrivent en rupture de continuité par rapport aux processus qui étaient en cours chez le sujet au moment de leur apparition.

Cette définition est bien sûr beaucoup plus précise que les définitions données par les dictionnaires généralistes. En analysant cette définition, nous trouvons qu’une émotion :

- est un ensemble de **réponses** : une émotion émerge en réponse à un ou plusieurs stimuli, qu’ils soient internes (provenant de l’individu même) ou externes (provenant du monde qui l’entoure).
- est une **constellation** de réponses : Le choix du terme “constellation” laisse entendre que ces réponses sont interconnectées entre elles.
- ces réponses sont de **forte intensité** : le corps et l’esprit réagissent fortement.
- comporte des **manifestations expressives** : Nous exprimons nos émotions par divers canaux : les expressions du visage, les intonations de notre voix, notre gestuelle.
- comporte des **manifestations physiologiques** : par exemple, le cœur s’accélère ou ralentit, les vaisseaux sanguins se rétractent pour mieux irriguer les muscles : nous pâlissons.
- comporte des **manifestations subjectives** : Selon le même dictionnaire, “subjectif” est défini par : 1) relatif à un sujet, 2) Qui n’est accessible, en tant que connaissance ou affect, qu’à un seul sujet. La manifestation subjective est le “ressenti” de l’émotion.
- Ces manifestations sont **typiques** : Elles sont propres à chaque émotion.

La seconde phrase de la définition nous informe qu’une émotion nous laisse enclins à agir d’une certaine façon : par exemple, la colère nous prépare à l’attaque, la peur à la fuite. Enfin, les manifestations de l’émotion ne sont pas dépendantes des états précédents de l’individu. Bien que considérées comme telles pendant longtemps, les émotions n’ont cependant pas un seul caractère interruptif ; bien souvent, elles viennent au contraire en support à l’action en cours.

[2] http ://www.askoxford.com

1.1.1 Taxonomie retenue

Picard, dans son livre pionnier *Affective Computing* [114], ne fait pas de distinction entre les termes "émotion" et "affect". Cependant, certains termes peuvent prêter à discussion. La joie, par exemple, est une émotion ; la nervosité ou l'intérêt sont déjà plus ambigus. Scherer établit une taxonomie des états affectifs. En particulier, et contrairement à Picard, il définit l'expression "état affectif" comme une notion générique impliquant tout ce qui a trait à l'affect. L'émotion est un ensemble inclus dans l'ensemble des états affectifs. Cette notion d'état affectif est introduite en 1984 dans [129], puis affinée en 2000 dans [132] pour être divisée en cinq catégories :

- les humeurs : de durée moyenne (de l'ordre de quelques heures), de faible intensité, présentant surtout des manifestations subjectives.
- les attitudes interpersonnelles : prises par rapport à une autre personne lors d'une interaction spécifique, modulant l'échange.
- les préférences ou attitudes : croyances modulées par l'affect, préférences, prédispositions envers des objets ou des personnes.
- les traits de personnalité : dispositions affectives stables, tendances de comportement.
- les émotions : épisodes brefs de réponses synchronisées à l'évaluation d'un évènement externe ou interne.

Design Features / *Types of Affect*	Intensity	Duration	Synchronization	Event focus	Appraisal elicitation	Rapidity of change	Behavior impact
Emotions: *angry, sad, joyful, fearful, ashamed, proud, elated, desperate*	●	•	●	●	●	●	●
Moods: *cheerful, gloomy, irritable, listless, depressed, buoyant*	●	●	•	•	•	●	•
Interpersonal stances: *distant, cold, warm, supportive, contemptuous*	●	●	•	●	•	●	●
Preferences/Attitudes: *liking, loving, hating, valuing, desiring*	●	●	•	•	•	•	●
Affect dispositions: *nervous, anxious, reckless, morose, hostile*	•	●	•	•	•	•	●

Tab. 1: Catégorisation des états affectifs. Table tirée de [136].

La table 1 illustre la catégorisation des états affectifs de Scherer grâce à sept critères de classement : l'intensité de l'état affectif, sa durée, la synchronisation des réactions de l'ensemble des sous-systèmes organiques (réactions physiologiques et motrices), la focalisation sur l'évènement (l'état affectif considéré est-il en relation directe avec un évènement déclencheur ?), l'évaluation cognitive du stimulus, la rapidité de changement entre deux états affectifs de cette catégorie, et enfin l'impact sur le comportement.

Selon cette caractérisation, les émotions sont donc des états affectifs de forte intensité et de courte durée. Une émotion est également caractérisée par une forte synchronisation : tout

	Intensité	Durée	Synchronisation	Focalisation sur l'évènement	Evaluation intrinsèque	Evaluation transactionnelle	Rapidité de changement	Impact sur le comportement
Em. utilitaires	+	-	++	++	o	++	++	++
Em. esthétiques	- / o	-	o	+	++	-		

TAB. 2: Caractérisations des Emotions utilitaires et esthétiques, en utilisant les mêmes critères que la figure 1. Légende : - : bas, o : moyen, + : haut, ++ : très haut.

le corps, et l'esprit, tendent à réagir à l'unisson. Une émotion est le fruit de l'évaluation selon certains critères de stimuli externes ou internes et est focalisée sur l'évènement déclencheur (le stimulus). Enfin, une émotion affecte lourdement le comportement de l'individu. Cette caractérisation rejoint la définition retenue de [9] (page 10).

En 2004, Scherer distingue deux catégories d'émotions [134] : les émotions utilitaires et les émotions esthétiques. Les émotions utilitaires sont des "épisodes relativement brefs de réponses synchronisées de tous ou de la plupart des systèmes organiques en réponse à l'évaluation d'un évènement interne ou externe étant d'une importance majeure pour des besoins ou pour des buts personnels"[3]. Les émotions esthétiques sont des "émotions de stimuli audio ou visuel en termes de qualités intrinsèques de forme ou de relation entre éléments"[4]. "L'expérience esthétique n'est pas déclenchée par la pertinence d'une perception par rapport aux besoins, valeurs sociales, ou buts"[5].

La table 2 précise la catégorie "Emotions" de la table 1 en émotions utilitaires et esthétiques et liste les différences que l'on peut trouver entre ces deux sous-catégories. Les émotions esthétiques sont dans l'ensemble moins intenses, ont une faible synchronisation et un faible impact sur le comportement. Le corps réagit en effet beaucoup moins à une émotion esthétique qu'à une émotion utilitaire. Les émotions esthétiques mettent en œuvre une évaluation des qualités intrinsèques du stimulus (par exemple une œuvre musicale). Elles n'évaluent pas le stimulus en termes de pertinence par rapport aux besoins personnels et d'adaptabilité à la situation. Les émotions utilitaires, au contraire, se focalisent sur les besoins et nécessités de l'individu et sa capacité à s'adapter à la situation. Le processus d'évaluation dans ces deux types d'émotions est donc différent.

[3] "Relatively brief episode of synchronized responses by all or most organismic subsystems to the evaluation of an external ot internal event as being of major significance (e.g. anger, sadness, joy, fear, shame, pride, elation, desperation)."

[4] "Evaluations of auditory or visual stimuli in terms of intrinsic qualities of form or relationship of elements (moved, awed, surprised, full of wonder, [...], rapture, solemnity)."

[5] "Aesthetic experience is one that is not triggered by concerns with the relevance of a perception to my bodily needs, my social values, or my current goals or plans."

1.1.2 Emotion : définition retenue

Dans ce mémoire, nous considérons les émotions comme présentées par Scherer à la figure 1, soit comme une catégorie d'état affectif, caractérisée par une forte intensité, une durée courte, une très forte synchronisation, une focalisation très forte sur l'évènement déclencheur, une évaluation intrinsèque moyenne, une très forte évaluation transactionnelle, une très grande rapidité de changement, et un impact très fort sur le comportement. Dans nos travaux nous ne nous préoccupons que des émotions, et nous focalisons sur les émotions utilitaires. Ainsi, nous utiliserons le terme "émotions" pour nous référer aux émotions utilitaires dans cette thèse. Lorsque nous considérerons les émotions esthétiques, nous l'expliciterons par l'expression "émotions esthétiques".

1.2 Théories de l'émotion

Historiquement, de nombreuses théories de l'émotion ont été formulées, parfois controversées et déclenchant des débats dont certains ne sont pas conclus à ce jour. Dans [132], Scherer décrit cette histoire des théories de l'émotion en l'organisant autour des débats principaux du domaine. Le paragraphe 1.2.1 résume cette description. Le but de ce résumé n'est pas de recenser les différents courants mais d'introduire les théories retenues présentées au paragraphe 1.2.2. Enfin, le paragraphe 1.2.3 décrit deux modèles de transition entre deux émotions.

1.2.1 Problématiques relatives aux théories de l'émotion

Cognition et émotion

Le débat sur les relations entre cognition et émotion a émergé dès les philosophes et la Grèce antique. Platon sépare l'âme en une structure de trois parties opposées : la cognition, la motivation et l'émotion. Cette idée de séparation en trois parties, et notamment de l'opposition entre cognition (raison) et émotion, a traversé les siècles pour être fermement implantée dans nos esprits. Elle est cependant source à controverse : Aristote, cinquante ans après Platon, réfutait déjà cette hypothèse et proposait au contraire une théorie où cognition, émotion et motivation étaient étroitement mêlés par de nombreuses interactions. De nombreux théoriciens modernes reprennent ce concept.

Corps et esprit

Descartes fut le premier à considérer les émotions comme un ensemble comprenant à la fois des processus mentaux et des processus physiques et physiologiques. Si l'existence de changements physiques et physiologiques dus à un état émotionnel et plus particulièrement à l'évaluation d'évènements précédents est admise, de nombreuses questions restent en suspens sur la nature de ces changements par rapport à un état émotionnel donné.

Descartes, dans [47], oppose l'émotion à la raison. Il sépare les émotions en deux catégories : les émotions primaires, pures, et les émotions secondaires, qu'il définit comme un mélange d'émotions primaires, et fait l'analogie avec la palette de couleurs. Cette catégorisation en

deux ensembles est acceptée de nos jours. Cependant, la définition des émotions secondaires comme mélanges d'émotions primaires est sujette à controverse.

Antonio Damasio, dans son livre *L'erreur de Descartes* [44], réfute quant à lui l'opposition entre raison et émotion et explore les implications de l'émotion dans divers processus cognitifs tels que le raisonnement et la prise de décision. Il étudie en particulier le cas de Phineas Gage, dont le cortex préfrontal fut endommagé dans un accident. Il en résulta des capacités intellectuelles intactes mais une incapacité à faire l'expérience d'émotions. Damasio soutient que l'incapacité de Gage à prendre des décisions simples et à se comporter convenablement en société à la suite de son accident est la conséquence de la perte de ses émotions.

Biologie et culture

Le livre de Darwin *Expression of Emotion in Man and the Animals* [45], écrit en 1872, décrit les émotions comme un résultat de l'évolution des êtres vivants. Les émotions permettent de préparer un être à agir en réponse à un stimulus, tant dans les processus mentaux que physiques et physiologiques. Darwin décrit d'ailleurs pour de nombreuses émotions les changements physiques et physiologiques qu'il a pu observer. A l'inverse, les théories socio-constructivistes décrivent les émotions comme issues de la socialisation de l'individu. Selon ces théories, les émotions et leurs expressions sont donc sujettes à la culture d'un individu.

L'antagonisme de ces deux courants théoriques (biologie et évolution contre construction socioculturelle) a suscité de nombreux débats. Paul Ekman, un psychologue américain, a mené plusieurs études interculturelles afin d'isoler des émotions dont l'expression et la reconnaissance sont indépendantes de la culture [53][55]. Il isole en particulier les "6 émotions de base" : la joie, la peur, la tristesse, la colère, le dégoût et la surprise. Ces six émotions, par l'uniformité de leurs expressions au travers des cultures, ont été largement étudiées en informatique.

Cerveau et réactions du corps

A la fin du XIXème siècle, William James définit l'émotion comme la perception des changements physiques et physiologiques. Dans cette théorie, le corps réagit à un stimulus donné (par exemple en irriguant mieux les muscles, en augmentant le rythme cardiaque). L'émotion n'est alors que la perception, dans notre cerveau, de tous ces changements. Cette théorie est largement réfutée telle quelle ; elle a cependant ouvert de nombreuses discussions sur la part du système nerveux central (cerveau) et périphérique (dans le reste du corps) dans le processus émotionnel.

1.2.2 Théories et modèles retenus

Du foisonnement théorique, nous retenons dans nos travaux deux théories complémentaires : la théorie de l'évaluation cognitive et celle de l'évolution. Nous retenons un modèle pour chacune de ces théories : le modèle à composants de Scherer, qui s'inscrit dans le courant théorique de l'évaluation cognitive et plus spécifiquement de la théorie des processus à composants (*Component Process Theory* ou CPM), et le modèle des émotions basiques d'Ekman,

qui prolonge la vision évolutionniste des émotions de Darwin.

Théorie de l'Evaluation cognitive : la *Component Process Theory* de Scherer

Selon cette théorie, une émotion est le fruit d'une évaluation cognitive, non forcément consciente, d'un stimulus ou d'une série de stimuli. Scherer [129] s'appuie sur une approche fonctionnelle ; il divise le processus émotionnel en cinq composants (voir table 3, page 16). Une émotion est une séquence de changements d'états parmi ces cinq composants. Une émotion est donc une suite d'états et non un état statique. Ces changements d'états sont déclenchés par une évaluation du stimulus selon cinq critères d'évaluation (*Stimulus Evaluation Check* ou SECs). Dans ce mémoire, nous considérons qu'un stimulus est un évènement quelconque perçu par le sujet. Nous présentons tout d'abord les SECs déclenchant les changements d'état dans les composants du processus émotionnel, que nous décrivons ensuite.

Le premier critère est celui de la **nouveauté** de l'évènement, c'est à dire à quel degré cet évènement a été attendu. Cette évaluation peut se faire de façon très basique (par exemple dans le cas d'un bruit soudain, initiant une réponse immédiate comme un sursaut), mais également faire appel à un processus incluant des fonctions cognitives de plus haut niveau (par exemple l'annonce d'une mauvaise note à un examen *a priori* réussi).

Le deuxième critère est la **valence intrinsèque** de l'évènement, c'est-à-dire si l'évènement est considéré comme intrinsèquement plaisant ou déplaisant. Ce critère de valence n'est pas à mettre en relation avec une notion de but ; par exemple, un agriculteur recevant une pluie froide évaluera l'évènement comme intrinsèquement déplaisant, même si cette pluie sert son but de produire une récolte abondante.

Le troisième critère est celui de **l'adéquation au but** de l'évènement, c'est-à-dire dans quelle mesure l'évènement va-t-il pouvoir aider ou gêner les buts ou les envies de l'individu. En reprenant l'exemple de l'agriculteur ci-dessus, l'adéquation de l'évènement au but de l'individu peut déclencher une émotion positive malgré l'évaluation d'une pluie froide comme étant un évènement intrinsèquement déplaisant.

Le quatrième critère est celui du **potentiel d'adaptation** d'un organisme par rapport aux conséquences d'évènements passés ou futurs. L'évaluation de ce critère permet de déterminer la meilleure réaction à avoir face à un évènement, et donc de déclencher l'émotion adéquate. Ce critère est divisé en quatre sous-critères. Le critère de *causation* sert à déterminer l'agent ou la cause origine de l'évènement. Le critère de *contrôle* détermine la capacité de l'individu à influer sur les conséquences de l'évènement. Le critère de *puissance* évalue la capacité de l'individu à surmonter des obstacles ou des ennemis. Enfin, le critère d'*ajustement* évalue la facilité avec laquelle l'individu peut s'adapter aux conditions changées par les évènements survenus.

Fonctions	*Composants*
Evaluation de l'environnement	Traitement cognitif du stimulus
Régulation du système	Processus neurophysiologiques
Préparation à l'action	Motivation et tendances d'actions
Communication d'intention	Expression motrice
Retour et surveillance du système	Ressenti, "feeling"

TAB. 3: Fonctions de l'émotion et les composants qui y sont liés. Tableau tiré de [129].

Le cinquième critère est celui de la **compatibilité avec la norme et soi-même** et est propre aux humains. Il consiste en la comparaison de réactions à l'évènement avec des standards internes et externes comme la norme sociale et la conception de soi.

Ces cinq critères sont évalués consécutivement et chaque évaluation permet d'affiner l'émotion qui sera déclenchée. Certaines évaluations peuvent prendre le pas sur des évaluations antérieures. De plus, l'évaluation de certains critères peut générer des réponses avant que tous les critères ne soient évalués : par exemple, l'évaluation d'une nouveauté extrêmement forte (surprise) déclenche un sursaut de recul avant l'évaluation des autres critères.

Toujours selon la théorie de Scherer, l'émotion remplit cinq fonctions : l'évaluation de l'environnement, la régulation du système organique, la préparation à l'action, la communication d'intention et le retour et la surveillance du système organique (qui est modifié par le processus de régulation) [129]. Cinq composants sont asservis à ces cinq fonctions de l'émotion (voir table 3 page 16). Le premier composant est le composant *traitement cognitif du stimulus*. Il prend en charge l'évaluation de l'environnement. Le composant *processus neurophysiologique* régule le système organique, c'est-à-dire les changements de température, de sudation, d'irrigation des muscles par les vaisseaux sanguins, etc. Le composant *motivation et tendances d'action* prépare à une action de l'individu face à la nouvelle situation, après l'évènement déclencheur de l'expérience émotionnelle. Le composant *expression motrice* remplit la fonction de communication d'intention ; par exemple exprimer la colère communique une intention (feinte ou non) d'attaquer. Enfin, le composant *feeling* ou *ressenti* permet un retour conscient de l'état général de l'organisme. Les différents critères sont évalués par le composant de traitement du stimulus. Ce traitement et les évaluations qui sont faites pour chaque critère déclenche des changements d'états dans les différents composants.

Chacun de ces composants peut prendre plusieurs états. Les résultats des différents SECs vont provoquer des changements d'états dans un ou plusieurs de ces composants. La séquence de SECs induit donc une séquence de changements d'états dans les différents composants. De plus, chaque changement d'état dans un composant peut influer sur le comportement d'un composant et déclencher un changement dans ce composant. Pour résumer, le comportement d'un composant dépend à la fois des évaluations des SECs et des comportements des différents composants (y compris lui-même). Les composants ne sont donc pas indépendants. La séquence complète de changement d'état est un processus émotionnel, c'est-à-dire, dans la théorie de Scherer, une émotion.

La théorie évolutionniste : les émotions basiques d'Ekman

Pour Darwin, les émotions sont des réactions à des stimuli, préparant le corps à agir d'une certaine façon. Issues de l'évolution, elles se sont développées pour offrir une réponse extrêmement rapide et adaptée à la perception de la situation. Il est ainsi possible de voir les émotions comme des réactions préprogrammées à certains évènements. Par exemple, la peur prépare le corps à la fuite ; la colère, à l'attaque.

Dans [49], le psychologue Paul Ekman reprend la théorie évolutionniste des émotions et extrait entre six et sept émotions basiques selon une série de onze critères. Pour être considérée comme basique, c'est-à-dire indubitablement issue de l'évolution des espèces animales, une émotion doit donc :

E_1 générer des signaux universels distincts ;

E_2 générer des réactions physiologiques distinctes ;

E_3 présenter une évaluation automatique de l'évènement déclencheur ;

E_4 être déclenchée par des stimuli distinctifs ;

E_5 aborder un développement distinct ;

E_6 être présente chez d'autres primates que l'homme ;

E_7 apparaître rapidement ;

E_8 être de courte durée ;

E_9 être capable d'apparaître de façon interruptive par rapport aux états précédents ;

E_{10} générer des pensées et images mémorielles distinctes ;

E_{11} générer une expérience subjective distincte.

Ces critères permettent à Ekman de déterminer sept émotions dites basiques, universelles et présentes chez d'autres espèces que l'homme. Ekman propose donc **la joie**, **la colère**, **la peur**, **le dégoût**, **la surprise**, **la tristesse** et **le mépris** [53]. Les six premières sont connues comme les "*basic six*" d'Ekman et servent de base à de nombreuses études en informatique affective. Ekman étend par la suite les émotions considérées comme basiques et pose le concept de familles d'émotions [49]. Ces familles d'émotions sont un raffinement des sept émotions basiques évoquées ci-dessus : l'amusement, la colère, le mépris, le contentement, le dégoût, l'embarras, l'excitation, la peur, la culpabilité, la fierté de la réussite, le soulagement, la tristesse, la satisfaction, le plaisir sensoriel et la honte[6]. Ces différents termes représentent donc des familles d'émotions et donnent le thème de chaque famille. Les différentes émotions d'une même famille sont des variations d'intensité de l'émotion principale, représentant la famille. Par exemple, la colère est déclinée en une famille regroupant d'autres labels représentant ses différentes intensités, comme la rage (forte intensité) ou l'agacement (faible intensité).

[6] "...amusement, anger, contempt, contentment, disgust, embarrassment, excitement, fear, guilt, pride in achievement, relief, sadness/distress, satisfaction, sensory pleasure, and shame."

Les deux théories vues dans cette section et leurs modèles correspondants peuvent à notre sens être vues comme complémentaires. En effet, la théorie de Scherer se penche sur l'émergence d'une émotion chez un individu, tandis que la théorie d'Ekman se penche sur l'apparition des émotions dans l'espèce humaine. Ainsi, nous retenons chacune des deux théories et modèles correspondants. De plus, les différents critères $\{E_1 \dots E_{11}\}$ d'Ekman peuvent se rapporter aux critères de Scherer pour la caractérisation d'une émotion (voir partie 1.1.1, paragraphe page 11), ainsi qu'aux composants vus au paragraphe 1.2.2 (page 15). Les onze critères d'Ekman servent en effet trois buts. Le premier est d'identifier un état affectif donné comme étant une émotion : les critères E_7 et E_9 peuvent se rapporter au critère de "Rapidité de changement" de la table 1 de caractérisation d'une émotion (page 11). Le critère de durée (E_8) se retrouve tel quel dans ces critères, de même que celui concernant le déclenchement d'une évaluation (E_3). On retrouve également une idée du composant expression motrice (voir le modèle à composant partie 1.2.2, page 15) dans le critère E_1, et du composant d'activation physiologique dans le critère E_2. Le deuxième but est de caractériser chaque nouvelle émotion candidate au titre d'émotion basique comme suffisamment distincte des autres émotions basiques. Enfin, le troisième but est de tester le caractère universel et inter-espèce de l'émotion candidate, afin d'écarter les émotions qui pourraient venir de la personnalité ou de la culture d'un individu : l'on retrouve donc ces émotions chez d'autres animaux, notamment les primates. Malheureusement, ces émotions basiques sont très rarement exprimées de façon pure, du moins dans un contexte d'expérience [110].

Le modèle à composants de Scherer s'inscrit dans la continuité de la définition et caractérisation d'une émotion retenue à la section 1.1. Elle offre également une vision englobante de l'émotion en décrivant les processus internes de l'expérience émotionnelle. Elle connaît un succès grandissant dans le domaine de l'informatique affective. Elle décrit des mécanismes *a priori* adaptables à l'informatique. Le modèle des émotions basiques d'Ekman quant à lui fournit six émotions de base dont l'universalité est prouvée. Ceci offre aux informaticiens six catégories distinctes, ce qui rend possible la classification d'expressions émotionnelles parmi ces catégories. De plus, l'universalité de ces émotions permet, dans des cas d'expressions spontanées et intenses, de s'affranchir des biais posés par la teinte personnelle et culturelle de l'expression émotionnelle. Ainsi, pour ces six émotions de base nous retenons la théorie de Scherer pour leur apparition chez un être humain, tout en gardant leur propriété d'universalité démontrée par Ekman.

1.2.3 Évolution d'une émotion au cours du temps et transitions entre émotions

Dans notre travail de recherche, nous nous sommes focalisés sur la reconnaissance d'émotions à un instant ou sur une période de temps donnés. Nous nous sommes limités à ne pas considérer le problème de l'évolution d'une émotion au cours du temps ni de la modélisation de la transition entre deux expériences émotionnelles. Ces problématiques sont cependant intéressantes à aborder dans le cadre de travaux futurs. Nous présentons donc ici la métaphore de la cloche de Picard [114], qui propose une modélisation de l'évolution d'une émotion au cours du temps, et la transition en hystérésis de Scherer [131]. Ces modèles datent de la fin des années 90, et ouvrent des perspectives sur la façon de considérer l'émotion d'un point de vue temporel dans un système informatique.

Évolution d'une émotion au cours du temps : la métaphore de la cloche de Picard [114]

Picard utilise la métaphore de la cloche pour illustrer les propriétés d'un système émotionnel et en particulier l'évolution de ces émotions au cours du temps. Ces propriétés sont les suivantes :

- **Affaiblissement de la réponse au cours du temps.** Une cloche est activée par un choc et produit alors un son qui augmente rapidement en intensité avant de s'affaiblir progressivement. De même, une émotion est provoquée par un stimulus. Une réponse émotionnelle est de courte durée ; elle monte rapidement en puissance avant de s'affaiblir progressivement jusqu'à un niveau imperceptible (voir figure 2).
- **Coups répétés.** Si l'on frappe une cloche de façon répétée, le son produit augmente en intensité. Une série de chocs d'importance moyenne peuvent provoquer un son bien plus fort qu'un unique choc fort. Un sujet subissant des stimuli répétés provoquant la même émotion verra l'intensité de cette émotion augmenter (voir figure 2).
- **Activation et Saturation.** Un choc trop faible ne fera pas sonner la cloche ; de plus, une cloche ne peut sonner qu'au dessous d'une certaine intensité qu'elle ne peut dépasser. De même, des stimuli émotionnels de trop faible intensité ne provoqueront pas de réactions émotionnelles chez un sujet ; un stimulus de très forte intensité, ou une série de stimuli, ne pourront provoquer une réponse émotionnelle d'intensité supérieure à un certain seuil.
- **Influence de la personnalité et du tempérament.** Le tempérament et la personnalité d'une personne influencent ses réponses émotionnelles et les transitions entre ces réponses. Dans la métaphore de la cloche, les propriétés physiques de la cloche influencent l'intensité du son produit par un choc et les seuils d'activation et de saturation.
- **Linéarité.** Le système émotionnel humain est non-linéaire. Cependant, Picard suppose la possibilité de trouver des relations de linéarités entre certaines informations d'entrée (caractéristiques de stimuli émotionnels) et informations de sortie (caractéristiques de l'expression émotionnelle).
- **Invariance selon le temps.** Sous certaines conditions, le système émotionnel humain peut être considéré comme temporellement invariant. En général, un même stimulus provoquera une réponse émotionnelle similaire. Le critère E_4 d'Ekman (paragraphe 1.2.2 page 17) impose d'ailleurs à des émotions différentes d'être élicitées par des stimuli différents. Cependant, un effet d'habitude peut se mettre en place chez le sujet ; pour provoquer une réponse émotionnelle similaire, deux stimuli doivent satisfaire aux mêmes critères de nouveauté. Définir les conditions et les limites dans lesquelles on peut considérer un système émotionnel comme étant temporellement invariant et linéaire semble extrêmement difficile, mais présente l'avantage d'offrir un cadre de modélisation simplifié.
- **Retour physique et cognitif.** Les stimuli provoquant des réponses émotionnelles peuvent être des processus internes, qu'ils soient cognitifs ou physiques. Ainsi, l'expression physique d'une émotion peut agir comme un stimulus provoquant une autre réaction émotionnelle : en effet, la théorie de Scherer stipule que les composants sont interdépendants et qu'un changement d'état dans un composant peut influer les autres composants.
- **Impact de l'humeur.**Tous les paramètres perçus par le sujet, qu'ils soient internes ou externes, influent sur l'humeur du sujet à un instant donné. L'humeur est un état

affectif de durée moyenne (de l'ordre de quelques heures), également influencée par le tempérament et la personnalité du sujet. L'humeur d'un sujet influence ses réactions émotionnelles.

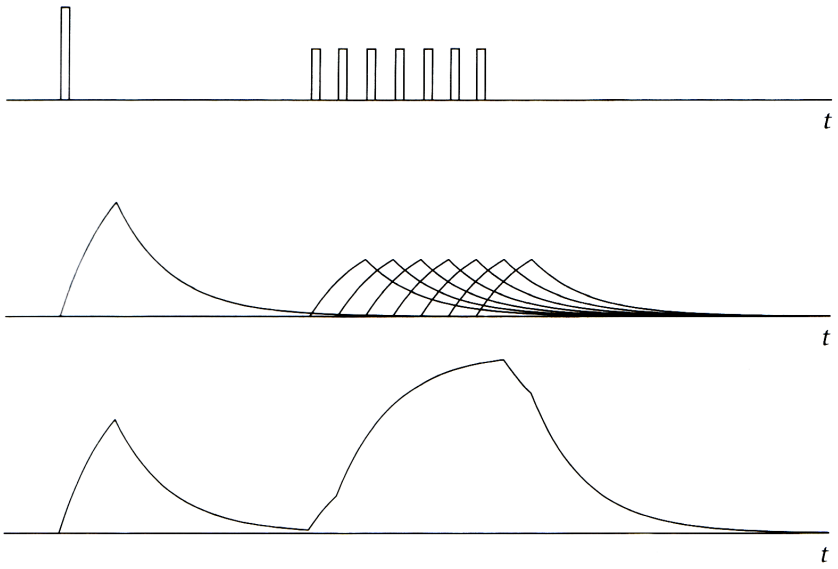


FIG. 2: Métaphore de la cloche. En haut : "coups" ou stimuli émotionnels de différentes intensités. Au milieu : réponses émotionnelles résultantes. En bas : somme des réponses émotionnelles. Figure tirée de [114].

Picard propose ainsi une équation de type "non-linéarité sigmoïdale" (voir figure 3) pour représenter la transition vers une réponse émotionnelle. Cette courbe illustre les différentes propriétés évoquées ci-dessus. En particulier, la sigmoïde élimine les réponses aux stimuli de trop faible intensité et sature les réponses aux stimuli de trop forte intensité. La zone de la courbe entre activation et saturation est la zone de transition.

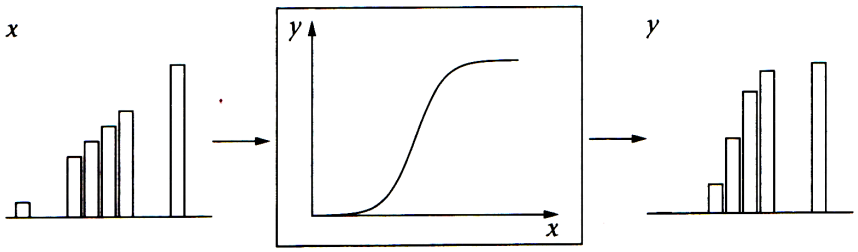


FIG. 3: Sigmoïde représentant la génération d'une émotion et son intensité en fonction de l'intensité du stimulus. Figure tirée de [114].

La transition en hystérésis de Scherer [131]

Scherer propose une autre courbe illustrant les transitions entre émotions. Il s'agit d'une hystérésis (voir figure 4). L'hystérésis implique que le chemin pour transiter d'une émotion à une autre est différent selon le sens dans lequel il est parcouru. Nous considérons la courbe en deux dimensions seulement afin d'illustrer le concept de façon simplifiée ; Scherer propose également un modèle de transition surfacique que nous n'abordons pas ici.

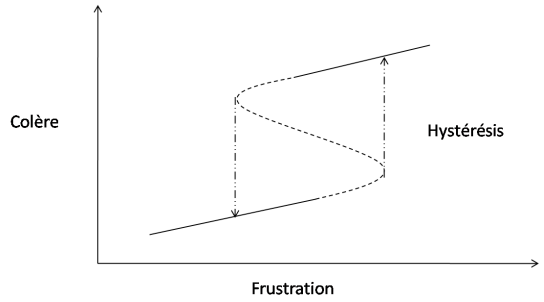


Fig. 4: Illustration d'une hystérésis dans une relation entre frustration et colère. Figure tirée de [131].

Cette hystérésis est présente pour les émotions à forte activation. Compatible avec la sigmoïde précédente, cette hystérésis souligne néanmoins que les chemins sont différents selon le sens de la transition. Cette courbe présente donc une zone de "repli" (en pointillé sur la figure 4) dans laquelle il est impossible de se trouver. En prenant l'exemple d'un sujet soumis à une frustration croissante :

- en faisant monter progressivement la frustration du sujet, on augmente sa colère. A un certain stade S1 de frustration, on arrive à une cassure dans l'intensité : le sujet explose de colère. L'intensité de celle-ci monte jusqu'à atteindre son maximum.
- on fait maintenant redescendre la frustration du sujet. Celui-ci est toujours dans une colère intense, qui décroît petit à petit. Arrivé au stade de frustration S1, sa colère est largement supérieure au cas précédent. Il faut continuer de réduire la frustration jusqu'à un stade S2, inférieur à S1, pour observer finalement la colère du sujet retomber et revenir au niveau de départ (c'est-à-dire calme).

1.2.4 Conclusion sur les théories de l'émotion

Dans le cadre de ce mémoire, nous retenons donc particulièrement le modèle à composants (CPM) développé par Scherer ainsi que les émotions basiques d'Ekman. Selon le modèle de Scherer, un stimulus, qu'il soit interne ou externe, est séquentiellement évalué selon une série de critères (*Stimulus Evaluation Checks* ou SECs). La séquence d'évaluation va ensuite provoquer des micro-changements d'états dans cinq composants : le composant "traitement cognitif du stimulus", le composant "processus neurophysiologiques", le composant "motivation et tendances d'actions", le composant "expression motrice", et le composant "ressenti-feeling". Le modèle à composants de Scherer offre une modélisation du fonctionnement interne de

l'expérience émotionnelle en mettant en avant la notion de processus émotionnel. Cette modélisation s'affranchit en outre de l'utilisation de mots pour décrire l'émotion, évitant ainsi les problèmes inhérents à ce genre de représentation que nous verrons dans la partie 1.3.1. Elle propose par contre une représentation complexe de l'émotion : une émotion y est en effet définie par une séquence de changements d'états dans les composants. Cette représentation est bien plus complexe à utiliser en informatique que les émotions basiques d'Ekman. Une solution est de considérer, comme Lisetti *et al.*, que les réponses aux SECs caractérisent totalement une expérience émotionnelle [89], et de représenter une émotion par les réponses aux SECs au lieu des changements d'états dans les composants.

La théorie évolutionniste considère que les émotions sont des processus physiques, physiologiques et mentaux nous permettant de mieux réagir à une situation donnée. Paul Ekman a isolé plusieurs familles d'émotions selon une série de critères permettant d'établir le caractère interculturel voire inter-espèce de certaines émotions. Nous retrouvons des similarités entre les critères posés par Ekman pour isoler les émotions basiques et les critères de Scherer de catégorisation d'un état affectif en tant qu'émotion. Le modèle des émotions basiques d'Ekman fournit un cadre moins complexe à appréhender que celui offert par Scherer puisque chacune des émotions basiques est représentée par une catégorie. Ces émotions basiques sont universelles. Se baser sur ce modèle permet donc de s'affranchir des biais amenés par les différences entre individus et les différences culturelles, propres à ce type de représentation.

Enfin, nous avons présenté deux modèles de transition entre émotions : la métaphore de la cloche de Picard se traduit par une sigmoïde permettant de déterminer l'intensité d'une émotion par rapport à l'intensité d'un stimulus. La transition en hystérésis de Scherer propose un repli dans la transition, ayant pour effet des chemins différents lors des passages d'une émotion à une autre et vice-versa.

En conclusion, nous considérons donc que les théories et modèles de Scherer et Ekman sont complémentaires. Le modèle à composant de Scherer décrit un processus liant un stimulus à l'apparition d'une émotion. La théorie d'Ekman et son modèle à base de familles d'émotions permet d'isoler des émotions basiques universelles. L'apparition de ces émotions basiques suite à un stimulus peut être décrite grâce au processus proposé par Scherer. Scherer, dans sa théorie, n'exclue pas l'existence d'émotions universelles dont l'évaluation est presque "automatique" : par exemple, une évaluation forte du critère de nouveauté (pouvant être provoquée par un bruit fort et soudain comme une détonation) entraine une réaction de surprise (sursaut) avant que la situation ne soit totalement évaluée. De même, Ekman inclue l'évaluation du stimulus dans ses critères discriminant les émotions basiques.

1.3 Modèles de représentation applicables aux systèmes informatisés

Scherer propose dans [132] une description des différents modèles de représentation de l'émotion. Nous décrivons ici les trois types utilisés en reconnaissance d'émotions par informatique : les modèles discrets, les modèles dimensionnels continus, et les modèles à composants. Les deux premiers types de modèles sont ceux que l'on rencontre dans la quasi-totalité des systèmes de reconnaissance affective par ordinateur. L'utilisation de modèles discrets est de loin majoritaire dans ces études ; certaines se basent sur des modèles dimensionnels. Le modèle à composants (CPM) ne propose pas directement de représentation de l'émotion, ce qui freine son adoption en reconnaissance.

1.3.1 Modèles de représentation discrets

Les modèles discrets représentent un ensemble d'émotions comme un ensemble discret, où chaque type d'émotion est désignée par un label spécifique (c'est-à-dire un mot - par exemple "joie", "peur", etc.). Un exemple courant de ce type de modèle est l'ensemble des émotions basiques proposées par Ekman [49] (décrit au paragraphe 1.2.2 page 17). Chaque émotion, en tant que réponse typique de l'organisme et de notre système nerveux central, est ici catégorisée par le terme qui lui correspond. Cette approche catégorielle propose un avantage certain en informatique. En effet, reconnaître une émotion d'un ensemble discret revient à choisir une catégorie (i.e. une émotion) parmi celles proposées dans l'ensemble ; et la catégorisation de signaux est un problème largement étudié en informatique, qui dispose d'outils adéquats à sa résolution (comme les réseaux de neurones).

Les modèles discrets comportent cependant des défauts. Premièrement, la labellisation des émotions sous-entend qu'à chaque label correspond un état émotionnel défini. Or, nous avons vu en 1.2.2 qu'une émotion peut être vue comme un processus émotionnel dynamique plutôt que comme un état émotionnel statique. L'utilisation de labels ne permet pas de qualifier certaines subtilités de l'émotion (par exemple son intensité) ni de relier entre elles les différentes variantes d'une émotion (par exemple, la colère, la rage, l'agacement), d'où l'introduction de familles d'émotions (voir page 17). Un recensement aboutit à un grand nombre de labels, relatifs à un langage propre : Scherer dénombre ainsi plus de 500 termes relatifs à une émotion ou un état affectif en anglais, et plus de 200 en allemand [129]. Enfin, cette labellisation est sujette à un idiome particulier. La traduction peut parfois porter à confusion et ne pas refléter le même état interne. En effet de nombreuses émotions (hormis les émotions basiques) sont potentiellement sujettes à la culture et aux différences entre individus.

Comme nous l'avons dit précédemment, les modèles discrets d'émotions, de par leur nature, se prêtent particulièrement bien à la reconnaissance d'émotions par l'ordinateur. Ils sont donc largement utilisés dans le domaine de la reconnaissance d'émotions [132]. En 2009, cette assertion est toujours largement vérifiée [152].

1.3.2 Modèles de représentation continus

L'ensemble des émotions peut être considéré comme un espace continu, comportant *a priori* un nombre élevé de dimensions. Les efforts se tournent donc vers l'identification d'un sous-ensemble de dimensions, de façon à ce que les dimensions retenues permettent de définir chaque émotion sans ambiguïté, en minimisant la corrélation entre ces différentes dimensions. Les espaces obtenus les plus célèbres sont bidimensionnels (plan) ou tridimensionnels (volume).

Le modèle bidimensionnel le plus connu est l'espace "plaisir-activation" proposé par Russell dans [125] (*valence-arousal space*)(voir figure 5). Le premier axe est celui du plaisir. Plus on s'éloigne à droite de l'origine (valeurs "positives"), plus l'émotion est plaisante. Les valeurs négatives transcrivent des émotions déplaisantes. Le deuxième axe est celui de l'activation : il correspond à une tendance à l'action dans la réponse émotionnelle de l'individu. Les émotions sont donc distribuées dans cet espace bidimensionnel dans lequel on peut distinguer les quatre quadrants : les émotions à valence positive et forte activation (par exemple la joie), les émotions à valence positive et faible activation (par exemple la relaxation), les émotions à valence négative et faible activation (par exemple l'ennui) et les émotions à valence négative et forte activation (par exemple la peur). Dans cet espace, il s'avère que les émotions basiques sont distribuées selon un cercle de centre l'origine du repère. Cet espace est donc qualifié de circomplexe. La distance à l'origine qualifie l'intensité de l'émotion. Une représentation discrète des émotions peut ainsi être plongée dans l'espace continu "plaisir-activation". Ce modèle bidimensionnel a cependant été critiqué sur les ambiguïtés qu'il induit dans les positions de certaines émotions sur les axes. En particulier, la colère et la peur sont deux émotions à valence fortement négative et à forte activation : elles se retrouvent donc dans la même zone de l'espace.

Le modèle tridimensionnel a été proposé pour remédier à ce problème. L'axe "dominance-soumission" véhicule la notion de contrôle sur les évènements extérieurs. Ainsi, la colère est une émotion à forte dominance, alors que la peur est caractérisée par une forte soumission. Mehrabian introduit ainsi l'espace Plaisir- Activation- Dominance (*Pleasure- Arousal- Dominance* ou PAD) comme modèle tridimensionnel représentant l'espace des émotions [96]. Cet espace tridimensionnel est découpé en octants. Une liste de labels d'émotions est proposée dans [126], avec leurs coordonnées dans l'espace PAD, permettant ainsi de plonger ces émotions dont les coordonnées sont identifiées dans l'espace continu à trois dimensions PAD. L'espace PAD a ainsi été utilisé dans diverses applications de test, comme prédire l'attraction physique, la désirabilité d'un nom ou des préférences de produits. L'espace PAD est également proposé comme espace de représentation des tempéraments [95].

1.3.3 Modèle de représentation basé composants

Le modèle à composants de Scherer offre une modélisation des processus internes d'une expérience émotionnelle. Cette modélisation plus complexe, issue du domaine de la psychologie, ne propose pas de représentation d'une émotion directement applicable à la reconnaissance d'émotions en informatique. Cette complexité des évaluations d'un évènement, les paramètres à prendre en compte pour connaître les réactions d'un composant et cette absence de représentation directe constituent un frein à une adoption plus large du modèle à composants en

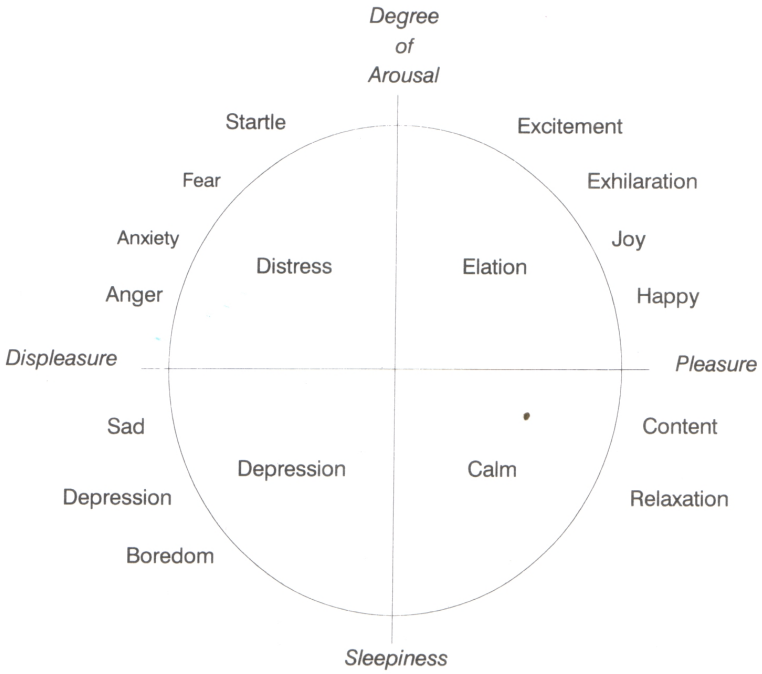


Fig. 5: L'espace plaisir-activation, un espace circomplexe des émotions. Figure tirée de [120].

informatique. L'objectif serait ici de pouvoir représenter toute la séquence de changements d'états dans les divers composants pour représenter une émotion.

1.3.4 Intérêt de ces modèles de représentation

Nous avons vu dans cette section trois façons de représenter les émotions. Les modèles discrets se basent sur des labels pour représenter chaque émotion et chaque variation. Le problème majeur de ce type de modèle vient du nombre de labels utilisés pour couvrir un maximum d'états émotionnels ; de plus, chaque label étant un mot d'un langage particulier, il est difficile d'assurer que la traduction d'un terme d'une langue à une autre définit bien la même émotion. Il peut également être difficile d'inférer des transitions entre les différentes émotions. Les modèles continus à deux ou trois dimensions permettent de lever cette limitation en proposant un espace dans lequel on peut se déplacer librement. L'effondrement de l'espace multidimensionnel des émotions en un espace bi- ou tridimensionnel suppose par contre une perte d'information et n'est pas adapté à toutes les situations. En particulier, Scherer argumente l'inadaptation d'un tel modèle à l'évaluation d'émotions esthétiques [134]. Enfin, le modèle à composants complexifie la représentation d'une émotion.

Ces trois modèles de représentation de l'émotion sont utilisés dans les travaux de reconnaissance automatique des émotions dans la littérature, travaux présentés au chapitre 2. Dans le cadre ce mémoire, nous prenons en compte la prépondérance de ces trois modèles en reconnaissance d'émotions pour la conception de notre modèle d'architecture présenté au

chapitre 5.

1.4 Expression de l'émotion

Dans cette section, nous nous attachons à la fonction de communication de l'émotion. La communication émotionnelle est dérivée de deux fonctions et de leurs composants correspondants : la fonction "régulation du système" (composant "processus neurophysiologiques" et la fonction "communication d'intention" (composant "expression motrice") (voir table 3, page 16). Chaque expérience émotionnelle va déclencher des signaux moteurs et physiologiques. Dans sa caractérisation des émotions basiques, Ekman impose d'ailleurs comme critère que chaque émotion doit avoir une expression physique et physiologique distincte. Selon Mehrabian [97], 93% de la communication est non-verbale et la façon la plus expressive de faire passer des émotions est par les expressions faciales et la gestuelle. Les expressions émotionnelles non-verbales peuvent être si subtiles qu'elles ne sont ni codées ni décodées de façon intentionnelle ni même consciente [3].

1.4.1 Les canaux de l'expression

Picard évoque dans [114], les notions de bande passante affective (*affective bandwidth*) et de canal de communication émotionnelle. Notre corps est en effet capable de transmettre de l'information par divers canaux ; la communication émotionnelle est effectuée via ces différents canaux :

- les expressions faciales ;
- les intonations de la voix ;
- les réactions du Système Nerveux Autonome (*Autonomous Nervous System* ou ANS) ;
- les positions et mouvements du corps.

Les expressions faciales constituent le moyen le plus évident de l'expression émotionnelle. Lors de l'expérience d'une émotion par un individu, des muscles spécifiques sont activés permettant par exemple de sourire ou de froncer les sourcils. Ces réactions ont pour but principal de communiquer l'émotion ressentie. Les signaux de la voix peuvent provenir à la fois d'une expression motrice (modulations, fréquence utilisée) et d'un processus neurophysiologique (par exemple, lorsqu'on a la gorge serrée par la tristesse). Les réactions du système nerveux autonome sont les réactions neurophysiologiques. Il s'agit, par exemple, d'une augmentation du rythme cardiaque, de la rétractation des vaisseaux pour mieux irriguer les muscles, ou de la sudation. Enfin, les positions prises par l'individu et les mouvements effectués sont également porteurs d'information émotionnelle. De l'extérieur, nous sommes capables de percevoir ces signaux chez un individu faisant l'expérience d'une émotion. Bien que la communication émotionnelle passe par chacun de ces canaux, ces derniers ne sont naturellement pas exclusivement dédiés à la communication émotionnelle : par exemple, parler et articuler déforme notre visage sans que ce soit à vocation de communication émotionnelle. Cela pose la difficulté de percevoir quelles caractéristiques expressives sont relatives à l'émotion. De plus, nous sommes capables de cacher nos émotions ou de contrefaire des émotions, c'est à dire d'exprimer des émotions que nous ne ressentons pas forcément. Morris a établi dans [101] une hiérarchie de confiance dans ces différents canaux de communication émotionnelle, du plus crédible au moins crédible :

1. signaux d'expression physiologique ;
2. positions et gestuelle ;
3. expressions faciales ;
4. expressions verbales.

Les signaux physiologiques sont très difficiles voire impossibles à contrefaire, bien que nous puissions être conscients de tels signaux. Par exemple, un individu rougissant de honte aura conscience de son rougissement mais ne pourra l'arrêter volontairement. De même, le rythme cardiaque est quasiment impossible à contrôler. Certaines intonations de la voix peuvent se situer dans cette catégorie, la réaction émotionnelle mettant en jeu les muscles utilisés pour parler ainsi que les cordes vocales. Ainsi, il peut être très difficile de cacher les effets d'une profonde tristesse dans notre voix.

Le canal de communication de la position et de la gestuelle peut encore être subdivisé selon le critère de crédibilité. Les mains sont les indicateurs les moins crédibles, ensuite le buste, et enfin les jambes et les pieds. En effet, ces derniers sont loin du centre d'attention lors d'une communication émotionnelle (le visage). Nous sommes donc moins habitués à les contrôler lorsqu'il s'agit d'exprimer une émotion non ressentie. Le buste reflète le tonus musculaire du corps ; il est difficile de le maintenir dans un état contraire à ce qui est ressenti (par exemple se tenir droit pour simuler l'intérêt lorsqu'on est ennuyé). Enfin, les mains croisent sans arrêt notre regard lorsque nous communiquons : nous sommes donc plus à même de penser à les contrôler.

Enfin les expressions faciales et verbales sont des indicateurs peu dignes de confiance. Les expressions faciales étant le principal canal de communication émotionnelle, nous apprenons rapidement à les contrôler. L'expression verbale (ce que nous disons) est tout à fait contrôlable. Cette expression verbale est à différencier de l'expression non verbale de la voix : intonations, fréquences, etc., qui peuvent relever de processus neurophysiologiques.

1.4.2 Les variables captées pour reconnaître les émotions

La reconnaissance d'émotions consiste à se baser sur une perception des signaux émotionnels émis par un individu pour en inférer son expérience émotionnelle. Dans la communication humaine, nous le faisons chaque jour, de façon non forcément consciente [3]. La recherche s'applique donc à identifier les caractéristiques pertinentes pour l'émotion. Les variables non-verbales sont classées en trois catégories : les variables motrices, les variables distales et les variables proximales [137]. Les variables motrices sont les mesures de l'activité corporelle d'un sujet. Par exemple, la tension de certains muscles, l'orientation des différents segments du corps humain, ou les battements du cœur. Les variables motrices peuvent être difficiles à capturer, et souvent par des dispositifs intrusifs qui limitent leur utilisation aux expériences en laboratoire. Les variables distales sont la mesure des signaux produits par le corps et perçus par un observateur avant toute évaluation cognitive de la part de cet observateur. On peut par exemple enregistrer le son de la voix grâce à un microphone, ou enregistrer des mouvements grâce à une caméra vidéo. Enfin, les variables proximales sont liées à l'interprétation d'un

observateur, et sont relevées par exemple grâce à des entretiens ou des questionnaires. Lors de la récolte de variables non verbales dans le cadre d'une expérimentation, il est possible de croiser des variables motrices, mesurant l'activité corporelle et physiologique, des variables distales, mesurant le sujet de l'extérieur mais de façon objective, et des variables proximales, passant par le filtre cognitif d'un évaluateur.

Il existe sur chacun des canaux d'expression de l'émotion une "bande passante émotionnelle", sur laquelle sont transmises les informations relatives à l'émotion que nous ressentons ou à l'émotion que nous souhaitons exprimer. Lorsque nous voulons cacher une émotion ressentie, la synchronisation émotionnelle lutte contre le contrôle que nous avons de nos expressions faciales, de nos mouvements, etc. Lorsque nous voulons feindre une émotion non ressentie, nous ne pouvons atteindre par le contrôle conscient de notre corps la même synchronisation que celle apportée par une émotion réellement ressentie (nous mettons ici de côté les techniques d'acteurs visant à jouer une émotion le plus fidèlement possible). Ces deux cas de figure induisent une dissonance entre les différents signaux émotionnels des différents canaux ; l'échelle de confiance présentée dans le paragraphe 1.4.1 permet de prendre en compte des signaux plus crédibles.

1.5 Conclusion

Nous avons présenté, dans ce chapitre, les bases du domaine de la psychologie concernant l'émotion. Nous avons tout d'abord choisi dans la littérature, la terminologie que nous adoptons dans nos travaux. Nous choisissons le terme "état affectif" comme terme général se rapportant à tout ce qui a trait à l'affect, et comprenons l'émotion comme une catégorie d'état affectif. Nous nous plaçons dans le cadre de la théorie de l'évaluation cognitive dans notre choix de définition, caractérisation, et théorie de l'émotion. Cependant, la théorie évolutionniste et en particulier les émotions basiques d'Ekman présentent un intérêt certain de par leur universalité, largement prouvée. Le modèle à composants de Scherer est un modèle de fonctionnement de l'émotion, et non un modèle de représentation. Il s'agit là de l'un des problèmes posés par ce modèle en informatique : comment représenter les données d'une émotion ? Dans la suite de ce mémoire, nous ferons appel à la théorie du modèle à composant dans les chapitres 4 et 5. Le modèle des émotions basiques sera utilisé dans les chapitres 4, 5 et 6.

Après avoir vu d'un point de vue psychologique ce qu'est une émotion, comment elle est générée, et comment elle peut être représentée, nous adressons dans le chapitre suivant un état de l'art des travaux en reconnaissance automatique des émotions. Ces travaux se basent sur les définitions, théories et modèles présentés dans ce chapitre. Comme nous l'avons vu, une émotion génère en particulier des réactions motrices et neurophysiologiques ; le but de la reconnaissance d'émotions est de partir de ces signaux générés par une émotion pour retrouver l'émotion exprimée, voire ressentie. En effet, nous avons vu que l'expression d'une émotion ne correspond pas forcément au ressenti du sujet, dans les cas d'une émotion feinte ou cachée. En tant qu'observateur, un système automatique n'a accès qu'aux expressions émotionnelles et non à la composante subjective de l'émotion. Cependant, certains capteurs permettent de

détecter des signaux impossibles à contrefaire (comme le rythme cardiaque). En théorie, on se rapproche donc de la reconnaissance d'émotions ressenties.

Chapitre 2

Systèmes de reconnaissance d'émotions

L'informatique affective cherche à accorder une place aux émotions et à l'affect dans le domaine de l'informatique et, plus largement, des machines contenant une partie logicielle comme les robots. En Interaction Homme-Machine, Picard scinde cette informatique affective en deux grands domaines : la synthèse et la reconnaissance d'émotions.

La synthèse d'émotions a pour but de générer des émotions virtuelles chez une machine à but de communication avec l'utilisateur. Ainsi, de nombreuses recherches ont été menées sur les avatars expressifs (voir par exemple [62], [77], [111]). Ces avatars sont des personnages en trois dimensions capables d'exprimer des émotions selon les différents canaux de communication affective, afin d'améliorer la communication avec un utilisateur humain.

La reconnaissance d'émotions, quant à elle, s'attache à déceler l'émotion exprimée voire ressentie chez un utilisateur humain afin de prendre en compte cet état émotionnel. Cette prise en compte peut être au cœur du système (par exemple dans ce mémoire nous présentons nos travaux de reconnaissance d'émotions sur la gestuelle d'un danseur - la reconnaissance étant le but primaire du système) ou à des fins d'amélioration de l'interaction (par exemple en détectant l'intérêt chez un apprenant [102]).

Picard propose dès 1997 une série d'étapes à suivre pour effectuer une reconnaissance affective [114] :

1. acquérir le signal en entrée : pour cela il est nécessaire de mettre en place des dispositifs d'acquisition ou de capture des données du monde réel. Par exemple, des microphones pour capter le signal sonore de la voix, ou une caméra pour le visage.
2. reconnaître des formes dans le signal, c'est-à-dire extraire des caractéristiques ou des variations typiques d'une émotion ou d'un état affectif.
3. raisonner, c'est-à-dire être capable, d'après les caractéristiques et formes reconnues, d'inférer l'émotion plus probablement exprimée par l'utilisateur.
4. apprendre, c'est-à-dire entraîner la machine à reconnaître et classer une émotion.

5. évaluer automatiquement le biais dans la reconnaissance.
6. délivrer l'émotion finalement interprétée.

Une version simplifiée de cette méthode se retrouve dans la majorité des études et systèmes de reconnaissance d'émotions. Classiquement, cette reconnaissance est effectuée en trois étapes. Tout d'abord, la capture de données depuis le monde extérieur, que ce soit un flux sonore, un flux vidéo du visage, des informations de positions du corps ou de réactions du système nerveux autonome (étape 1) ; ensuite, l'extraction de caractéristiques depuis les différentes données acquises (étape 2) ; enfin, la classification des valeurs de caractéristiques en une catégorie d'émotion (voir les états de l'art sur la reconnaissance d'émotions multicanaux [139] et [152]) (étape 3). La classification en catégories est une méthode d'interprétation que l'on retrouve souvent dans la littérature. En effet, le modèle discret des émotions, présenté au paragraphe 1.3.1 du chapitre 1, permet de considérer chaque émotion comme une catégorie ; un système de classification est ainsi capable de fournir l'émotion correspondant à un vecteur de valeurs de caractéristiques. La majorité des systèmes logiciels de reconnaissance d'émotions sont donc basés sur le modèle discret des émotions et en particulier sur les six émotions basiques d'Ekman (voir chapitre précédent). Une fois conçu, les performances d'un système de reconnaissance d'émotions sont évaluées face aux performances en reconnaissance d'un groupe d'humains.

Ce chapitre est structuré selon les différentes étapes de la reconnaissance. Dans la première section nous présentons deux axes permettant de situer les travaux en reconnaissance d'émotions. Dans la section suivante, nous présentons des dispositifs matériels spécifiques à la reconnaissance d'émotions ; nous focalisons en particulier sur les dispositifs portés ou ambiants. La troisième section traite des études basées sur un unique canal de communication émotionnelle, puis des études se basant sur plusieurs canaux. Nous présentons en quatrième section les questions posées par l'interprétation d'une émotion, c'est-à-dire sur l'inférence d'une émotion à partir de caractéristiques. Enfin, nous abordons la problématique de l'évaluation d'un système de reconnaissance d'émotions.

2.1 Deux axes pour les systèmes de reconnaissance d'émotions

Nous organisons les travaux sur les systèmes de reconnaissance d'émotions selon deux axes : l'axe reconnaissance générique / reconnaissance personnalisée et l'axe reconnaissance active / reconnaissance passive. Notons de suite que la majorité des systèmes de reconnaissance d'émotions sont de type générique et passif.

2.1.1 Axe reconnaissance générique / personnalisée

La reconnaissance d'émotions peut être générique ou personnalisée. Chaque choix apporte ses questions et ses problèmes. Le parallèle est souvent fait avec la reconnaissance vocale : pour s'adapter au plus grand nombre, un système doit être générique ; pour obtenir de meilleurs résultats, il est préférable qu'il soit paramétré en fonction de l'application et/ou de l'utilisateur, ou qu'il apprenne de cet utilisateur.

A l'exception de [67], la totalité des travaux que nous avons répertoriés en reconnaissance d'émotions s'orientent vers des systèmes génériques. De nombreux psychologues se sont penchés sur l'identification et la validation de caractéristiques émotionnelles dans la voix (par exemple [135]), le visage (par exemple [50]), ou les mouvements (par exemple [147], [46]). Toutes ces études proposent des caractéristiques génériques : celles-ci sont évaluées et validées par un nombre suffisant d'humains pour en faire une étude statistique. La reconnaissance générique consiste alors à traiter les informations acquises par les dispositifs de capture (caméra, microphone) pour en extraire des caractéristiques émotionnelles ne dépendant pas d'une personne en particulier. Dans ces études, les caractéristiques testées et leurs interprétations ne sont pas forcément interculturelles ; elles sont cependant génériques dans le sens où elles ne reflètent pas la manière de réagir d'un seul individu.

Une reconnaissance personnalisée signifie que l'on s'attache à rendre un système dépendant d'un unique utilisateur, et capable d'apprendre de cet utilisateur pour mieux détecter et reconnaître les émotions qu'il exprime. Potentiellement, un tel système est sans conteste plus précis et mieux adapté à l'utilisateur. Nous n'avons répertorié qu'une seule étude [67] visant l'adaptation du mécanisme de reconnaissance à l'utilisateur. En effet, notre compréhension des émotions, de l'impact de la personnalité et surtout de la construction de la personnalité à partir d'un entraînement préalable est très limitée dans le domaine de l'informatique. Au niveau technique, entraîner un système à reconnaître la personnalité d'un utilisateur nécessite que le système accompagne l'utilisateur dans diverses situations de vie génératrices d'émotions. L'informatique portée et mobile permet de remplir ce rôle. Picard et Healey proposent plusieurs dispositifs portés et mobiles pour une reconnaissance d'émotions [117], mais ne détaillent pas leur utilisation dans un système complet de reconnaissance, ni si le système est générique ou personnalisé.

2.1.2 Axe reconnaissance active / passive

Nous reprenons du domaine de l'interaction multimodale les notions d'interaction active et passive décrites dans [15], où une modalité est dite active si elle requiert une action explicite de l'utilisateur pour communiquer ou percevoir les données (taper une commande, regarder un écran mobile) ; elle est dite passive si l'interaction ne requiert pas l'attention ni d'action explicite de l'utilisateur (localisation GPS). Nous étendons donc ces notions à celles de reconnaissance active et passive d'émotions.

La reconnaissance active sous-entend une action initiatrice de la part du système ; simplement, le système demande à l'utilisateur son état affectif courant. L'utilisateur répond alors par l'interaction proposée par le système. Reynolds *et al.* [123] proposent ainsi un système où l'utilisateur peut à tout moment, à l'aide d'un curseur (*slider*), signifier au système son degré de frustration. Ce mécanisme est intéressant au niveau interaction entre l'homme et la machine, car il montre que signifier sa colère ou sa frustration à la machine permet également d'en faire baisser l'intensité (mécanisme de *venting*). Ce mécanisme est réutilisé par Klein *et al.* en 2002 pour permettre à l'utilisateur de se remettre d'états émotionnels négatifs [83].

La reconnaissance passive, au contraire, ne sollicite pas l'utilisateur. Un système de reconnaissance passive observe continuellement l'utilisateur grâce à des dispositifs de capture divers (caméra, microphone, combinaison de capture du mouvement, capteurs musculaires...). Les données sont acquises et traitées (à la volée ou en différé) pour en extraire des caractéristiques et en inférer des émotions. Comparés aux systèmes à reconnaissance active, ces systèmes sont cognitivement non-intrusifs puisqu'ils ne sollicitent pas activement l'utilisateur (bien sûr, ils peuvent se révéler intrusifs selon d'autres aspects, en particulier par les dispositifs de capture déployés). La majorité des systèmes existants reposent sur une reconnaissance passive.

A quelques exceptions près [67] [123] [83], la quasi-totalité des systèmes que nous avons rencontrés sont des systèmes passifs et génériques. Nous nous limitons donc à ce type de systèmes pour la conception de notre modèle d'architecture. Le fait de ne considérer que les systèmes passifs implique en particulier certaines hypothèses que nous utilisons aux chapitres 4 et 5.

2.2 Capteurs utilisés pour la reconnaissance d'émotions

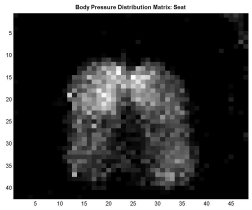
La machine dispose d'une pléthore de capteurs permettant de mesurer les comportements de l'utilisateur. Le dispositif de capture le plus couramment utilisé pour la reconnaissance d'émotions est la caméra vidéo. En effet, la caméra présente l'avantage d'être non intrusive et peu onéreuse ; le domaine de l'informatique dispose en outre de toute une batterie d'outils de traitement d'image permettant le suivi d'un visage, la détection des points d'intérêt, et un suivi du corps dans l'espace. La caméra vidéo est ainsi parfaitement adaptée à la reconnaissance d'expressions faciales. Elle est également utilisée pour la reconnaissance à partir du corps (voir par exemple [20]). Le microphone est bien sûr constamment utilisé pour la capture de la voix. Il existe de nombreux autres capteurs : combinaisons de capture du mouvement pour enregistrer les mouvements d'un individu, capteurs de pression sanguine, électrocardiogrammes, électro-encéphalogrammes, électromyogrammes, capteurs d'expansion de la cage thoracique pour mesurer la respiration... etc.

Parmi tous ces dispositifs, nous étudions ici ceux qui s'inscrivent dans l'informatique ubiquitaire et portée. En effet, dès son livre pionnier *Affective Computing*, Picard défend l'idée que l'informatique portée et pervasive peut clairement apporter au domaine de la reconnaissance d'émotions. L'informatique pervasive, par l'ajout de capteurs à des objets existants, permet une mesure non intrusive dans des conditions de vie réelle. L'informatique portée possède également cet avantage, mais y ajoute celui d'accompagner l'utilisateur dans de nombreuses situations de la vie réelle. Le but de ce paragraphe est également de montrer que les capteurs mis en œuvre pour la reconnaissance d'émotions ne sont ni forcément complexes ni intrusifs ; des capteurs simples peuvent être conçus, qui à eux seuls ne fournissent pas assez d'information mais qui, utilisés en combinaison, peuvent permettre d'effectuer une reconnaissance.

La *sentic mouse* [82] et la *pressure chair* [102] sont deux exemples de dispositifs ubiquitaires pour la reconnaissance d'émotions. La *sentic mouse* est une souris sur laquelle est rajouté un capteur de pression directionnelle sur le bouton gauche (figure 6a). En effet, il apparaît que

la direction du clic donne une indication sur la valence de l'état émotionnel de l'utilisateur : un clic où le bouton est attiré à soi tend à indiquer une valence positive, tandis qu'un clic où le bouton est repoussé tend à indiquer une valence négative. La *pressure chair* est un fauteuil de bureau où sont placées des matrices de capteurs de pression dans l'assise et le dossier. Les matrices de valeurs de pression ainsi obtenues permettent de déterminer la position de l'utilisateur lorsqu'il est assis dans la chaise (figure 6b). De là, une analyse peut être faite pour détecter un état affectif comme l'intérêt.

(a) La *Sentic Mouse*



(b) Matrice de pression de la *pressure chair*

Fig. 6: Exemples de dispositifs ambiants pour la capture de signaux portant de l'information émotionnelle. Figures tirées de [82] et [102].

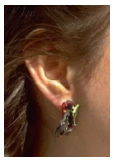
De tels dispositifs illustrent la possibilité de transformer certains objets de tous les jours pour leur permettre de capter de signaux permettant de reconnaître des émotions. Cette approche présente de nombreux intérêts. Tout d'abord, il n'est pas nécessaire de modifier l'aspect de ces objets, ni leurs affordances, c'est-à-dire les fonctions que leur forme suggère (de façon intrinsèque ou en se basant sur l'apprentissage préalable de l'utilisateur). Par exemple, il est possible de cacher le capteur de pression directionnelle dans la *sentic mouse* ; elle se présente alors comme une souris traditionnelle, et l'utilisateur la manipulera comme telle, sans savoir que la valence de ses états affectifs est mesurée (ce qui impacte l'évaluation en y enlevant un biais comme nous le décrivons dans la prochaine section).

Picard *et al.* décrivent dans [119] plusieurs dispositifs portés pour la capture d'émotion. Les auteurs y présentent une chaussure munie d'un capteur mesurant la conductivité de la peau (figure 7a), une boucle d'oreille munie d'un capteur de pression sanguine (figure 7b), et une paire de lunettes munie de capteurs de tension musculaire, permettant ainsi de détecter les froncements de sourcils (figure 7c).

Les systèmes portés (comme la boucle d'oreille) offrent l'avantage d'accompagner potentiellement l'utilisateur sur des périodes de plusieurs heures, dans diverses situations de vie, non forcément relatives à une tâche en particulier, et permettent donc de relever des expressions propres de l'émotion ; il est alors possible d'imaginer des systèmes pouvant apprendre de la personnalité de leur possesseur (systèmes personnalisés).

En conclusion, les dispositifs de capture propices à la reconnaissance d'émotions sont nombreux et variés. Notre modèle d'architecture doit pouvoir intégrer tous ces dispositifs, et

(a) Chaussure avec capteur de conductivité de la peau.



(b) Boucle d'oreille captant la pression sanguine.

(c) Lunettes avec capteurs musculaires.

FIG. 7: Exemples de dispositifs portés pour la capture de signaux portant de l'information émotionnelle. Figures tirées de [119]

prendre en compte le fait qu'il n'y a pas de liste arrêtée de dispositifs de capture pour la reconnaissance d'émotions. Notre architecture se doit donc d'être suffisamment ouverte pour permettre l'intégration de n'importe quel dispositif.

2.3 Canaux de communication émotionnelle

Nous avons vu au chapitre précédent que l'émotion s'exprime au travers de quatre canaux de communication émotionnelle : les expressions faciales, la voix, les positions et mouvements, et les réactions du système nerveux autonome (voir section 1.4, page 26). De nombreuses études ont été effectuées en informatique et en psychologie, sur la caractérisation de diverses émotions selon un unique canal de communication émotionnelle [152] ; les travaux sur les expressions faciales sont largement majoritaires. Néanmoins, de plus en plus de travaux considèrent simultanément plusieurs canaux. Dans la littérature, ces systèmes sont appelés systèmes multimodaux. Afin de garder une cohérence avec la suite de notre mémoire et en particulier les chapitres concernant la multimodalité dans le cadre d'une application interactive intégrant la reconnaissance d'émotions, nous choisissons d'appeler ces systèmes *systèmes multicanaux* (de communication émotionnelle).

Il est à noter que les résultats donnés dans les prochains paragraphes ne peuvent être comparés entre eux. D'une manière générale, les systèmes de reconnaissance d'émotions sont testés en prenant pour référence un groupe d'évaluateurs humains observant les mêmes stimuli que ceux donnés à la machine. En l'absence d'un corpus de test et d'une méthodologie expérimentale uniques, chaque système est testé individuellement et différemment des autres systèmes. Ces différences de protocoles induisent un trop grand biais inter-étude pour permettre une comparaison fiable des résultats.

2.3.1 Reconnaissance par expressions faciales

Le visage est le premier site de l'expression émotionnelle [72]. En effet, le visage est le canal communiquant le plus d'information émotionnelle [43]. Il est à noter qu'en se plaçant dans le

modèle circomplexe des émotions, le visage communique plutôt la valence (plaisir-déplaisir) d'une émotion, alors que la voix permet de mieux juger de son activation [116].

L'étude statique des expressions faciales fournit de nombreux renseignements (par exemple le sourire, le plissement des yeux ou la forme des sourcils) et les systèmes de reconnaissance d'émotions depuis les expressions faciales statiques (voir par exemple [151]) se basent sur des corpus de photographies d'expressions faciales (par exemple le corpus JAFFE[1] - *JApanese Female Facial Expression database*). Il apparaît cependant que la dynamique temporelle du comportement facial est un facteur critique pour la distinction entre un comportement spontané ou acté [152].

Quelques caractéristiques faciales et méthodes d'extraction

Paul Ekman, outre avoir isolé les émotions basiques, a également poursuivi les observations de Darwin quant à l'expression des émotions pour développer un système de codage des expressions faciales : le *Facial Action Coding System* ou FACS [52]. FACS est un système d'encodage s'appuyant sur l'anatomie du visage, basé sur la définition d'unités d'action (*Action Units* ou AU) causant des mouvements sur le visage. Une unité d'action n'est pas relative à un muscle mais à un point d'intérêt pour la dynamique du visage ; le mouvement d'une unité d'action peut ainsi mettre en jeu plusieurs muscles du visage. A l'inverse, les différents mouvements d'un seul muscle peuvent avoir une action sur plusieurs unités d'action. Le FACS a inspiré la définition des paramètres d'animation faciale (les *Facial Animation Points*) dans la norme MPEG-4. Les unités d'actions servent donc de base à de nombreux systèmes de reconnaissance d'émotions par le visage. On y trouve par exemple pour la région des yeux [50] (entre parenthèses est donné le numéro de l'unité d'action) :

- l'élévation de l'intérieur des sourcils (AU1) (figure 8b)
- l'élévation de l'extérieur des sourcils (AU2) (figure 8c)
- l'abaissement des sourcils (AU4)
- l'élévation de la paupière supérieure (AU5)
- la fermeture des paupières (AU7)

(a) Neutre. (b) AU1 : élévation intérieure des sourcils (c) AU2 : élévation extérieure des sourcils

Fig. 8: Illustration des unités d'action 1 et 2 du FACS. Images tirées de [50]

A chaque unité d'action correspond une description des changements physiques induits. L'action est également codée selon son intensité, variant de A à E ; par exemple, l'AU7A

[1] http ://www.kasrl.org/jaffe.html

correspond à une fermeture très faible des paupières, à peine remarquable, alors que l'AU7E dénote des yeux presque fermés, laissant à peine entrevoir l'iris.

Etudes de reconnaissance sur le visage

Nous proposons au lecteur de se référer à trois états de l'art de la littérature recensant les avancées dans le domaine de la reconnaissance d'émotions basée sur des expressions faciales [139] [73] [152]. Les méthodes utilisées dans les travaux de ces états de l'art reposent principalement sur le système FACS, sur des visages vus de face. Hammal *et al.* proposent une méthode de classification basée sur les distances entre points d'intérêt du visage [66]. Quelques exceptions existent quant au point de vue utilisé pour la reconnaissance : ainsi Pantic *et al.* [109] proposent une recherche sur des visages vus de profil.

La précision et la robustesse de la reconnaissance varie selon les systèmes. Pantic *et al.* établissent dans [110] un état de l'art de systèmes de reconnaissance basé sur les expressions faciales. Les systèmes décrits effectuent une reconnaissance sur des modèles discrets d'émotions (entre trois et sept catégories d'émotions). Ces émotions sont exprimées par des sujets humains (entre cinq et quarante selon les systèmes). Pantic *et al.* reportent des résultats d'entre 64% et 98% de reconnaissance selon les systèmes. Il est cependant à noter qu'en l'absence d'un protocole expérimental et de matière première (corpus de photos par exemple) unique, il est impossible d'utiliser ces pourcentages de reconnaissance pour classer ces différents systèmes et en faire émerger le plus précis.

2.3.2 Reconnaissance par la voix

Outre la verbalisation, la voix communique également notre état émotionnel selon plusieurs caractéristiques. Dans une représentation circomplexe des émotions, la voix permet de mieux communiquer l'activation d'une émotion que sa valence [116].

Quelques caractéristiques vocales

Des caractéristiques de son pour la reconnaissance des six émotions basiques sont évoquées dans [139], comme la vitesse du discours, la hauteur moyenne, l'intervalle de hauteur, l'intensité et la qualité de voix. La colère chez un individu sera par exemple caractérisée par une vitesse plus grande que pour un état émotionnel neutre, une hauteur largement supérieure, un intervalle de hauteur plus large et une plus forte intensité. La voix laisse alors également mieux entendre les respirations. Les intervalles de son et de silence sont également des caractéristiques pertinentes pour la reconnaissance d'émotions par la voix, ainsi que la mesure de l'énergie du signal vocal [137]. Pour cette dernière caractéristique cependant, la mesure de l'énergie du signal de la voix relatif à l'effort vocal de celui qui parle n'est pas forcément corrélée à l'impression de celui qui écoute : il y a biais de perception. Zeng *et al.* notent [152] que les efforts récents en reconnaissance par l'audio dérivent des émotions basiques vers des émotions plus spécifiques à un certain contexte d'usage (par exemple un contexte d'apprentissage ou ludique). Sur ces divers contextes il est plus facile d'obtenir des enregistrements spontanés (enregistrements dans des centres d'appels par exemple). Les travaux sur des enregistrements

d'émotions spontanées montrent que les seuls paramètres acoustiques sont insuffisants pour déterminer un état affectif ; une analyse linguistique du discours est nécessaire. Les deux analyses sont conjointement utilisées pour reconnaître l'état affectif d'un sujet.

Etudes de reconnaissance sur la voix

Selon Pantic *et al.* [110], les résultats de reconnaissance d'émotions basés sur la voix en 2005 n'étaient pas réellement probants. Les caractéristiques de voix testées ne permettaient pas de suffisamment discriminer différentes émotions entre elles pour obtenir des résultats convenables. Un état de l'art datant de 2009 [152] cite une quinzaine d'études dans ce domaine, notant que si la recherche en reconnaissance par la voix est largement influencée par la théorie des émotions basiques, des travaux se penchent sur la reconnaissance d'états affectifs dépendants d'un contexte d'usage et d'une application. Les travaux se basent donc toujours sur des modèles discrets d'émotions mais permettent la reconnaissance d'états affectifs autres que les émotions basiques, comme le stress ou la frustration.

2.3.3 Reconnaissance par le mouvement

Les mouvements permettent également l'expression émotionnelle. Pour une machine, cette reconnaissance est ardue, car l'expression émotionnelle n'est pas la fonction primaire du geste. Il faut donc déterminer, lors de l'analyse d'un mouvement, les composantes émotionnelle des composantes fonctionnelles du mouvement, comme prendre un objet ou faire un signe de la main. Certaines caractéristiques traduisant un certain état émotionnel ont cependant pu être isolées [37] [46] [147].

La reconnaissance d'émotions par le mouvement est le coeur de notre cas d'application. Nous avons donc dédié un chapitre aux travaux relatifs à ce domaine (voir chapitre 3).

2.3.4 Reconnaissance par le Système Nerveux Autonome

Selon les critères de confiance que l'on peut accorder aux différents canaux de communication émotionnelle, les réactions du système nerveux autonome (*Autonomous Nervous System* ou ANS) sont le canal le plus fiable (voir section 1.4). En effet, ces réactions n'ont pas pour but principal la communication de l'état émotionnel aux autres individus ; Scherer sépare d'ailleurs le composant "réactions neurophysiologiques" du composant "expression motrice" qui se rapporte à la fonction de communication (voir paragraphe 1.2.2 au chapitre précédent). Néanmoins, les réactions de l'ANS présentent des signaux qu'il est possible de percevoir. A ce titre, la technologie permet de capter certains de ces signaux grâce à des dispositifs adaptés, alors que ce sont parfois les plus difficiles, voire impossibles à remarquer pour un observateur humain. Le rythme cardiaque et ses changements ne sont ainsi pas perceptibles par un humain dans une situation normale, alors qu'un électrocardiogramme permet à une machine de relever l'activité cardiaque de l'individu. De nombreux signaux des réactions du système autonome sont ainsi captables par des dispositifs dédiés.

Quelques caractéristiques du système nerveux autonome

Le système nerveux autonome regroupe le système nerveux central et le système périphérique.

Le système central est le cerveau : l'expérience d'une émotion induit une activité cérébrale mesurable et analysable. Chanel *et al.* [28] évaluent ainsi l'activation d'une émotion en se basant sur les électroencéphalogrammes de quatre individus.

Concernant le système périphérique, plusieurs types de signaux peuvent être utilisés pour reconnaître l'émotion. Chaque signal n'est pas étudié seul mais en conjonction avec d'autres signaux du système central ou périphérique. Le rythme cardiaque, la conductivité de la peau, l'activité musculaire (électromyogrammes), les variations de température de la peau, les variations de pression sanguine sont des signaux régulièrement utilisés pour la reconnaissance d'émotions.

Etudes de reconnaissance sur le système nerveux autonome

Chanel *et al.* [28] se basent sur un électroencéphalogramme pour évaluer l'activation de quatre individus. Ils mesurent également la conductivité de la peau, la pression sanguine, les mouvements abdominaux et thoraciques dus à la respiration, et la température de la peau. L'électroencéphalogramme seul produit des performances de reconnaissance supérieures à 50%. L'intégration des autres signaux rend la reconnaissance plus robuste. Kim *et al.* [81] utilisent un électrocardiogramme, les variations de température de la peau et l'activité électrodermale pour reconnaître la tristesse, la colère, le stress et la surprise. Une étude sur cinquante individus donne des résultats de 61,8% dans la reconnaissance des quatre émotions : la tristesse, la colère, le stress et la surprise. Vyzas et Picard [144] utilisent l'activité musculaire de la mâchoire, la pression sanguine et la conductivité de la peau des doigts, et mesurent l'expansion thoracique pour la respiration chez une actrice exprimant huit états affectifs : le neutre, la colère, la haine, le chagrin, l'amour platonique, l'amour romantique, la joie et le respect. Le système exposé offre un taux de reconnaissance de 81,5% pour ces huit émotions.

2.3.5 Reconnaissance multicanaux

La reconnaissance d'émotions chez l'humain est intrinsèquement multimodale. Nous nous basons instinctivement et de façon non forcément consciente sur la totalité des signaux émotionnels que nous percevons chez un individu pour en déterminer l'état émotionnel : son visage, sa voix, ses positions et mouvements. Nous percevons également certaines réactions physiologiques, comme le rougissement. Cette évaluation simultanée des différents canaux de communication affective nous permet de déceler des émotions feintes ou dont l'expression est volontairement bridée. Il semble donc naturel pour les systèmes informatiques de s'orienter vers une reconnaissance considérant plusieurs canaux. La reconnaissance multicanaux permet en effet de multiplier les sources de données, les analyses faites sur ces données, et d'effectuer une interprétation finale tenant compte de bien plus d'informations. En particulier, nous avons vu que certains signaux ou canaux permettaient de mieux véhiculer une certaine composante

de l'émotion. Par exemple, le visage communique particulièrement la valence d'une émotion ; la conductivité de la peau, le degré d'activation [144]. Analyser ainsi plusieurs canaux permet d'affecter des facteurs de confiance aux différents signaux reçus, par exemple en donnant une grande confiance aux expressions faciales pour obtenir la valence d'une émotion et en combinant cette mesure avec l'activation trouvée via des mesures physiologiques.

Pantic *et al.* [110] définissent un système idéal de reconnaissance d'émotions comme étant (entre autres critères) multicanaux. En 2004, peu de systèmes, malgré les techniques avancées en traitement audio et vidéo, cherchaient à proposer une reconnaissance utilisant ces deux canaux [139]. Un état de l'art plus récent (2009) cite plus d'une dizaine de systèmes bi-canaux utilisant la voix et le visage pour reconnaître les émotions [152].

Zeng *et al.* [153] classent des signaux audio et visuels en onze états affectifs relatifs à l'interaction homme-machine. Le test du système sur un panel de 38 évaluateurs approche un résultat de 90%. Kapoor *et al.* [76] proposent un système détectant l'intérêt chez un enfant résolvant un puzzle sur ordinateur en combinant les informations du visage, de posture, et les actions effectuées sur le jeu lui-même. Balomenos *et al.* [5] proposent un système se basant sur le visage et les gestes des mains interprétant les six émotions basiques d'Ekman. Ce système atteint une reconnaissance de 85% et utilise l'information de gestuelle pour améliorer la robustesse de la reconnaissance faciale. Gunes *et al.* [64] proposent une approche vers la réalisation d'un système se basant sur le visage et la gestuelle. En 2007, Castellano *et al.* [26] proposent un système intégrant les expressions faciales, les mouvements des mains et la voix capable de reconnaitre huit états affectifs : la colère, le désespoir, l'intérêt, le plaisir, la tristesse, l'irritation, la joie et la fierté, couvrant ainsi chaque quadrant de l'espace plaisir activation par deux états affectifs. Cette étude permet de comparer les résultats de reconnaissance pour chaque canal et les résultats de la reconnaissance multicanaux, les différents résultats provenant tous de la même expérience. La reconnaissance multicanaux arrive à un taux de reconnaissance de plus de 74%, là où les reconnaissances selon les expressions faciales, les mouvements et la voix sont respectivement de 48%, 67%, et de 57%. Peu de travaux sont réalisés sur des systèmes multicanaux incluant les réponses physiologiques. Dès 1999 cependant, Healey *et al.* étudient le stress d'un conducteur grâce à certaines réactions physiologiques et à une analyse du visage [68].

En conclusion, pour le modèle d'architecture visé, il convient de noter la variété des caractéristiques manipulées dans les systèmes de reconnaissance d'émotions. Outre cette variété, nous soulignons également la reconnaissance multicanaux, impliquant la gestion de plusieurs caractéristiques.

Dans ce travail de thèse nous avons choisi la danse comme cadre applicatif. Nous avons donc porté une attention particulière à la reconnaissance d'émotions par la gestuelle, qui fera l'objet d'un chapitre dédié.

2.4 Interprétation

Dans le processus de reconnaissance d'émotions en trois étapes, la capture des données depuis le monde réel est suivie par l'analyse de ces données pour en extraire des caractéristiques pertinentes de l'émotion, puis par l'interprétation de ces caractéristiques. Nous avons vu dans la section précédente quelques caractéristiques utilisées par les canaux de communication émotionnelle pour reconnaître les émotions d'un utilisateur. Dans cette section, nous étudions les techniques d'interprétation mises en œuvre.

Afin de mettre en place un système d'interprétation de caractéristiques en émotions, il convient de définir deux points. Il est tout d'abord nécessaire de définir le modèle des émotions qui sera utilisé. D'un point de vue technique ensuite, il est nécessaire de définir par quel algorithme l'interprétation automatique sera effectuée.

2.4.1 Modèles pour l'interprétation

Les modèles discrets permettent de considérer chaque émotion comme une catégorie (voir la section 1.3.1, page 23) ; reconnaître une émotion revient alors à classer un vecteur de caractéristiques dans l'une des catégories. Les modèles continus permettent de décomposer une émotion selon deux ou trois axes (voir section 1.3.2, page 24). Le modèle à composants de Scherer suscite des adeptes mais est difficilement applicable au domaine de la reconnaissance d'émotions en informatique.

Modèles discrets

Le choix d'un modèle discret des émotions représente une large majorité des systèmes de reconnaissance d'émotions [152]. La reconnaissance se fait dans ce cas sur un ensemble fini de catégories d'émotions. Bien souvent, l'ensemble ou un sous-ensemble des émotions basiques d'Ekman est inclus dans l'ensemble de reconnaissance [152] (par exemple, [153], [5], [81], [151] ainsi que la plupart des travaux du laboratoire Infomus de Gênes [143] [25]). Le caractère universel largement testé des émotions basiques (permettant de minimiser les modulations d'expression liées à la culture et à la personnalité) apportent un côté pratique et simple à la reconnaissance. Elles posent cependant deux fortes limitations, de par leur nombre restreint d'une part, et par le fait que les expressions intenses de ces émotions sont rarement observées dans un contexte hors laboratoire d'autre part.

Nous avons vu au chapitre 1 qu'il existait, en anglais, plus de 500 termes relatifs à l'affect. Cette large base pose un inconvénient certain pour qui voudrait créer un système de reconnaissance capable de discerner toutes ces émotions ; Il permet par contre de constituer un sous-ensemble d'émotions relatif à un certain domaine. Par exemple, Zeng *et al.* [153] basent leur système de reconnaissance sur les six émotions basiques d'Ekman, un état neutre, et quatre émotions pertinentes dans le cadre de l'interaction homme machine : l'intérêt, l'ennui, la confusion et la frustration. El Kaliouby *et al.* [56] proposent une étude sur six états mentaux : l'accord, la concentration, le désaccord, l'intérêt, la réflexion et l'incertitude[2]. Bien que

[2]"agreeing, concentrating, disagreeing, interested, thinking and unsure"

ces états mentaux ne soient pas des états affectifs (excepté l'intérêt, qui peut être vu comme un état affectif), la méthode de reconnaissance utilisée est la même que la reconnaissance d'émotions ; seules les caractéristiques diffèrent.

Modèles Continus

Les espaces continus d'émotions à deux ou trois dimensions sont également largement utilisés dans le contexte de l'informatique affective, bien que dans une moindre mesure que les modèles discrets. L'espace valence-activation de Russel est largement utilisé ; la 3ème dimension de soumission-dominance est parfois prise en compte [152]. L'espace circomplexe n'est pas un espace précis ; les espaces 2D sont découpés en quadrants (valence positive/négative et activation faible/forte), les espaces 3D en octants (valence positive/négative, activation faible/forte, soumission/dominance). Les systèmes de reconnaissance se basant sur ces modèles cherchent donc généralement plutôt à classer un vecteur de caractéristiques dans l'un des quadrants ou octants de l'espace considéré [74]. En effet, une telle discrétisation permet de ramener l'interprétation à un problème de classification. De plus, certaines caractéristiques sont évaluées comme particulièrement porteuses d'information selon un axe particulier. Par exemple, Chanel *et al.* évaluent l'activation émotionnelle d'un sujet grâce à son électroencéphalogramme [28].

Les espaces continus classiques des émotions ont été largement étudiés notamment dans la décorrélation de leurs axes. Ils ne sont par contre pas adaptés à toutes les tâches. Scherer souligne l'inadaptation de tels espaces dans l'évaluation de stimuli artistiques [134]. Kort *et al.* [84] proposent un modèle continu adapté à l'apprentissage selon deux dimensions principales. La première dimension est celle de la valence des émotions. La deuxième dimension donne leur caractère constructif ou destructif de l'apprentissage. Cet espace est donc formé de quatre quadrants (voir figure 9). Selon les auteurs, le processus d'apprentissage passe idéalement par chacun de ces quadrants : tout d'abord la curiosité devant un nouveau centre d'intérêt, puis la confusion devant les problèmes qu'il pose. Viennent ensuite la frustration et les conceptions erronées engendrées par la réflexion, pour arriver enfin à la résolution, qui peut engendrer une nouvelle curiosité, propulsant l'apprenant à nouveau dans le premier quadrant.

Modèle à composants

Le modèle à composants de Scherer suscite un intérêt grandissant en informatique affective. En reconnaissance, la réalisation logicielle de ce modèle permettrait d'obtenir potentiellement tous les états affectifs possibles en permettant de représenter directement les processus internes d'une émotion. Cependant, représenter des émotions selon les processus intra- et inter-composants est une tâche d'une grande complexité. Cette modélisation des composants et des processus correspondants convient mieux à la génération d'émotions. En effet, créer un système de reconnaissance basé sur la théorie de Scherer reviendrait à créer un organisme émotionnel virtuel se forgeant ses propres expériences émotionnelles à partir des stimuli perçus (ces stimuli pouvant être des mesures capturées du monde ou des caractéristiques déjà extraites de ces mesures).

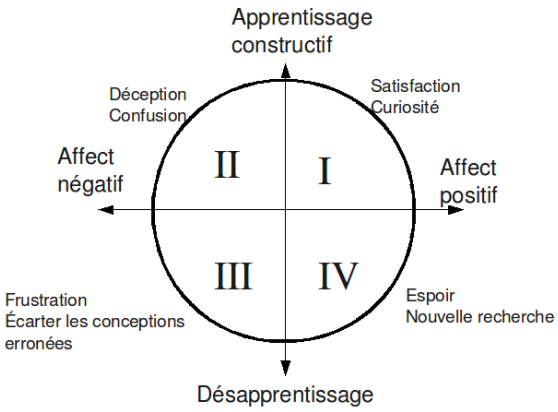


FIG. 9: Proposition de modèle affectif adapté à l'apprentissage. Figure tirée de [84].

Nous n'avons donc relevé qu'un seul travail en reconnaissance d'émotions exploitant explicitement le modèle à composants de Scherer. Lisetti *et al.* proposent ainsi le système MAUI (Multimodal Affective User Interface), basé sur la théorie de l'évaluation cognitive et en particulier sur le modèle à composants de Scherer [88]. Ce système se fonde principalement sur les réponses aux critères d'évaluation (SECs) pour représenter une émotion. Ces vecteurs-réponses sont ensuite directement injectés dans un système de génération d'émotions pour, par exemple, la synthèse d'expressions faciales d'un avatar [89].

Il existe donc principalement trois types de modèles d'émotions utilisés en informatique. Les modèles discrets caractérisent chaque émotion par un label ; les modèles continus par un couple ou triplet de coordonnées (bien qu'en général soient considérés les quadrants ou octants du modèle choisi) ; enfin le modèle à composants de Scherer définit une émotion par l'ensemble des réponses aux SECs. Notre modèle d'architecture pour la reconnaissance d'émotions doit permettre le choix de chacun d'entre eux.

2.4.2 Algorithmes pour l'interprétation

Le deuxième choix à effectuer lors de la mise en place de l'interprétation des caractéristiques émotionnelles dans un système de reconnaissance est un choix technique : il s'agit de choisir l'algorithme permettant d'inférer une émotion depuis les vecteurs de caractéristiques. Dans la grande majorité des systèmes, l'interprétation renvoie en sortie des labels d'émotions (dans le cas de systèmes se basant sur des espaces discrets) ou des régions (quadrants ou octants) d'espaces continus. Dans tous ces cas l'interprétation revient à classer chaque vecteur de caractéristiques en une catégorie (un label ou une région de l'espace). Le domaine de l'informatique propose de nombreux algorithmes permettant d'effectuer une telle classification. La première option est celle d'un système de règles. Sebe *et al.* préconisent cependant de mimer le fonctionnement humain en utilisant des algorithmes d'apprentissage automatique [139].

Bien que massivement répandues, les techniques de classification ne sont pas les seules existantes pour interpréter un ensemble de caractéristiques en une émotion. Le processus émotionnel tel que défini par Scherer par exemple (voir paragraphe 1.2.2, page 15) ne permet pas une classification en états émotionnels. Ainsi le système MAUI+ASIA de Lisetti *et al.* [89], construit autour des travaux de Scherer, propose un système d'interprétation qui ne classifie pas les caractéristiques en catégories émotionnelles. Le système proposé utilise en effet les valeurs des caractéristiques pour évaluer les réponses aux SECs (voir paragraphe 1.2.2) ; l'ensemble de ces réponses sont considérées comme définissant un processus émotionnel, directement réutilisé comme entrée par un agent expressif.

Exemples d'algorithmes

Les réseaux de neurones sont couramment utilisés pour la classification de vecteurs de caractéristiques en catégories émotionnelles. Ioannou *et al.* proposent ainsi un réseau de neurones adaptatif permettant de combiner des mesures selon plusieurs canaux de communication affective [70]. L'analyse conjointe des expressions faciales et des caractéristiques vocales permet à l'algorithme de ré-entraîner l'interprétation selon l'un des canaux lorsque l'analyse de celui-ci ne permet plus de résultats satisfaisants. Les modèles de Markov cachés sont également largement utilisés. Zeng *et al.*, par exemple, proposent un tel algorithme pour la reconnaissance audio-visuelle [154], Kapoor *et al.* pour la détection de l'intérêt chez un utilisateur [76]. Ghamen et Caplier proposent un système de reconnaissance d'expressions faciales se basant sur la théorie de l'évidence [61]. Wong *et al.* utilisent des arbres de décision pour la reconnaissance par des expressions faciales [151]. Certains travaux s'attachent à tester plusieurs algorithmes d'interprétation. Grâce à un système de reconnaissance dans lequel seul l'algorithme d'interprétation varie, Vyzas *et al.* comparent trois algorithmes de classification : la *Sequential Floating Forward Search* (SFFS), la projection de Fisher, et une méthode hybride [145].

Nous retenons qu'il existe un large choix de techniques algorithmiques permettant l'interprétation d'un vecteur de caractéristiques en une émotion (notamment des techniques de classification en catégories). Ces différentes méthodes offrent des résultats comparables. Aussi, en l'absence d'un algorithme unique, notre modèle d'architecture pour la reconnaissance d'émotions doit prendre en compte la diversité des choix possibles et proposer un système suffisamment ouvert pour permettre l'intégration de chacun des ces algorithmes.

2.4.3 Fusion et synchronisation de données

Le fait de considérer plusieurs canaux de communication émotionnelle amène le problème de la fusion des données. Dans le cadre de l'informatique affective, le terme *fusion* est moins cerné que dans le domaine de l'interaction multimodale [86]. Tout comme en interaction multimodale, on distingue cependant généralement trois types de fusion : la fusion au niveau signal, au niveau caractéristiques, et au niveau décision. La fusion au niveau signal consiste à fusionner les données issues de différents dispositifs. La fusion au niveau des caractéristiques consiste à synchroniser et à regrouper toutes caractéristiques extraites des différents dispositifs et selon les différents canaux de communication émotionnelle avant de les interpréter. L'algorithme d'interprétation reçoit ainsi, à chaque pas de temps, l'ensemble des caractéristiques. La fusion au niveau décision est ici une fusion des émotions et consiste à interpréter les caractéristiques

de chaque canal séparément : on obtient ainsi autant d'émotions que de canaux étudiés. Les émotions ainsi obtenues doivent alors être synchronisées et amalgamées pour n'obtenir qu'une émotion finale. Lié à la fusion se pose donc le problème de la synchronisation. Les dispositifs de capture ne fournissent pas les données à la même fréquence. De plus, les caractéristiques ne peuvent pas forcément être extraites de façon régulière.

Chez l'homme, la fusion se fait apparemment au niveau des caractéristiques [152]. En mimant le comportement humain, une solution consiste à fournir toutes les caractéristiques extraites à un algorithme d'interprétation. La plupart des systèmes propose cependant une fusion des émotions et non des caractéristiques, la première étant plus facile. Ainsi, Zeng *et al.* choisissent une fusion au niveau émotion afin d'éviter le problème de synchronisation des caractéristiques des expressions faciales et de la voix [153], fournies par les systèmes de capture de manière asynchrone. Cette fusion est effectuée selon un mécanisme de vote parmi les solutions des interprétations de chaque canal. De plus, la fusion au niveau émotion permet de combiner plusieurs systèmes existants (produits lors de travaux antérieurs par exemple) bien plus facilement que de ré-entraîner un algorithme de classification. Enfin, les bénéfices d'une fusion au niveau caractéristiques ne sont pas clairement établis ; ainsi Busso *et al* [17] montrent que pour leur système de reconnaissance par les expressions faciales et la voix, la fusion au niveau caractéristiques est à peu près équivalente à la fusion au niveau émotions. Une manière d'obtenir une meilleure fusion est de proposer plusieurs mécanismes. Ainsi Lisetti *et al.* proposent un système incluant plusieurs mécanismes de fusion aux trois niveaux signal, caractéristique et décision ainsi que le choix automatique du type de fusion offrant le meilleur résultat [89].

Pour la fusion dans un système multi-canaux, nous retenons qu'il est nécessaire de prendre en compte la synchronisation des données. En effet, les données recueillies par le système peuvent différer au niveau de la fréquence de capture, mais également au niveau de la phase, c'est-à-dire que les dispositifs de capture n'effectuent pas forcément leur première mesure au même moment. De plus, certaines caractéristiques ne peuvent être mesurées de façon régulière. Par exemple, si l'on veut extraire la hauteur moyenne d'une phrase parlée (en supposant que cette caractéristique véhicule de l'information émotionnelle), rien ne peut être calculé avant la fin de la phrase : le système est bloqué jusqu'à ce que le sujet finisse sa phrase. Il est également nécessaire de considérer la fusion des données ; cette fusion peut être effectuée au niveau des caractéristiques et/ou au niveau d'émotions déduites selon les différents canaux de communication émotionnelle considérés indépendants. Dans le cadre de nos travaux sur l'ingénierie de la reconnaissance d'émotions, il est donc nécessaire de prendre en compte ces diverses possibilités et contraintes au sein de notre modèle d'architecture pour la reconnaissance des émotions.

2.5 Evaluation d'un programme de reconnaissance

En combinant la littérature de la reconnaissance d'émotions en psychologie et en informatique, nous identifions deux types d'expérimentations. La première est une expérimentation exploratoire visant à identifier et valider des caractéristiques porteuses d'information émotionnelle et potentiellement à calculer l'impact de chacune des caractéristiques testées dans

l'expression de chaque émotion étudiée. Ce type d'expérimentation se retrouve principalement en psychologie et ne fait alors pas intervenir de système informatique. La deuxième consiste à valider un système complet de reconnaissance, en comparant ses résultats avec une référence humaine. Ce type d'expérimentation est une évaluation d'un système automatique par rapport à des performances humaines et ne se retrouve donc que dans le domaine informatique. Nous ne traitons ici que de ce dernier cas, visant à évaluer un système de reconnaissance automatique des émotions.

Pour évaluer un système de reconnaissance d'émotions, il est nécessaire de confronter les capacités de reconnaissance du système à une référence. Les évaluations actuelles prennent l'humain comme référence. L'évaluation nécessite donc une expression émotionnelle de la part d'un sujet, ainsi qu'une évaluation humaine à confronter aux résultats du système.

2.5.1 Critères d'évaluation

Picard propose dans [114] une série de critères caractérisant la collection de données pour l'évaluation de systèmes de reconnaissance d'émotions. Le premier critère est celui du caractère acté ou spontané de l'expression émotionnelle. Le jeu d'acteur introduit potentiellement un biais du fait que les émotions exprimées ne sont pas forcément ressenties (bien que la méthode de l'Actor's studio préconise de se plonger dans l'état émotionnel recherché afin de l'exprimer de manière réaliste). Or nous avons vu au chapitre 1 que certaines expressions de l'émotion sont extrêmement difficiles à contrefaire. L'expression d'une émotion actée ne reprend donc pas forcément les caractéristiques de l'expression spontanée d'une émotion ressentie. De plus, la qualité de l'expression dépend directement de la compétence de l'acteur et de sa perception de l'émotion qui lui a été demandé de jouer. La personnalité de l'acteur teinte son jeu, ce qui peut introduire un biais supplémentaire. Le biais de l'interprétation d'une émotion peut être partiellement résolu en adoptant une approche scénario (utilisée par exemple dans [137]). Dans le cas où l'on cherche à produire un système générique, les expressions spontanées sont donc préférables. Elles sont cependant plus difficiles à obtenir. Pour disposer d'expressions spontanées comme matériau de base à une évaluation, il est possible de recourir à un corpus existant de photos ou vidéos d'émotions spontanées comme EmoTV1 [1], qui regroupe des vidéos de reportages ou d'informations montrant des personnes émotionnellement expressives dans un contexte réel. Une deuxième option est de déclencher une émotion chez un sujet par un protocole d'induction. Cette méthode pose alors le problème du biais entre l'émotion que l'on cherche à déclencher et l'émotion effectivement ressentie par le sujet.

Le deuxième critère concerne ce que l'on cherche à mesurer et à évaluer. Lors d'une communication émotionnelle entre un sujet expressif et un observateur humain, il convient de distinguer l'émotion ressentie par le sujet, l'émotion qu'il exprime, et l'émotion interprétée par l'observateur. Ces différences sont importantes : en effet l'expression d'une émotion ne correspond pas forcément à l'émotion ressentie : nous tentons souvent de minimiser nos expressions affectives. De même, le jugement d'un observateur est un jugement subjectif et peut donc ne pas correspondre à l'émotion exprimée par le sujet. Si le but est de tester la reconnaissance d'une émotion exprimée, il faut prendre des observateurs dont le jugement va servir de référence. Prendre un groupe suffisamment grand d'observateurs permet de lever, par une

analyse statistique, le biais d'erreur de jugement. Si le but est de mesurer une émotion ressentie, alors le sujet expressif est en même temps observateur ; il est d'ailleurs le seul. Une verbalisation du ressenti (méthode du *self-assessment*) par des entretiens ou questionnaires peuvent permettre de relever l'appréciation du sujet de sa propre émotion mais cette méthode est également sujette à la capacité du sujet à s'analyser et à verbaliser.

Les trois critères suivants se rapportent au biais mis en œuvre lorsque l'on cherche à mesurer une émotion spontanée. Ils influent l'expression du sujet. Le premier critère est l'environnement dans lequel est menée l'expérience impacte directement l'expression émotionnelle d'un sujet : celui-ci n'aura pas la même expression émotionnelle en laboratoire ou sur le terrain. Les deux autres critères sont liés à la conscience que le sujet peut avoir de l'expérimentation. Si le sujet est conscient que l'expérimentation cherche à analyser son expression émotionnelle, ladite expression en sera affectée. De même, la présence évidente de matériel d'enregistrement ou de mesure peut impacter l'expressivité du sujet. Afin d'obtenir des expressions les plus pures possibles et les plus proches d'une situation réelle, il est donc préférable de mener une expérience sur le terrain, sans que le sujet ne sache ni le but de l'expérience ni qu'il est enregistré. Ces conditions peuvent être difficiles à réunir pour des raisons pratiques mais certains contextes peuvent les rendre dispensables voire caduques. Par exemple, la reconnaissance d'émotions en danse ne peut s'effectuer que sur des séquences dansées ; il ne s'agit donc pas d'expressions spontanées. De l'avis des danseurs avec lesquels nous travaillons, l'enregistrement, le lieu de capture et le but de l'expérimentation n'influent peu ou pas leurs performances.

Selon ces critères, une évaluation d'un système de reconnaissance d'émotions devrait être effectuée sur des émotions spontanées dans un contexte hors laboratoire, et où le sujet expressif ne sait ni que l'expérience a pour cadre ses émotions ni qu'il est enregistré par des appareils de mesure. Un panel d'observateurs est nécessaire pour confronter le jugement humain au jugement du système automatique. De plus, un système automatique est à une place d'observateur. Il n'a donc pas accès au ressenti de l'émotion car par définition, le ressenti est subjectif donc seulement accessible par le sujet (voir section 1.1, page 9). Un système de reconnaissance d'émotions ne peut que mesurer l'émotion exprimée par le sujet. Il convient alors de noter les capacités de mesure d'un système informatique alliées à l'impossibilité de contrefaire ou restreindre certaines réactions physiologiques. Un système se basant sur un électro-encéphalogramme ou le rythme cardiaque par exemple mesurera toujours l'émotion exprimée, mais par un canal que le sujet ne peut maîtriser, se rapprochant ainsi le plus possible de l'émotion ressentie.

2.5.2 Protocoles d'induction

Il existe dans la littérature quelques protocoles [58] [149] permettant l'induction d'émotions. Ces protocoles se situent en laboratoire et le sujet sait qu'il est enregistré ; leur but est de cacher au sujet le but de l'expérimentation (i.e. recueillir ses émotions) et de déclencher des émotions spontanées. Cependant ces protocoles sont limités à un contexte et à un environnement particulier, les rendant inadaptés à une généralisation. Fevrier *et al.* proposent ainsi un protocole où le sujet prend la place d'un vendeur de places de théâtre et où un compère joue le rôle d'un client au téléphone [58]. Ce protocole permet de déclencher la satisfaction, l'amusement, l'embarras, l'incompréhension et la surprise. Le sujet est placé sur une chaise devant

un bureau. Le sujet est filmé sous prétexte d'une évaluation ergonomique de l'application de réservation des places. Ce protocole peut donc difficilement être utilisé tel quel pour d'autres systèmes de reconnaissance. Un système de reconnaissance par la gestuelle serait limité par l'environnement (le sujet est sur une chaise) ; une reconnaissance par réactions physiologiques par le contexte (quelle excuse trouver pour le placement de capteurs physiologiques ?). Wang *et al.* proposent un jeu de rôle inducteur de la joie, l'ennui, la surprise, la colère et la déception [149]. De même que pour Février *et al.*, le contexte (jeu vidéo) et l'environnement (joueur assis devant un bureau) en limitent l'utilisation.

Nous retenons que la constitution d'un corpus de données affectives et l'évaluation d'un système de reconnaissance d'émotions peut présenter de nombreux biais. Picard propose des recommandations permettant d'éviter ces biais mais celles-ci sont en compétition avec des aspects pratiques de méthodologie expérimentale. Par exemple, il est souvent préférable de collecter des expressions spontanées ; surviennent alors les difficultés techniques et éthiques sur le déclenchement de ces émotions chez un sujet. De même les émotions seront exprimées naturellement hors laboratoire ; cela implique cependant de ne pas pouvoir observer le sujet aussi bien que dans un cadre contrôlé. Certaines de ces difficultés peuvent néanmoins disparaître si l'on considère le but du système de reconnaissance à évaluer. Par exemple, l'évaluation d'un logiciel cherchant à reconnaître des expressions dansées (voir chapitre 6) de l'émotion pourra s'astreindre de la plupart des recommandations décrites dans cette section. En effet, un danseur ne cherche pas à ressentir l'émotion mais se concentre sur le mouvement. La mesure entre donc plutôt dans le cadre des "émotions actées". Les autres critères définis ci-dessus ont alors une importance bien moindre. Pour la prise de vue, la différence entre un laboratoire et un studio de danse est minime. Le danseur avec qui nous travaillions avait également l'habitude d'être enregistré. Le contexte d'application que nous avons choisi dans ce travail nous affranchit donc des contraintes liées à l'utilisation d'émotions spontanées.

2.6 Conclusion

Dans ce chapitre nous avons étudié les différentes facettes d'un système de reconnaissance d'émotions. Nous avons tout d'abord présenté le niveau capture en décrivant des capteurs permettant la reconnaissance d'émotions. Par rapport à un humain, les capteurs présentent des défauts et des inconvénients : les caméras sont bien moins précises que les facultés visuelles humaines et ne permettent pas une analyse aussi fine que celle que nous faisons lorsque nous observons quelqu'un. Par contre, il existe de nombreux capteurs permettant de mesurer des signaux auxquels un observateur n'a normalement pas accès (comme le rythme cardiaque). Nous avons ensuite étudié le niveau analyse, par les canaux de communication émotionnelle et les caractéristiques observées pour chacun de ces canaux. L'analyse de plusieurs canaux simultanément renforce la reconnaissance et permet de passer outre certaines caractéristiques de faible confiance pour se focaliser sur des caractéristiques à forte confiance, améliorant ainsi la robustesse voire permettant la reconnaissance d'émotions cachées. Nous avons ensuite traité le niveau interprétation, se basant sur les caractéristiques extraites pour en déduire une émotion, en se concentrant sur ses deux composantes principales : le modèle d'émotion utilisé pour l'interprétation et l'algorithme d'interprétation utilisé. Nous avons enfin étudié l'intégration de plusieurs canaux de communication émotionnelle, et en particulier les aspects

de synchronisation et de fusion des données.

Pour la suite de ce mémoire, nous retenons donc qu'un modèle d'architecture pour la reconnaissance d'émotions doit être capable d'intégrer des capteurs extrêmement variés mesurant des signaux selon plusieurs canaux de communication émotionnelle ; qu'il doit permettre la gestion de caractéristiques variées, dont certaines encore non identifiées, caractéristiques potentiellement extraites de chaque canal de communication émotionnelle ; au niveau interprétation, de permettre l'utilisation d'au moins trois types de modèles (discrets, continus, et à composant) et de plusieurs techniques algorithmiques d'interprétation. Il doit également permettre la synchronisation de données potentiellement asynchrones, irrégulières, et parfois temporisées (par exemple, lorsqu'on veut analyser une phrase dans son entier ; il est alors nécessaire d'attendre la fin de la phrase pour en extraire des caractéristiques). Le modèle d'architecture que nous avons développé et que nous détaillerons en deuxième partie de ce mémoire se veut enfin capable d'intégrer des applications existantes autant que de permettre d'en créer de nouvelles.

Nous avons traité en dernière section de ce chapitre de l'évaluation de systèmes de reconnaissance automatique des émotions. Nous avons souligné les difficultés inhérentes à l'évaluation de systèmes de reconnaissance d'émotions. Dans ce mémoire, nous illustrons nos propositions par la conception d'une application de reconnaissance d'émotions basée sur la gestuelle, appliquée à la danse. Notre cas applicatif se situe donc dans un contexte d'émotions actées (dansées) par le sujet. La mise en place d'une évaluation sera donc, dans notre cas, plus simple car le système a pour but de reconnaître des émotions dansées et exprimées, et donc ni ressenties ni spontanées.

Dans le prochain chapitre, nous nous focalisons sur la reconnaissance d'émotions par les positions et mouvements du corps. Nous y approfondissons les notions introduites dans ce chapitre en nous focalisant sur la gestuelle.

Chapitre 3

Reconnaissance d'émotions par la gestuelle

La recherche effectuée dans le cadre de cette thèse s'articule autour de deux axes principaux : premièrement, définir un modèle d'architecture pour les applications interactives prenant en compte l'émotion de l'utilisateur ; deuxièmement, appliquer ce modèle d'architecture pour concevoir une application de reconnaissance d'émotions basée sur la gestuelle. Notre objectif n'est pas d'apporter une contribution par l'identification et la validation de caractéristiques de position et mouvement permettant une reconnaissance automatique des émotions du sujet, mais d'identifier les modèles de la littérature les mieux adaptés à ce type d'application. Dans ce chapitre, nous dressons donc un état de l'art concernant la reconnaissance d'émotions par la gestuelle. Nous débutons ce chapitre en offrant une base théorique à l'identification de caractéristiques de gestuelle : par des définitions et taxonomies relatives à la gestuelle d'une part, et par la présentation de la théorie de l'effort du chorégraphe Laban sur l'expressivité du mouvement d'autre part. De cette base théorique nous passons ensuite à l'identification de caractéristiques de gestuelle pour l'émotion. Enfin, ces deux parties nous permettent d'aborder les travaux en informatique sur la reconnaissance d'émotions par la gestuelle.

3.1 Définitions du geste

Deux approches permettent d'aborder le geste. L'approche "modalités" considère les mouvements des différentes parties du corps. L'approche "fonctionnelle" considère le geste selon les fonctions qu'il remplit.

3.1.1 Le geste : un mouvement

Le geste est un mouvement dans les trois dimensions d'une certaine partie du corps. Des catégories de mouvement sont définies pour les parties du corps. Le mouvement est donc décrit par son affiliation à ces catégories. Par exemple dans [147], Wallbott définit le mouvement des épaules : celles-ci peuvent être en arrière ou en avant, la tête peut être penchée en avant, en arrière, tournée ou penchée sur les côtés gauche ou droit. De même, de Meijer [46] considère deux catégories pour le mouvement du tronc, étiré ou voûté :

> *Mouvements du tronc :* les mouvements étirés devraient être accomplis avec les jambes et le dos droit, les mouvements voûtés devraient être accomplis en voûtant le tronc et la tête et en pliant légèrement les genoux (sans qu'ils ne touchent le sol)[1].

Coulson [37] considère le corps comme un assemblage mécanique décrit par sept degrés de liberté. Les positions véhiculant de l'émotion sont donc précisément décrites en termes d'angles de rotation des sept degrés de liberté considérés.

3.1.2 Le geste : ses fonctions

Comme souligné dans [46] [147], les travaux visant à établir une taxonomie de la gestuelle et du mouvement se focalisent sur des situations de conversation et souvent plus particulièrement sur les gestes des bras et des mains dans de telles situations. Wallbott et Scherer, dans [137], adoptent une approche fonctionnelle pour catégoriser les expressions non-verbales (non limitées à la gestuelle) en situation de conversation. Ils identifient quatre fonctions : les fonctions sémantiques, syntaxiques, pragmatiques et dialogiques. Les caractéristiques sémantiques remplissent les fonctions de modification et modulation du sens du discours. Les caractéristiques syntaxiques ont pour rôle de structurer le discours et d'en annoncer les différentes parties (comme une position amorçant un départ signifie la fin de la conversation), et de synchroniser l'expression non-verbale avec le discours. Les caractéristiques pragmatiques renseignent sur l'individu en termes d'identité propre et d'identité sociale. Les caractéristiques dialogiques permettent l'orchestration de la prise de tours dans le dialogue. Ekman et Friesen [51] proposent une autre taxonomie en cinq catégories : les emblèmes (gestes se substituant à un mot ou une phrase, comme le V de victoire), les illustrateurs (illustrant le discours, comme le fait de pointer en donnant des directions), les régulateurs (permettant l'interaction du discours), les démonstrations affectives et enfin les adaptateurs (gestes n'appuyant pas le discours, comme fumer une cigarette).

Kendon [78] définit un geste comme une action visible à intention de communication. Il propose une segmentation du geste en unités de gestuelle. Une unité de gestuelle dure d'une pause à une autre pause et est séparée en différentes phases. La première phase est celle de préparation, et vient avant la phase de *stroke*, phase pendant laquelle l'expressivité du mouvement est à son maximum. La phase de *recovery* suit la phase de *stroke*.

Kendon [79] propose également une taxonomie fonctionnelle du geste sous la forme d'un continuum. La première catégorie est la gesticulation : des mouvements spontanés des mains et des bras qui accompagnent le discours. Viennent ensuite les "para-langages". Ceux-ci ne sont pas réellement des langages, mais plutôt des systèmes de codage (par exemple pour l'arbitrage de certains sports) [90]. Les pantomimes visent à représenter, par la gestuelle, un objet, évènement, ou état d'esprit. Les emblèmes sont, tout comme la catégorie du même nom

[1] *Trunk movement :* stretched movements should be performed with straight trunk and legs, bowed movements should be performed by bowing the trunk and head and bending the knees a little (but the knees do not touch the floor).

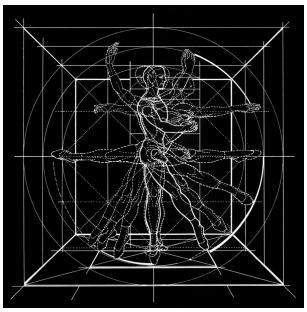
d'Ekman, des gestes symboliques se substituant à un mot ou une phrase. Enfin, les langages de signes se réfèrent aux véritables langages (comme la langue des signes française).

Les approches fonctionnelles se basent non pas sur une observation objective du mouvement mais sur ses fonctions dans un contexte particulier d'interaction et de communication [147]. Notre but est de reconnaître l'émotion sans se préoccuper du contexte. Dans le cadre de ce mémoire, nous ne nous intéressons pas aux gestes pour l'interaction. Ainsi, nous considérons les seules gesticulations (au sens de Kendon), qui peuvent parfois être des démonstrations affectives. Les approches fonctionnelles présentent cependant un intérêt pour des travaux futurs mêlant reconnaissance d'émotions et interaction par la gestuelle.

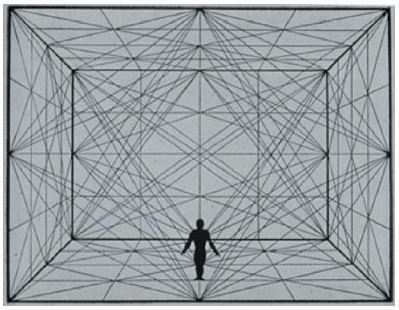
3.2 Expressivité du geste : l'analyse du mouvement de Laban

L'analyse du mouvement de Laban (*Laban Movement Analysis* ou LMA) a été développée par le chorégraphe Laban et est désormais l'un des piliers de la conception moderne de la danse. La LMA fournit des méthodes d'analyse dans quatre catégories : le corps, l'effort, la forme et l'espace [69]. La catégorie du corps propose des analyses sur le mouvement des différentes parties du corps. La catégorie de forme se focalise sur la forme du corps à un instant donné et aux transitions entre les différentes formes que peut prendre le corps. La catégorie de l'espace offre une analyse des espaces : Laban y introduit la kinésphère, un espace centré sur le corps délimité par les mouvements des membres (figure 10a). Laban étudie donc l'occupation de l'espace par le danseur dans sa kinésphère, et la position de cette kinésphère dans l'espace général (figure 10b). Enfin, la catégorie de l'effort explore la dynamique du mouvement et se décompose en quatre dimensions : le poids (de ferme à léger), le temps (de soutenu à soudain), l'espace (de direct à indirect) et le flux (de restreint à libre). Les trois premières dimensions forment un espace permettant de caractériser la dynamique d'un geste. La dimension de poids s'apparente à la force que l'on met dans un geste. Un coup sera ferme, un salut timide sera léger. La dimension de temps est explicite et caractérise la longueur du geste. Un coup sera soudain, une lente révérence soutenue. La dimension d'espace représente l'espace "occupé" par le geste dans son exécution. Un mouvement en ligne droite comme un coup de poing sera direct, un mouvement avec de multiples changements de directions - une révérence exagérément compliquée par exemple - sera indirect. La notion de flux s'apparente au contrôle mis sur le mouvement. Un geste minutieux est restreint ; un geste est libre lorsqu'on abandonne le contrôle sur le mouvement, comme dans le cas d'un tournoiement effréné.

L'analyse du mouvement de Laban a grandement influencé le domaine de la danse mais on la retrouve également comme base à de nombreux travaux en psychologie et informatique sur les gestes expressifs, et en particulier sur la reconnaissance et la synthèse d'émotions par la gestuelle. Par exemple, Chi *et al.* [29] s'appuient sur l'Analyse du Mouvement de Laban pour générer des animations expressives chez un avatar.



(a) La kinésphère de Laban.



(b) L'espace général de Laban.

FIG. 10: La kinésphère et l'espace général, caractérisation de l'espace par Laban.

3.3 Gestuelle et émotion

Notre gestuelle permet, entre toutes ses autres fonctions (voir paragraph 3.1.2 page 52), de communiquer et d'exprimer une émotion. Boone et Cunningham montrent que les enfants sont capables dès quatre ans de reconnaître des émotions discrètes en observant des mouvements expressifs [11]. Les mêmes auteurs montrent trois ans plus tard que ces enfants sont également capables d'encoder de l'émotion par la gestuelle en faisant danser un ours en peluche [12]. Cette capacité à reconnaître les états affectifs par la gestuelle gagne en maturité avec l'âge mais décline chez les personnes âgées, en particulier pour la reconnaissance d'émotions négatives [100]. Dans cette section, nous présentons des travaux issus du domaine de la psychologie et ayant pour but d'identifier des caractéristiques de mouvements et de positions du corps permettant de communiquer un état émotionnel. La plupart des études dans ce domaine se base sur le même schéma suivant. Les auteurs font tout d'abord l'hypothèse de plusieurs caractéristiques de mouvement ou de position. Du contenu visuel est alors produit, présentant chaque combinaison des caractéristiques considérées. Ce contenu visuel peut prendre la forme de vidéos ou d'images d'acteurs ou de mannequins 3D. Chaque combinaison de caractéristiques est ensuite présentée à un groupe d'évaluateurs. Ces derniers reportent alors l'émotion qu'ils pensent être exprimée. Les réponses des observateurs sont ensuite analysées afin de déterminer la contribution de caractéristiques ou de combinaisons de caractéristiques à la reconnaissance d'une émotion. Ce schéma d'analyse présente de nombreux biais comme ceux présentés à la section 2.5 (page 46). Les protocoles expérimentaux mis en place sont donc conçus pour s'affranchir le plus possible de ces biais. Dans cette section, nous présentons principalement trois études dont l'intérêt est de présenter des caractéristiques adaptables à l'informatique, et les paramètres de leurs interprétations vers des ensembles donnés d'émotions.

De Meijer [46] établit l'influence de sept caractéristiques générales de mouvement dans l'expression de douze états affectifs, dont les six émotions basiques d'Ekman (voir paragraphe 1.2.2 page 17). De Meijer s'appuie entre autres sur l'analyse du mouvement de Laban (voir partie 3.2 page 53) pour déterminer sept dimensions de mouvement : le mouvement du tronc (étiré ou voûté), le mouvement des bras (ouverture ou fermeture), la direction verticale du mouvement (vers le haut ou vers le bas), sa direction sagittale (vers l'avant ou vers l'arrière), sa force (fort ou léger), sa vélocité (rapide ou lent), et enfin sa directivité (direct ou indirect). 90

sujets ont évalué des vidéos de chaque combinaison de valeurs de ces sept caractéristiques. De Meijer en tire l'influence de chaque dimension de mouvement à l'attribution d'une émotion par les évaluateurs. Chaque caractéristique seule ne permet pas de déterminer l'émotion exprimée ; c'est la combinaison des caractéristiques qui permet la reconnaissance. Par exemple, la joie est caractérisée par le tronc droit, une ouverture des bras, et un mouvement orienté vers le haut. Les autres caractéristiques sont non significatives pour cette émotion. Chaque combinaison n'est pas forcément représentative d'une émotion ; certaines combinaisons peuvent exprimer une émotion mais semblent "dissonantes". C'est le cas lorsque l'une des caractéristiques ne correspond pas à la valeur attendue. Par exemple, le mouvement du tronc est primordial pour la reconnaissance de chacune des émotions étudiées. Un mouvement fait avec le tronc voûté ne sera pas reconnu comme de la joie, même si les autres caractéristiques expriment la joie.

Coulson [37] se focalise sur des positions statiques et l'impact du point de vue de l'observateur au lieu de considérer le mouvement. Une originalité de ce travail est que Coulson considère le corps comme un assemblage articulé, et utilise un mannequin 3D pour représenter les positions à tester, ce qui lui permet d'obtenir des images neutres en dehors de la position et de définir formellement les positions étudiées (par des angles de rotation entre les différents segments du corps). Coulson simplifie le squelette humain pour ne retenir que sept degrés de liberté : la translation du centre de gravité vers la caméra ou en s'éloignant d'elle, l'inclinaison de la tête, l'inclinaison du tronc, la rotation de l'abdomen, les deux rotations de l'épaule, et la pliure du coude. Les deux bras sont positionnés symétriquement. Pour chaque degré de liberté, entre deux et trois valeurs d'angle sont choisies, ainsi que deux valeurs de translation (vers l'avant ou vers l'arrière) du centre de gravité. Des images de toutes les combinaisons possibles sont ensuite générées selon trois points de vue (face, profil et vue en plongée arrière gauche). Le point de vue ne semble cependant pas jouer dans la reconnaissance : les mêmes postures sont évaluées comme la même émotion quel que soit le point de vue. Les résultats obtenus sont cohérents par rapport aux autres études. Ainsi, la joie est caractérisée par le tronc et la tête érigés voire en arrière (rotation de 20°vers l'arrière), des bras vers les côtés ou le haut, et tendus ou pliés à 50°au niveau du coude, et une translation vers l'avant.

Wallbott [147] focalise sur les mouvements de la partie supérieure du corps. Tandis que De Meijer et Coulson partent de caractéristiques et évaluent leur impact sur la reconnaissance d'émotions, Wallbott adopte l'approche inverse en extrayant des caractéristiques communes à des mouvements expressifs. Des acteurs jouent quatorze émotions selon un scénario donné. Les vidéos sont ensuite analysées par des observateurs experts pour en extraire des catégories de mouvement. Les catégories extraites sont le mouvement du tronc, des épaules (vers le haut, l'avant ou arrière), la position de la tête, les mouvements des bras et des mains, ainsi que trois caractéristiques plus générales du mouvement : l'activité, l'expansion dans l'espace, et l'énergie du mouvement. Outre les mouvements eux-mêmes, Wallbott considère aussi les fonctions de ces mouvements comme les manipulateurs, les illustrateurs (en particulier le pointage) et les emblèmes (voir partie 3.1.2, page 52). Wallbott choisit donc d'étudier des expressions de l'émotion pour trouver les caractéristiques de mouvement correspondantes, alors que De Meijer cherchait à tester des caractéristiques posées en hypothèse. Il en résulte une caractérisation des différentes émotions par des mouvements typiques. Ainsi, la joie se caractérise par un tronc érigé, les épaules tournées vers le haut, la tête penchée en arrière, des bras tendus vers l'avant ou le haut, des mains actives, une utilisation intensive d'illustrateurs

et une activité, une expansion et une énergie fortes.

Ces travaux en psychologie posent une base robuste pour la reconnaissance d'émotions par la gestuelle en informatique. En effet, ils identifient et valident des caractéristiques de position et de mouvement propres à chaque émotion, en suivant une méthodologie expérimentale visant à minimiser les biais. Il est cependant difficile de réutiliser ces travaux tels quels pour créer un programme de reconnaissance d'émotions par la gestuelle. Les principales difficultés viennent de la définition et de la captation des caractéristiques : ce qui est facilement visible par un humain peut être très difficile à retrouver de façon automatique. Pour la définition des caractéristiques, des définitions suffisamment précises ou l'adoption d'une description formalisée dès le départ sont nécessaires à une réutilisation directe. Les travaux de De Meijer [46] ou ceux de Coulson [37] proposent ainsi des descriptions précises (de Meijer) voire formelles (Coulson), permettant ainsi une réutilisation en limitant l'interprétation nécessaire à la traduction vers une méthode d'extraction automatique. Pour la captation, certaines caractéristiques nécessitent des dispositifs dédiés. De Meijer considère par exemple la force du mouvement, et définit un mouvement "fort" comme effectué avec les muscles contractés. Cette caractéristique n'est donc captable qu'avec un dispositif adapté (électromyogramme).

La mise en œuvre informatique de toutes les caractéristiques d'un travail existant dans le but d'une réutilisation est donc attrayante mais difficile. Les travaux en psychologie s'appuient sur des méthodologies expérimentales solides pour identifier et valider des caractéristiques. Transposer de tels travaux à l'informatique requiert cependant une certaine prudence : les caractéristiques considérées doivent être captables et décrites suffisamment précisément pour être réutilisées. Certaines caractéristiques peuvent être délaissées lors de la transposition d'une étude à un système automatique ; il est alors nécessaire de s'assurer que les variables délaissées sont non significatives pour l'expression des émotions à reconnaître.

Parmi les trois études présentées, nous retenons les travaux de De Meijer [46] pour notre cas applicatif (voir chapitre 6) . En effet, les caractéristiques identifiées sont à la fois décrites précisément et relativement simples à extraire par des algorithmes. De plus, la contribution de chaque caractéristique à l'expression des douze émotions étudiées y est évaluée et quantifiée.

3.4 Mise en œuvre informatique de la reconnaissance d'émotions par la gestuelle

Pour étudier les systèmes de reconnaissance d'émotions vus au chapitre 2, nous reprenons notre schéma d'analyse du chapitre précédent, reposant sur les grandes étapes que sont la capture, l'analyse et l'interprétation des comportements de l'utilisateur. La conception d'un système de reconnaissance d'émotions basé sur la gestuelle nécessite de répondre aux questions inhérentes à chacune de ces étapes : Quels senseurs utiliser ? Quelles caractéristiques extraire ? Enfin, quel modèle d'émotion utiliser et quel algorithme d'interprétation mettre en place ?

3.4.1 Spécificités du niveau capture

Au niveau capture, on trouve principalement dans la littérature deux types de capteurs : les caméras vidéo et les combinaisons de capture du mouvement. Les systèmes basés caméras sont plus accessibles et bénéficient d'algorithmes connus d'analyse d'image. Les combinaisons de capture sont dans l'ensemble plus précises mais financièrement moins accessibles. Ces deux types de dispositifs permettent une capture complète du corps. D'autres types de capteurs ne proposent qu'une capture partielle de la posture ou des mouvements, comme la *pressure chair* [102] présentée en section 2.2. Nous nous concentrons cependant ici sur les systèmes considérant le corps dans son intégralité. Caméra et combinaisons délivrent des données différentes (respectivement images et coordonnées) et induisent donc des algorithmes différents pour l'extraction de caractéristiques au niveau analyse.

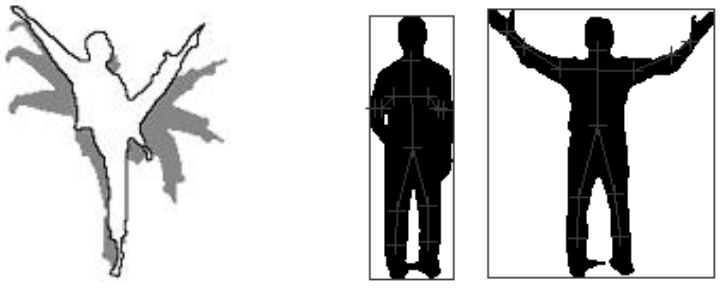
3.4.2 Spécificités du niveau analyse

Au niveau analyse, on note une prédominance de travaux ne considérant que la partie supérieure du corps, et des bras et des mains en particulier. Ces travaux se focalisent généralement sur des situations d'interaction et de communication. Les informations de gestuelles viennent alors suppléer les informations émotionnelles tirées des expressions faciales et de la voix. Ainsi Balomenos *et al.* [5] considèrent diverses catégories de mouvements des mains (par exemple "applaudissement rapide", "applaudissement lent", "levé de la main") pour améliorer la reconnaissance par les expressions faciales. Parfois seules quelques caractéristiques adaptées au contexte d'usage sont utilisées pour la reconnaissance. Ainsi, Mota et Picard [102] déterminent l'intérêt d'un enfant en situation d'apprentissage par ordinateur grâce à sa seule position sur une chaise munie de senseurs de pression. Dans cette étude, les auteurs choisissent de ne pas restreindre les mouvements de l'enfant dans le protocole expérimental. Le système permet de reconnaître l'intérêt que porte l'enfant à la tâche d'apprentissage à plus de 75%. La connaissance complète de la position et du mouvement n'est donc pas forcément nécessaire. Pollick *et al.* [121] montrent que des vidéos de mouvements ne représentant que des points aux articulations du bras sont suffisantes pour que des observateurs humains, non conscients de ce que représente la vidéo, puissent évaluer la dimension d'activation de l'émotion jouée. Cet exemple montre qu'une représentation même dégradée d'un mouvement peut encore véhiculer de l'information émotionnelle.

Dans le cadre de notre mémoire, nous nous intéressons à l'étude du corps au complet, sans dégradation de l'information. Nous n'avons relevé que peu de travaux en informatique prenant cette même direction, à l'exception de l'équipe du laboratoire Infomus de Gênes, créé et dirigé par Antonio Camurri. Ce laboratoire s'est attaché depuis ces quelques dernières années à étudier l'expressivité du geste et à reconnaître les émotions par la gestuelle dans diverses situations. Concernant les caractéristiques de mouvement à étudier, Castellano distingue dans [25] le "quoi" du "comment", c'est-à-dire la distinction entre la reconnaissance par identification de gestes connus et la reconnaissance par identification de la manière dont ces gestes sont réalisés. La différence peut être mieux appréhendée en considérant le "quoi" comme une expression verbale ("lever la main", "pencher la tête") et le "comment" comme un adverbe ("rapidement", "directement"). L'équipe du laboratoire Infomus s'intéresse au "comment" là où les approches modalités et fonctionnelles abordent plutôt le "quoi" (section 3.1 page 51).

Les premiers travaux du laboratoire dans ce domaine datent de 1997 [18]. Camurri *et al.* y présentent les prémisses d'une analyse basée image du mouvement pour extraire l'expressivité du mouvement qui s'appuie entre autres sur l'Analyse du Mouvement de Laban (section 3.2). Ces travaux sont poursuivis en 2000 dans [23], où Camurri et Trocca introduisent le calcul d'expansion et de contraction du corps, en rapport avec la kinésphère de Laban. Plus tard, en 2004, Volpe identifie dans [143] des caractéristiques de mouvement relatives à la danse extraites automatiquement pour la reconnaissance d'émotions. Le système créé extrait des caractéristiques de séquences dansées. Ce système est divisé en cinq processus successifs.

- Le premier processus consiste en la capture du danseur au moyen d'enregistrements vidéo.
- Le deuxième processus est un traitement bas niveau du signal vidéo en vue d'extraire la silhouette et de suivre le mouvement. En particulier, Volpe extrait les images de mouvements de silhouettes (*Silhouette Motion Images* ou SMI) : Ces SMI sont le résultat de la juxtaposition des silhouettes des n dernières images de la vidéo, desquelles on a retiré la silhouette de la dernière image (figure 11a). Une SMI ne représente donc plus que le mouvement qui a été effectué dans les n dernières images.
- Le troisième processus effectue une analyse du mouvement de plus haut niveau. Tout d'abord, Volpe calcule la quantité de mouvement (*Quantity of Movement* ou QoM) du danseur. Cette quantité de mouvement se rapporte directement à une SMI ; de fait la QoM est égal à l'aire d'une SMI sur les n dernières images, divisée par l'aire de la silhouette de la dernière image. Cette division permet une normalisation par rapport à la taille du danseur à l'image (taille réelle du danseur et distance à la caméra). Volpe calcule également un index de contraction (*Contraction Index* ou CI) permettant de savoir si le mouvement est expansif (par exemple bras et jambes écartés) ou contracté (danseur replié sur lui-même). Le CI dans une image i est égale au rapport entre l'aire de la silhouette dans i et son rectangle englobant (figure 11b). Le CI est donc une mesure comprise entre 0 et 1 ; une valeur proche de 1 indique une position contractée (la silhouette remplit le rectangle englobant), une mesure proche de 0 une position expansive (l'expansion agrandit le rectangle englobant sans changer l'aire de la silhouette). Enfin, Volpe considère également dans ce niveau de traitement la trajectoire du centre de gravité au cours du temps.



(a) Exemple d'une SMI sur 4 images.

(b) Calcul du CI par carré englobant.

FIG. 11: Extraction de caractéristiques de gestuelle par la vidéo. Figures tirées de [143].

– Le quatrième processus se base sur les quantités de mouvement pour établir une segmentation du geste dans une danse. En dessous d'un certain seuil (trouvé empiriquement) de QoM, l'auteur considère que le danseur est en phase de pause. Il considère alors un mouvement comme étant ce qu'il y a entre deux phases de pause. Volpe obtient ainsi des "cloches de mouvement" (*motion bells*) (voir figure 12). L'analyse de ces cloches permet d'en tirer la fluidité et l'impulsivité d'une séquence dansée. Une danse est fluide si elle présente de longues cloches de mouvement. L'impulsivité se calcule sur la forme de la cloche. Une cloche en pic montre un mouvement impulsif, une cloche écrasée un mouvement coulé.

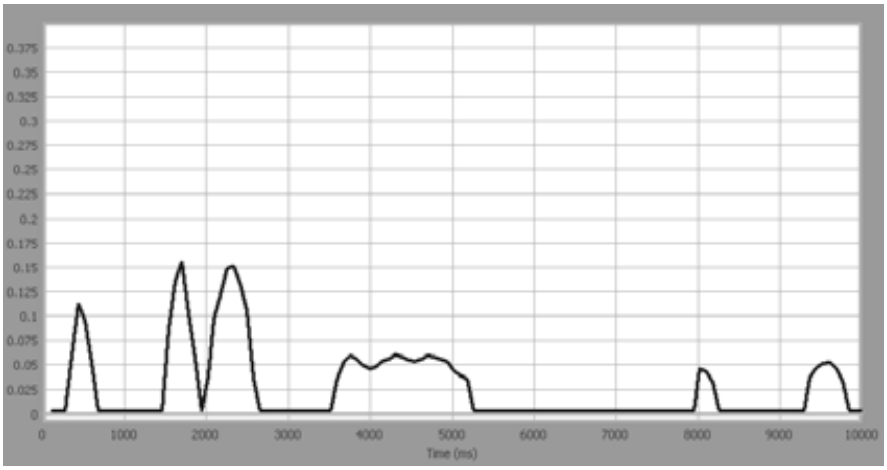


Fig. 12: Segmentation du mouvement en cloches de mouvement et phases de pause. Figure tirée de [143].

– Enfin, le cinquième processus est celui de l'interprétation. Volpe utilise des arbres de décision classant les valeurs des différentes caractéristiques extraites en quatre émotions basiques : la colère, la peur, la tristesse et la joie. Pour chaque séquence de l'ensemble d'entraînement, on donne comme solution au système l'émotion que le danseur a voulu exprimer. Les résultats sont ensuite comparés à l'évaluation d'observateurs humains. Globalement le système reconnaît très bien la colère (60% de reconnaissance par le système contre 60,6% pour l'humain) ; les résultats oscillent ensuite, le plus mauvais étant la reconnaissance de la peur (9,15% de reconnaissance correcte pour la machine contre 55% pour le groupe humain).

Castellano, en 2008, poursuit les travaux de Volpe dans [25] et se concentre sur la segmentation du geste et sur l'analyse des "cloches de mouvement" obtenues grâce à la mesure de la QoM. Elle définit ainsi un modèle mathématique de la dynamique du mouvement expressif. Ce modèle comprend seize caractéristiques calculables sur le mouvement. Ces caractéristiques sont évaluées sur plusieurs cas d'études, notamment sur l'expressivité gestuelle d'un joueur de piano [27]. Ce modèle mathématique offre un cadre pour une analyse fine du mouvement au niveau du geste.

De ces travaux, nous soulignons tout d'abord la nécessité de pouvoir traiter des caractéristiques de façon séquentielle, c'est-à-dire une série de caractéristiques élevant à chaque fois le niveau d'abstraction. Dans le système de Volpe par exemple, la détection des SMI permet le calcul de la QoM, qui est utilisée pour segmenter une danse en phases de pause et de mouvement. Cette segmentation permet ensuite l'étude de chaque phase de mouvement de manière fine, comme le propose Castellano. Nous retenons donc que les caractéristiques peuvent se situer à plusieurs niveaux d'abstraction, et que des caractéristiques de plus haut niveau (par exemple la fluidité d'un geste) peuvent être calculées non pas sur des signaux de capture (flux vidéo) mais sur des caractéristiques de plus bas niveau (calcul de QoM et segmentation du geste). Le fait qu'une information de base puisse subir diverses transformations pour en extraire des informations de haut niveau est donc à prendre en compte dans la suite de ce mémoire. Nous soulignons ensuite la nécessité de pouvoir traiter des caractéristiques bloquantes ou *temporisées*. La méthode de segmentation d'une danse en phases de mouvement et phases de pause illustre parfaitement le problème de ces caractéristiques temporisées. Pour calculer la fluidité d'un geste par exemple, il est nécessaire d'attendre que le geste soit terminé afin de disposer de l'information dans son intégralité. La caractéristique ne peut donc pas être calculée tant que le geste n'est pas terminé, ce qui est entièrement sous le contrôle du danseur. Il est donc impossible d'obtenir une mesure continue de la fluidité. D'un point de vue système, les QoM calculées à chaque image sont donc mises en mémoire jusqu'à ce que le geste se termine et que le danseur soit en phase de pause. A ce moment seulement, l'extraction de la caractéristique de fluidité peut être lancée. L'existence et l'utilisation de telles caractéristiques temporisées constitue un requis logiciel que nous traitons au chapitre 5 dans le cadre de notre modèle d'architecture.

3.4.3 Spécificités du niveau interprétation

Enfin, le niveau interprétation ne propose pas de spécificité propre à la reconnaissance par la gestuelle. Les notions générales abordées au chapitre 2 (section 2.4, page 42) s'appliquent de la même façon à ce domaine.

3.5 Conclusion

Ce chapitre a exposé les travaux sur la reconnaissance des émotions par la gestuelle. Nous avons discuté des différentes approches et taxonomies du geste et nous sommes penchés tout particulièrement sur les différentes caractéristiques de mouvement véhiculant de l'information émotionnelle en prenant comme exemples des travaux en psychologie, qui nous ont permis de souligner les problèmes pour la réutilisation de ces caractéristiques et des informations pour leur interprétation. Nous avons également décrit les travaux du laboratoire Infomus de Gênes, explicitant ainsi l'existence de caractéristiques à plusieurs niveaux d'abstraction et le problème des caractéristiques temporisées, susceptibles de bloquer un système.

Nous retenons pour nos travaux une approche reposant sur les mouvements de tout le corps. Sans considérer un contexte particulier, nous adoptons pour les caractéristiques à considérer une approche basée sur les mouvements et non les fonctions de communication du geste. Nous considérons trois niveaux dans le processus de reconnaissance d'émotions : les niveaux

Capture, Analyse et Interprétation. Nous posons comme base de notre travail deux des travaux présentés dans ce chapitre : la publication de De Meijer [46] et le mémoire de thèse de Volpe [143]. En effet, ces deux travaux fournissent des éléments réutilisables tant au niveau analyse qu'au niveau interprétation.

Ce chapitre conclut la première partie de ce mémoire, dans laquelle nous avons présenté le contexte de notre recherche. Dans le premier chapitre nous avons choisi définition, théories et modèles pour l'émotion. Dans le second chapitre, nous avons présenté le domaine de la reconnaissance d'émotions en informatique, avant de se focaliser sur la reconnaissance par la gestuelle dans ce chapitre trois. A partir des éléments retenus dans cette partie, nous développons notre contribution dans la partie suivante. Nous y présentons une identification des concepts de l'interaction multimodale à la reconnaissance d'émotions au chapitre 4, et décrivons un modèle conceptuel pour la conception d'applications interactives sensibles à l'émotion au chapitre 5.

Deuxième partie

Un modèle d'architecture pour la reconnaissance d'émotions dans les applications interactives

Chapitre 4

Modèle d'architecture : un point de vue interaction homme-machine de la reconnaissance d'émotions

Le domaine de la reconnaissance d'émotions est encore un domaine relativement jeune. Le modèle de Gaines [59] sur l'évolution du niveau d'apprentissage d'une technologie découpe l'apprentissage en six étapes. La percée technologique consiste en l'essai de résolution d'un problème soldé par un échec, menant à des avancées créatives. Cette phase est suivie par la phase de réplication de la percée technologique, amenant une expérience croissante du domaine. Cette expérience permet de passer à la phase d'empirisme, où des règles basées sur l'expérience sont tirées pour la conception. Ces règles sont ensuite étendues et des hypothèses formulées pour fournir une théorie du domaine : c'est la phase de théorisation. Dans la phase d'automatisation, les théories sont employées automatiquement, prédisent l'expérience et produisent des règles de conception. Enfin, dans la phase de maturité, les théories sont acceptées et utilisées sans remises en cause. En ce qui concerne la reconnaissance d'émotions, nous considérons que nous nous situons entre les phases de réplication et d'empirisme. En effet, de nombreux travaux consistent encore à étendre l'expérience que nous pouvons avoir en reconnaissance d'émotions. Quelques travaux explicitent des problématiques liées à la conception de systèmes de reconnaissance (par exemple [89]). Ces quelques avancées montrent la maturité de la phase de réplication et un besoin grandissant d'un espace de conception pour la construction d'applications interactives intégrant la reconnaissance d'émotions. Notre objectif est donc de combler ce besoin en fournissant un modèle intégrateur pour la conception de systèmes de reconnaissance d'émotions. Pour atteindre cet objectif, nous puisons nos solutions d'un domaine plus avancé en termes d'ingénierie : l'interaction homme-machine (IHM). En particulier, nous considérons la reconnaissance d'émotions comme une interaction multimodale et nous inspirons donc des travaux en interaction multimodale.

Dans ce chapitre, nous présentons les travaux en IHM que nous avons réutilisés avant de présenter notre solution architecturale pour la reconnaissance d'émotions par un système informatique. Le cadre architectural établi, nous nous focaliserons au chapitre suivant sur l'architecture de la partie dédiée à la reconnaissance d'émotions. Nous résumons les requis de la solution architecturale qui sont issus des états de l'art des chapitres précédents mais aussi

les limitations que nous nous sommes fixés et définissons le cadre conceptuel de notre modèle.

4.1 Modèle d'architecture : définition et exemples

Face au foisonnement des systèmes conçus, notre objectif est la conception d'un modèle d'architecture unificateur pour la conception d'applications interactives pouvant reconnaître des émotions.

4.1.1 Définition d'un modèle d'architecture

Nous nous basons sur [39] pour définir la notion de modèle d'architecture. Nous n'expliciterons, dans ce paragraphe, que les notions utiles à notre travail. Coutaz et Nigay définissent une architecture logicielle de la manière suivante :

> L'architecture d'un système informatique est un ensemble de structures comprenant chacune : des composants, les propriétés extérieurement visibles de ces composants et les relations que ces composants entretiennent. [...] Un composant est [ici] une unité d'abstraction, dont la nature dépend de la structure considérée dans le processus de conception architecturale. Ce peut être un service, un module, une bibliothèque, un processus, une procédure, un objet, une application, etc.

Un composant est alors caractérisé par ce qu'il encapsule (structure et comportement internes) mais également par ses possibilités d'interactions avec d'autres composants (propriétés extérieurement visibles et comportements observables). L'architecture d'un système peut avoir plusieurs structures, correspondant à plusieurs points de vue.

Le processus de conception architecturale (illustré à la figure 13) débute par le choix d'un modèle architectural de référence. Les requis et les spécifications externes du logiciel à concevoir permettent une décomposition fonctionnelle du besoin et la conception d'une architecture conceptuelle adaptée au besoin. Les quatre étapes suivantes sont la décomposition modulaire de l'architecture conceptuelle, l'identification des processus, la mise en correspondance des modules et des processus, et enfin la mise en correspondance des processus avec les processeurs. Ces quatre étapes permettent d'arriver à une architecture implémentationnelle, c'est-à-dire la façon dont doit être implémenté le logiciel. Le modèle de référence et les architectures conceptuelle et implémentationnelle sont influencées par des styles et des motifs architecturaux.

La décomposition fonctionnelle consiste à exprimer les besoins du système. Dans le cadre de la reconnaissance d'émotions, il est donc nécessaire d'exprimer les besoins du domaine en termes de fonctions. Un modèle architectural de référence est une décomposition fonctionnelle normalisée pour un problème connu. En IHM, les modèles ARCH et PAC-Amodeus sont des exemples de modèles architecturaux de référence. Le choix du modèle architectural de référence est donc dépendant des besoins généraux du système. L'architecture conceptuelle est obtenue après le choix (optionnel) d'un modèle architectural de référence et des requis et besoins du système. Elle permet de répondre aux requis du système. De cette architecture conceptuelle va émerger l'architecture implémentationnelle.

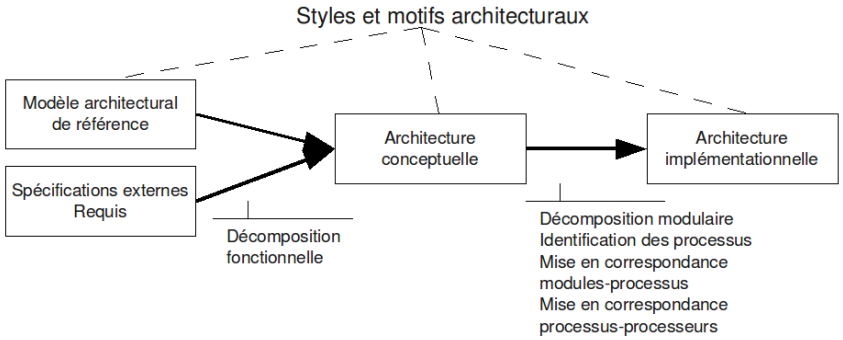


Fig. 13: Processus de conception architecturale. Figure tirée de [39].

Les motifs et styles architecturaux sont des notions orthogonales à la conception d'une architecture logicielle. Un style architectural est une aide générique à l'élaboration de structures architecturales. Il est défini par un vocabulaire d'éléments (concept de machine abstraite, concept d'objet), de règles de configuration entre les éléments (par exemple, deux états dans une machine à états doivent être reliés par une transition), et une sémantique fournissant un sens à la description structurelle. Un motif est une micro-architecture répondant à un problème récurrent. Un motif au niveau conceptuel est un micro-modèle de référence. En tant que "micro-architecture", un motif répond à un style donné.

Nos travaux en ingénierie logicielle pour la reconnaissance d'émotions s'articulent autour des ces différentes notions. Dans ce chapitre, nous proposons tout d'abord une décomposition fonctionnelle de la reconnaissance d'émotions en trois unités d'où émerge un motif de référence pour la conception d'un système sensible à l'émotion. Dans le chapitre 5 nous proposons un modèle d'architecture conceptuelle pour les applications de reconnaissance d'émotions.

4.1.2 Exemples de modèles de référence

Nous présentons ici quatre modèles de référence en Interaction Homme-Machine, que nous utilisons ensuite dans ce mémoire.

Le modèle MVC

Le modèle Modèle-Vue-Contrôleur ou MVC [85] propose une architecture sous forme d'agents composés de trois facettes (figure 14). La facette Modèle est le cœur fonctionnel de l'agent. La facette vue prend en charge l'affichage et capte les actions de l'utilisateur (entrée de texte, clics de souris, etc.). Enfin, la facette contrôleur a pour rôle de contrôler les données provenant de la vue et d'observer les changements d'états du modèle.

Lorsqu'un utilisateur effectue une action *via* la vue, un évènement est envoyé au contrôleur. Celui-ci va alors demander au modèle d'effectuer un traitement. Le modèle notifie la vue de son changement d'état. La vue est capable de lire les données directement depuis le modèle.

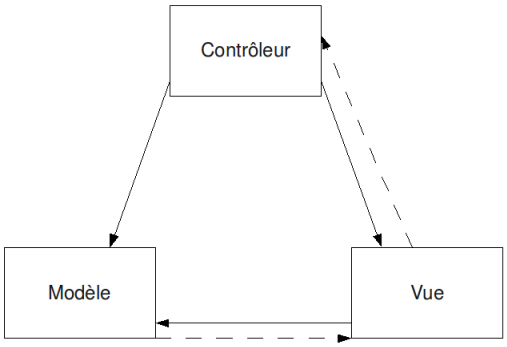
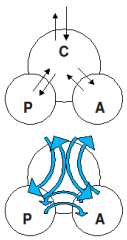


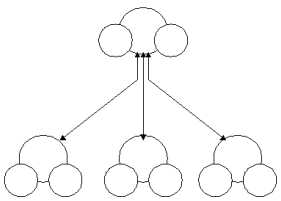
FIG. 14: Le modèle MVC [85]. En trait plein : appels de méthodes, en traits pointillés : évènements.

Le modèle PAC

Le modèle PAC (Présentation, Abstraction, Contrôle) est un modèle d'architecture en interaction homme-machine basé sur des agents comportant trois facettes [38] représentées figure 15 : la facette Présentation (P), la facette Abstraction (A), et la facette Contrôle (C). Un agent PAC est à la fois un système interactif et un constituant de système interactif. La facette Présentation prend en charge l'interaction avec l'utilisateur en entrée et sortie. La facette Abstraction est le noyau fonctionnel du composant, comprenant ses différentes fonctionnalités. Enfin, la facette Contrôle permet le dialogue entre les facettes Présentation et Abstraction (voir figure 15a). Les agents PAC sont organisés en une hiérarchie (figure 15b). Le rôle des facettes Contrôle est donc également d'établir un dialogue avec des agents PAC pères ou fils, par l'intermédiaire de leurs facettes Contrôle.



(a) Communication dans un agent PAC.



(b) Hiérarchie d'agents PAC.

FIG. 15: L'architecture Présentation Abstraction Contrôle.

Le rôle de dialogue de la facette Contrôle se scinde en deux fonctions : tout d'abord l'arbitrage, recouvrant la synchronisation et la coordination à la fois entre les facettes Abstraction et Présentation d'un agent mais aussi entre les différents agents ; ensuite, la traduction des

représentations utilisées entre les facettes Abstraction et Présentation.

Le modèle ARCH

Le modèle ARCH [6] est une extension du modèle de Seeheim [113] et propose une modélisation en cinq composants indépendants permettant une grande modifiabilité du système (voir figure 16). Le contrôleur de dialogue (CD) est la clé de voûte du système et sert de lien entre les branches fonctionnelles et de présentation.

La branche fonctionnelle est constituée d'un Noyau Fonctionnel (NF) et d'une Interface au Noyau Fonctionnel (INF). Le noyau fonctionnel regroupe les fonctions abstraites relatives au concept. Ces fonctions sont totalement orientées informatique et indépendantes de la représentation du concept par l'utilisateur. L'interface au noyau fonctionnel sert de lien entre le noyau fonctionnel et le contrôleur de dialogue. Son rôle est de traduire les concepts purement informatiques du noyau fonctionnel en concepts plus orientés tâches utilisateur.

La branche de présentation est constituée d'un Composant d'Interaction Physique (CIP) et d'un composant d'Interaction Logique (CIL). Le premier est l'interface au plus bas niveau et dépend du système utilisé. Le deuxième traduit ces informations pour les envoyer au contrôleur de dialogue.

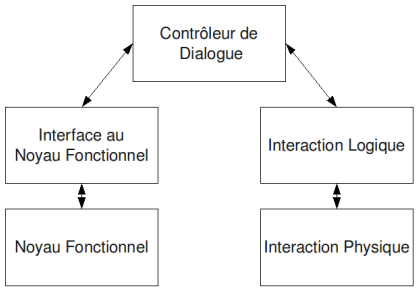


Fig. 16: Le modèle ARCH.

Le modèle PAC-Amodeus

Le modèle se fonde sur le modèle en 5 composants Arch [103]. Le modèle PAC-Amodeus affine le modèle Arch en définissant le contrôleur de dialogue comme une hiérarchie d'agents PAC (figure 17). Le modèle PAC-Amodeus est complètement présenté dans [103].

4.2 Interaction Homme-Machine et multimodalité

Dans cette section, nous énonçons les travaux existants en IHM que nous avons utilisé dans notre approche inspirée de l'interaction multimodale. L'interaction multimodale fait intervenir plusieurs modalités d'interaction pour interagir avec un système. En transposant une

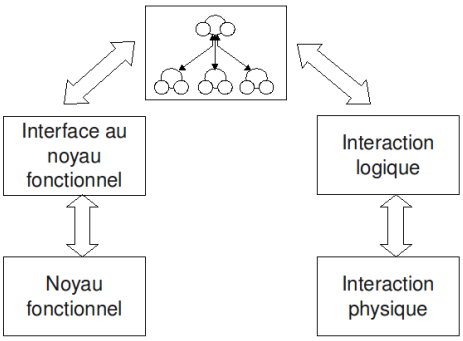


FIG. 17: Le modèle PAC-Amodeus [103]

modalité d'interaction par une modalité d'expression émotionnelle, le lien avec l'interaction multimodale est immédiat. L'intérêt d'adopter ce point de vue est la possibilité d'exploiter et d'infléchir les résultats en interaction multimodale au cas de la reconnaissance d'émotions. Or, le domaine de l'interaction multimodale est certainement plus établi en termes d'ingénierie logicielle que celui de la reconnaissance d'émotions, qui est beaucoup plus récent. L'interaction multimodale est née en 1980 du paradigme "mets çà là" de Richard Bolt [10]. Dans le paragraphe suivant, nous rappelons les points les plus importants de l'interaction multimodale qui ont influencés notre proposition.

4.2.1 Définition d'une modalité

Le paradigme de la multimodalité en interaction homme machine se caractérise par l'utilisation de plusieurs moyens de communication pour communiquer avec une machine, comme l'illustre le célèbre exemple du "mets çà là" combinant parole et geste [10]. Nous reprenons la définition d'une modalité donnée en introduction du mémoire de thèse de Bouchet sur l'ingénierie de l'interaction multimodale en entrée [14] :

> L'utilisateur interagit avec le système informatique pour réaliser des tâches. Les moyens d'action et de perception, appelés "modalités d'interaction", sont les médiateurs matériels et logiciels permettant à un utilisateur d'agir sur le système informatique ou de percevoir son état. Les modalités d'interaction composent le module interface du système informatique.

Dans le cadre de ce travail, nous nous basons sur la définition d'une modalité donnée par Nigay dans [106]. Une modalité y est définie par la relation suivante :

$$modalité =< d, sr > | < modalité, sr > \quad (1)$$

où :
– d est un dispositif physique d'interaction : souris, caméra, microphone, capteurs de mouvement, GPS, écran...

– *sr* est un système représentationnel, c'est à dire un système conventionnel structuré de signes assurant une fonction de communication.

Cette définition permet de caractériser une interaction en entrée en prenant en compte et en reliant deux niveaux d'abstraction à la fois du point de vue humain et du point de vue système :

– Du point de vue humain, le dispositif est à un bas niveau d'abstraction. L'humain agit sur le dispositif. Le système représentationnel est au niveau de la cognition de l'utilisateur.
– D'un point de vue système, le couple renseigne sur le dispositif mis en œuvre et sur le domaine et le format des données échangées entre l'homme et la machine.

Modalité d'entrée et modalité de sortie

Une modalité permet l'interaction avec la machine. Une interaction peut se produire de l'utilisateur vers la machine : la modalité utilisée est alors une modalité d'entrée. Dans ce cas, l'utilisateur utilise un dispositif physique d'entrée pour interagir avec la machine, comme un clavier, une souris, une caméra ou un microphone. A l'inverse, une interaction peut s'effectuer de la machine vers l'utilisateur. La modalité utilisée est alors une modalité de sortie. L'information est transmise selon un certain système représentationnel (code couleur, texte, synthèse vocale) via un dispositif physique de sortie (écran, enceintes).

Dans nos travaux, nous ne considérons que les modalités en entrée du système puisque nous adoptons le parallèle entre modalités d'interaction et modalités d'expression émotionnelle pour la reconnaissance. Dans le reste de ce mémoire, nous spécifierons donc "modalité de sortie" pour désigner les modalités en sortie du système ; le terme "modalité" se rapportera aux modalités d'entrée.

Modalité active et modalité passive

Une modalité peut être passive ou active. Une modalité est dite *active* lorsqu'elle requiert un effort conscient de l'utilisateur pour utiliser un dispositif physique en vue de spécifier une commande au système [15]. Par exemple, les couples <*clavier, commandes Unix*> ou <*microphone, langage naturel*> sont des modalités actives. A l'inverse, une modalité est dite *passive* lorsqu'elle ne requiert pas d'action consciente de l'utilisateur. Dans ce cas, le système ne fait que scruter ("monitorer") l'utilisateur. Les informations obtenues sont injectées dans le système tout comme les informations obtenues *via* une modalité active. Par exemple, tous les systèmes présentés au chapitre 2 proposent une interaction en entrée *via* des modalités passives : les divers capteurs utilisés (caméras, microphones, senseurs de l'ANS) ne font qu'enregistrer les informations de l'utilisateur, sans le solliciter. Par exemple, un récepteur GPS suit la position d'un utilisateur sans que celui-ci n'ait à le lui spécifier continuellement.

Transfert de modalités

La définition (1) est une définition récursive. En développant cette définition, on obtient qu'une modalité est constituée d'un dispositif physique et d'une suite de 1 à n systèmes représentationnels. En développant la définition, on obtient donc :

$$modalité =< ... << d, sr_1 >, sr_2 > ...sr_n > \quad (2)$$

Cette écriture explicite la possibilité de transfert des systèmes représentationnels. Cette notion représente le fait qu'une même information peut être traduite selon plusieurs systèmes représentationnels et utilisée par d'autres modalités avant d'obtenir une commande ou tâche complète.

4.2.2 Multimodalité

La multimodalité est la multiplicité des modalités, c'est à dire des dispositifs et des systèmes de représentation utilisés afin d'agir sur le système ou d'avoir des informations sur ce système.

Pratiquement, la multimodalité en sortie représente la multiplicité des moyens mis en œuvre pour rendre perceptible une information à l'utilisateur. Un exemple de multimodalité en sortie est le jeu vidéo : les consoles de salon récentes permettent aux jeux de mettre en œuvre un couplage entre retour visuel à l'écran, retour sonore par les enceintes et retour tactile par les vibrations de la manette. Par exemple, un atterrissage après une haute chute sera notifié par une couleur rouge envahissant l'écran, un bruit sourd de chute et une forte vibration de la manette, renforçant ainsi l'immersion du joueur. Les modalités de sortie ont tout comme les modalités d'entrée fait l'objet de travaux en ingénierie logicielle [93].

Dans le cadre de la reconnaissance d'émotions par ordinateur, nous nous intéressons à l'entrée du système, et considérons donc la multimodalité en entrée. La multimodalité en entrée signifie l'utilisation de plusieurs moyens de communication pour spécifier des commandes ou tâches élémentaires à la machine. En reprenant le paradigme du "mets çà là" de Bolt [10], l'utilisateur peut formuler vocalement une commande ("mets ça là"), où les inconnues ("ça" et "là") sont spécifiées par le biais d'une autre modalité d'interaction : l'utilisateur pointe du doigt l'objet à déplacer en prononçant le "ça" puis la place qu'il doit occuper en prononçant le "là". Ce paradigme permet de favoriser une interaction plus naturelle avec la machine mais aussi plus efficace car la bande passante entre l'utilisateur et le système est élargie par la disponibilité de plusieurs modalités.

Caractérisation d'un système multimodal : l'espace MSM

L'espace MSM est un espace de caractérisation d'un système interactif multimodal. Il comprend six axes explicités figure 18. Le premier axe "fusion-fission" concerne la fusion (rassemblement de données selon plusieurs modalités) et la fission (éclatement des données issues d'une modalité en plusieurs) des données. La fusion est abordée plus en détail au paragraphe 4.2.4. Les deux axes suivants sont le nombre et le sens (entrée ou sortie) de canaux

pour le système considéré. Un canal de communication regroupe un dispositif physique capable d'émettre ou de recevoir des données. Le quatrième axe est celui du niveau d'abstraction. Chaque canal de communication peut abstraire l'information à des niveaux différents, du niveau le plus bas (signal) au niveau le plus haut (signification). Le cinquième axe est celui du contexte et caractérise la capacité du système à prendre en compte le contexte. Enfin, le sixième axe est celui du parallélisme. Il caractérise la capacité du système à gérer le parallélisme aux niveaux physique (utilisation de plusieurs dispositifs en même temps), élémentaire (utilisation de plusieurs canaux de communication en même temps) et tâche (construction de plusieurs tâches de manière concurrente).

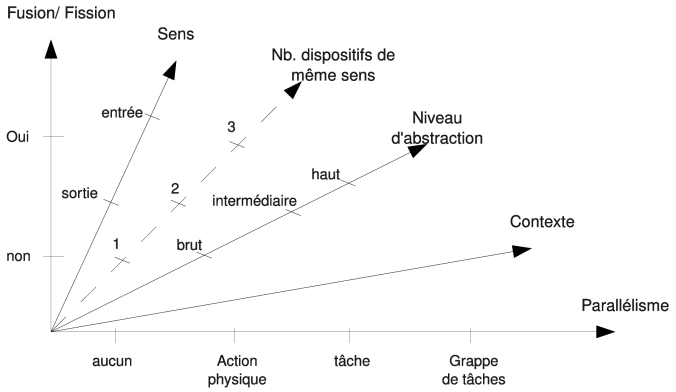


Fig. 18: Les 6 axes de l'espace MSM.

Dans cet espace, une modalité d'interaction correspond à un canal de communication avec un niveau haut sur l'axe "Niveau d'abstraction". Un système multimodal implique au moins deux modalités de même sens (entrée ou sortie).

4.2.3 Relations entre modalités : espace TYCOON et propriétés CARE

Pour caractériser la multimodalité, nous retenons deux espaces de conception qui caractérisent les relations entre modalités d'interaction : l'espace TYCOON [94] et les propriétés CARE de la multimodalité [40].

Espace TYCOON

L'espace TYCOON présente un cadre théorique à l'étude d'interfaces multimodales, un langage de spécification et un module multimodal intégrant les évènements provenant de plusieurs modalités. Le cadre théorique d'étude d'interfaces multimodales se présente comme un espace à deux dimensions (voir figure 19). La première dimension est celle du type de coopération de modalités. Martin identifie cinq types de coopération : le transfert, l'équivalence, la spécialisation, la redondance et la complémentarité. La seconde dimension est celle des buts que l'on cherche à atteindre lors de la conception de l'interaction : par exemple, l'apprentissage ou une interaction rapide. Chaque coopération de modalités en abscisse peut permettre

d'atteindre des objectifs parmi ceux listés en ordonnée.

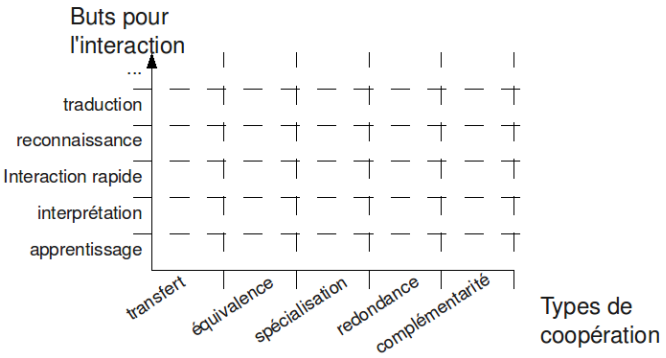


FIG. 19: Cadre d'étude théorique pour la multimodalité. Figure tirée de [94].

La coopération par **transfert** signifie qu'une unité informationnelle produite par une modalité est utilisée par une autre modalité. Par exemple, un utilisateur peut effectuer des gestes à la souris pour effectuer certaines commandes (par exemple, tracer un trait de droite à gauche pour aller à la page précédente dans un navigateur internet). Dans ce cas, la modalité <*souris, couple de coordonnées (x,y)*> est transférée vers une modalité <*souris, geste*> ; les coordonnées de la souris sont utilisées pour retrouver les gestes effectués par l'utilisateur.

La coopération par **équivalence** implique qu'une unité informationnelle peut être obtenue selon plusieurs modalités de façon alternative. Un exemple d'équivalence est un système où l'utilisateur peut énoncer une commande au système grâce au clavier ou en l'énonçant à voix haute. L'équivalence permet d'améliorer la reconnaissance et/ou la vitesse d'interaction en permettant de choisir la modalité la plus efficace et/ou la plus rapide ; elle permet également de s'adapter à l'utilisateur qui peut choisir la modalité d'interaction.

La **spécialisation** implique que l'information est analysée par la même modalité. Martin différencie la spécialisation de type modalité et la spécialisation de type données. Par exemple, le son est spécialisé dans la notification d'erreurs (un bip est émis par un dispositif dédié lors d'une commande interdite). Il s'agit d'une spécialisation de type modalité si les sons émis par ce dispositifs ne sont utilisés que pour ces erreurs. Il s'agit d'une spécialisation de type données si une commande interdite ne déclenche qu'une notification sonore. La spécialisation permet d'aider l'utilisateur à interpréter les informations de sorties ; elle permet également d'aider à la reconnaissance, par exemple en forçant l'entrée d'un paramètre de commande au clavier, plus facile à reconnaître que par la voix.

La coopération par **redondance** signifie qu'une unité informationnelle est traitée par plusieurs modalités équivalentes, en même temps : par exemple, en énonçant une commande à la fois au clavier et à la voix. La redondance permet selon les cas une interaction plus rapide et un meilleur apprentissage.

	Information identique	Information différente
Pas de fusion	équivalence	spécialisation
Fusion	redondance	complémentarité

Tab. 4: Comparaison de l'équivalence, spécialisation, redondance et complémentarité dans l'espace TYCOON. Tableau tiré de [94].

Enfin, la coopération par **complémentarité** signifie que plusieurs unités informationnelles différentes sont produites par plusieurs modalités mais doivent être fusionnées. Le paradigme "mets ça là" est un exemple de complémentarité des modalités. La complémentarité permet une meilleure reconnaissance de la commande et une interaction plus rapide en attribuant les différents éléments d'une commande aux modalités les plus adéquates.

A l'exception du transfert, Martin compare les différentes coopérations entre modalités selon qu'elles requièrent une fusion ou non et la transmission d'informations différentes ou identiques. La table 4 illustre cette comparaison.

Propriétés CARE

Coutaz *et al.* proposent quatre propriétés permettant de caractériser l'interaction multimodale : les propriétés CARE, pour Complémentarité, Assignation, Redondance, et Equivalence. Cet ensemble de propriétés se décline en propriétés D-CARE pour les dispositifs, et L-CARE pour les systèmes représentationnels.

Soit un système informatique s permettant d'accomplir une tâche t. Les propriétés D-CARE (resp. L-CARE) sont définies de la façon suivante.

La complémentarité entre n dispositifs $\{d_1, ..., d_n\}$ (resp. n systèmes représentationnels $\{sr_1, ..., sr_n\}$) signifie que tous ces dispositifs (resp. systèmes représentationnels) sont nécessaires à l'accomplissement de la tâche t. t ne peut être effectuée s'il manque l'information de l'un des dispositifs ou système représentationnel. Le "mets ça là" de Bolt [10] que nous avons évoqué précédemment est un exemple de complémentarité entre la commande vocale et le geste de pointage.

L'assignation d'un dispositif à un système représentationnel désigne l'obligation d'utiliser ce dispositif pour ce système représentationnel ; l'assignation d'un système représentationnel à une tâche désigne l'obligation d'utiliser ce système représentationnel pour cette tâche. L'assignation indique donc l'absence de choix.

L'équivalence entre n dispositifs $\{d_1, ..., d_n\}$ (resp. n systèmes représentationnels $\{sr_1, ..., sr_n\}$) signifie que tous ces dispositifs (resp. systèmes représentationnels) permettent un même système représentationnel (resp. d'accomplir une même tâche) en produisant les même données. Par exemple, sous Windows, il est possible de sauvegarder un document à la

souris par le menu "Fichier→Sauvegarder" ou par le raccourci clavier "Ctrl+S". L'équivalence permet donc le choix des modalités.

Enfin, la redondance de n dispositifs $\{d_1, ..., d_n\}$ (resp. n systèmes représentationnels $\{sr_1, ..., sr_n\}$) signifie que ces dispositifs (resp. systèmes représentationnels) sont équivalents mais que leur utilisation de concert (de façon séquentielle ou parallèle) est forcée. Cela correspond à l'assignation de modalités équivalentes et ne permet donc plus le choix de la modalité. La redondance permet en revanche d'améliorer la robustesse du système.

Conclusion

Les types de coopération de TYCOON et les propriétés CARE de la multimodalité sont des propriétés bien établies de l'interaction multimodale. Par rapport à CARE, TYCOON propose le transfert de modalités et une granularité plus fine de la spécialisation (assignation dans CARE). Cependant, la notion de transfert a également été implémentée par la récursivité de la définition d'une modalité présentée au paragraphe 4.2.1. Les propriétés CARE et l'espace TYCOON sont donc semblabes. Dans nos travaux, nous avons retenu les propriétés CARE. Celles-ci proposent un certain formalisme qui nous est utile dans la définition de notre modèle d'architecture et l'intégration de la reconnaissance d'émotions en tant qu'interaction multimodale. De plus, les propriétés CARE ont été la base d'un modèle d'architecture conceptuelle pour la conception d'applications interactives multimodales : le modèle ICARE (pour Interaction Complémentarité Assignation Redondance Equivalence) [15]. Ce modèle pour l'interaction multimodale fournit une base sur laquelle nous nous appuyons pour développer notre propre modèle d'architecture décrit au chapitre 5.

4.2.4 Fusion de données multimodales

Le premier axe de l'espace MSM est celui de la fusion/fission des données. La fusion permet de considérer comme un tout des données véhiculées selon plusieurs modalités (comme le "mets ça là"). La fission permet à l'inverse de morceler une information afin de la présenter selon plusieurs modalités. Fusion et fission des données peuvent être présentes aussi bien dans les modalités d'entrées que dans les modalités de sortie d'un système. La fission de modalité d'entrée correspond à l'éclatement d'une information en diverses modalités. La fusion des données issues de différentes modalités d'entrée peut s'effectuer aux niveaux lexical, syntaxique et sémantique [104]. La fusion au niveau lexical consiste en l'agrégation de signaux directement issus de dispositifs physiques (par exemple, la combinaison des touches "contrôle" et "s" s'est démocratisée comme le raccourci clavier pour sauvegarder). La fusion syntaxique permet de fusionner diverses informations afin de construire une commande (exemple du "mets ça là"). La fusion sémantique consiste en la fusion des résultats de plusieurs commandes spécifiées selon plusieurs modalités d'interaction : par exemple, en dessinant une forme à la souris tout en changeant la couleur par une commande parlée.

Lalanne *et al.* ont publié en 2009 un état de l'art [86] sur les moteurs de fusion de 1980 (travail séminal de Bolt en interaction multimodale) à 2009. Cet état de l'art montre que l'interaction multimodale et les technologies de fusion multimodale sont dans la phase de

maturité du modèle de Gaines (voir introduction de ce chapitre) : les produits commerciaux comme la Wii ou l'iPhone proposent nativement une interaction multimodale.

Nous détaillons ici un mécanisme générique (c'est-à-dire indépendant des modalités) pour la fusion des données multimodales [105], basé sur un conteneur structuré (le melting pot) et des algorithmes pour la fusion syntaxique. Le melting pot est un conteneur dont la structure peut être spécifiée en dehors du système de fusion (par exemple par un fichier XML ou par des informations données par l'application). Un melting pot est un tableau à deux dimensions où chaque ligne contient des *unités informationnelles*, notées UI_{in}, où i est une information complète (par exemple une commande) composée de n unités. Chaque unité informationnelle est également définie par la période de temps dans laquelle elle est acquise. Par exemple, une commande vocale ("Ouvrir Notepad") contient deux unités informationnelles (l'action d'ouvrir et ce qu'il y a à ouvrir) dont la période est le temps de prononciation de la commande. Chaque colonne représente l'ensemble des données acquises à un temps t (voir figure 20).

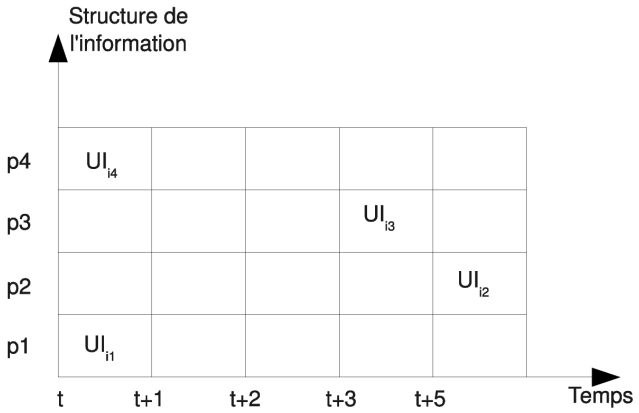
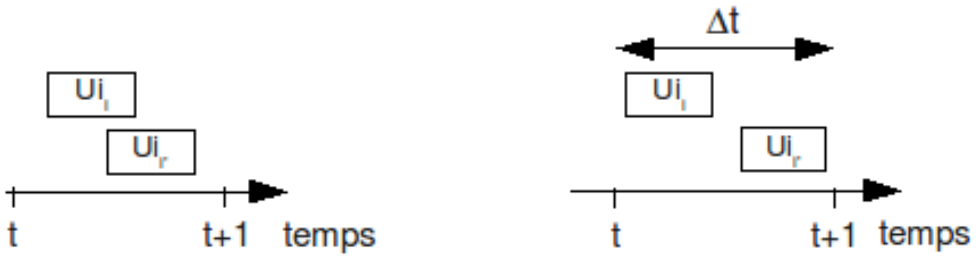


Fig. 20: Le melting pot.



(a) Fusion microtemporelle.

(b) Fusion macrotemporelle

Fig. 21: Fusions temporelles des unités informationnelles.

Le but de la fusion est de remplir, à un temps t, l'ensemble des cases de la colonne afin de reconstituer une commande complète depuis les informations issues des différentes modalités. Pour cela, trois fusions sont effectuées : la fusion *microtemporelle*, la fusion *macrotemporelle*, et la fusion *contextuelle*. La fusion microtemporelle fusionne tout d'abord les unités informationnelles se chevauchant temporellement (voir figure 21a). Cette fusion est effectuée en premier lieu afin de permettre le parallélisme au niveau de l'utilisateur. Chaque unité informationnelle possède en effet une estampille de temps permettant de connaître le temps de début d'expression de l'information et son temps de fin. La fusion macrotemporelle consiste ensuite à remplir les cases vides de la colonne avec des unités informationnelles d'autres types, ne se chevauchant pas mais restant à l'intérieur d'une fenêtre temporelle (définie lors du développement du système) (voir figure 21b). Enfin, la fusion contextuelle désigne comme candidates des unités informationnelles appartenant au même contexte (c'est-à-dire à la commande active). Cette fusion contextuelle permet de s'affranchir des contraintes temporelles, dont elle peut se défaire grâce à un mécanisme permettant de défaire une fusion temporelle déjà accomplie.

4.3 IHM et reconnaissance d'émotions

Dans cette section, nous reprenons les différents points abordés précédemment dans ce chapitre et instancions les concepts abordés au domaine de la reconnaissance d'émotions. Dans un premier temps, nous définissons donc les requis généraux de notre modèle d'architecture, ainsi que ses limitations ; nous définissons ensuite un motif architectural pour la conception de systèmes sensibles à l'émotion, motif sur lequel s'appuiera notre modèle conceptuel présenté au chapitre 5.

4.3.1 Décomposition fonctionnelle : requis et limitations de notre modèle d'architecture

L'étude des travaux existants nous a permis de dégager de nombreux points à prendre en compte dans la conception de notre modèle architectural. Les besoins identifiés reposent sur l'état de l'art et notre objectif d'intégration et capitalisation des travaux existants. Ces besoins sont tous motivés par l'aspect générique de la solution architecturale à identifier. Nous présentons également le cadre de notre modèle et les limites que nous lui avons posées.

Requis guidant la conception de notre modèle d'architecture

Tout d'abord, le modèle ne doit pas dépendre d'une théorie ou d'un modèle de représentation de l'émotion. Bien qu'il y ait une tendance affirmée à l'utilisation d'émotions basiques et du modèle discret dans les applications actuelles de reconnaissance de l'émotion, le modèle d'architecture doit pouvoir permettre l'appui sur d'autres modèles de représentation de l'émotion comme le modèle à composants ou les modèles continus.

D'un point de vue technique ensuite, le modèle d'architecture doit permettre l'intégration de plusieurs dispositifs de capture, caractéristiques à extraire, et de plusieurs algorithmes d'interprétation. Le fait que de nouveaux capteurs puissent apparaître ou de nouvelles caractéristiques puissent être identifiées est à prendre en compte.

Le modèle d'architecture doit enfin permettre de modéliser le cheminement de l'information d'un bas niveau d'abstraction à un haut niveau d'abstraction, ainsi que la fusion et la synchronisation des différentes informations, notamment dans le cas de variables temporisées (voir partie 3.4.2, page 57).

Cadre de notre modèle : ses limites

Premièrement, nous focalisons sur la reconnaissance passive des émotions. En effet, l'expression volontaire et verbalisée d'un état émotionnel à une machine, bien que d'une utilité prouvée (voir paragraphe 2.1.2, page 33), n'est pour le moment que très peu développée dans les travaux existants et s'apparente directement à l'envoi d'une commande. Une telle reconnaissance est donc directement régie par les principes de l'interaction homme-machine. Nous nous limitons donc à une reconnaissance passive, ne requérant aucun effort cognitif de la part de l'utilisateur.

Deuxièmement, nous focalisons sur le cas de la reconnaissance de l'émotion exprimée, au sens littéral, par l'utilisateur. Nous ne cherchons pas à inférer l'émotion ressentie (et encore moins la perception que pourrait en avoir une autre personne). Cependant, l'utilisation de capteurs mesurant les réactions du système nerveux autonome ou l'analyse de caractéristiques de gestuelle permet de se rapprocher de la reconnaissance d'émotions ressentie, ces signaux étant difficiles voire impossibles à feindre ou à cacher (voir conclusion du chapitre 1, page 28).

4.3.2 Un motif en trois niveaux pour la reconnaissance d'émotions

La reconnaissance d'émotions se base sur un cadre classique dans la littérature en trois niveaux successifs : les niveaux de Capture, Analyse et Interprétation (voir figure 22), recoupant ainsi également le fonctionnement humain. Ces trois niveaux sont une vue fonctionnelle d'un système de reconnaissance d'émotions. Ils permettent de scinder le système en délimitant trois rôles fonctionnels.

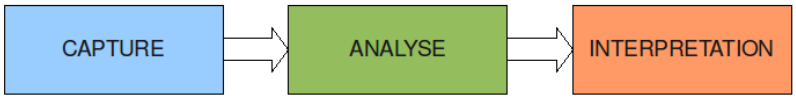


Fig. 22: Les trois niveaux de la branche émotion.

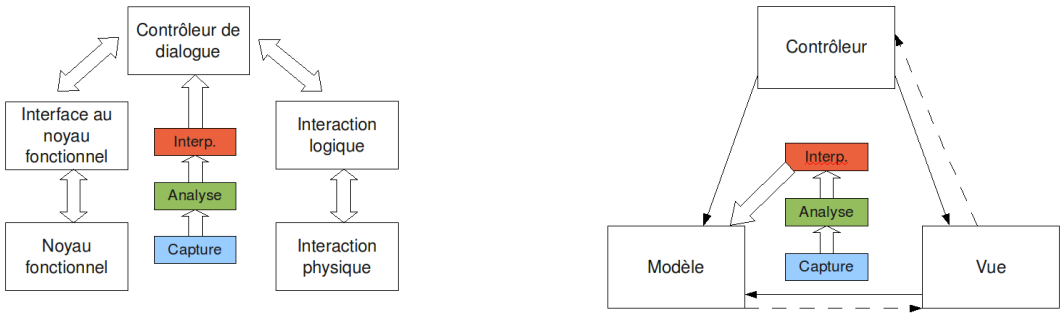
Le niveau Capture a pour rôle d'acquérir des données du monde réel et en particulier de l'utilisateur. Il englobe les dispositifs physiques d'acquisition (caméra, microphone, capteurs physiologiques en sont des exemples) ainsi que leurs interfaces logicielles avec le système (pilotes et traitements préalables). Le niveau Analyse regroupe les traitements permettant d'extraire, depuis les données capturées, des caractéristiques pertinentes pour la reconnaissance d'émotions. Le niveau Interprétation se base sur les valeurs des caractéristiques trouvées pour en déduire une émotion.

Terminologie

Dans la littérature, nous trouvons fréquemment une autre dénomination : les niveaux Signal, Caractéristique, et Décision. De notre étude de l'existant (chapitres 1, 2 et 3), nous avons conclu que la reconnaissance d'émotions se fait par l'obtention d'un ou plusieurs signaux provenant d'un ou plusieurs dispositifs. Des caractéristiques sont extraites de ces signaux. Dans la grande majorité des cas, le problème de reconnaissance de l'émotion est ensuite réduit à un problème de classification ou de décision. La plupart de ces travaux exploitent un modèle discret des émotions, ou un modèle continu segmenté en régions. Cette dénomination ne reflète pas la variété des techniques physiques ou logicielles mises en œuvre, en particulier pour l'interprétation. Par exemple, certains travaux se basant sur d'autres modèles que le modèle discret commencent à émerger. Ainsi, Lisetti *et al.* [88] s'appuient sur le modèle componentiel de Scherer, et représentent une émotion par l'ensemble des réponses aux critères d'évaluation des stimuli (Stimulus Evaluation Checks ou SECs). Il n'y a donc ici pas de classification de l'état émotionnel de l'utilisateur. Nous avons donc choisi d'utiliser le terme "niveau Interprétation" pour ce niveau plutôt que "Décision", car il n'induit aucune connotation sur l'algorithme ou le modèle d'émotions utilisés. De même, nous avons choisi le terme de "niveau Capture" au lieu de "niveau Signal", qui sous entend un degré très bas d'abstraction ; or, nous avons souligné la possibilité de traitements à ce niveau. Pour la même raison, nous avons choisi le terme "niveau Analyse", plus générique que "niveau Caractéristique".

4.3.3 Intégration de la branche émotion dans des applications interactives

La branche émotion peut être intégrée dans des modèles d'architecture classiques en interaction homme-machine. La figure 23 illustre deux modèles clés de l'IHM : le modèle ARCH [6] et le modèle MVC [85], présentés au paragraphe 4.1.2 (page 67). Nous utilisons le mécanisme de *branching* du modèle ARCH pour y intégrer la branche émotion (voir figure 23a). La branche émotion s'intègre de la même façon dans le modèle PAC-Amodeus. Dans le cas du modèle MVC (figure 23b), nous considérons une nouvelle facette (la branche émotion) constituée des trois niveaux Capture, Analyse et Interprétation.



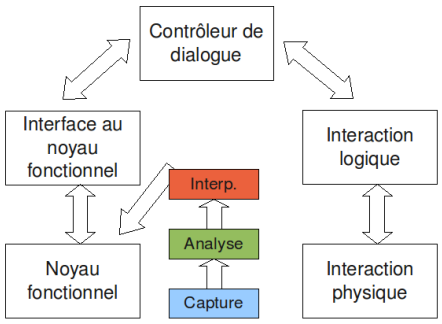
(a) Intégration au modèle ARCH.

(b) Intégration au modèle MVC.

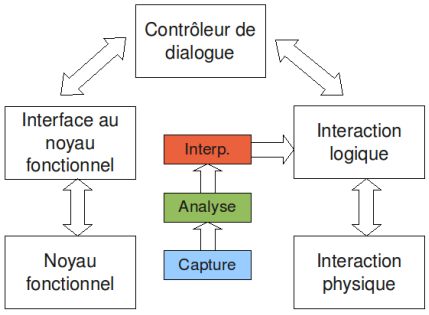
FIG. 23: Intégration de la branche émotion dans les modèles classiques de l'interaction homme-machine : ARCH et MVC.

Dans la figure 23, la branche émotion est connectée au contrôleur de dialogue de ARCH et à la facette Modèle de MVC. D'autres connexions sont cependant possibles. Nous avons identifié trois cas correspondant aux différents rôles que l'émotion peut jouer dans le cadre d'un système interactif.

Comme illustré figure 23, les émotions de l'utilisateur peuvent avoir un impact direct sur le contrôleur de dialogue (CD). Le contrôleur de dialogue est responsable du séquencement au niveau tâche. Chaque tâche ou objectif correspond à un fil de dialogue. Dans ce premier cas, où la branche émotion est connectée au contrôleur de dialogue, les tâches et leur séquencement peuvent être modifiés selon l'émotion reconnue. Par exemple, dans un système interactif d'apprentissage, reconnaître la tristesse ou la colère de l'apprenant pourrait déclencher l'apparition d'une boîte de dialogue offrant une aide sur l'exercice en cours.



(a) Manipulation de l'émotion par le noyau fonctionnel.



(b) Manipulation de l'émotion par l'interaction logique.

Fig. 24: Intégration de la branche émotion dans ARCH.

Le deuxième cas est celui où la branche émotion est manipulée par la branche fonctionnelle (regroupant le noyau fonctionnel et l'interface au noyau fonctionnel), présenté à la figure 24a. L'émotion reconnue est alors un objet du domaine. C'est le cas, par exemple, de l'application de ballet augmenté que nous présentons au chapitre 7, où l'émotion exprimée par le danseur au travers de ses mouvements est présentée au public.

Le troisième cas est celui où l'émotion détectée a un impact sur la branche interaction (figure 24b). Par exemple, l'émotion reconnue pourrait déclencher le changement de modalités de sortie. Pour l'interaction en entrée, la reconnaissance d'une émotion comme le stress chez un pilote de ligne pourrait par exemple induire un changement dynamique d'une modalité vocale vers une modalité mettant à contribution une interface physique (manette).

4.4 Intégration de la reconnaissance d'émotions aux principes de la multimodalité

Nous rappelons que nous faisons le parallèle entre la reconnaissance affective et l'interaction en entrée d'un système informatique. Nous caractérisons donc la reconnaissance affective dans l'espace MSM (voir partie 4.2.2). Tout d'abord, nous ne considérons que l'entrée comme sens d'interaction. L'axe "nombre de dispositifs" est considéré dans son entier : nous ne posons pas de limite au nombre de dispositifs mis en œuvre pour la reconnaissance. Il en est de même pour l'axe "niveau d'abstraction". De fait, l'émotion se situe à un haut niveau d'abstraction, tout système de reconnaissance d'émotions se place donc au niveau "haut" de cet axe. Nous prenons en compte la fusion des données. Nous choisissons par contre d'ignorer le contexte d'interaction, en nous focalisant sur l'expression pure des émotions. Enfin, nous considérons le parallélisme jusqu'au niveau tâche. Une émotion est une réponse immédiate à un stimulus, à courte durée. Le parallélisme au niveau "grappe de tâche" n'a donc pas de sens dans notre contexte.

4.4.1 Définition d'une modalité dans le cadre de la reconnaissance passive d'émotions

Nous avons vu au chapitre 1 que l'homme exprime ses émotions par plusieurs canaux : les expressions faciales, les tonalités de la voix, la gestuelle et les postures. Les réactions du système nerveux autonome peuvent également prendre part à cette expression. Ces différents canaux sont, dans les travaux existants, assimilés à des modalités. Un système est alors dit multimodal s'il effectue une reconnaissance selon au moins deux de ces canaux. Dans ce paragraphe, nous appliquons tout d'abord cette définition d'une modalité pour en montrer les limites ; nous appliquons ensuite la définition vue au paragraphe 4.2.1.

Limites d'une traduction littérale d'une modalité au domaine de la reconnaissance d'émotions

La nature même de la reconnaissance d'émotions rend la définition d'une modalité difficile à établir. Principalement, la définition donnée par l'équation (1) permet d'évoquer à la fois les aspects humains (utilisateur) et techniques d'une interaction (voir paragraphe 4.2.1). En adaptant littéralement cette définition à celle donné dans l'équation (1) (page 70), on obtient comme définition d'une modalité :

$$modalité =< dispositif,\ canal\ de\ communication\ émotionnelle > \tag{3}$$

La première limite de cette définition est que la séparation en cinq canaux de communication émotionnelle n'est pas forcément soutenue. Ainsi Scherer distingue les composants "expression motrice" et "processus neurophysiologiques" (voir partie 1.2.2, page 15), et ne distingue pas les différents canaux. La gestuelle et les expressions faciales se rattachent au composant "expression motrice", les réactions de l'ANS au composant "processus neurophysiologiques". La voix fait intervenir les deux composants. Les réactions de l'ANS sont parfois divisées en deux modalités distinctes : les réactions du système central (le cerveau) et les réactions du système

périphérique (rythme cardiaque, sudation, etc.). Cette catégorisation en quatre canaux n'est donc pas clairement établie et ne semble pas, dans le cadre de l'émotion, refléter la cognition de l'utilisateur.

La deuxième limite est celle de la granularité trop grossière au niveau technique. Dans cette définition, le dispositif et le canal de communication émotionnelle de la modalité considérée indiquent sans les spécifier les caractéristiques à extraire et les dispositifs à utiliser. Le dispositif dispose certes d'un format de sortie des données captées, mais la notion de canal de communication émotionnelle est trop floue pour que cette définition d'une modalité puisse induire une représentation ou un format des données. La définition (3) offre donc une précision insuffisante d'un point de vue technique.

Enfin, tout comme la définition d'une modalité dans le cadre général (équation (1)), cette définition propose un point de vue utilisateur : quelle modalité adopter pour interagir avec la machine ? Cette question, essentielle en interaction, est caduque dans le cadre de la reconnaissance passive des émotions. En effet, l'expression de l'émotion est intrinsèquement multimodale et n'implique pas un choix conscient de l'utilisateur de la construction de sa communication émotionnelle pour être compris par la machine.

En conclusion, la définition (3), utilisant la définition d'une modalité telle que classiquement vue en reconnaissance d'émotions, présente des limitations techniques tout en offrant un point de vue utilisateur inutile. Nous avons donc choisi, comme dit au chapitre 2 (voir introduction de la section 2.3 page 36) de nommer "canaux de communication affective" ou "canaux de communication émotionnelle" les canaux que sont le visage, la voix, le corps et les ANS, faisant ainsi écho à la notion d'*affective channels* proposée par Picard dans [114].

Spécialisation de la définition d'une modalité pour la reconnaissance d'émotions

Dans le contexte de la reconnaissance d'émotions, nous identifions les niveaux Capture, Analyse et Interprétation aux niveaux articulatoire, syntaxique et sémantique introduits par Vernier dans [142] pour la multimodalité en sortie de systèmes interactifs. Les systèmes représentationnels peuvent appartenir aux niveaux Capture, Analyse ou Interprétation. Dans le cadre de ce mémoire, et contrairement à la définition classique dans la littérature en reconnaissance d'émotions, nous considérons qu'une application de reconnaissance d'émotions est multimodale si elle met en œuvre plusieurs modalités telles que définies par l'équation (1) (page 70).

Le point de vue utilisateur donné par la définition d'une modalité dans le cadre général est inutile dans notre cadre de recherche. Nous adoptons donc un point de vue système uniquement. Dans ce cadre, la définition (1) peut être étendue et précisée. Tout particulièrement, il est possible de distinguer les trois niveaux de Capture, Analyse et Interprétation dans la séquence des différents systèmes représentationnels des données au cours du processus de reconnaissance. Typiquement, la donnée est tout d'abord capturée du monde réel grâce à un dispositif. Elle est ensuite susceptible de subir plusieurs transformations dans ce niveau Capture. La donnée est ensuite envoyée au niveau Analyse, où des caractéristiques sont extraites.

Enfin, cette donnée analysée passe au niveau Interprétation. Une fois encore, elle peut être sujette à une séquence d'interprétations. Nous proposons donc le développement de la définition d'une modalité de la manière suivante.

Soit $modalité =< d, sr > | < modalité, sr >$. On réécrit alors

$$< ... << d, sr_1^C >, sr_2^C > ...sr_n^C >, sr_1^A > ... >, sr_m^A >, sr_1^I > ...sr_p^I > \quad (4)$$

où la séquence des systèmes représentationnels explicite les transferts subis par une donnée depuis le dispositif jusqu'à son interprétation finale. On définit ainsi une **modalité de capture** comme une modalité dont le dernier système représentationnel est défini au niveau Capture. Une **modalité d'analyse** est une modalité dont le dernier système représentationnel est défini au niveau Analyse. Une **modalité d'interprétation** est une modalité dont le dernier système représentationnel est défini au niveau Interprétation.

Contrairement à la définition de la multimodalité dans le contexte d'un système interactif, notre définition ne prend pas en compte l'aspect humain de l'interaction. Cet aspect humain, nécessaire dans le cadre d'une interaction active (construire une commande par exemple), devient inutile dans notre cadre de reconnaissance passive des émotions. Cette définition d'une modalité adopte donc le point de vue de la conception d'un système de reconnaissance d'émotions. La précision de cette définition est nécessaire dans ce contexte et sera exploitée au chapitre suivant.

Le niveau Capture a pour rôle de transformer l'information du monde réel en données exploitables par l'ordinateur. Le système représentationnel utilisé est dépendant du capteur utilisé. Au niveau Analyse, nous considérons que chaque caractéristique extraite et les différentes valeurs qu'elle peut prendre forment un système représentationnel au niveau Analyse. Nous obtenons donc un système représentationnel par caractéristique exploitée pour l'interprétation. Enfin, le système représentationnel du niveau Interprétation définit le format de données qui encode l'émotion reconnue. Ce format est totalement dépendant du modèle d'émotions choisi et donc de son mode de représentation ; nous ne proposons donc pas de format "standard" pour la communication de l'émotion à une application interactive.

Systèmes représentationnels au niveau Capture

Soit $sr_{1..n}^C$ la séquence des systèmes représentationnels à l'intérieur du niveau Capture. Les données peuvent subir de profondes transformations dans leurs natures et leurs représentations. Par exemple, considérant un système reconnaissant l'émotion à partir de coordonnées 3D du corps, on peut imaginer une capture par caméra de l'information réelle. Des algorithmes de suivi seront alors appliqués pour obtenir en sortie du niveau Capture des coordonnées 3D.

Tout système contient une séquence $sr_{1..n}^C$ d'au moins un élément ($n = 1$). En effet, le premier système représentationnel est le flux de données directement fourni par le dispositif.

Systèmes représentationnels au niveau Analyse

Soit $sr^A_{1..m}$ la séquence des systèmes représentationnels du niveau Analyse. Une caractéristique peut être extraite de caractéristiques de plus bas niveau.

Il est possible que la séquence $sr^A_{1..m}$ soit vide. Cela correspond au cas où l'interprétation est directement effectuée sur un système représentationnel du niveau Capture. Par exemple, certains appareils photo du commerce sont capables de détecter un sourire dans un visage[1]. Cette information, fournie par le dispositif, peut être envoyée telle quelle au niveau Interprétation. Dans le cas général cependant, cette séquence comportera au moins un élément.

Systèmes représentationnels au niveau Interprétation

Soit $sr^I_{1..p}$ la séquence des systèmes représentationnels du niveau Interprétation. Dépendamment des besoins de l'application en termes de format de représentation de l'émotion, les données liées à l'émotion reconnue peuvent être transformées par une succession de systèmes représentationnels.

La séquence $sr^I_{1..p}$ des systèmes représentationnels peut être vide. Cela correspond au cas d'un hypothétique dispositif dont le pilote délivrerait directement une émotion reconnue. Il n'y a pas de limite supérieure au nombre de systèmes représentationnels successifs dans le niveau Interprétation. Généralement cependant ce niveau ne comporte qu'un seul élément : une unique interprétation de caractéristiques extraites. Une succession de systèmes représentationnels au niveau Interprétation peut se trouver dans le cas d'une fusion au niveau Interprétation. Par exemple, Castellano [26] propose un système multicanaux où une interprétation est faite pour la voix, le visage et la gestuelle. Chacun de ces canaux dispose donc d'un interpréteur. Les émotions trouvées sont fusionnées ensuite.

4.4.2 Adaptation et extension de la hiérarchie des niveaux d'abstraction

Vernier propose dans [142] de considérer trois niveaux d'abstraction pour une modalité : les niveaux lexical, syntaxique, et sémantique. Nous identifions ces trois niveaux aux niveaux Capture, Analyse et Interprétation de la reconnaissance d'émotions. Du point de vue de la tâche de reconnaissance d'émotions en effet, les informations de capture fournissent l'information de base (lexicale). L'extraction de caractéristiques permet d'obtenir un "vocabulaire" émotionnel. L'arrangement des caractéristiques et leurs combinaisons permettent d'inférer l'émotion, que nous identifions au niveau sémantique. Nous utilisons ces termes - syntaxique, lexicale, sémantique et vocabulaire - au sens large, à des fins d'identification : en effet, rien n'indique que l'émotion soit un langage.

Nous proposons dans le cas de l'informatique affective d'affiner cette hiérarchie de niveaux d'abstraction. Nous partons de la définition d'une modalité explicitée dans l'équation (4). Cette équation montre que la donnée part du dispositif et suit une séquence de transferts de

[1] par exemple, le Sony t200.

systèmes représentationnels avant d'arriver au dernier système représentationnel de l'émotion qui sera utilisé dans l'application interactive. Nous supposons qu'il n'existe pas de boucle dans ces transferts. Soient sr_1 et sr_2 deux systèmes représentationnels à n'importe quel niveau de la branche émotion, où sr_1 est utilisé pour former sr_2. Nous écartons alors le cas où sr_2 serait utilisé pour former sr_1, pour deux raisons : tout d'abord, il nous apparaît inutile d'introduire une boucle de transfert des systèmes représentationnels. Si sr_3 est un système représentationnel issu de sr_1, alors il semble naturel de produire sr_2 et sr_3 en parallèle plutôt que d'introduire une boucle de type $sr_1 \rightarrow sr_2 \rightarrow sr_1 \rightarrow sr_3$. Ensuite, et par ailleurs, nous n'avons jamais rencontré ce cas dans les divers travaux étudiés aux chapitres 2 et 3.

Ceci nous permet de qualifier une modalité comme étant de plus bas niveau qu'une deuxième modalité si la première est utilisée pour former la seconde. Par exemple, si sr_1 permet de former sr_2, alors sr_1 est de plus bas niveau que sr_2. Cette définition ne nous permet pas de qualifier absolument le niveau d'abstraction d'un système représentationnel mais permet d'établir une séquence des systèmes représentationnels d'une même modalité. Nous obtenons donc deux mesures : une mesure absolue à gros grains en trois niveaux Capture, Analyse et Interprétation ; et une mesure relative mais plus fine, permettant de hiérarchiser les différents systèmes représentationnels d'une même modalité.

4.4.3 Multimodalité et composition des modalités, relecture des propriétés CARE

Nous avons identifié la branche émotion en entrée comme une interaction en entrée avec le système. En tant que telle, la branche émotion peut donc être multimodale et combiner plusieurs modalités telles que définies dans le paragraphe 4.4.1.

Il existe de nombreux dispositifs permettant de capter l'information du monde réel et de la transmettre à un ordinateur. Ces dispositifs permettent même de capter des informations sur un utilisateur qu'un humain ne peut percevoir, ou difficilement (par exemple, le rythme cardiaque ou la sudation). Ces dispositifs peuvent être utilisés de concert pour obtenir des données plus nombreuses et/ou plus robustes sur l'utilisateur. De la même façon et au niveau Analyse, il est extrêmement restrictif de se limiter à l'extraction d'une seule caractéristique ou séquence de caractéristiques. Plus souvent, de nombreuses caractéristiques sont extraites en parallèle afin de bénéficier d'un maximum d'information pour l'interprétation. Enfin, il est possible d'avoir plusieurs interprétations en parallèle dans un système, afin d'obtenir une reconnaissance finale plus robuste.

Considérer la branche émotion comme une interaction mettant en œuvre plusieurs modalités permet directement d'appliquer les propriétés CARE ou TYCOON de la multimodalité pour identifier les relations entre un dispositif et un système représentationnel, ou entre deux systèmes représentationnels. Dans ce qui suit, nous présentons les implications des propriétés CARE dans le contexte de la reconnaissance d'émotions.

Propriétés CARE dans le cadre de la reconnaissance d'émotions

En considérant la définition étendue de la multimodalité (4) (page 84), nous identifions le premier système représentationnel sr_1^C comme le flux de données directement émis par le dispositif. En tant que tel, le dispositif est toujours assigné à sr_1^C. Un système représentationnel de niveau Capture sr_i^C peut être le produit de l'assignation d'un système représentationnel précédent ou de plusieurs systèmes représentationnels complémentaires, redondants, ou équivalents. Plusieurs sr^C peuvent donc être utilisés pour former un nouvel sr^C.

N'importe quel système représentationnel de la branche émotion peut être produit par la combinaison de n'importe quelle modalité de plus bas niveau. Un sr^A peut être formé de sr^C, de sr^A de plus bas niveau, ou d'un mélange des deux. Un sr^I peut être formé de sr^I de plus bas niveau, de sr^A, de sr^C, ou d'une quelconque combinaison de la réunion de ces trois ensembles.

Les propriétés CARE de la multimodalité dans le cadre de la reconnaissance d'émotions s'appliquent de la même façon que dans le cadre général de l'interaction multimodale. Nous avons vu qu'un dispositif est toujours assigné au premier système représentationnel de capture, celui-ci correspondant au flux de données émis par le dispositif. En considérant une relation d'ordre $<$ dans l'ensemble $\{C, A, I\}$ telle que $C < A < I$, on a :

- $\forall X \in \{C, A, I\}$, $\forall Y \in \{C, A, I\}$, $X \leqslant Y$, $\forall i \in \mathbb{N}$, sr_i^X est assigné à sr_{i+1}^Y si sr_i^X est nécessaire à la formation de sr_{i+1}^Y. Ceci représente un transfert de modalités.
- $\forall X \in \{C, A, I\}$, $\forall Y \in \{C, A, I\}$, $\forall Z \in \{C, A, I\}$, $X \leqslant Y \leqslant Z$, $\forall i \in \mathbb{N}$, $\forall j \in \mathbb{N}$, sr_i^X et sr_j^Y sont équivalents pour sr_{i+1}^Z si l'un ou l'autre permettent de former sr_{i+1}^Z.
- $\forall X \in \{C, A, I\}$, $\forall Y \in \{C, A, I\}$, $\forall Z \in \{C, A, I\}$, $X \leqslant Y \leqslant Z$, $\forall i \in \mathbb{N}$, $\forall j \in \mathbb{N}$, sr_i^X et sr_j^Y sont complémentaires pour sr_{i+1}^Z s'ils ne sont pas équivalent et sont tous deux nécessaires pour former sr_{i+1}^Z.
- $\forall X \in \{C, A, I\}$, $\forall Y \in \{C, A, I\}$, $\forall Z \in \{C, A, I\}$, $X \leqslant Y \leqslant Z$, $\forall i \in \mathbb{N}$, $\forall j \in \mathbb{N}$, sr_i^X et sr_j^Y sont redondants pour sr_{i+1}^Z s'ils sont équivalents et sont tous deux nécessaires pour former sr_{i+1}^Z.

La notion d'ordre entre niveaux Capture, Analyse et Interprétation permet d'interdire des boucles dans la séquence de systèmes représentationnels comme énoncé au paragraphe 4.4.2.

Les différentes modalités implémentées dans un système de reconnaissance d'émotions sont assignées, complémentaires ou redondantes. La redondance en particulier permet d'améliorer la robustesse d'un système. Plus généralement, l'identification des propriétés CARE à chaque niveau Capture, Analyse et Interprétation offre un cadre génératif permettant d'imaginer des combinaisons n'étant pas encore apparues dans les systèmes actuels. La figure 25 illustre les combinaisons de modalités proposées par les propriétés CARE utilisées dans les travaux existants en reconnaissance d'émotions.

La complémentarité, tout d'abord, est obligatoire au niveau Analyse (les caractéristiques sont toutes interprétées ensemble). Au niveau Capture, il arrive que plusieurs dispositifs non équivalents soient utilisés pour extraire une caractéristique (par exemple, l'utilisation de deux

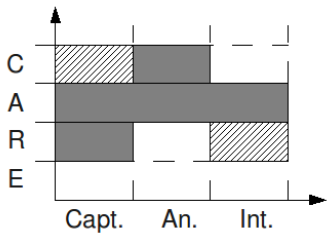


FIG. 25: Fréquence d'utilisation de la Complémentarité, Assignation, Redondance et Equivalence des modalités dans les travaux existants en reconnaissance d'émotions. En gris plein : utilisation fréquente, en hachuré : utilisation occasionnelle, en blanc : non utilisé.

caméras filmant sous des angles différents). La complémentarité au niveau Interprétation serait observée dans le cas où deux modalités permettraient des interprétations non directement compatibles (par exemple, utilisant deux modèles d'émotion différents). Ce cas de figure n'est pas apparu dans notre étude de l'existant.

L'assignation est observée à chaque niveau Capture, Analyse et Interprétation. En effet, aucun des travaux de notre étude de l'existant ne permettait un choix à l'utilisateur ou au système.

La redondance est fortement utilisée au niveau Capture : il n'est pas rare que plusieurs dispositifs équivalents soient mis en place pour améliorer la robustesse de la capture. Au niveau Analyse, la redondance est inexistante. Notre étude de l'existant n'a pas relevé de travaux où une même caractéristique était extraite depuis des dispositifs différents. Enfin, la redondance au niveau Interprétation est de plus en plus utilisée car les travaux sur la reconnaissance multicanaux se basent souvent sur plusieurs interprétations (une pour chaque canal de communication émotionnelle) basées sur les mêmes modèles d'émotions. Ces différentes interprétations fournissent donc des flux de données redondants.

Enfin, l'équivalence, qui signifie un choix entre deux modalités similaires au lieu de forcer leur utilisation en parallèle, n'existe pas encore dans les systèmes de reconnaissance d'émotions. Cette notion de choix de l'utilisation d'une modalité est cependant intéressante car elle permet une meilleure robustesse. Ce choix peut être effectué par l'utilisateur ou par le système. Par exemple, un système reconnaissant les expressions du visage grâce à une caméra peut décider de basculer sur une reconnaissance vocale si la luminosité est trop faible.

4.4.4 Conséquences sur la fusion des données

La fusion des données multimodales est une question majeure du domaine de l'interaction multimodale. Plus particulièrement, la définition d'une modalité en reconnaissance d'émotions ne faisant pas intervenir l'aspect humain, la fusion des données devient la problématique majeure. Nous avons vu en section 4.2.4 (page 76) le moteur de fusion proposé par Nigay

et Coutaz [105]. Le cadre de la reconnaissance d'émotions tel que nous l'avons décrit stipule que :

- nous nous limitons à la reconnaissance passive des émotions. Cette reconnaissance passive entre dans le cadre de l'interaction passive telle que définie dans [15]. L'on ne fait que capturer des données de l'utilisateur sans que celui-ci n'ait besoin d'en être conscient.
- les émotions sont des réactions rapides et hautement synchronisées : lorsque l'on ressent une émotion, tout le corps réagit à l'unisson peu de temps après la perception du stimulus déclencheur.

En prenant en compte ces deux faits, nous arrivons à la conclusion que le système de fusion des données peut être simplifié pour le domaine de la reconnaissance d'émotions. En nous limitant à une reconnaissance passive, nous ne considérons pas les cas où l'utilisateur exprime activement son émotion par rapport à une situation donnée à la machine ; la fusion contextuelle n'a plus lieu d'être, puisque les diverses réactions sont simultanées et correspondent au stimulus précédent directement ces réactions. De plus, ce dernier point nous permet de ne considérer que la fusion micro-temporelle : seules les caractéristiques simultanées sont considérées.

Nous avons vu au paragraphe 4.2.4 qu'il existait trois niveaux de fusion en interaction multimodale : le niveau lexical, le niveau syntaxique et le niveau sémantique. En reconnaissance d'émotions, ces trois niveaux sont identifiés comme la fusion au niveau Capture, au niveau Analyse et au niveau Interprétation. La fusion au niveau Capture consiste à fusionner les signaux des différents dispositifs ayant potentiellement déjà subis certains traitements. Ce bas niveau de fusion relève particulièrement du domaine du traitement du signal. Une fusion au niveau Analyse consiste à fusionner les différentes caractéristiques extraites et à les soumettre ensemble à une unique interprétation. La fusion au niveau Interprétation consiste à fusionner les émotions issues de diverses interprétations. Les résultats ne présentent pas un écart important [88], avec cependant un taux de reconnaissance légèrement meilleur (entre cinq et dix pour cent) pour une fusion au niveau Analyse [25]. Permettre une fusion au niveau Interprétation présente cependant l'avantage de pouvoir faire cohabiter deux systèmes de reconnaissance d'émotions.

Le moteur de fusion présenté au paragraphe 4.2.4 (page 76) offre une structure et des mécanismes génériques permettant la fusion de différents types de données. Nous adaptons donc ce moteur de fusion à la reconnaissance d'émotions et en particulier à la fusion aux niveaux Analyse et Interprétation. Les mécanismes résultant de cette adaptation sont décrits dans le chapitre 5.

4.5 Conclusion

Nous avons vu, dans ce chapitre, les principaux concepts de la multimodalité dans le cadre d'applications interactives, que nous avons ensuite appliqués et instanciés au domaine de la reconnaissance d'émotions. Dans ce chapitre, nous défendons plusieurs points, en nous limitant au cadre d'une reconnaissance passive et directe des émotions (et non pas des autres états affectifs).

Tout d'abord, l'expression émotionnelle est hautement synchronisée. Elle peut être contrôlée dans une optique de conformation à des règles personnelles ou sociales, mais pas contrôlée en vue d'être mieux comprise par un système informatique. Ceci implique que la division en canaux de communication émotionnelle n'est pas une base solide pour la définition d'une modalité. Une définition de la multimodalité faisant intervenir les canaux de communication émotionnelle permet plutôt de refléter le choix du concepteur de limiter la reconnaissance à un ou plusieurs canaux spécifiques de communication émotionnelle.

La reconnaissance d'émotions dans le cadre que nous avons développé pour ce mémoire est donc une forme d'interaction multimodale avec la machine. Elle présente cependant certaines caractéristiques spécifiques lorsqu'on la considère sous l'angle des concepts de la multimodalité. Nous avons proposé dans ce chapitre (équation (4)) une définition de la multimodalité axée sur son aspect système. Cette définition permet de considérer une modalité de façon suffisamment précise pour en appliquer les principes de transfert de système représentationnel et les propriétés CARE de la multimodalité. Notre définition fait également ressortir les trois niveaux Capture, Analyse et Interprétation et offre une hiérarchisation des niveaux d'abstraction dans le système. Cette intégration de la reconnaissance d'émotions dans les principes de la multimodalité permet de mettre en place les fondations de la conception d'un modèle d'architecture pour la reconnaissance d'émotions que nous abordons au prochain chapitre. De plus, cette intégration permet d'aborder la conception d'un système multimodal de reconnaissance d'émotions d'une façon similaire à la conception d'un système interactif multimodal. Notamment, les propriétés CARE permettent de concevoir des combinaisons de modalités lors de la conception d'un système de reconnaissance d'émotions, combinaisons pour le moment écartées dans la littérature.

Nous proposons, dans le prochain chapitre, un modèle d'architecture se basant sur cette vue de la modalité et de la multimodalité.

Chapitre 5

La branche émotion

Dans ce chapitre, nous détaillons la branche émotion introduite au chapitre précédent. Nous proposons un modèle conceptuel pour la conception de systèmes de reconnaissance d'émotions, basé sur une approche à composants [34]. Pour cela, nous définissons dans la première section les architectures à composants. Nous présentons dans la section suivante trois systèmes existants dont nous nous inspirons pour la conception de notre modèle. Dans la section trois, nous présentons une conception générale de la branche émotion où nous définissons les différents composants qui la constitue. La section quatre couvre la spécification des divers composants ainsi définis. Dans la section cinq, nous présentons le moteur de synchronisation, mécanisme sous-jacent permettant la communication entre des composants instanciés selon nos spécifications. Nous discutons en section six de cas particuliers d'implémentation et de validation des composants. Enfin, nous évaluons en dernière section notre modèle, en modélisant deux travaux existants : une étude issue du domaine de la psychologie et un système informatique de reconnaissance des émotions.

5.1 Architecture à composants

5.1.1 Présentation

Pour définir un composant, nous reprenons la définition donnée par Bass [7] (traduction de [14]) :

> "Un **composant** logiciel est une implémentation de fonctionnalités. Le composant est réutilisable dans différentes applications et il est accessible via une interface de développement logiciel. Il peut, mais ce n'est pas une nécessité, être vendu comme un produit commercial. Un composant logiciel est généralement implémenté par et pour une technologie particulière.[1]"

[1] "A software component is an implementation, in software, of some functionality. It is reused as-is in different applications, and accessed via an application-programming interface. It may, but need not be, sold as a commercial product. A software component is generally implemented by and for a particular component technology."

Un composant logiciel est donc vu comme une boîte noire définie par ses ports d'entrée (réceptacles et puits d'évènements) et de sortie (facettes et sources d'évènements). Les réceptacles et facettes sont les interfaces d'entrée et de sortie des composants. Les puits d'évènements sont des points de connexion permettant à un composant de recevoir des évènements de l'extérieur ; les sources d'évènements sont des points de connexion permettant à un composant d'émettre des évènements vers l'extérieur.

D'un point de vue extérieur, un composant n'est donc défini que par ses ports d'entrée/sortie. L'intérieur d'un composant regroupe des fonctions permettant de traiter les données reçues en entrée et d'émettre de nouvelles données en sortie.

Dans le contexte de la reconnaissance d'émotions, l'information provenant du monde extérieur est enregistrée en continu puis passe par une séquence de traitements, chaque traitement élevant le niveau d'abstraction des données véhiculées. La continuité du flux d'enregistrement au cours d'une capture peut être interrompue lors de traitements, notamment lors de l'extraction de caractéristiques temporisées (voir section 3.4, page 59). D'une manière générale, nous considérons cependant des flux continus de données. Un flux de données est constitué de blocs de données valides pendant une certaine durée (typiquement, l'intervalle de temps entre deux mesures).

5.1.2 Atouts et limitations

Les principaux intérêts d'une architecture à composants sont la réusabilité des composants et la modifiabilité de l'application. La réutilisation des composants permet de réduire les coûts et la durée de développement. Dans une application, il est donc possible de remplacer un composant par un autre pourvu qu'ils présentent des interfaces semblables. Une technologie à composants permet enfin de construire une application en assemblant entre elles des briques de base et en les connectant. Ce travail d'assemblage ne nécessite pas de compétences approfondies du développement des composants manipulés.

Les technologies à composants présentent également certains inconvénients [14]. Tout d'abord, elles nécessitent le développement d'une infrastructure de déploiement et d'assemblage plus lourde que les paradigmes plus classiques de développement. Ensuite, il n'existe pour l'instant pas de standard ou d'environnement clairement établis pour créer des applications basées composants.

5.2 Architectures et applications existantes

Dans cette section, nous présentons un modèle d'architecture et deux outils de conception d'applications basées composants. Le modèle MAUI+ASIA [89] est un modèle d'architecture pour l'informatique affective : le système s'étend de la reconnaissance à la synthèse d'émotions. EyesWeb [19] est une application basée composants permettant de créer des systèmes de reconnaissance d'émotions. Nous nous focalisons ici sur la partie reconnaissance du modèle.

Enfin, nous présentons ICARE [14], un outil basé composants pour la création d'applications interactives multimodales.

5.2.1 MAUI+ASIA, une architecture pour l'informatique affective

Le système MAUI (Multimodal Affective User Interface) [108] décrit une architecture pour des agents sociaux capables de reconnaître les émotions exprimées par l'utilisateur (voir figure 26). Ces agents modélisent ensuite l'état de l'utilisateur par ses émotions, ses buts connus, ses traits de personnalité, et sa connaissance. L'agent décide alors d'un état à adopter selon ces mêmes critères. L'agent interagit alors avec l'utilisateur de façon expressive. Le bloc ASIA a en charge de mettre en relation les différents blocs modélisant l'utilisateur (but, émotion, personnalité et connaissance) et de déclencher des changements dans les blocs similaires relatifs à l'agent. Le modèle MAUI se base sur la théorie à composants de Scherer (voir partie 1.2.2 sur la CPM, page 15).

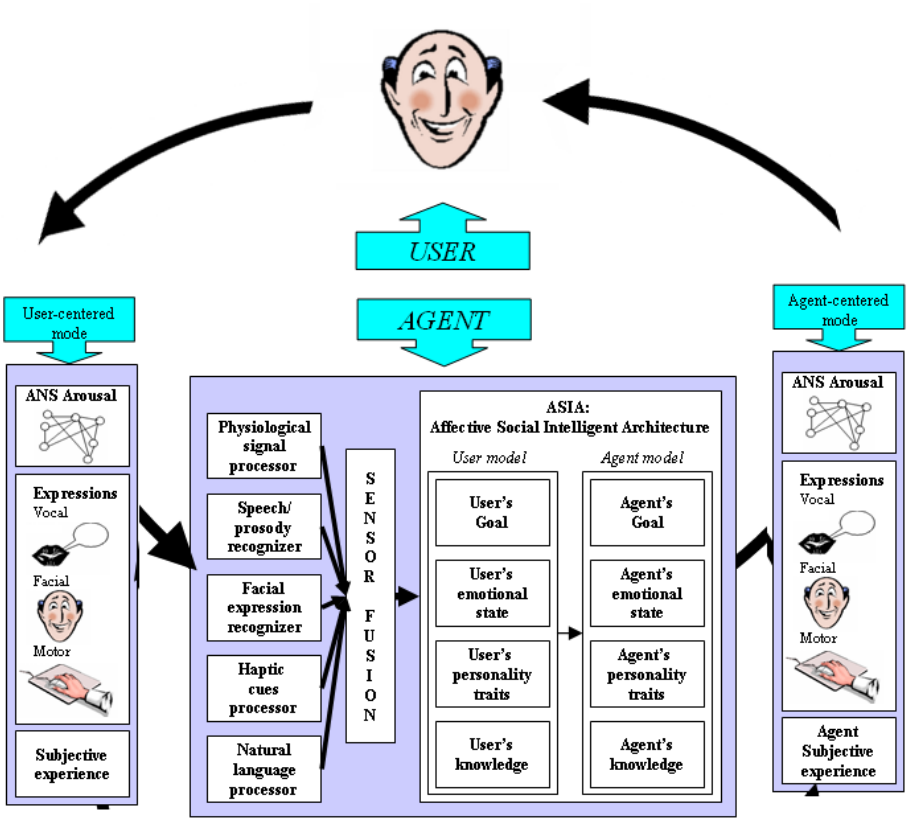


Fig. 26: Le paradigme MAUI+ASIA. Figure tirée de [108]

Le système MAUI+ASIA propose donc un cadre pour la création d'agents sociaux et expressifs. La partie "synthèse de comportement expressif" de ce cadre est prépondérante et est décrite par Lisetti dans [89]. Nous n'aborderons pas cette partie mais nous focalisons sur la partie "reconnaissance d'émotions" du système, qui propose certaines particularités et

fonctionnalités.

Concernant la multimodalité, Paleari et Lisetti proposent dans [108] un cadre architectural partiellement implémenté pour la fusion des données dans le contexte de la reconnaissance d'émotions par ordinateur. Notamment, les auteurs considèrent la fusion aux trois niveaux *signal* (capture), *feature* (analyse), et *decision* (interprétation). Cette fusion est basée sur une synchronisation des différentes données. Les auteurs distinguent également les reconnaissances en temps réel et celles qui ne peuvent l'être (cas des caractéristiques temporisées). Leur architecture est donc scindée en deux parties, la première dédiée au temps réel, la seconde dédiée aux traitements temporisés, comportant des *buffers* afin de réaliser des temporisations (voir figure 27). L'une des particularités les plus intéressantes de ce cadre de fusion multimodale est qu'il propose plusieurs algorithmes de fusion. L'algorithme le plus stable est considéré comme le meilleur et est donc sélectionné automatiquement après une séquence d'entraînement sur un cas particulier. De plus, cette automatisation du choix d'un mécanisme de fusion générique permet au système de s'adapter automatiquement à l'ajout de nouvelles modalités.

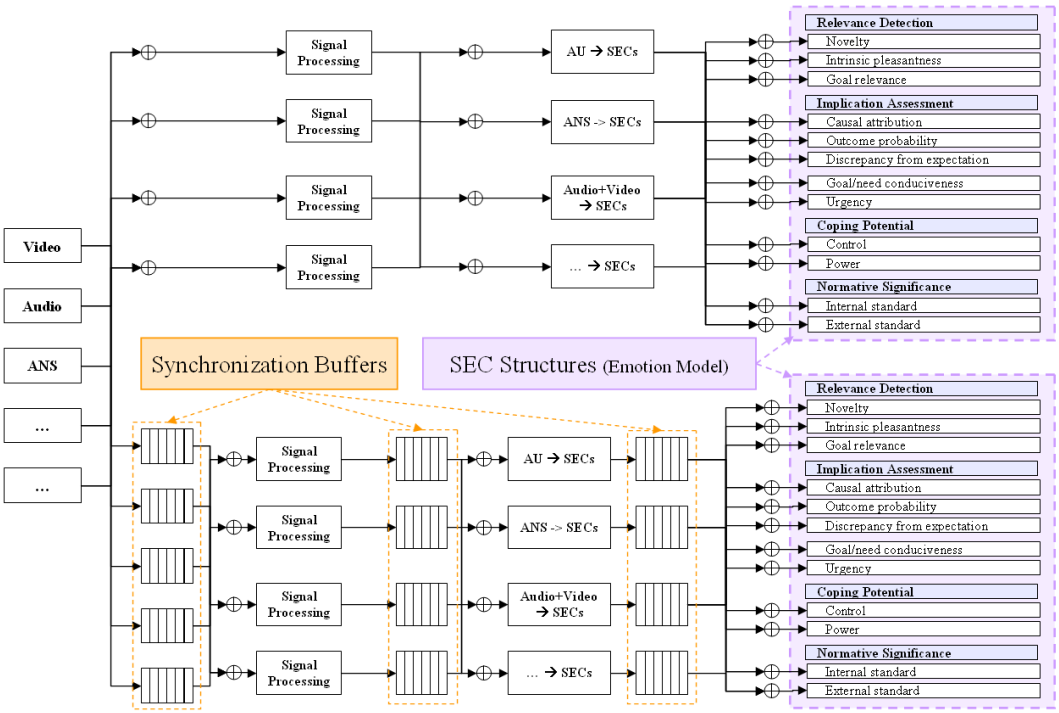


FIG. 27: MAUI : deux chemins pour la fusion de données temps réel ou temporisées. Figure tirée de [108]

Le système MAUI+ASIA est entièrement basé sur la CPM de Scherer. Les SECs sont donc évaluées. Cependant, les auteurs ne considèrent pas les composants proposés par Scherer ; l'émotion reconnue étant directement injectée dans les blocs relatifs à l'agent virtuel, Lisetti choisit d'utiliser l'ensemble des réponses aux SECs comme représentation de l'émotion ; cet ensemble de réponses est directement interprété par l'agent pour déclencher des changements

d'état dans ses propres composants. Ainsi, le système complet, et en particulier le mécanisme de fusion présenté au paragraphe précédent, sont complètement dépendants de la théorie de l'évaluation cognitive et la CPM de Scherer, ainsi qu'à cette représentation de l'émotion.

5.2.2 Eyesweb, une plate-forme basée composants pour la reconnaissance d'émotions

L'application EyesWeb[2] [22] est le résultat du projet du même nom. L'application EyesWeb a été développée principalement pour l'étude des signaux audio et l'analyse de mouvements expressifs. L'application EyesWeb comporte un environnement de développement et des librairies de composants. Ces composants sont regroupés en sept catégories : capture de l'information, filtres et opérations mathématiques, algorithmes d'analyse d'images, son et MIDI, analyse du mouvement, génération audio et visuelle, et enfin canaux de sortie. EyesWeb regroupe ainsi une base importante de composants permettant la construction d'applications. EyesWeb est d'ailleurs utilisé pour tous les travaux du laboratoire Infomus concernant l'analyse du mouvement expressif et la reconnaissance d'émotions via le mouvement[3].

L'application présente une interface permettant l'assemblage des composants. L'assembleur dispose en effet d'un espace sur lequel il glisse et dépose des composants. Il peut ensuite connecter ces composants entre eux en reliant les points de connexion pour créer un patch[4], c'est-à-dire une chaîne de composants fournissant un traitement global. Un patch exécuté dans Eyesweb est l'équivalent d'une application complète. La figure 28, tirée du site d'EyesWeb[5], illustre l'interface d'assemblage ainsi que les résultats de l'extraction d'une silhouette au cours du temps.

EyesWeb contient en particulier une librairie de traitement du geste expressif[6] [21]. Cette librairie est divisée en trois sous-librairies. La librairie d'analyse du mouvement fournit des composants permettant d'extraire des caractéristiques de mouvement bas-niveau et générales comme la QoM et le CI (voir partie 3.4), mais aussi l'orientation et la forme générale de la silhouette. La librairie d'analyse de l'espace permet d'étudier les mouvements d'une personne dans l'espace général (voir l'Analyse du Mouvement de Laban à la section 3.2, page 53). Enfin la librairie d'analyse de trajectoire fournit des algorithmes de suivi 2D de points dans une vidéo et des outils d'analyse des trajectoires de ces points dans l'espace 2D (permettant par exemple de calculer la directivité d'un geste).

Camurri *et al.* présentent dans [19] la version 4.0 d'EyesWeb et les fonctionnalités ajoutées au système. La nouvelle version d'Eyesweb est donc une application basée composants, ce qui permet une bonne réusabilité : réusabilité des composants eux-mêmes au sein de l'application, mais également réusabilité des patches de composants, qui peuvent être réutilisés pour former

[2]Le site d'EyesWeb propose une description ainsi qu'un téléchargement de l'application à l'adresse http ://www.infomus.org/EywMain.html

[3]Liste disponible à l'adresse http ://www.infomus.org/Publications.html

[4]Nous avons ici gardé le terme patch, terme anglais original des publications

[5]http ://www.infomus.org/EywMain.html

[6]expressive gesture processing library

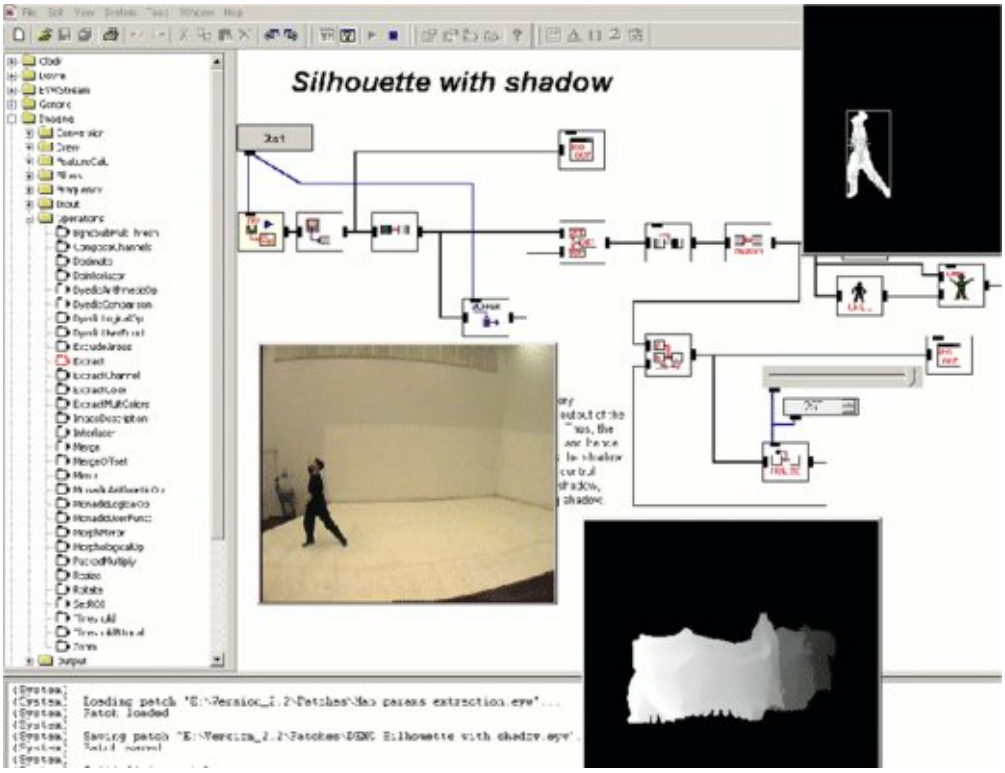


FIG. 28: Capture d'écran de l'application EyesWeb présentant un patch de composants et les résultats d'extraction de silhouette au cours du temps. Figure tirée du site web d'EyesWeb.

un patch plus important. La modifiabilité d'un patch est également assurée, puisqu'il suffit d'échanger des composants pour modifier le patch souhaité. Le système et les patches sont parallélisables sur différentes machines grâce à des composants réseaux. La multimodalité est prise en compte au niveau capture (divers dispositifs sont à disposition), analyse (diverses caractéristiques sont extractibles, et selon divers canaux, notamment le mouvement et la musique), et au niveau interprétation (bien qu'il fasse pour cela créer des composants *ad hoc*). La fusion de données multimodales est un point qui n'est pas abordé ; EyesWeb ne propose *a priori* pas de système générique de fusion des émotions. La synchronisation des flux est par contre présente et explicitée dans [19]. En particulier, EyesWeb gère les variables temporisées par l'intermédiaire des composants : chaque composant temporisant le flux de données possède un buffer interne permettant de stocker l'information jusqu'à ce qu'elle puisse être traitée. Enfin, Eyesweb permet l'interprétation des caractéristiques émotionnelles par l'implémentation de composants *ad hoc*. Les travaux de Castellano sur la reconnaissance multicanaux des émotions [26] montrent qu'il est possible d'implémenter plusieurs interprétations avant de les fusionner. Dans les travaux effectués avec EyesWeb, seul le modèle discret des émotions, et en particulier les émotions basiques d'Ekman, est considéré. Rien ne semble cependant interdire l'utilisation d'autres modèles d'émotions.

EyesWeb est un outil propriétaire et gratuit pour la recherche et l'éducation. Le code du système ou des librairies n'est pas disponible. Ainsi, l'analyse du système que nous venons

de faire s'appuie sur des publications et essais de l'outil. Il existe par contre une série de spécifications permettant à une tierce partie de développer ses propres composants. Enfin, elle ne permet l'intégration d'un système existant que via une communication réseau ou via une retraduction complète de l'application tierce en composants EyesWeb.

En conclusion, l'application EyesWeb offre une base très complète à la fois au concepteur d'un composant et au développeur par toutes les spécifications proposées pour la création de nouveaux composants. Elle offre à l'assembleur un système proposant une interface pour l'assemblage des composants et un moteur permettant l'exécution d'un patch. Du point de vue de la reconnaissance d'émotions, EyesWeb propose une large collection de composants permettant d'extraire le geste expressif, collection extensible à d'autres canaux de communication émotionnelle. EyesWeb offre la synchronisation des flux ; or nous avons vu au chapitre précédent que dans le cadre de la reconnaissance d'émotions, l'on pouvait limiter la fusion des données multimodales à une synchronisation microtemporelle. L'application et ses librairies ne sont pas open source et ne permettent donc pas de connaître précisément les mécanismes de synchronisation du moteur, ni les algorithmes utilisés pour l'extraction des caractéristiques.

5.2.3 ICARE, un outil basé composants pour les applications interactives multimodales

L'outil ICARE (Interaction Complémentarité Assignation Redondance Equivalence) propose un modèle à composants pour la création d'applications interactives multimodales décrit par Bouchet dans [14]. ICARE se base sur la définition d'une modalité (voir partie 4.2.1, page 70) et les propriétés CARE (voir partie 4.2.3 page 75). Notons qu'une autre plate-forme logicielle adoptant les mêmes principes par assemblage et les propriétés CARE a été développée dans le cadre du projet européen OpenInterface[7]. Cette plate-forme OpenInterface est open source est permet la manipulation de composants de tous types et développés dans des langages de programmation différents, contrairement à ICARE qui ne manipule que des composants JavaBeans. Pour nos travaux, nous nous intéressons au modèle conceptuel à composants et nous basons sur les composants d'ICARE, proches de ceux d'OpenInterface. ICARE définit donc trois types de composants : les composants dispositif, système représentationnel (appelé *langage d'interaction* dans [14]) et la superclasse combinaison, dont héritent les composants Complémentarité, Redondance, et Redondance/Equivalence. Nous avons choisi de nous inspirer des spécifications de chacun de ces composants pour renforcer l'intégration de notre modèle d'architecture dans les principes de l'interaction multimodale ; nous présentons pour cette raison plus en détail les différents composants mis en jeu, en particulier les attributs. Bouchet offre une description précise des différents attributs dans [14]. Les spécifications des composants ICARE servent d'appui à nos propres spécifications pour la branche émotion et sont illustrées à la section 5.4.

Le composant dispositif

Le composant dispositif fournit une interface avec le dispositif physique. Il peut s'agir d'une simple encapsulation du pilote ou d'une couche supplémentaire permettant de compléter les

[7] http ://www.openinterface.org

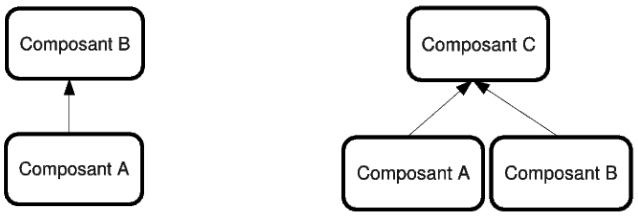
données émises par le dispositif physique par des propriétés supplémentaires utiles d'un point de vue génie logiciel (estampilles de temps, version, précision, domaine des valeurs de sortie... etc.).

Le composant système représentationnel

Le composant système représentationnel (appelé "composant langage d'interaction" lors de la publication de [14]) a pour rôle de transformer les données venant des composants dispositifs en une forme idéale et compréhensible par le système.

Les composants de composition

ICARE se fonde sur les concepts de la multimodalité et en particulier sur les propriétés CARE décrites dans le chapitre précédent (partie 4.2.3, page 75). L'assignation et l'équivalence ne nécessitent pas de composants particuliers. En effet, l'assignation est représentée par la connexion d'un composant dispositif à un composant système représentationnel ou entre deux composants système représentationnel. L'équivalence exprime le choix et correspond donc à deux composants (dispositif ou système représentationnel) connectés à un même composant système représentationnel. La figure 29 illustre ces cas de figure.



(a) Assignation dans ICARE.

(b) Equivalence dans ICARE

FIG. 29: ICARE : Assignation et Equivalence se font par la connexion des composants. Figures tirées de [14].

ICARE propose trois composants pour la composition de modalités. Le composant complémentarité gère les modalités complémentaires, c'est-à-dire le cas où chaque modalité est nécessaire à un système représentationnel. Ce composant reprend les principes du moteur de fusion proposé par Nigay et Coutaz dans [105] et présenté au chapitre précédent (partie 4.2.4, page 76), mais ne reprend que les principes de fusions micro- et macrotemporelles ; la fusion contextuelle des données est laissée à des travaux futurs. Le composant redondance fonctionne sur le même principe mais au lieu d'envoyer les deux données, il écarte l'information au facteur de confiance moindre pour ne renvoyer qu'un seul bloc informationnel. Enfin, le composant équivalence/redondance fonctionne de la même façon que le composant redondance à la différence près qu'il est capable de s'adapter à l'absence de l'une des modalités d'entrée. Si les deux modalités équivalentes sont présentes, alors le composant agit comme un composant redondance ; si une modalité est absente (par exemple, l'utilisateur a éteint le dispositif), alors

le composant fonctionne en équivalence. Le composant redondance/équivalence propose deux types de stratégie de fusion. La stratégie précoce permet une fusion efficace (car rapide, les premières données reçues étant envoyées), la stratégie différée une fusion plus sûre (le système attends toutes les données et ne fusionne que celles dotées du meilleur facteur de confiance). Ce composant est donc plus souple que le composant redondance. Ce dernier reste néanmoins nécessaire pour les cas de redondance pure.

Les flux de données

L'outil ICARE établit une communication entre ses différents composants par flux de données. Ces données sont envoyées grâce au système d'évènements sous forme de blocs de données. Chaque bloc de données contient les informations à transmettre ainsi qu'une série d'attributs.

5.2.4 Conclusion

Dans cette section, nous avons présenté trois systèmes sur lesquels nous nous appuyons pour construire la branche émotion. EyesWeb propose un produit fini permettant le développement et l'assemblage de composants, ainsi que l'exécution des patches de composants. Nous retenons de ce logiciel sa capacité d'ouverture par la création de nouveaux composants, et son environnement d'assemblage et d'exécution. Cependant, son format propriétaire empêche une analyse approfondie du logiciel. Le système MAUI+ASIA propose un système englobant, allant de la reconnaissance d'émotions à la synthèse de comportements pour un agent social et expressif. Nous retenons en particulier de ces travaux le moteur de fusion multimodale dans le cadre de la reconnaissance d'émotions. Enfin, nous avons décrit plus en détail l'outil ICARE, un modèle d'architecture basé composants pour l'interaction multimodale. Dans le cadre de ce mémoire, nous nous inspirons fortement d'ICARE pour mettre en place la branche émotion, un modèle d'architecture basée composants pour la reconnaissance d'émotions. En effet, EyesWeb est avant tout un outil de prototypage rapide de systèmes (entre autres) de reconnaissane d'émotions. Il ne permet pas de créer des systèmes indépendants (par exemple en exportant les patches créés). EyesWeb doit nécessairement être lancé pour exécuter un patch. ICARE propose quant à lui une modélisation de ses différents composants dans le cadre de l'interaction multimodale. Nous avons donc préféré reprendre cette modélisation et l'étendre à la reconnaissance d'émotions, intégrant ainsi la reconnaissance d'émotions comme une interaction multimodale.

5.3 La branche émotion : conception globale

Notre approche est de nous appuyer sur les concepts de la multimodalité en interaction homme-machine pour créer un modèle d'architecture adapté à la reconnaissance d'émotions. Dans la suite de ce chapitre, nous décrivons donc la branche émotion, un modèle d'architecture à base de composants pour la reconnaissance multimodale des émotions. Notre approche est basée sur l'observation des systèmes existants et l'identification dans ces systèmes des méthodes récurrentes de conception et des contraintes auxquels de tels systèmes doivent faire face. Dans cette section, nous nous penchons sur la conception globale de la branche émotion,

c'est à dire l'identification des différents composants nécessaires, et la manière de les agencer entre eux.

5.3.1 Fondations

Nous énonçons les principes sous-jacents à notre modèle d'architecture. Ces principes sont issus du domaine de la reconnaissance d'émotions et du domaine de l'IHM.

Fondations issues du domaine de la reconnaissance d'émotions

Nous avons vu au chapitre 2 que la reconnaissance d'émotions suivait un processus en trois étapes que nous avons appelé niveaux Capture, Analyse, et Interprétation. Le niveau Capture est le niveau le plus bas du système. Comme son nom l'indique, c'est à ce niveau que sont capturées les données provenant du monde réel, et en particulier celles issues de l'utilisateur. Il regroupe donc les différents capteurs utilisés. Le niveau Analyse englobe l'extraction des différentes caractéristiques porteuses d'information émotionnelle. Enfin le niveau interprétation englobe les différentes interprétations des caractéristiques pour en inférer des émotions. En tant que modèle d'architecture pour la reconnaissance d'émotions, notre modèle d'architecture doit permettre l'intégration :

- des dispositifs de capture. Une liste exhaustive des dispositifs existants est impossible. De plus, notre modèle d'architecture doit pouvoir permettre la création d'architectures permettant l'intégration de dispositifs non encore conçus.
- des caractéristiques selon n'importe quel canal de communication affective, y compris des caractéristiques non encore identifiées ou validées.
- des méthodes d'interprétation, en considérant notamment la possibilité de l'implémentation des différentes théories et modèles de représentation de l'émotion présentés dans le chapitre 1.

Le modèle d'architecture doit également permettre la fusion aux niveaux Capture, Analyse, et Interprétation. Des concepts de la reconnaissance d'émotions nous extrayons donc trois types de composants, relatifs aux trois niveaux Capture, Analyse et Interprétation : le composant **unité de capture** (UC), le composant **extracteur de caractéristiques** (EC), et le composant **interpréteur** (I).

Concepts relatifs à l'interaction multimodale

D'après la définition de la multimodalité que nous proposons en(4) (page 84), une modalité est définie par un dispositif et une séquence de systèmes représentationnels aux niveaux Capture, Analyse et Interprétation. Nous disposons de l'outil ICARE, un modèle basé composants se basant sur les concepts de la multimodalité et en particulier les propriétés CARE. Par analogie avec ICARE, nous considérons donc l'unité de capture comme une spécialisation du composant dispositif d'ICARE, et l'extracteur de caractéristiques et l'interpréteur comme des spécialisations du composant système représentationnel d'ICARE.

Nous avons choisi de ne pas reprendre les composants complémentarité, redondance et redondance/équivalence d'ICARE dans le cadre de la reconnaissance d'émotions. En effet, dans ICARE ces composants partagent un même algorithme de synchronisation et ne diffèrent que dans les politiques de fusion des données, présentées au paragraphe 5.2.3. La synchronisation d'ICARE fait intervenir une fenêtre temporelle pour la fusion macrotemporelle. Dans le cadre de la reconnaissance d'émotions, nous limitons la fusion à une synchronisation micro-temporelle. Nous avons choisi de centraliser cette synchronisation en ne gardant des composants que les algorithmes de fusion. De cette façon, la complémentarité est automatiquement gérée par le système de synchronisation, les données temporellement superposées étant fusionnées de façon centralisée. La redondance, l'équivalence, et la redondance équivalence nécessitent par contre des algorithmes spécifiques à la fusion qui dépendent de plus de la politique de fusion choisie.

Nous proposons en section 5.5 un moteur générique gérant la synchronisation des données. Grâce à ce moteur, la complémentarité est directement prise en compte et ne nécessite pas de composant supplémentaire. La redondance et l'équivalence sont également synchronisées. Nous proposons un composant permettant de gérer le choix des données et l'alternance des modalités sources : le **concentrateur**. L'assignation est gérée de la même façon que dans ICARE, en connectant des composants entre eux.

Enfin, nous notons que les trois composants unité de capture, extracteur de caractéristiques et interprétation sont des composants dédiés à des fonctions particulières. Ils ne permettent pas de faire face à d'autre problématiques de changement de système représentationnel. Par exemple, des prétraitements sur une image (passage en niveaux de gris, application de filtres) peuvent être nécessaires à l'extraction d'une caractéristique mais ne remplissent pas eux-mêmes une fonction d'extraction de caractéristique. Un flux de données peut véhiculer les bonnes données pour le fonctionnement d'un composant, mais dans un mauvais format ; il est alors nécessaire de reformater le flux en changeant le format des données véhiculées. Nous proposons donc un cinquième composant, l'**adaptateur**, dont le rôle est d'effectuer ces divers traitements non directement liés à la reconnaissance d'émotions.

5.3.2 Définition des composants de la branche émotion

Nous avons donc introduit cinq types de composants pour la reconnaissance d'émotions (figure 30). L'unité de capture, l'extracteur de caractéristiques et l'interpréteur sont des composants remplissant des fonctions directement liées à la reconnaissance d'émotions. L'adaptateur et le concentrateur sont des composants génériques. Ils ne reprennent pas de concept particulier à la reconnaissance d'émotions. Leur présence est nécessaire pour la multimodalité. En effet, le composant unité de capture est une spécialisation du composant dispositif proposé par ICARE. Les composants extracteur de caractéristiques et interprétation permettent, d'un point de vue strictement système, de modifier le système représentationnel utilisé. Ils se rapportent donc aux composants système représentationnel d'ICARE. Néanmoins ces composants ne couvrent pas tous les cas de changement de système représentationnel : le composant adaptateur permet de pallier à cette lacune. Enfin le composant concentrateur permet de gérer la redondance et l'équivalence dans notre système.

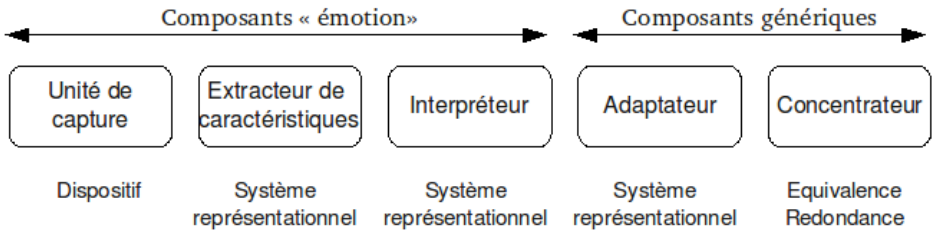


FIG. 30: Les composants de la branche émotion.

L'unité de capture

L'unité de capture a pour rôle de faire l'interface avec le dispositif. Ce composant peut encapsuler le pilote du dispositif considéré ou fournir une couche supplémentaire afin d'ajouter certaines informations (estampilles de temps par exemple) aux données du dispositif pour le bon fonctionnement du système. L'unité de capture peut présenter des ports d'entrée lui permettant de recevoir des communications du dispositif physique. L'unité de capture délivre un flux de données. Ce flux étant émis depuis le niveau capture, nous le nommons **flux de capture**.

Ainsi, une unité de capture permet d'intégrer n'importe quel dispositif au système. Au chapitre 6 par exemple, nous utilisons pour capter la gestuelle deux unités de capture gérant pour la première une paire de capteurs à six degrés de liberté et pour la seconde une combinaison de suivi de mouvement. Dans cette application, la première unité de capture est un composant encapsulant le pilote ; la deuxième est basée sur une socket réseau recevant les données de la combinaison ; son seul rôle est donc de retraduire les coordonnées du corps pour les envoyer aux composants suivants.

L'extracteur de caractéristiques

L'extracteur de caractéristiques est le type de composant principal du niveau Analyse. Son rôle est de traiter un ou plusieurs flux complémentaires en entrée pour en extraire une ou plusieurs caractéristiques porteuses d'information émotionnelle. Idéalement, afin de séparer au maximum le code et assurer ainsi une meilleure réusabilité, il est préférable qu'un extracteur ne s'abonne qu'à un minimum de flux et n'extraie qu'une caractéristique. Cependant, pour des raisons d'optimisation, il est parfois préférable de regrouper certains calculs et l'extraction de plusieurs caractéristiques.

Typiquement, un extracteur de caractéristiques s'abonne à un flux de capture, en extrait une caractéristique et produit un flux des valeurs de cette caractéristique au cours du temps, appelé **flux de caractéristiques**. Par exemple, dans notre système de reconnaissance d'émotions basé sur la gestuelle présenté au chapitre 6, une combinaison de capture du mouvement nous donne un flux continu (à 100 Hz) des coordonnées de chaque segment du corps. Un extracteur de caractéristiques calcule l'expansion des bras dans les valeurs Indéterminé, Fermé,

Ouvert. Cet extracteur s'abonne au flux de capture des coordonnées du corps, et seuille la distance entre les poignets pour déterminer la valeur de sortie. Dans ce cas, à chaque mesure correspond une valeur de caractéristique. Il existe également des variables temporisées, nécessitant l'extraction sur un ensemble de données dans le temps. C'est le cas par exemple de la directivité d'un geste calculée par Volpe dans [143] (voir section 3.4, page 59).

Un extracteur peut s'abonner à un ou plusieurs flux de capture. Il peut également s'abonner au flux produit par un autre extracteur. Par exemple, dans [143], Volpe calcule la quantité de mouvement de l'utilisateur. Cette valeur peut-être utilisée telle quelle car la quantité de mouvement est déjà représentative de l'émotion. Elle est également utilisée pour segmenter le mouvement en phases de pause et de mouvement, par seuillage. En transposant cet exemple dans une architecture à composants, on obtient un extracteur de caractéristiques calculant la quantité de mouvement, et un extracteur de caractéristiques déterminant l'état de la phase courante {pause, mouvement}. Ce dernier prend en entrée le flux de quantité de mouvement.

L'interpréteur

La brique de base du niveau interprétation est l'interpréteur. Il s'abonne à un ou plusieurs flux de caractéristiques émis par le composant Analyse. Son rôle est de déduire une émotion des valeurs qu'il reçoit au temps t pour ces caractéristiques.

Un interpréteur se définit ainsi :

$$f_{\{C \to E\}}(\{p\}) \tag{5}$$

Où :

1. C, ensemble de départ, est l'ensemble de caractéristiques extraites au niveau Analyse, sur lequel va se baser l'interprétation ;
2. E, ensemble d'arrivée, est l'ensemble d'émotions considéré par le système et le modèle de représentation utilisé pour cet ensemble ;
3. f, fonction d'interprétation, va fournir une émotion appartenant à E à partir des valeurs des caractéristiques de C ;
4. Enfin l'ensemble $\{p\}$ est l'ensemble des paramètres de la fonction f.

L'interpréteur délivre un flux de données décrivant les émotions interprétées que nous appelons **flux d'émotions**. Les composants "interpréteur" sont dépendants du modèle d'émotion (modèle discret, continu ou à composants - voir section 1.3). Selon le modèle adopté, l'émotion n'est pas représentée de la même manière aussi nous ne figeons pas dans notre modèle d'architecture le format des sorties des composants interpréteurs.

L'adaptateur

En plus des composants liés aux fonctions successives d'un système de reconnaissance d'émotions, chaque niveau peut également posséder des composants de type "adaptateur". Un tel

composant a pour rôle de retranscrire les données dans les flux selon un autre format. Il est générique dans le sens où il est indépendant d'un modèle d'émotion particulier. Les adaptateurs peuvent regrouper tout type de traitement n'ayant pas un rapport direct avec la tâche d'extraire l'émotion. Un adaptateur effectue un transfert de système représentationnel qui n'est pas en relation avec la tâche de reconnaissance d'émotions. Ainsi, un adaptateur peut être utilisé pour :

- transformer la manière dont sont structurées les données.
- effectuer un traitement préalable sur les données : par exemple, passer une image en niveaux de gris, ou extraire les informations spatiales du corps de l'utilisateur sous forme de coordonnées en prenant de la vidéo en entrée.
- recalculer l'information : par exemple, replacer les coordonnées acquises par deux capteurs dans deux repères différents dans un repère unique.

Le concentrateur

Une solution pour améliorer la robustesse d'un système est de le doter de plusieurs dispositifs de mesure et de méthodes pour un même calcul, puis de comparer et combiner les données ou les résultats obtenus. Ainsi dans notre architecture, des unités de capture ou des extracteurs de caractéristiques peuvent produire des flux similaires, dont la finalité est d'être fusionnés afin de ne fournir qu'un seul flux plus fiable.

Le terme "fusion" utilisé ici désigne la fusion des données au sens traitement du signal : la redondance de deux flux permet de pallier aux erreurs de l'un d'entre eux ; deux flux peuvent être fusionnés en favorisant le flux le plus fiable et en n'utilisant l'autre que comme complément. Cependant, afin d'éviter d'utiliser le terme "fusion", dont la signification est différente selon les domaines de traitement du signal et de l'interaction multimodale, nous parlons ici de **concentration**. Dans le cadre de la multimodalité, un concentrateur a pour rôle d'effectuer le dernier traitement des composants redondance et redondance/équivalence d'ICARE. En effet, notre système permet la synchronisation constante des flux de données, de façon générique. Le seul traitement ad hoc à fournir dans le cas d'équivalence, redondance ou redondance/équivalence est la fusion des données au sens signal (amalgamer les deux flux en un seul) et l'alternance entre plusieurs modalités sources.

Par exemple, en considérant notre application de reconnaissance d'émotions par la gestuelle (voir chapitre suivant) permet l'utilisation redondante d'une combinaison de capture de mouvement et d'une paire de capteurs à six degrés de libertés. Les deux capteurs sont placés sur les poignets. Un adaptateur permet de recaler le système de coordonnées des capteurs dans celui de la combinaison. Un concentrateur permet de recalculer la position précise des poignets en utilisant les deux capteurs.

La concentration de deux flux en un seul a donc pour rôle d'améliorer la fiabilité des données obtenues. Cette concentration est décidée au moment de la conception du système. De la même manière que le concepteur d'interaction multimodale décide d'utiliser deux modalités de façon complémentaire (par un assemblage de composants au sein d'ICARE), le concepteur décide

ici d'utiliser deux sources d'information pour rendre la reconnaissance plus fiable en utilisant un composant concentrateur.

5.3.3 Caractérisation des flux de données

Nous considérons les données sous forme de flux émis par les composants. Ces flux sont constitués de blocs de données. Nous identifions deux types de composants selon les flux qu'ils émettent. Les composants à **flux tendus** permettent un traitement continu des données. Le passage d'un filtre sur une image d'une vidéo est un exemple de composant à flux tendus : chaque image de la vidéo est filtrée puis renvoyée. Ces composants peuvent induire un retard dans le traitement. Ainsi, une accélération se calculant sur trois positions, le calcul d'une accélération sur un flux de coordonnées induit automatiquement un retard de trois pas de temps par rapport au flux de coordonnées. Cependant, un tel retard n'est pas bloquant pour le système. Une nouvelle accélération est calculée à chaque arrivée d'une nouvelle coordonnée. Les composants à **flux temporisés** présentent par contre le problème d'être bloquant. Par exemple, Volpe [143] segmente le mouvement en phases de pause et de mouvement. Une phase de mouvement est analysée dans son entier pour en extraire la caractéristique de fluidité du mouvement. Dans cet exemple, la segmentation produit un flux temporisé : l'information arrivant en continu est bloquée et gardée en mémoire jusqu'à un changement d'état. L'information d'une phase de mouvement est alors envoyée dans son ensemble au calcul de fluidité.

5.3.4 Propriétés CARE de la multimodalité appliquées aux composants

Dans cette partie, nous explicitons les propriétés CARE vues au chapitre précédent (partie 4.2.3, page 75).

D-propriétés CARE

Nous explicitons les propriétés CARE de la multimodalité au niveau Capture, c'est-à-dire entre les unités de capture et les extracteurs de caractéristiques.

Soit S un système de reconnaissance d'émotions et soit $\{UC_1, ..., UC_n\}$ un ensemble d'unités de capture et $\{EC_1, ..., EC_m\}$ l'ensemble des extracteurs de caractéristiques de ce système. On définit alors les propriétés suivantes :

- Assignation : L'assignation d'une unité de capture UC_i à un extracteur de caractéristiques EC_j signifie l'obligation d'utiliser ce dispositif pour alimenter l'extracteur.
- Equivalence : $\{UC_1, ..., UC_p\}$ unités de capture sont équivalentes pour un extracteur de caractéristiques EC_j si EC_j ne nécessite qu'un seul élément de $\{UC_1, ..., UC_p\}$.
- Redondance : $\{UC_1, ..., UC_p\}$ unités de capture équivalentes pour un extracteur de caractéristiques EC_j sont redondantes pour cet extracteur si EC_j nécessite d'être alimenté par tous les éléments de $\{UC_1, ..., UC_p\}$.
- Complémentarité : $\{UC_1, ..., UC_p\}$ unités de capture sont complémentaires pour un extracteur de caractéristiques EC_j si toutes ces unités de capture sont nécessaires à EC_j.

L-propriétés CARE

De même, on redéfinit les propriétés L-CARE de la multimodalité. Ces propriétés s'appliquent à chaque changement de système représentationnel. Elles s'appliquent donc sans distinction aux extracteurs de caractéristiques, aux interpréteurs, et aux adaptateurs. Soit S un système de reconnaissance d'émotions et soient $\{C_{SR1}, ..., C_{SRn}\}$ l'ensemble des composants permettant de modifier le système représentationnel (extracteurs de caractéristiques, interpréteurs et adaptateurs). On définit alors les propriétés suivantes :

- Assignation : L'assignation d'un composant C_{SRi} à un composant C_{SRj} signifie l'obligation d'utiliser C_{SRi} pour alimenter C_{SRj}.
- Equivalence : les composants $\{C_{SRi}, ..., C_{SRp}\}$ sont équivalents pour un composant C_{SRj} si C_{SRj} ne nécessite qu'un seul élément de $\{C_{SRi}, ..., C_{SRp}\}$.
- Redondance : les composants $\{C_{SRi}, ..., C_{SRp}\}$ sont redondants pour un composant C_{SRj} si C_{SRj} nécessite d'être alimenté par tous les éléments de $\{C_{SRi}, ..., C_{SRp}\}$.
- Complémentarité : les composants $\{C_{SRi}, ..., C_{SRp}\}$ sont complémentaires pour un composant C_{SRj} si $\{C_{SRi}, ..., C_{SRp}\}$ sont non équivalents et tous nécessaires à C_{SRj}.

Résumé

La figure 31 résume les propriétés CARE dans le cadre de la reconnaissance d'émotions. Une modalité de reconnaissance émotionnelle délivrant une émotion est utilisée pour informer le système interactif de l'état émotionnel de l'utilisateur. Comme toute autre modalité, de telles modalités peuvent être assignées, complémentaires, équivalentes ou redondantes pour une tâche t_i.

Dans les deux cas (D-propriétés CARE et L-propriétés CARE) le composant concentrateur permet de gérer les cas de redondance et équivalence en proposant un cadre pour l'implémentation de politique de choix du système représentationnel (équivalence) ou de concentration de flux redondants pour améliorer la robustesse (redondance).

5.3.5 Exemple

Nous présentons un exemple des différents types de composants décrits ci-dessus sur un système de reconnaissance d'émotions basé sur la gestuelle (figure 32, page 108).

Deux unités de capture sont utilisées. La première permet de faire l'interface avec une paire de capteurs à six degrés de liberté, attachés aux poignets du sujet. Les données envoyées sont donc un couple de coordonnées et de rotations par rapport à un repère fixe dans l'espace. La deuxième unité de capture fait l'interface avec une combinaison de capture du mouvement. Cette unité de capture renvoie les coordonnées et informations de rotations de 23 segments du corps dans l'espace par rapport à un repère fixe. Les deux repères fixes sont différents. Un adaptateur est donc utilisé pour traduire les coordonnées données par la paire de capteurs dans le repère de la combinaison. Un concentrateur est ensuite utilisé. Au niveau du concentrateur, on a donc une redondance partielle des données : les coordonnées des poignets sont données à la fois par les capteurs et par la combinaison. Le concentrateur choisit la donnée ayant

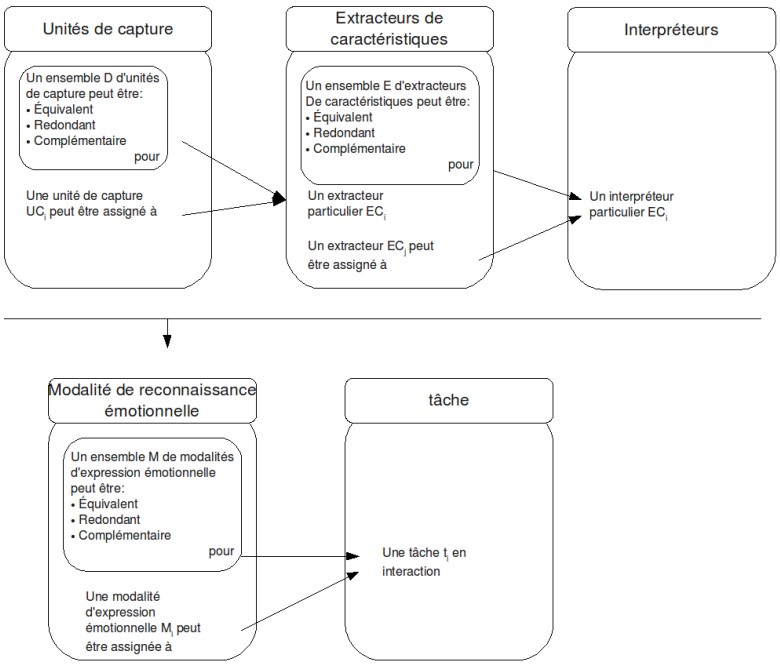


Fig. 31: Récapitulatif des propriétés CARE dans la reconnaissance d'émotions.

le plus haut facteur de confiance et si nécessaire, remplace dans la description du corps les coordonnées des poignets donnés par la combinaison par celles données par les capteurs.

Le flux de données passe alors au niveau Analyse ; son format est la liste des coordonnées des 23 segments du corps. Ce flux est envoyé à chaque extracteur de caractéristiques. Le premier calcule la position du tronc (droit ou voûté). Le deuxième calcule l'écartement des bras (ouverts ou fermés). Enfin le dernier calcule la vitesse du bassin (lent ou rapide).

Enfin, chaque caractéristique est envoyée à l'interpréteur. Un interpréteur simple pourrait se baser sur les trois caractéristiques extraites pour en inférer, par exemple, la joie (tronc droit, mouvement rapide, bras écartés) et la tristesse (tronc voûté, mouvement lent) grâce à un système de règles, se basant ainsi sur un modèle discret d'émotions.

5.4 Spécifications des composants de la branche émotion

Dans cette section nous décrivons les spécifications pour chaque type de composant proposé (unité de capture, extracteur de caractéristiques, interpréteur, adaptateur et concentrateur). Nous spécifions également le format d'un bloc de données échangé entre deux composants.

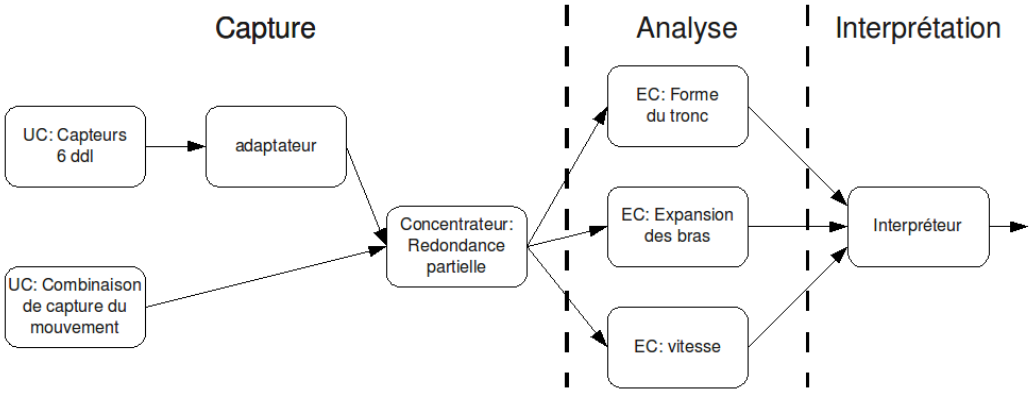


FIG. 32: Système illustratif de reconnaissance d'émotions par la gestuelle.

5.4.1 L'unité de capture

Le type de composant unité de capture est dédié à capturer de l'information du monde réel. La reconnaissance d'émotions étant une forme d'interaction, nous nous basons sur les spécifications du composant dispositif d'ICARE [14] (partie 5.2.3, page 97) et en faisons hériter (au sens de la programmation objet) leurs attributs à notre type de composant unité de capture (voir figure 33).

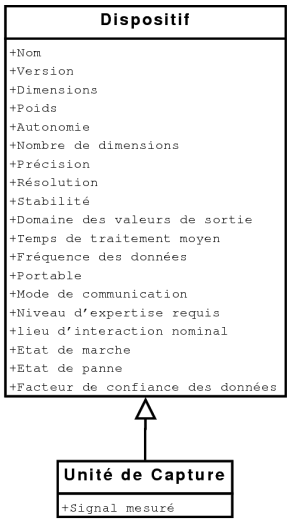


FIG. 33: L'unité de capture hérite du composant dispositif.

Les attributs *Niveau d'expertise requis* et *Lieu d'interaction nominal* concernent le cas de l'interaction active et ont donc été supprimés dans notre cadre de reconnaissance passive. Parmi les attributs hérités figurent notamment le *Domaine des valeurs de sortie*. Cette propriété définit la nature des données transmises par le dispositif. Elle spécifie donc entièrement

le ou les flux de sortie. Les flux de sortie des dispositifs n'étant généralement pas standardisés, il est important de pouvoir communiquer le format utilisé pour transmettre les données. En effet, pouvoir annoncer les formats d'entrées-sorties permet la mise en place d'une gestion des connexions de flux de données par un agent extérieur.

A ces attributs nous en rajoutons un autre : le *signal mesuré*. Il s'agit d'un texte décrivant le type de signal mesuré : battements du cœur, coordonnées du corps, signal vidéo, etc. Cet attribut vient en complément du domaine des valeurs de sortie afin de permettre l'identification de modalité équivalentes ou redondantes.

5.4.2 L'extracteur de caractéristiques

De même que le type de composant unité de capture hérite des propriétés des dispositifs explicitées ci-dessus, l'extracteur de caractéristiques hérite des attributs définis pour les langages d'interaction. De même que pour l'unité de capture, certains attributs sont à préciser dans notre cadre de recherche (figure 34, page 110).

Nous évinçons les attributs *Caractère arbitraire* et *Caractère linguistique* du composant langage d'interaction pour construire notre composant "extracteur de caractéristiques". Tout d'abord, un extracteur de caractéristiques a pour rôle d'extraire une caractéristique pertinente pour la reconnaissance d'émotions. En tant que tel, le caractère de la caractéristique ne peut être arbitraire. Ensuite, en l'absence dans la littérature que nous connaissons de structuration sous forme de langage des caractéristiques d'expression émotionnelle, nous considérons que les émotions n'ont pas de caractère linguistique.

De même que pour les unités de capture, les attributs *Domaine des valeurs d'entrées* et *Domaine des valeurs de sortie* permettent au composant d'annoncer les données et leur format qu'il attend en entrée et qu'il délivre en sortie. Ceci permet la mise en place d'une gestion des connexions par un agent extérieur. L'attribut de *Dimension temporelle* permet de classifier les caractéristiques en caractéristiques statiques ou dynamiques. Ce caractère a un impact fort sur la synchronisation des données.

Nous ajoutons à l'extracteur de caractéristiques un attribut permettant de caractériser la dynamique de la caractéristique : la temporisation. Cet attribut est un booléen. Un extracteur temporisé calcule une caractéristique temporisée et donc bloquante (voir partie 3.4.2, page 57).

5.4.3 L'interpréteur

L'interpréteur est, comme l'extracteur de caractéristiques, un composant permettant un transfert de système représentationnel. Il hérite donc également du composant système représentationnel du système ICARE [14] (voir figure 34). Comme nous l'avons vu à la section 5.3.2, le type de composant interpréteur se base principalement sur quatre paramètres :

- le modèle d'émotion considéré et les émotions reconnues ;
- les caractéristiques interprétées pour en déduire une émotion ;

– l'algorithme utilisé permettant l'interprétation ;
– les paramètres de cet algorithme.

Nous créons donc des attributs correspondants à chacun de ces paramètres afin de permettre à un interpréteur d'annoncer sa structure interne.

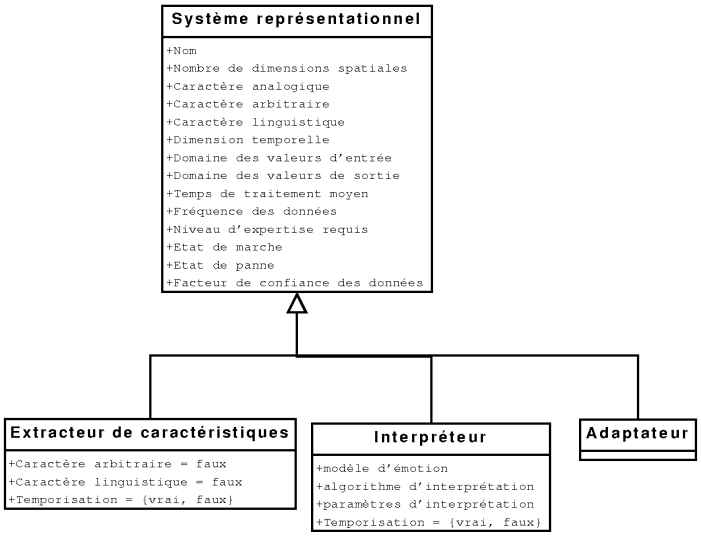


FIG. 34: Les composants extracteur de caractéristiques et interpréteur héritent du composant système représentationnel.

De même que pour les autres composants, nous ajoutons certains attributs au composant interpréteur :

– *Modèle d'émotion* : le modèle d'émotion choisi pour l'interprétation. Le choix du modèle conditionne les traitements effectués et les valeurs de sortie de l'interpréteur.
– *Algorithme d'interprétation* : un texte décrivant l'algorithme utilisé.
– *Paramètres d'interprétation* : les paramètres de l'algorithme d'interprétation.
– *Temporisation* : Cette propriété indique si l'interpréteur est temporisé ou non. Un interpréteur est considéré comme temporisé s'il fournit une interprétation en se basant sur au moins une caractéristique temporisée. Dans le cas contraire, l'interpréteur est considéré comme non temporisé.

5.4.4 L'adaptateur

L'adaptateur est équivalent à un composant système représentationnel et en tant que tel reprend les attributs décrits à la figure 34.

5.4.5 Le concentrateur

Le concentrateur a pour rôle de finaliser la redondance en amalgamant des flux ou l'équivalence en alternant entre plusieurs modalités sources équivalentes. Son rôle peut cependant

être élargi à la gestion de flux partiellement redondants ou équivalents. Le composant concentrateur hérite du composant combinaison d'ICARE (voir figure 35).

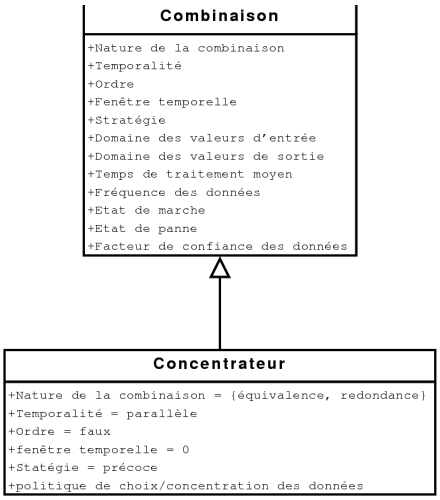


Fig. 35: Le composant concentrateur hérite du composant combinaison.

Au sein d'ICARE, quatre attributs spécifiques aux composants de combinaison de modalités sont identifiés : la temporalité, la fenêtre temporelle, l'ordre et la stratégie. La temporalité décrit l'usage temporel que doit effectuer l'utilisateur sur les différentes modalités mises en jeu dans la combinaison. Elle peut prendre quatre valeurs : aucune, parallèle, séquentiel, parallèle et séquentielle. Les émotions déclenchant des expressions hautement synchronisées, la temporalité pour la reconnaissance d'émotions est forcément parallèle. La fenêtre temporelle permet de spécifier l'intervalle de temps pour une fusion macrotemporelle et contextuelle. Nous ne considérons pas cet attribut, étant donné que nous ne considérons que la fusion microtemporelle dans notre cadre de la reconnaissance d'émotions. L'attribut "ordre" spécifie l'ordre dans lequel les modalités doivent être utilisées pour effectuer la fusion. Ici également, la synchronisation de l'expression émotionnelle implique une absence d'ordre. Enfin, la stratégie spécifie la stratégie de fusion (précoce ou différée). La stratégie différée est utile lors de fusion contextuelle et macrotemporelle, l'attribut sera donc constamment à la valeur "précoce" en reconnaissance d'émotions.

Nous recensons donc, en plus des attributs donnés pour le composant combinaison, un attribut supplémentaire permettant de caractériser la politique de choix entre systèmes représentationnels (cas d'une équivalence) ou de concentration de données redondantes (cas de la redondance). Dans les cas triviaux, la stratégie du concentrateur peut être relativement simple : par exemple, le choix selon des facteurs de confiance ou le calcul d'une moyenne entre deux mesures. Les stratégies de concentration peuvent être cependant bien plus complexes. Notamment, au niveau interprétation, la concentration de plusieurs flux d'émotions fournis par divers interpréteurs nécessite souvent la mise au point et la validation d'une stratégie *ad hoc*.

5.4.6 Le bloc de données

Dans le cadre de la branche émotion, nous considérons qu'un flux est composé de blocs de données. En effet, au niveau de la capture, l'information est de toute façon temporellement discrétisée. Un flux est donc le canal qui transmet les blocs de données d'un composant à un autre. Un bloc de données décrit un état que l'on considère constant entre le temps t_0 de mesure et le temps t_1 de la mesure suivante.

Nous reprenons la spécification d'un bloc de données ICARE [14] (figure 36). Nous y ajoutons cependant plusieurs attributs principalement exploités par notre moteur de synchronisation décrit dans la prochaine section.

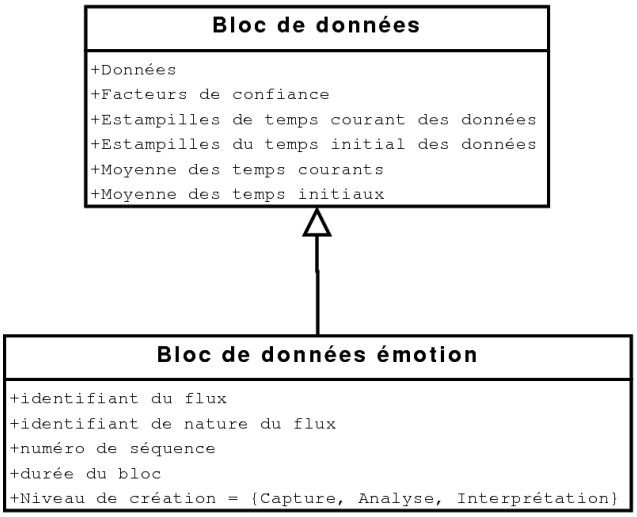


FIG. 36: Le bloc de données dans notre modèle pour la reconnaissance d'émotions hérite du bloc de données ICARE.

Des attributs de la classe mère, nous retenons particulièrement les estampilles de temps. L'estampille de temps initial des données donne le moment où la donnée est arrivée dans le système au niveau de la capture. Il est préférable que cette date soit la plus proche possible du moment de l'expression émotionnelle par l'utilisateur. Elle est initialisée au niveau capture et maintenue dans les niveaux Analyse et Interprétation afin de tenir compte du temps de traitement et recaler l'émotion interprétée au plus proche du moment de son expression. L'estampille de temps courant des données correspond au dernier traitement de l'information initial par un composant. Elle est affectée par le dernier composant ayant traité l'information. Les données sont envoyées à l'intérieur du bloc, accompagnées de leurs facteurs de confiance.

Nous définissons donc les attributs d'un bloc de données :
- *Identifiant* : un texte ou un numéro permettant d'identifier à quel flux le bloc appartient.
- *Identifiant de nature du flux* : un texte ou un numéro permettant d'identifier la nature du flux ; deux flux de même nature sont considérés comme pouvant être concentrés en

un seul grâce à un concentrateur.
- *Numéro de séquence* : ce numéro permet de séquencer les blocs dans un flux.
- *Durée du bloc* : une estampille décrivant le temps écoulé entre deux mesures : l'état décrit par le bloc est considéré valide pendant ce temps.
- *Niveau de création du bloc* : à titre informatif, un texte ou un numéro permettant de savoir si le flux est issu d'un composant du niveau Capture, Analyse, ou Interprétation.

5.5 Moteur de synchronisation

Nous avons présenté dans les sections précédentes de ce chapitre le détail de notre modèle conceptuel pour la conception de systèmes de reconnaissance d'émotions. Ce modèle conceptuel est basé sur des composants logiciels dont les spécifications ont été fournies à la section 5.4. Ces divers composants ne gèrent pas par eux même la synchronisation des différents flux de données. Nous décrivons ici un système de synchronisation des données placé en arrière-plan des composants.

Dans le cadre de la reconnaissance d'émotions, nous avons assimilé la fusion multimodale des données à une synchronisation, avec des traitements dans les cas de la redondance et de l'équivalence. Du point de vue de la multimodalité, utiliser plusieurs caractéristiques (pouvant toutes être du même canal de communication émotionnelle) pour une même interprétation revient à considérer plusieurs systèmes représentationnels complémentaires. Une fusion des données complémentaires est donc nécessaire. Contrairement à ce qui a été fait dans ICARE, nous avons donc choisi de concevoir un moteur unique et générique de synchronisation au lieu de disperser la fusion dans plusieurs composants dédiés. Ce choix est motivé par le fait que nous ne considérons que la synchronisation micro-temporelle, c'est-à-dire que nous fusionnons des blocs de données se chevauchant. Ne pas avoir à considérer une fenêtre temporelle pouvant être différente pour chaque composant (cas de la fusion macro-temporelle) ni le contexte d'interaction (cas de la fusion contextuelle) permet de s'affranchir des spécificités de chaque composant et de n'utiliser qu'un mécanisme générique de synchronisation, qui peut donc être déporté en arrière-plan. Dans cette section, nous présentons donc ce moteur de synchronisation. Il s'agit d'une entité sous-jacente du système, capable de communiquer avec les différents composants. La figure 37 reprend la figure 32 du paragraphe 5.3.5 et y fait apparaître le moteur de synchronisation.

Notre moteur dispose d'une structure de stockage appelée conteneur et de trois algorithmes : une synchronisation des données en entrée du conteneur, une synchronisation des données avant l'envoi de données complémentaires à un composant, et un algorithme de gestion de la mémoire (ramasse-miettes) permettant de libérer la mémoire correspondant aux données déjà consommées.

5.5.1 Structure : le conteneur

Nous nous inspirons fortement du *melting pot* [105] décrit à la partie 4.2.4 (page 76) pour construire notre moteur de synchronisation. Ainsi, nous définissons tout d'abord une structure d'accueil des données : le conteneur, illustré à la figure 38.

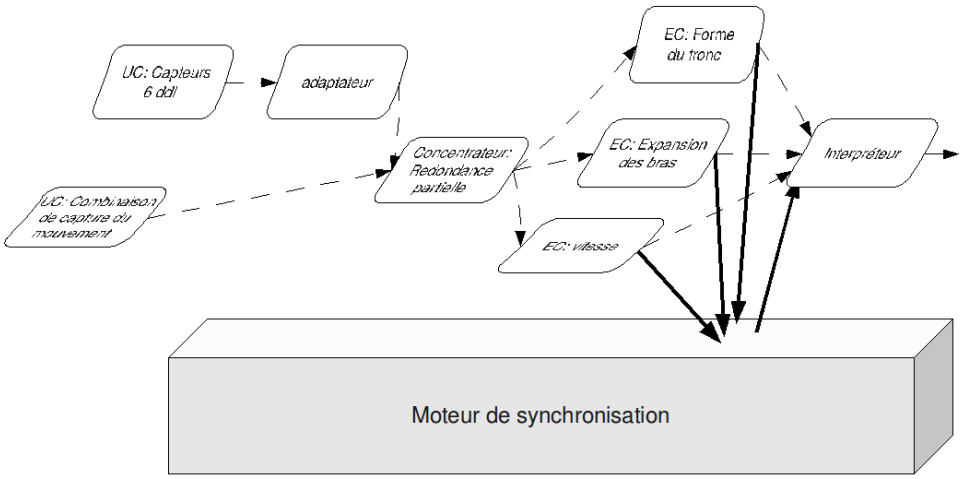


FIG. 37: Le moteur de synchronisation en arrière-plan de la structure en composants. Les flèches interrompues représentent les flux d'informations entre composants tels que définis par notre modèle ; les flèches en gras représentent le trajet réel de l'information. Seuls les trajets réels des flux de caractéristiques sont ici représentés.

Le conteneur est composé d'une ligne de temps (*timeline*) et de pistes. Chaque piste correspond à un composant producteur de données (tous les types de composants que nous avons présentés sont producteurs de données). Chaque piste a donc pour rôle de stocker les blocs de données issus du composant qui lui correspond en les alignant sur la ligne de temps. Par exemple, la figure 38 illustre le cas d'une application à cinq composants : deux unités de capture (UC_1 et UC_2), deux extracteurs de caractéristiques (EC_1 et EC_2), et un interpréteur (I_1).

Le conteneur est une structure dépendant uniquement des composants présents dans l'application. De plus, les algorithmes de synchronisation présentés dans les parties suivantes ne travaillent que sur des blocs de données, sans avoir à connaître les données transportées par ces blocs. La structure du conteneur peut donc être spécifiée sans manipuler les algorithmes, par exemple par une description XML. En perspective, et dans le cas d'un assemblage dynamique, il est également possible d'imaginer une construction "à la volée" du conteneur, sachant les composants présents dans l'application.

5.5.2 Synchronisation en entrée du moteur de synchronisation

Chaque bloc de donnée émis par l'un des composants du système est envoyé au moteur de synchronisation pour être recalé dans la ligne de temps du conteneur. Chaque bloc de données comporte une estampille de temps initial des données (voir le paragraphe 5.4.6 sur les spécifications du bloc de données), qui marque le moment ou la donnée est initialement capturée.

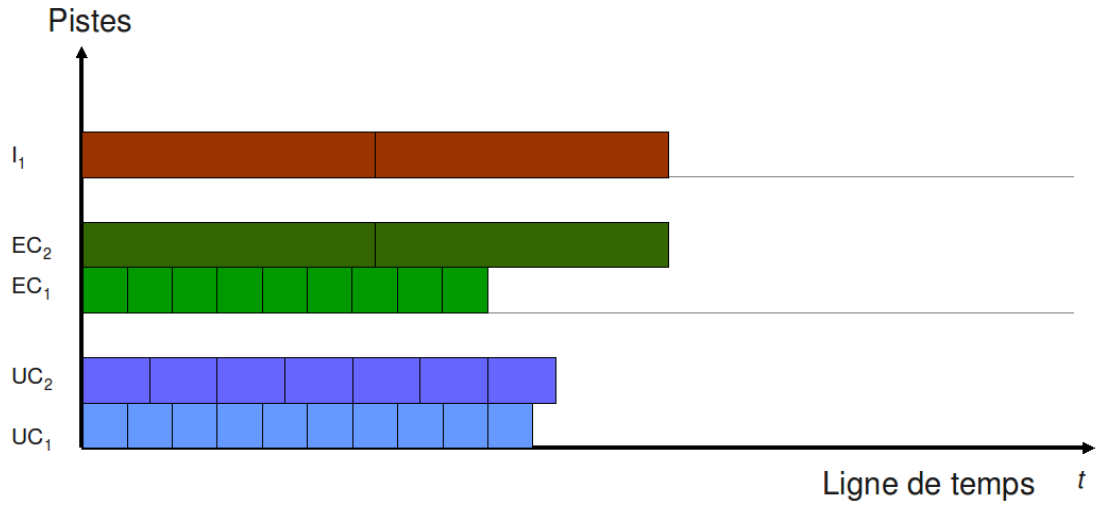


Fig. 38: Le conteneur : division en pistes.

Lorsqu'une donnée est capturée et émise par un dispositif, elle est prise en charge par l'unité de capture correspondant à ce dispositif. Une estampille de temps initial lui est alors apposée. Cette estampille peut être apposée directement par le dispositif si celui-ci le permet (par exemple, la combinaison de capture du mouvement présentée au chapitre 6 appose une estampille de temps à chaque mesure). Si le dispositif ne marque pas le temps de capture, alors l'unité de capture doit apposer cette estampille de temps initial à la création du bloc de données. L'unité de capture calcule et affecte également au bloc de données sa durée de validité. Dans la figure 38, cette durée de validité est représentée par les différentes longueurs de blocs dans les différentes pistes. Cette donnée de validité peut être calculée dans le cas de données produites régulièrement ou mesurée : la fin de la validité d'une mesure est alors marquée par l'arrivée dans le système de la mesure suivante.

Chaque composant reçoit des blocs à traiter et crée de nouveaux blocs en sortie. Un composant recopie l'estampille de temps initial et la durée de validité d'un bloc d'entrée dans le bloc de sortie correspondant, permettant ainsi la propagation de ces valeurs.

Le rôle de la synchronisation en entrée est d'aligner les blocs de données selon leur estampille de temps initial. Ainsi, les différentes données relatives à un même moment de la capture sont regroupées selon la ligne de temps. Cette méthode permet de regrouper toutes les données, qu'elles soient parallèles (plusieurs caractéristiques ou dispositifs de capture par exemple) ou correspondant à divers niveaux d'abstraction (une information de capture après son analyse et son interprétation).

Par exemple, à la figure 39, l'unité de capture UC_1 génère un premier bloc de données correspondant à l'information capturée au temps t_0. Ce bloc est donc estampillé t_0 comme valeur d'estampille de temps initial. Le bloc est aligné sur la ligne de temps à t_0. Ce bloc est ensuite utilisé par EC_1 pour en extraire une caractéristique. EC_1 crée donc un nouveau bloc

contenant la valeur de cette caractéristique. Ce bloc est également estampillé à t_0. Ainsi, le bloc produit par EC_1 est aligné avec le bloc produit par UC_1. Toutes les données relatives à la même expression émotionnelle sont ainsi alignées sur la ligne de temps du conteneur.

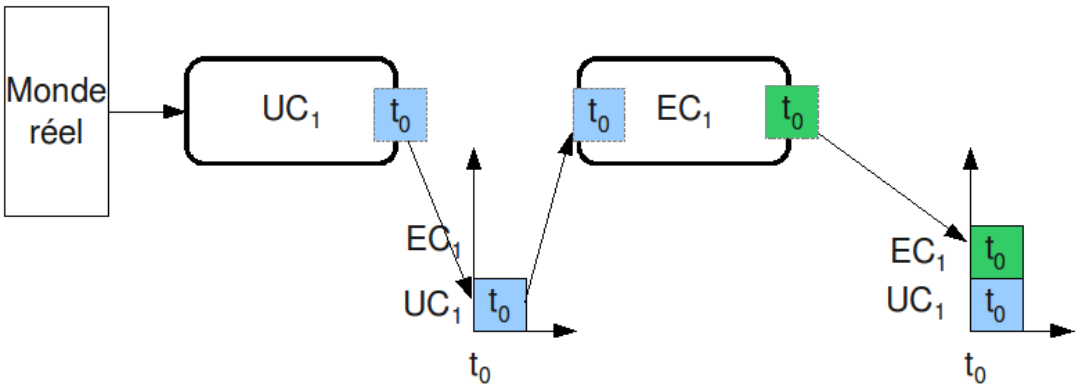


FIG. 39: Synchronisation en entrée du moteur de synchronisation

La synchronisation en entrée du moteur permet donc de synchroniser toutes les données relatives à une même expression émotionnelle lors de l'insertion de ces données dans le conteneur. Ainsi, il suffit de considérer un temps t pour obtenir toute l'information émotionnelle capturée ou calculée. Cet alignement permet, grâce à une synchronisation en sortie du conteneur, de récupérer des données complémentaires pour un composant.

Chaque bloc de données possède une deuxième estampille de temps : l'estampille de temps courant des données. Cette estampille n'est pas utilisée pour la synchronisation. Elle est apposée par le composant créateur du bloc. Cette estampille permet donc de connaitre, pour un bloc, le temps écoulé entre l'arrivée initiale de la donnée dans le système et la création du bloc.

5.5.3 Synchronisation en sortie du conteneur : les pots de synchronisation

En dehors des unités de capture, tous les composants que nous avons introduits sont producteurs mais également consommateurs de blocs de données. Un composant consommateur peut consommer des blocs d'une ou plusieurs pistes. Par exemple, les interpréteurs se basent sur plusieurs caractéristiques (donc plusieurs pistes) pour en inférer une émotion. Ce cas correspond à la complémentarité de plusieurs modalités : un composant s'appuie sur plusieurs flux complémentaires. Dans ce paragraphe, nous présentons l'algorithme des pots de synchronisation, inspirés du *melting pot* [105]. Ces pots de synchronisation s'appuient uniquement sur la fusion microtemporelle, c'est-à-dire lorsque les données selon plusieurs modalités se chevauchent temporellement.

Le principe du pot de synchronisation part de l'hypothèse que le moteur de synchronisation connaît les différents flux requis par chacun des composants. Cette information peut être

donnée par le composant lui-même (grâce à ses attributs *domaine des valeurs d'entrée*), par une source extérieure (fichier XML) ou lors de la construction du moteur lui-même (codage "en dur"). Pour chaque composant, le moteur de synchronisation va créer un "pot" : une sorte de conteneur miniature ne contenant que les pistes nécessaires au composant considéré. Le conteneur va alors scruter (*monitorer*) ces pistes et recopier les blocs dans le pot dès leur apparition. Pour cela, le conteneur doit connaître les domaines de valeurs d'entrées (c'est-à-dire les pistes utilisées en entrée) de chaque composant. Nous avons souligné cet attribut dans notre spécification à la section 5.4 (page 107). Lorsque chaque piste du pot contient au moins un bloc, le chevauchement de tous les blocs est vérifié. En effet, il existe dans le pot une piste ne contenant qu'un seul bloc (le plus long). Tous les blocs ne chevauchant pas ce bloc sont éliminés. Enfin, le pot de synchronisation est envoyé au composant auquel il est lié.

La figure 40 illustre le fonctionnement d'un pot de synchronisation. L'extracteur de caractéristiques EC_1 nécessite les données de deux unités de capture UC_1 et UC_2. Ces unités de capture délivrent des données à des fréquences différentes. Au temps t_0, un pot de synchronisation est créé pour le composant EC_1. Ce pot est constitué de deux pistes permettant d'accueillir les données provenant d'UC_1 et UC_2. Le pot scrute les pistes du conteneur relatives aux unités de capture UC_1 et UC_2. Au temps t_1, un premier bloc est inséré sur la piste UC_1. Ce bloc est copié dans le pot à la piste correspondante (étape (1)). Au temps t_2, un bloc est inséré sur la piste UC_2. Ce bloc est également copié dans le pot (étape (2)). Ce dernier dispose maintenant d'au moins un bloc dans chacune de ses pistes ; il est alors envoyé en entrée du composant EC_1, qui traite les données reçues pour produire un bloc. Ce bloc est inséré au temps t_0, et le pot de EC_1 est vidé (étape (3)). Au temps t_3, un nouveau bloc de UC_1 est inséré dans le conteneur. Ce bloc est copié dans le pot de EC_1 (étape (4)). Au temps t_4 deux blocs de UC_1 et UC_2 sont simultanément insérés dans le conteneur et copiés dans le pot (étape (5)). Le pot contient au moins un bloc sur chacune de ses deux pistes et est donc envoyé à EC_1 puis vidé. EC_1 produit un nouveau bloc de données qui est inséré dans le conteneur (étape (6)).

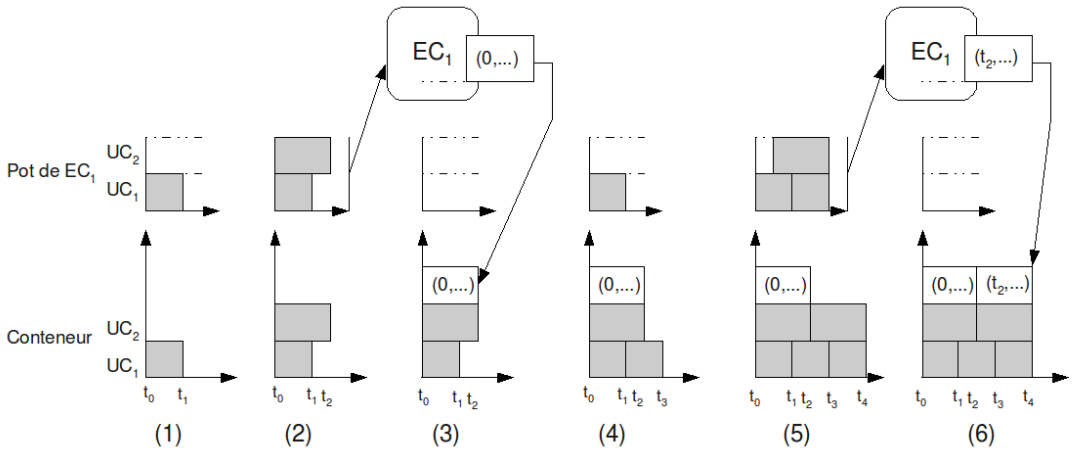


Fig. 40: Fonctionnement du pot de synchronisation d'un extracteur de caractéristique

Limitations du pot de synchronisation

La première limitation du pot de synchronisation tel que décrit dans cette section est qu'il ne restreint pas une piste à contenir un nombre fixe de blocs. Ceci oblige les composants à savoir manipuler, pour chaque piste en entrée, une liste de longueur variable de blocs de données. La seconde limitation est le délaissement de certains blocs. Dans certains cas particuliers, le pot de synchronisation ne prend pas en compte des blocs jusqu'alors non utilisé (pas de chevauchement avec les autres blocs). Ceci conduit à un sous-échantillonnage des mesures à la plus haute fréquence. L'impact d'un tel sous-échantillonnage reste à évaluer par des tests expérimentaux.

5.5.4 Gestion de la mémoire : le ramasse-miettes

Le dernier algorithme de notre moteur de synchronisation est le "ramasse-miettes". Son rôle est de surveiller la consommation des blocs afin de pouvoir éliminer les blocs complètement consommés. En effet, sans un tel mécanisme, les blocs de données seraient insérés de façon continue dans le conteneur le long de la ligne de temps. Nous considérons deux stratégies possibles.

La première stratégie se base sur le fait que la branche émotion a pour rôle de délivrer en sortie un flux d'émotions. Ce flux est également inséré dans le conteneur de la même façon que les autres blocs et représente le dernier niveau d'abstraction que l'information atteint au sein de la branche. Cela signifie qu'un bloc représentant une émotion de ce flux final, portant une estampille t d'entrée dans le système, est le produit du traitement de blocs portant une estampille d'entrée dans le système d'au plus tard t. Ainsi, la première stratégie consiste à scruter la piste correspondant au flux d'émotions (ou aux flux, si plusieurs flux d'émotions sont envoyés hors de la branche). Lorsqu'un bloc est inséré provenant de ce flux, il est alors possible d'effacer tous les blocs des autres pistes dont l'estampille de temps d'entrée dans le système est inférieure à t.

La deuxième stratégie consiste à permettre au moteur de synchronisation de recenser les flux d'entrées de tous les composants du système. Il est ainsi possible, lors de l'insertion d'un bloc dans le conteneur, d'y initialiser un compteur, notant le nombre de composants devant consommer ce bloc. Chaque consommation décrémente le compteur ; un compteur à 0 indique un bloc totalement consommé. Il est alors possible de l'effacer de la piste.

5.5.5 Exemples de fonctionnement

Nous illustrons le moteur de synchronisation par un exemple considérant des données acquises à des fréquences différentes et impliquant des phases distinctes (décalage le long de la ligne de temps). Pour cela nous considérons un cas tiré de [143] permettant d'illustrer deux aspects temporels : les variables d'état et les flux temporisés ou bloquants.

Volpe [143] se base sur un flux vidéo pour effectuer une reconnaissance d'émotions par le mouvement. En particulier, le système calcule la quantité de mouvement d'un danseur

(figure 41). Cette quantité de mouvement est ensuite seuillée pour obtenir des phases de pause (QoM en dessous d'un certain seuil) et des phases de mouvement ("cloches" entre les phases de pause). Chaque "cloche" de mouvement est ensuite analysée dans son ensemble pour extraire la fluidité du geste. Dans la figure 41, les pics abrupts correspondent à une augmentation rapide de la quantité de mouvement, c'est-à-dire à une forte accélération, suivie d'une forte décélération : ils traduisent des saccades dans le mouvement. La cloche plus régulière ne montre pas de pics : le mouvement est fluide. Dans le système de Volpe, il est nécessaire de connaître le mouvement dans son entier pour en extraire la fluidité. Or, la segmentation d'un mouvement n'est possible que si celui-ci est terminé et qu'une nouvelle phase de pause est détectée. Un mouvement continuel, sans pause, interdirait donc le calcul de la fluidité.

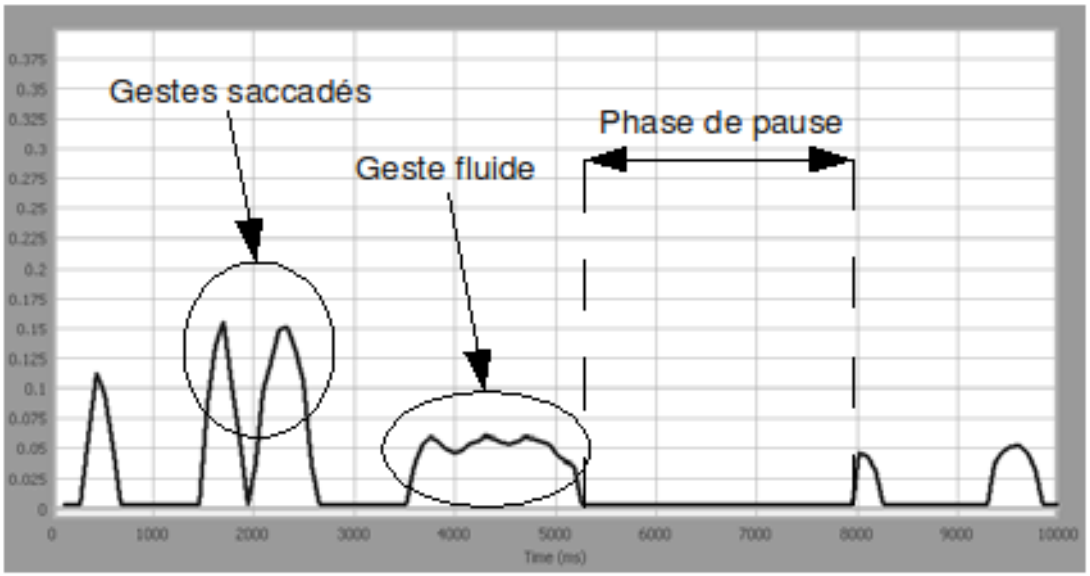


Fig. 41: Courbe de la QoM dans le temps, tirée de [143].

La segmentation en phases de pause et de mouvement est bloquante. Le sujet doit à nouveau entrer en phase de pause pour que la dernière phase de mouvement prenne fin et puisse être analysée pour en extraire la fluidité. Le calcul est donc temporisé jusqu'à ce que le mouvement se finisse. Nous avons modélisé la sous-partie du système de Volpe (du calcul de la QoM au calcul de la fluidité) à la figure 42 en utilisant les composants de la branche émotion.

Nous modélisons le sous-système par une séquence de trois composants. Le premier calcule la quantité de mouvement depuis un flux vidéo. Le deuxième composant a pour rôle de segmenter le mouvement en phases de pause et en phases de mouvement. Le quatrième composant prend en entrée une phase complète de mouvement (c'est-à-dire une liste à taille variable d'informations du corps) pour fournir la fluidité du geste. Les composants sont reliés entre eux (flèches en pointillé dans la figure 42) et échangent des données. Ces blocs de données échangés sont enregistrés sur des pistes dans le conteneur, puis synchronisés avant d'être renvoyés au composant suivant.

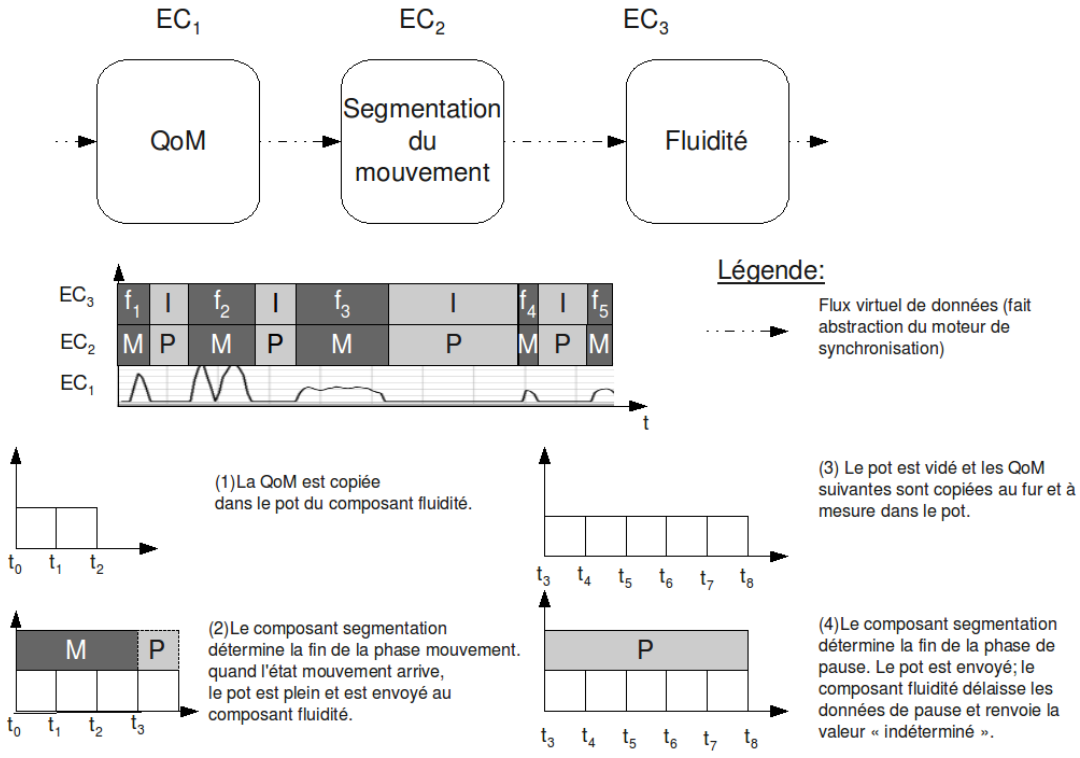


FIG. 42: Représentation selon notre modèle d'une partie du système de [143].

Tout d'abord, le composant QoM produit des blocs de données de façon régulière : en effet, une QoM est calculée à chaque image vidéo fournie au composant. Ces blocs de QoM sont copiés dans le pot du composant fluidité au fur et à mesure de leur arrivée (étape (1) dans la figure 42). Ils sont également directement envoyés au composant segmentation. Ce composant segmentation possède un buffer interne dans lequel il va stocker la séquence de QoM jusqu'à ce qu'il détermine un changement d'état. De t_0 à t_3, la QoM est au dessus du seuil de changement d'état. A t_4, la QoM est au dessous du seuil. Le composant segmentation renvoie donc un bloc de données de longueur t_0 à t_3 porteur de l'état "phase de mouvement". Ce bloc de donnée comporte comme estampille de temps initial l'estampille du premier bloc de QoM utilisé, soit t_0. Le premier bloc "mouvement" est synchronisé en entrée du conteneur : il est aligné sur t_0. Le buffer interne du composant segmentation est alors vidé et recommence à stocker les blocs de QoM entrants.

Lorsque le bloc "phase de mouvement" entre dans le conteneur (piste EC_2), il est copié dans le pot du composant fluidité (étape (2) dans la figure 42). Le pot est alors complet et est envoyé au composant fluidité. Celui-ci évalue l'état de la phase (mouvement) et calcule donc la fluidité sur la séquence de QoM véhiculée dans le pot. Cette mesure de fluidité est ajoutée au conteneur (piste EC_3). Encore une fois, l'estampille de temps initial du bloc de données "fluidité" est copié du bloc "phase de mouvement". Le premier bloc "fluidité" convoie la valeur f_1 et est aligné sur t_0. Le pot du composant fluidité est vidé, et les blocs de QoM y sont à

nouveau copiés à partir du temps t_3 (étape (3) dans la figure 42).

Le processus est le même ensuite : le composant segmentation définit un état de pause de t_3 à t_8 (étape (4) dans la figure 42). Le composant fluidité évalue l'état de la phase (pause) et ne fait aucun traitement mais renvoie un bloc de données en sortie portant la valeur "indéterminé".

Lorsque l'interpréteur du système (non représenté sur la figure) envoie en sortie le bloc de données "émotion" entre t_0 et t_3, le système sait alors que toutes les données entre t_0 et t_3 ont été utilisées ; le ramasse-miette libère l'espace occupé par ces données.

5.6 Implémentation de la branche émotion

Nous avons défini, dans les sections précédentes, le rôle de chaque composant que nous proposons dans un cadre "classique" où l'on construit entièrement une application de reconnaissance d'émotions. Nous ne spécifions cependant pas l'implémentation de ces composants selon une technologie à composants spécifique (JavaBeans, Corba, etc.). Notre modèle d'architecture est conceptuel et laisse le choix pour son implémentation (architecture implémentationnelle). Utiliser OpenInterface[8] et son OIDE (OpenInterface Interaction Development Environment) est une option pour définir un outil d'assemblage, à condition que les composants implémentés vérifient les spécifications d'OpenInterface. De même l'éditeur ICARE peut être utilisé dans le cas où seuls des composants JavaBeans sont utilisés. L'extension la plus importante de ces outils résiderait alors dans le développement du moteur de synchronisation qui permettrait la communication entre les composants.

5.6.1 Intégration de systèmes tiers dans la branche émotion

Les composants que nous avons proposés sont ouverts en ce qui concerne leur implémentation ; nous en spécifions quelques attributs génériques. Cette ouverture permet d'intégrer des systèmes tiers existants en encapsulant le système à intégrer dans un composant proposant les attributs et sorties que nous spécifions dans ce chapitre. Le choix du type de composant encapsulant dépend du flux de sortie du système.

Par exemple, une application fournissant un ensemble de caractéristiques peut être encapsulée dans un composant extracteur de caractéristiques délivrant plusieurs flux de caractéristiques. Pour le moteur de synchronisation, ce composant ne consomme pas de données ; il ne fait qu'en produire. De la même façon une application existante de reconnaissance d'émotions peut être toute entière encapsulée dans un interpréteur. Ici également cet interpréteur ne consomme pas de données, il ne fait qu'en produire.

L'approche à composants permet donc l'encapsulation de systèmes existants. L'enjeu d'intégration de code existant concerne principalement le moteur de synchronisation. En effet,

[8] http ://www.openinterface.org/home/

celui-ci se base sur l'estampille de temps initial (voir paragraphe 5.5.2, page 114), apposée par le dispositif ou son unité de capture correspondante. Un système existant peut ne pas souscrire à cette règle d'affectation d'une estampille de temps initial. Dans ce cas, le moteur de synchronisation ne peut aligner les données produites par le système tiers avec les autres données du système. Le composant encapsulant le système tiers peut apposer une estampille au moment d'encapsuler les données produites par le système tiers dans des blocs de données. Cette solution présente le défaut de potentiellement décaler les données produites par le système tiers d'un temps égal au temps de traitement. Si possible, il est donc préférable que le composant encapsulateur puisse communiquer avec l'application tierce afin d'apposer une estampille à chaque bloc la plus proche possible du moment de la mesure.

5.6.2 Validation et simulation d'un composant ou d'un patch de composants

L'implémentation d'un extracteur de caractéristiques ou d'un interpréteur soulève le problème de la validation du composant. Dans le cas où le composant est tiré de la littérature (en psychologie ou en informatique), il est à noter que l'implémentation consiste en une traduction forcément inexacte de la description proposée, ce qui peut introduire un biais. Par exemple, nous avons vu à la partie 3.1 (page 52) la description donnée par De Meijer pour la caractéristique d'expansion du tronc. Traduire cette description en un algorithme d'extraction induit un biais qui peut jusqu'à invalider la caractéristique considérée. En définitive, seuls des tests peuvent permettre de valider un composant. Pour cela, l'approche à composants sous-jacente à notre modèle permet une mise au point rapide, par exemple en simulant des composants non encore disponibles. Nous identifions deux cas : la simulation et le magicien d'Oz.

Simulation de composant : senseurs et systèmes virtuels

Une unité de capture peut être développée comme un composant purement logiciel (senseur virtuel), émulant les capacités d'un véritable senseur. Cette émulation peut être complètement automatique : par exemple, un composant mimant un électrocardiogramme envoie des impulsions cardiaques avec un tempo aléatoire. Elle peut également permettre de traiter le signal en différé : par exemple, un composant lisant un fichier préenregistré de l'électrocardiogramme.

Par extension, il est possible de simuler les composants extracteurs de caractéristiques et les composants interprétation. Il s'agit alors de composants ne consommant pas de données ; l'implémentation du composant se charge de créer et d'envoyer des données correspondant au format de sortie. L'intérêt de cette forme d'implémentation est de pouvoir mettre au point une stratégie dans la création des données (aléatoire, selon un schéma prédéfini, etc.). Un fichier préenregistré peut être lu plusieurs fois, afin de fournir la même mesure lors de l'évaluation de composants particuliers.

Implémentation de composants "Magicien d'Oz"

Une unité de capture, un extracteur de caractéristiques ou un interpréteur peut être implémenté de façon à proposer une interface permettant à un être humain de décider du signal

à envoyer : par exemple, choisir le rythme cardiaque à envoyer grâce à un curseur plutôt que d'utiliser un électrocardiogramme, ou choisir par un menu déroulant la valeur courante d'une caractéristique particulière. Cette dernière option permet des expérimentations de type Magicien d'Oz, où un compère simule une ou plusieurs fonctions non encore disponibles [140].

5.7 Validation de la branche émotion : modélisation de l'existant

Une première façon de valider un modèle d'architecture conceptuelle est de modéliser des applications existantes ou de montrer l'architecture d'un système de reconnaissance d'émotions. Une autre façon de valider le modèle est de l'appliquer au développement d'un nouveau système de reconnaissance d'émotions. Ceci fait l'objet du chapitre suivant. Dans ce paragraphe, nous modélisons deux travaux existants. Premièrement nous modélisons l'étude de De Meijer [46] sur la contribution de caractéristiques générales du mouvement à l'attribution d'émotions. Comme nous l'avons vu à la partie 3.3, De Meijer considère sept dimensions de mouvement : le mouvement du tronc (étiré ou voûté), le mouvement des bras (ouverture ou fermeture), la direction verticale du mouvement (vers le haut ou vers le bas), sa direction sagittale (vers l'avant ou vers l'arrière), sa vélocité (rapide ou lent), sa force (fort ou léger), et enfin sa directivité (direct ou indirect). L'expérimentation décrite dans [46] lui permet de définir la contribution de chacune de ces caractéristiques à l'attribution de 14 émotions par des observateurs humain. Deuxièmement, nous modélisons par notre architecture les travaux de Volpe [143], que nous avons déjà utilisés comme exemple de fonctionnement de notre moteur de synchronisation au paragraphe 5.5.5 (page 118).

5.7.1 Modélisation d'une application basée sur l'existant en psychologie [46]

La figure 43 illustre l'architecture d'un système basé sur [46] selon notre modèle. Dans cette modélisation, nous choisissons d'utiliser deux dispositifs : le premier est une combinaison de capture du mouvement supposée "idéale" (précise, robuste, résistante aux occlusions... etc.) délivrant un flux continu des coordonnées de chaque segment du corps. Le deuxième est un gant de données fournissant les informations de conformation de la main droite du sujet observé. En effet, la caractéristique de force spécifie que les muscles sont contractés et les poings fermés. L'utilisation d'électromyogrammes dans le cadre d'une reconnaissance par le mouvement est compliquée par les fils reliés à la centrale de mesure ; nous choisissons donc de détecter la forme de mains par un gant de données sans fil.

Notre modélisation d'une application basée sur [46] met en jeu treize composants : deux unités de capture, dix extracteurs de caractéristiques, et un interpréteur. Chacune des unités de capture établit l'interface avec un des capteurs physiques utilisés. Ainsi UC_1 établit une interface avec la combinaison de capture du mouvement et délivre un flux continu de coordonnées du corps. UC_2 établit une interface avec le gant de données et délivre un flux continu de coordonnées des segments de la main droite du sujet, à une fréquence deux fois moindre que UC_1. Ces deux flux de capture pourraient être concentrés en un seul flux comportant à la fois les coordonnées du corps et de la main droite. Dans notre cas, une telle concentration

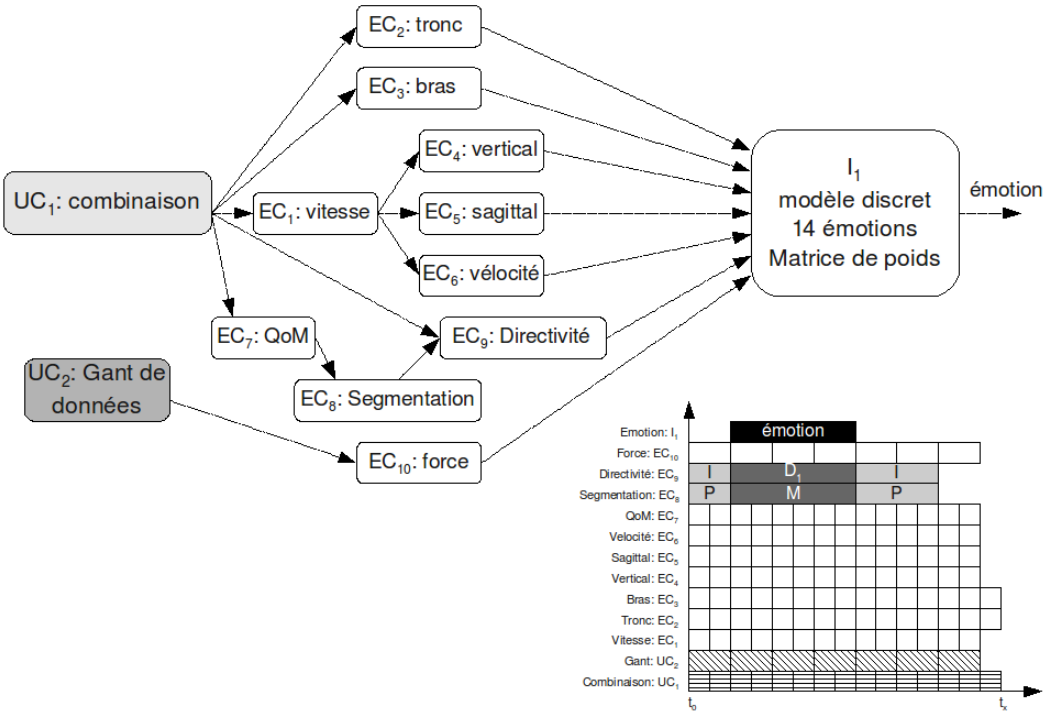


FIG. 43: Modélisation de [46] avec la branche émotion.

est inutile. En effet, seul un extracteur de caractéristiques se base sur le flux de coordonnées de la main. Les autres ne se basent que sur les coordonnées du corps.

Nous modélisons dix extracteurs de caractéristiques pour sept caractéristiques étudiées. En effet, nous traduisons le calcul de la composante sagittale du mouvement comme un seuillage de la composante sagittale du vecteur vitesse au bassin, et la composante verticale comme un seuillage de la composante verticale du vecteur vitesse du bassin. La vélocité est un seuillage de la norme de ce vecteur vitesse. Ainsi, nous localisons le calcul du vecteur vitesse dans le composant EC_1. Le vecteur vitesse ainsi calculé est ensuite envoyé aux trois composants d'extraction de la composante sagittale du mouvement (EC_4), de sa composante verticale (EC_5), et de sa vélocité(EC_6). Les extracteurs EC_2 et EC_3 calculent respectivement la forme du tronc et l'expansion des bras à chaque pas de temps.

Le calcul de la directivité se fait en trois étapes similaires au calcul de la fluidité décrit au paragraphe 5.5.5. La directivité d'un geste (par exemple du poing) est le ratio entre la longueur du chemin parcouru dans une phase de mouvement sur la longueur du chemin le plus direct. Il est donc nécessaire de segmenter la mesure en phases de pause et de mouvements. Comme au paragraphe 5.5.5, nous utilisons un composant QoM (EC_7) calculant la quantité de mouvement à chaque mesure de la combinaison, puis un composant permettant de segmenter le mouvement (EC_8). Le composant EC_9 "directivité" se base sur les données de la combinaison et sur l'état de la phase courante pour renvoyer une valeur de directivité dans les phases

de mouvements, et "indéfini" dans les phases de pause. Enfin, l'extracteur EC_{10} utilise les données issues du gant et de l'unité de capture UC_2 pour analyser la forme de la main et déterminer si le poing est fermé ou non.

Enfin, L'interpréteur I_1 s'appuie sur les caractéristiques délivrées par les extracteurs $EC_{2..6,9,10}$. Dans la modélisation proposée, nous avons choisi de reprendre complètement les résultats de De Meijer. L'interpréteur permet donc de discriminer 14 émotions. Pour cela, nous utilisons une matrice contenant les poids de contribution de chaque caractéristique de mouvement à chaque émotion donnés par de Meijer comme résultat de son expérimentation. Le composant I_1 renvoie un flux d'émotions. La directivité est ici une caractéristique bloquante ; l'interpréteur nécessite la valeur de la directivité pour inférer une émotion. Les blocs de données d'émotions correspondent donc à des phases de mouvement.

Bénéfices de la modélisation

Une telle modélisation permet de satisfaire aux critères de modifiabilité. Ainsi, il est possible de modifier les traitements effectués par un composant sans que cela n'impacte le reste du système. Par exemple, l'expansion des bras peut être calculée sur la distance entre les poignets ou en prenant également en compte la distance par rapport au tronc (afin d'éviter le cas où les bras tendus vers l'avant avec les mains jointes donne pour valeur "bras fermés", par exemple). Une telle modification requiert de modifier le composant EC_3 dans la figure 43. Les valeurs d'entrée et de sortie n'étant pas modifiées, le reste du système peut être conservé tel quel. De la même façon, il est possible de calculer la composante sagittale du mouvement dans un repère absolu (tel que c'est le cas dans l'architecture proposée) ou selon un repère relatif orienté par le corps. L'utilisation d'un repère relatif nécessite les coordonnées du corps afin de connaître sa position et son orientation. L'échange d'un repère absolu à un repère relatif est illustré à la figure 44. Il nécessite de replacer le composant EC_5 par le composant EC_{11} et d'établir les connexions nécessaires. Ce cas spécifique est implémenté dans notre application eMotion, décrite au chapitre suivant ; il y est explicité plus en détail.

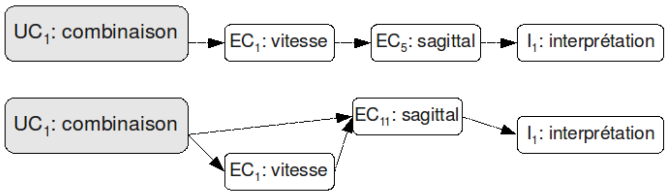


Fig. 44: Changement du calcul de la composante sagittale du mouvement.

De cette architecture nous tirons un exemple de simulation de composant. Nous avons vu que la caractéristique de force est extraite des données fournies par un gant de données. Si le dispositif n'est pas disponible, il est possible de simuler la force. Comme nous l'avons décrit au paragraphe 5.6.2, ce composant peut envoyer des données de façon automatique (préenregistrées, selon un modèle, etc.), ou par le biais d'un intervenant humain travaillant

sur interface graphique (par exemple en faisant glisser un *slider*). Dans ce cas, il suffit de déconnecter l'unité de capture correspondant au gant de données et de connecter l'extracteur simulé au composant interpréteur à la place de l'extracteur EC_{10}.

5.7.2 Modélisation d'une application existante de reconnaissance d'émotions [143]

La figure 45 illustre l'architecture d'un système simplifié de [143], décrit en partie au paragraphe 3.4.2 (page 57), selon notre modèle. L'application originale est développée sous EyesWeb (voir paragraphe 5.2.2).

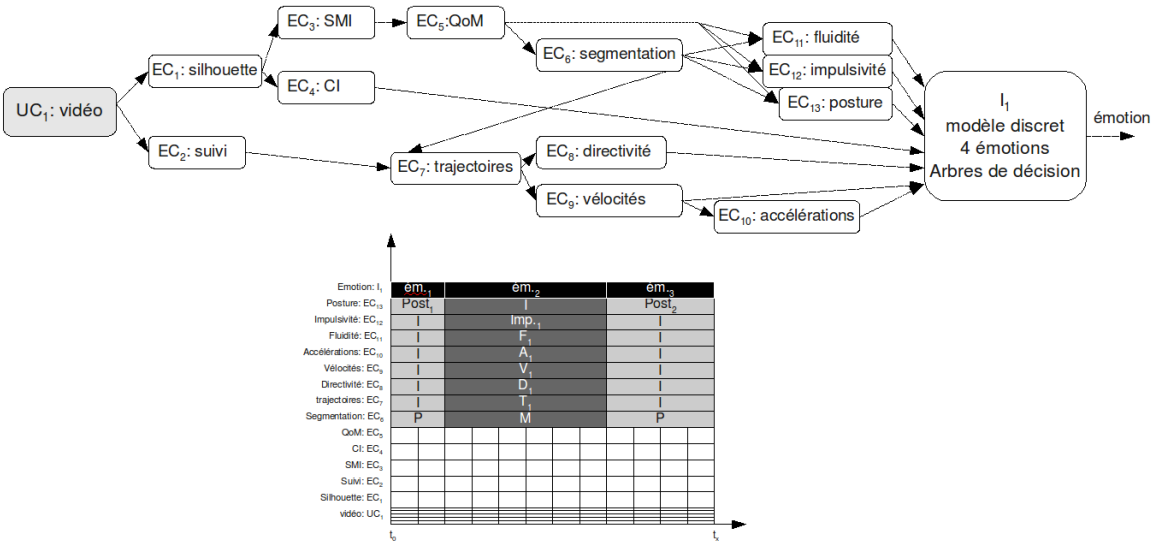


FIG. 45: Modélisation de [143] selon notre modèle conceptuel.

Le système se base sur un unique dispositif de capture : une caméra vidéo, délivrant un flux d'images (UC_1).

Des images du flux vidéo sont extraites les silhouettes du sujet (EC_1), ainsi que les coordonnées dans l'image de points d'intérêts grâce à un algorithme de suivi (EC_2). L'extraction de silhouette permet de générer les *Silhouette Motion Images* (SMIs) (EC_3) et de calculer l'index de contraction (CI) (EC_4). La génération de SMI permet de calculer la quantité de mouvement (QoM) 'EC_5). Ces caractéristiques sont décrites au paragraphe 3.4.2. La QoM permet la segmentation en phases de pause et de mouvement comme décrit au paragraphe 5.5.5 grâce à un composant de segmentation (EC_6). Les composants suivants utilisent cette segmentation et analysent les phases de mouvement (composants EC_7 à EC_{12}) ou les phases de pause (composant EC_{13}).

Tout d'abord, les trajectoires des points d'intérêt suivis par EC_2 sont calculées à chaque phase de mouvement. Une trajectoire permet de calculer la directivité d'un geste en faisant

le ratio du chemin effectué sur le chemin le plus court (EC_8). La vélocité moyenne (EC_9) et l'accélération moyenne (EC_{10}) du mouvement sont également calculées. La segmentation permet également d'analyser les phases de mouvement pour en extraire la fluidité (EC_{11}, décrite au paragraphe 5.5.5) et l'impulsivité (EC_{12}, obtenue en analysant la pente de la QoM au début d'une phase de mouvement - plus la pente est raide, plus le mouvement est impulsif). Enfin, le composant EC_{13} analyse la posture du sujet durant les phases de pause.

L'interprétation des caractéristiques est faite sur un modèle discret de quatre émotions : la colère, la peur, la tristesse et la joie grâce à des arbres de décision. Lors des phases de mouvement, l'interpréteur se base sur les caractéristiques de fluidité, d'impulsivité, d'index de contraction, de directivité, et sur les vélocités et accélérations. En phase de pause, l'interpréteur se base alors sur l'index de contraction et la posture pour déterminer l'émotion.

Bénéfices de la modélisation

L'application originale de cet exemple est développée sous EyesWeb : il s'agit d'un patch de composants développés et assemblés grâce à cet outil. Le patch original est donc modifiable et extensible dans le cadre de l'outil EyesWeb. Notre modèle conceptuel permet ici de structurer les différents composants selon les trois niveaux de capture, analyse et interprétation et d'y appliquer les principes de la multimodalité en IHM pour caractériser le système et concevoir de nouvelles extensions. Il est par exemple possible d'imaginer l'utilisation d'un combinaison de capture du mouvement dans le système venant remplacer la vidéo et l'algorithme de suivi la détermination des trajectoires. Le calcul des trajectoires nécessitant une segmentation en phases de pause et de mouvement, l'introduction d'une combinaison pour mesurer les trajectoires induirait la complémentarité des modalités <*combinaison, coordonnées du corps*> et <*caméra vidéo, flux vidéo, silhouette, SMI, QoM, segmentation*> pour le calcul de trajectoires.

Ainsi, le système [143] est une implémentation dans l'outil EyesWeb de l'architecture proposée dans ce paragraphe. Cette même architecture peut être implémentée grâce à d'autres outils (comme l'éditeur ICARE) ou en utilisant d'autres langages.

5.8 Conclusion

Dans ce chapitre nous avons présenté notre modèle d'architecture pour la reconnaissance d'émotions : la branche émotion. Après avoir montré son intégration dans l'architecture globale d'un système interactif au chapitre précédent (voir paragraphe4.3.3), nous avons détaillé dans ce chapitre les composants qui la composent. Ces composants communiquent entre eux grâce à un moteur de synchronisation qui travaille en tâche de fond de l'assemblage des composants.

Le processus de développement d'une architecture à composants suit cinq étapes, faisant appel à autant de métiers différents : la spécification par le concepteur, l'implémentation par le développeur, l'assemblage par l'assembleur, le déploiement par l'administrateur du site, et enfin l'exécution par l'utilisateur. Le concepteur a pour rôle de spécifier les différents

composants à mettre en place. Il définit les entrées/sorties du composant et les fonctions que le composant doit remplir. Des connaissances métier sont nécessaires à l'établissement de ces spécifications. Le développeur a ensuite pour rôle d'implémenter chaque composant en suivant les spécifications données par le concepteur. Il n'a donc pas besoin de connaissances métier, ni de connaître la façon dont les composants seront assemblés. L'assembleur réalise l'assemblage des composants entre eux. En effet, chaque composant peut être vu comme une brique logicielle capable de se connecter à d'autres composants. L'assemblage consiste donc à connecter entre eux les composants compatibles et adéquats par rapport à la tâche finale. L'administrateur déploie ensuite sur site le logiciel qui sera exécuté par l'utilisateur. Cette différenciation des métiers permet donc d'associer un "expert métier" (par exemple, un psychologue spécialisé en reconnaissance des émotions) aux phases de cahier des charges de l'application et d'assemblage de composants existants pour la production d'une nouvelle application de reconnaissance d'émotions.

Tandis que dans ce chapitre nous avons modélisé deux travaux existants, le chapitre suivant est consacré au développement d'un nouveau système de reconnaissance selon notre modèle, constituant ainsi une forme complémentaire de validation.

Troisième partie

Contributions pratiques : réalisations logicielles

Chapitre 6

eMotion, un canevas logiciel de reconnaissance d'émotions basée sur la gestuelle

Dans ce chapitre, nous présentons eMotion, un canevas logiciel de reconnaissance d'émotions basé sur la gestuelle. Ce système illustre la branche émotion, présentée au chapitre 5.

Nous abordons dans un premier temps les motivations qui nous ont poussées à développer eMotion, ainsi que les limitations du prototype actuel. Nous décrivons ensuite l'architecture du système, illustrant ainsi notre modèle présenté au chapitre 5. Nous illustrons ensuite le pouvoir génératif des propriétés CARE appliquées à la reconnaissance d'émotions en détaillant les processus de choix de modalités dans le système. Nous décrivons enfin le processus de validation de notre système, processus mettant en œuvre une expérimentation basée sur la danse.

6.1 Motivations et limitations

Nous avons développé le système eMotion en vue d'illustrer le modèle de la branche émotion présenté au chapitre 5. eMotion adopte donc une approche à composants. Pour les composants des niveaux Analyse et Interprétation, nous nous sommes principalement appuyés sur l'étude effectuée par de Meijer dans [46]. En effet, cette publication offre un ensemble de caractéristiques décrites avec une précision suffisante pour permettre leur implémentation, ainsi que les poids de chacune de ces caractéristiques à l'attribution d'émotions ; données que nous avons réutilisées pour implémenter l'interprétation des caractéristiques. Nous montrons ainsi le potentiel de réutilisation de l'existant dans notre architecture.

De par son architecture en tant que système interactif, couplée à son architecture fonctionnelle reposant sur la branche émotion, eMotion se définit comme un canevas logiciel extensible pour la reconnaissance d'émotions. L'application eMotion permet en effet l'ajout de composants présentant une facette de présentation pour la paramétrisation du modèle, proposant

ainsi au concepteur d'interagir avec l'application.

De plus, nous avons vu au chapitre 4 que l'intégration de la reconnaissance d'émotions dans les principes de l'interaction multimodale permettait d'imaginer des cas de combinaisons de modalités nouveaux dans les systèmes de reconnaissance d'émotions. La notion de choix d'une modalité par le concepteur ou par le système, absente dans les systèmes existants, est ici implémentée et illustrée.

eMotion est donc un canevas logiciel extensible et instanciant le modèle de la branche émotion. Dans le cadre de nos travaux, nous avons développés des composants spécifiques au mouvement, implémentant ainsi un système de reconnaissance d'émotions par le mouvement. Les différentes caractéristiques de mouvement et leur interprétation sont issues de la littérature et abordent le mouvement d'une façon générique (c'est-à-dire que les caractéristiques étudiées ne sont pas liées à la danse, notre cas d'application). Dans ce chapitre, nous décrirons eMotion selon ces deux points de vue : eMotion en tant que canevas logiciel extensible pour la reconnaissance d'émotions, et eMotion en tant que système de reconnaissance des émotions par le mouvement.

Le domaine applicatif dans lequel nous avons testé le canevas logiciel eMotion est la danse. Cet aspect applicatif concret et dédié, ainsi que ses contraintes et possibilités, sont développés dans le chapitre 7. En particulier, notre cadre imposait une contrainte de temps réel dans la reconnaissance d'émotions. Nous ne pouvions donc pas calculer de caractéristiques temporisées.

eMotion est un prototype n'implémentant pas toutes les fonctionnalités de notre modèle. En particulier, le moteur de synchronisation n'est pas intégré dans cette version d'eMotion ; des composants *ad hoc* permettent la synchronisation des flux de données lorsqu'elle est nécessaire.

6.2 Application des concepts de l'ingénierie logicielle pour la reconnaissance d'émotions

eMotion en tant que système de reconnaissance d'émotions se base sur les travaux de De Meijer, utilisés comme exemple illustratif à la partie 5.7. Dans cette section, nous reprenons les principes vus aux chapitres 4 et 5. Nous présentons tout d'abord la séparation fonctionnelle en trois niveaux du système et les différents composants développés aux niveaux Capture, Analyse, et Interprétation. Nous explicitons également les différentes modalités mises en œuvre. Nous présentons ensuite une modélisation PAC (paragraphe 4.1.2, page 68) d'eMotion en tant que système interactif pour le concepteur.

Deux unités de capture sont mises en jeu dans le niveau Capture d'eMotion, permettant d'utiliser une combinaison de capture du mouvement et une paire de capteurs à six degrés de liberté. Nous considérons cinq caractéristiques issues de [46] au niveau Analyse : l'expansion du tronc, l'écartement des bras, la vélocité du mouvement et les composantes sagittales et

verticales du mouvement. La force n'a pas été prise en compte car ne nous disposions pas de capteurs permettant une telle mesure ; la directivité a été laissée de côté car étant une caractéristique temporisée elle ne permettait pas de respecter la contrainte de temps réel qui nous était imposée par notre cadre applicatif. Au niveau Interprétation, un interpréteur combine les caractéristiques pour en inférer une émotion parmi les six émotions basiques d'Ekman (voir partie 1.2.2, page 17). Un composant joue le rôle d'adaptateur, de synchronisateur, et de concentrateur des données provenant des deux dispositifs utilisés. Un dernier composant permet la synchronisation des données des différentes caractéristiques avant qu'elles ne soient interprétées.

6.2.1 Architecture en trois niveaux fonctionnels

La figure 46 présente l'architecture en trois niveaux fonctionnels de notre système. Dans eMotion, les niveaux Capture, Analyse, et Interprétation sont concrétisés par une implémentation sous forme de composants Capture, Analyse et Interprétation. Cette implémentation en "sur-composants" est un choix implémentationnel permettant de gérer chaque niveau de façon séparée. Chacun de ces sur-composants encapsule les divers composants du niveau qu'il représente. La plupart des flux entre composants peuvent être activés ou désactivés par choix du concepteur ou du système ; seul les flux activés en permanence sont représentés dans le schéma.

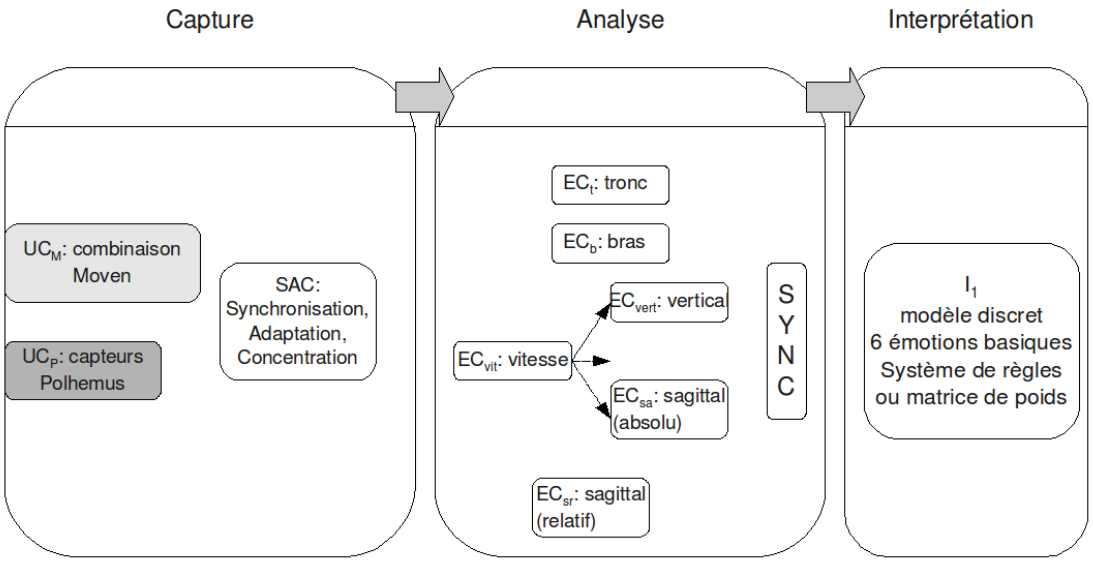
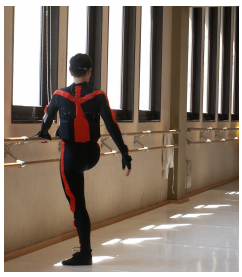


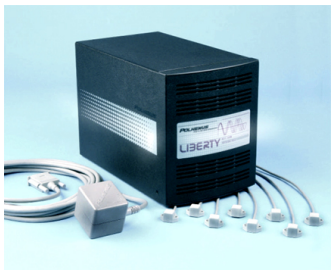
Fig. 46: Les composants d'eMotion

6.2.2 Présentation des composants et systèmes représentationnels mis en œuvre

Dans la figure 46, la première unité de capture permet d'établir une interface avec une solution commerciale de capture du mouvement : la combinaison Moven[1] (figure 47a) de l'entreprise XSens[2]. Le logiciel propriétaire Moven studio permet de recevoir des coordonnées mises en forme de 23 segments du corps et d'envoyer ces coordonnées sur le réseau. L'unité de capture UC_M permet de récupérer les données depuis le réseau, les mettre en forme, et les réinjecter dans le système. Le flux de données en sortie de UC_M est un flux structuré de blocs dont l'attribut données contient les coordonnées des segments du corps. La deuxième unité de capture fait l'interface avec le système Liberty[3] (figure 47b) de chez Polhemus[4]. L'unité de capture UC_P se présente comme une couche se plaçant au dessus du pilote du dispositif. UC_P renvoie un flux de donnée des coordonnées de chacun des capteurs. Les deux unités de capture représentent donc deux modalités : UC_M est une première modalité *<Moven, coordonnées de segments du corps>* ; UC_P est une deuxième modalité *<capteurs Polhemus, couple de coordonnées>*. Le dernier composant du composant Capture regroupe les fonctionnalités d'un synchronisateur, d'un adaptateur et d'un concentrateur. Lorsque les deux modalités sont utilisées en même temps, sont rôle est d'abord de synchroniser les données provenant de UC_M et UC_P. Les coordonnées données par le Polhemus sont ensuite adaptées au repère utilisé par la combinaison Moven. Enfin, le concentrateur remplace les coordonnées des poignets données par la combinaison par les coordonnées transformées du système Polhemus. Ainsi, on obtient en sortie de ce composant un flux de coordonnées du corps humain ; le système représentationnel est le même que celui fourni par UC_M. Les coordonnées des poignets sont par contre commandées par les Polhemus et non par la combinaison Moven : en effet les valeurs données pour les poignets proviennent alors des capteurs Polhemus (recalés sur la Moven) et non plus de la combinaison Moven.



(a) Combinaison Moven.



(b) Capteurs Polhemus Liberty.

FIG. 47: Capteurs utilisés par eMotion pour la capture de mouvement.

eMotion est composée de six extracteurs de caractéristiques. Deux d'entre eux possèdent deux méthodes de calcul de la caractéristique, selon le système représentationnel de capture

[1] http ://www.xsens.com/en/general/mvn
[2] http ://www.xsens.com/
[3] http ://www.polhemus.com/ ?page=Motion_Liberty
[4] http ://www.polhemus.com/

utilisé. En effet, le concepteur peut choisir de n'utiliser qu'un couple de coordonnées rattachées aux poignets (en utilisant les capteurs Polhemus seuls) ou les coordonnées complètes du corps (en utilisant la combinaison Moven ou les capteurs et la combinaison en parallèle). Ainsi L'extracteur EC_{vit} permet le calcul de la vitesse du mouvement soit en moyennant les vitesses des deux capteurs (utilisation des deux capteurs Polhemus seuls) ou en calculant la vitesse du bassin (utilisation de la combinaison Moven en conjonction ou non avec les capteurs Polhemus). De la même façon, l'extracteur EC_b seuille dans le premier cas la distance entre les poignets. Dans le second cas, il vérifie également la distance des poignets par rapport au tronc.

Ce choix de dispositif implique également deux composants utilisables seulement lorsqu'on dispose des coordonnées complètes du corps. L'extracteur EC_t calcule la forme du tronc en analysant la position des vertèbres par rapport au plan formé par les épaules et le bassin. L'extracteur EC_{sr} calcule la composante sagittale du mouvement de façon relative à l'orientation du tronc. Si le tronc est orienté dans le même sens que le mouvement, alors le mouvement est vers l'avant ; dans le sens inverse, vers l'arrière.

Le composant d'extraction de vitesse EC_{vit} possède deux flux de sorties. Le premier est un flux de vecteurs vitesse du mouvement. Le second est la caractéristique de vélocité du mouvement, calculée après seuillage de la norme du vecteur vitesse calculé. Les extracteurs EC_{vert} et EC_{sa} se basent sur le vecteur vitesse délivré par EC_{vit} pour connaître les composantes sagittales et verticale du mouvement. Ces tests sont effectués par simple test de signe de la composante verticale (axe des z) et sagittale (axe des x) du vecteur vitesse reçu. L'extracteur EC_{sa} délivre un flux identique à EC_{sr}. La différence entre les deux composants vient du fait que EC_{sa} calcule une composante sagittale dans un repère absolu. Par exemple, sur une scène munie d'un repère fixe avec l'axe des x orienté vers le public, EC_{sa} permettra de déterminer si le sujet de la reconnaissance se déplace vers le public ou en contraire s'en éloigne. Le composant EC_{sr} permettra de déterminer si ce sujet marche vers l'avant ou à reculons.

Le composant Interprétation ne contient qu'un interpréteur I_1. Cet interpréteur est basé sur la théorie des émotions basiques et un modèle discret des émotions (voir chapitre 1). Il identifie donc les émotions entre joie, peur, colère, tristesse, dégoût et surprise. Tout comme certains extracteurs de caractéristiques possèdent deux algorithmes différents afin de s'adapter au flux d'entrée, l'interpréteur I_1 met en jeu deux algorithmes d'interprétation différents. Le premier prend en entrée quatre caractéristiques : la vélocité (EC_{vit}), l'écartement de bras (EC_b), et les composantes verticales (EC_{vert}) et sagittale (EC_{sa}) du mouvement. Cette première interprétation est effectuée grâce à un système à base de règles. La deuxième interprétation prend en compte, en plus des caractéristiques susnommées, l'expansion du tronc (EC_t). De plus, elle permet le choix du calcul de la composante sagittale entre EC_{sa} et EC_{sr}. Cette interprétation repose sur un algorithme de calculs de poids dont les valeurs sont tirées de [46].

Chacun des composants présentés dans cette architecture se base sur des flux de données et produit un flux de données selon un système représentationnel qui lui est propre. A la liste de composants (UC_P, UC_M, EC_{vit}, EC_b, EC_t, EC_{sr}, EC_{vert}, EC_{sa}, I_1) correspond donc la liste de système représentationnels (sr^C_P, sr^C_M, sr^A_{vit}, sr^A_b, sr^A_t, sr^A_{sr}, sr^A_{vert}, sr^A_{sa}, sr^I_1).

6.2.3 Paramétrisation d'eMotion : architecture en agents PAC

Dans ce paragraphe nous abordons eMotion du point de vue d'un système interactif. En effet, le concepteur est capable de paramétrer, grâce à des *widgets* graphiques, chaque extracteur de caractéristiques du système. Le modèle PAC, initialement dédié aux systèmes interactifs, s'applique ici de façon originale : nous présentons ici l'architecture PAC du système interactif permettant de paramétrer le canevas eMotion (figure 48). Cette architecture est donc à concevoir de façon orthogonale à l'architecture fonctionnelle basée sur la branche émotion (figure 46). La solution adoptée a consisté à créer un agent PAC par composant (quel que soit son niveau fonctionnel - Capture, Analyse ou Interprétation). La facette Abstraction de l'agent est le noyau fonctionnel du composant tandis que la facette Présentation définit l'interface graphique qui permet au concepteur de paramétrer le composant. Outre ces agents feuille de la hiérarchie, nous avons créé un agent PAC par niveau fonctionnel. Ces trois agents médians sont contrôlés par un agent PAC unique, racine de la hiérarchie. Les agents médians et la racine ont pour rôle de gérer les connexions entre leurs agents fils et également de rassembler les interfaces graphiques de chacun de ces composants. Leur rôle est principalement un rôle de contrôle.

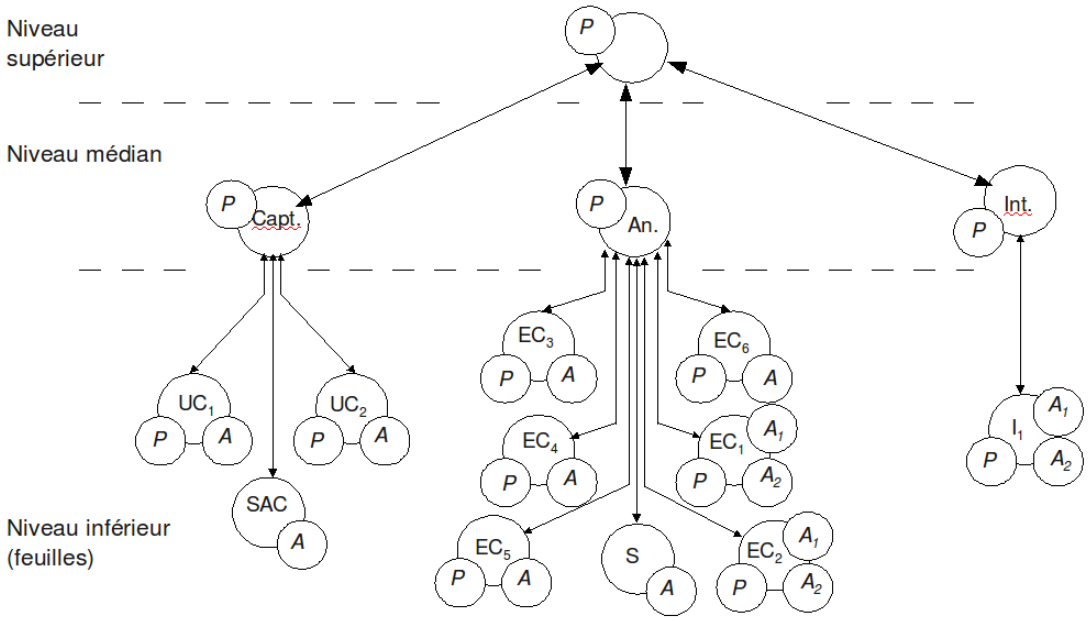


FIG. 48: Hiérarchie PAC d'eMotion, en trois niveaux. Au niveau médian sont présents les composants Capture ("capt"), Analyse ("An.") et Interprétation ("Int.").

Un composant typique possède donc une facette Présentation (*widget* graphique), une facette Abstraction et une facette Contrôle. Cependant, certains des composants présentés ne possèdent pas d'interface graphique et ne présentent donc pas de facette Présentation. De la même façon, certains composants ne servent qu'à assurer les communications dans la hiérarchie de composants et ne possèdent donc pas de facettes Abstraction.

Certains composants possèdent deux facettes Abstraction au lieu d'une : ce type d'agent est la contraction d'une hiérarchie d'un agent et de ses deux agents fils. Par exemple, eMotion laisse au concepteur et au système le choix de modalités à utiliser. Nommément, le concepteur peut choisir d'effectuer la reconnaissance sur le couple de coordonnées des poignets ou sur les coordonnées de tout le corps. Au niveau Analyse, ce changement se traduit par des traitements différents pour le calcul d'une caractéristique. Typiquement, une telle caractéristique devrait donc être extraite par deux composants différents et donc deux agents PAC feuilles (un pour chaque traitement) (figure 49). La facette Contrôle de la racine ($C0$) a alors pour rôle d'envoyer les données à l'agent feuille sachant traiter le système représentationnel utilisé. La contraction de cette hiérarchie en un unique agent PAC doté de 2 abstractions permet ainsi de factoriser les facettes présentation et de n'utiliser qu'un seul contrôle.

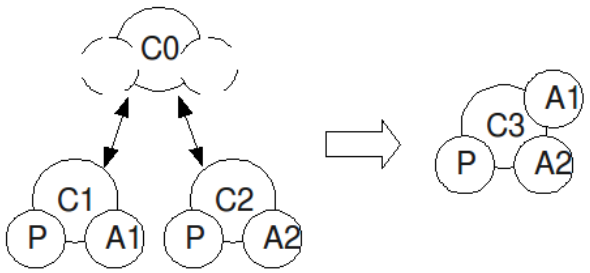


Fig. 49: Contraction d'une hiérarchie en un agent à deux abstractions.

La contraction d'une hiérarchie d'agents PAC nécessite que le changement de facette Abstraction soit transparent pour le reste du système. Cela implique donc que les agents feuilles concentrés produisent le même flux de sortie. Par exemple, le calcul de la vitesse présente dans tous les cas la même interface, et propose dans tous les cas les mêmes flux de sortie. La présence de deux facettes Abstraction permet d'alterner et de calculer la vitesse du bassin lorsque le système représentationnel "coordonnées du corps" est utilisé, et la vitesse moyenne des deux capteurs quand le système représentationnel "couple de coordonnées" est utilisé.

Les différents composants sont organisés dans une hiérarchie à trois niveaux. Au niveau le plus bas, chaque composant est indépendant et gère sa propre interface et ses propres communications internes. Au niveau médian, les composants Capture, Analyse et Interprétation manipulent chacun leurs composants fils. Enfin, un dernier composant au niveau supérieur agrège les différents *widgets* graphiques pour former une interface graphique complète, et gère les communications entre les composants Capture, Analyse, et Interprétation. Dans eMotion, les composants feuilles de la hiérarchie contiennent entièrement le code fonctionnel du système. Les composants des niveaux médian et supérieur n'ont que des rôles de gestion des communications et d'assemblage d'interface et ne disposent donc pas de facette Abstraction. De même, les composants SAC et S ne permettent pas une interaction graphique et ne présentent donc pas de facette Présentation.

Cette architecture en agents PAC permet à chaque composant de gérer sa structure interne

en agissant sur ses agents fils, et de ne prévenir le reste du système que lorsque cela est nécessaire. Le paragraphe suivant et la section 6.4 explicitent la communication entre composants lorsqu'un tel cas survient.

6.2.4 Résumé global de l'architecture : communication entre composants

Chaque composant d'eMotion est donc à la fois un composant de la branche émotion tel que défini au chapitre précédent et un agent PAC élément d'une hiérarchie. Cette hiérarchie se concrétise dans l'implémentation par l'encapsulation des composants par leur père. La figure 50 illustre l'architecture finale d'eMotion (les facettes Présentation ont été omises afin de clarifier la figure).

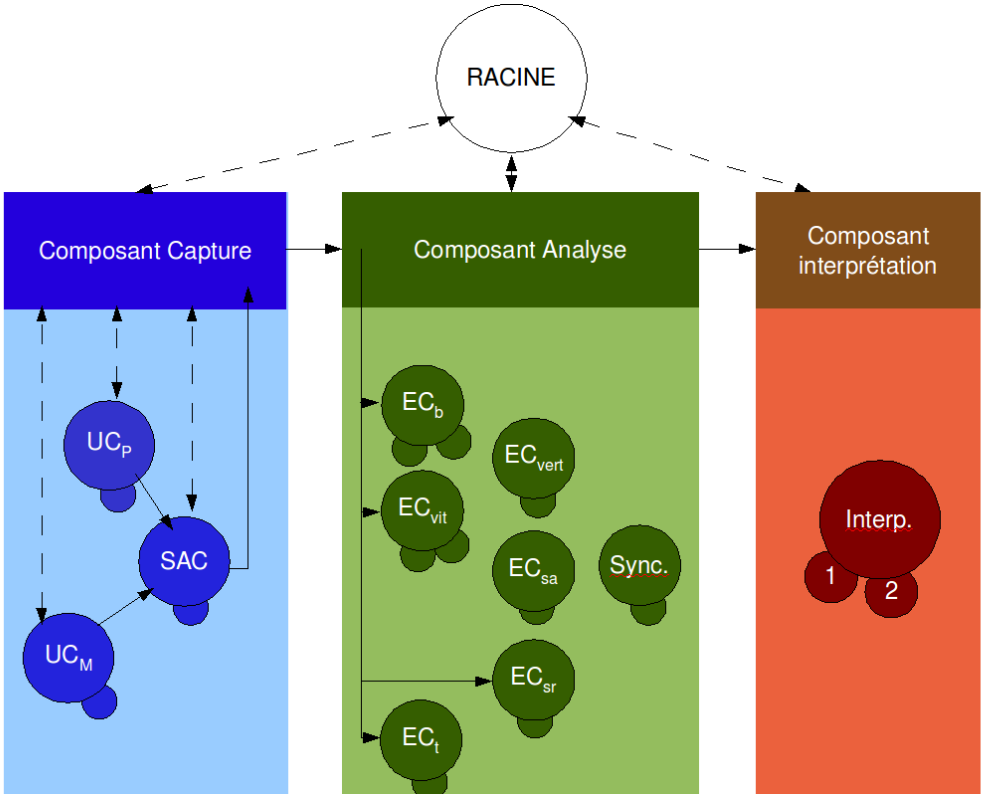


FIG. 50: Résumé de l'architecture d'eMotion. Illustration des communications de données (flèches continues) et de contrôle (flèches interrompues) dans le composant Capture.

Il existe deux types de communication dans eMotion. Les communications de données regroupent les différents flux de données émis par les composants. Ces flux sont illustrés dans la figure 50 pour les cas du composant Capture : les différents composants du niveau Capture produisent des données. Le flux final de l'assemblage au niveau Capture est envoyé à la facette Contrôle du composant Capture. Le flux est ensuite envoyé à la facette contrôle du composant Analyse, qui redistribue ce flux aux divers extracteurs de caractéristiques qu'il encapsule. Les communications de contrôle sont exclusivement verticales et permettent à un agent père de

gérer les connexions entre ses fils. Un exemple de cette gestion de la communication est le choix du dispositif à utiliser par le concepteur. Si la combinaison seule est activée, la facette Contrôle du composant Capture active l'unité de capture correspondant à la combinaison, et connecte son flux de sortie à son propre flux de sortie afin de l'envoyer au composant Analyse. En même temps, le composant Capture signale à l'agent racine le système représentationnel utilisé en sortie du niveau Capture. Le composant racine transmet l'information au niveau Analyse. Si le concepteur choisit alors d'utiliser la paire de capteurs Polhemus seule, la facette contrôle du composant Capture désactive l'unité de capture de la combinaison, déconnecte son flux de sortie de sa propre sortie, active l'unité de capture des capteurs Polhemus et connecte le flux de couples de coordonnées à sa propre sortie. Encore une fois, le composant racine est prévenu du changement de système représentationnel ; l'information est transmise au composant Analyse.

6.2.5 Interface utilisateur

Le système eMotion propose une interface graphique (voir figure 51) permettant de surveiller l'évolution des caractéristiques et des émotions reconnues. Elle est développée en C++ à l'aide de la librairie Qt[5], permettant un développement multiplateformes. Cette interface graphique permet également au concepteur de choisir le dispositif à utiliser ainsi que le calcul de la composante sagittale du mouvement.

A chaque composant correspond un *widget* graphique permettant au concepteur d'en ajuster les paramètres. Le *widget* graphique du composant EC_b, calculant l'expansion des bras, permet ainsi d'évaluer la distance entre les deux poignets et de voir la valeur de sortie choisie par le composant au temps courant. Deux curseurs glissants permettent d'ajuster les seuils minimaux et maximaux pour considérer les bras respectivement ouverts ou fermés. De même, les *widgets* graphiques des composants EC_{vit}, EC_{sa} et EC_{vert} illustrent les variations des valeurs sur lesquelles sont effectuées les seuillages et proposent des curseurs glissants permettant de modifier les seuils permettant l'attribution des valeurs de sortie. Les valeurs de sortie sont également affichées. Le *widget* graphique du composant EC_t affiche la valeur de sortie de ce composant. Le *widget* du composant Interprétation affiche quant à lui l'émotion reconnue.

Les composants Capture et Analyse disposent également d'une interface graphique permettant au concepteur d'influer sur le comportement du système. L'interface du composant Capture permet au concepteur de choisir son dispositif d'entrée : combinaison Moven, capteur Polhemus, ou la combinaison des deux. L'interface du composant Analyse permet au concepteur de choisir le calcul de la composante sagittale du mouvement : en absolu (composante selon X) ou en relatif (par rapport à la normale au plan formé par le tronc et le bassin).

[5]http ://qt.nokia.com/

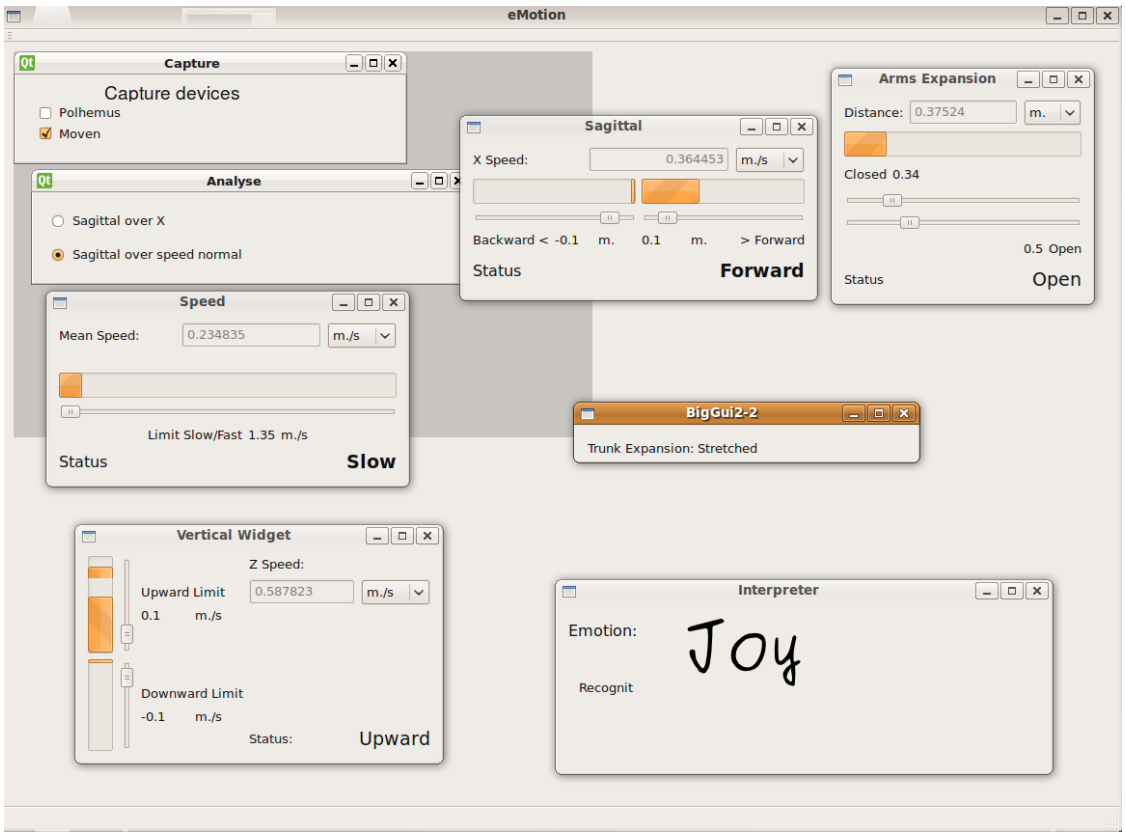


FIG. 51: Capture écran d'eMotion.

6.3 Analyse de l'architecture d'eMotion : extensibilité et modifiabilité

Le canevas logiciel eMotion permet l'ajout et l'échange de composants. Ces manipulations vont induire des modifications du reste du système. Nous identifions deux cas, liés aux modalités définies par les composants considérés. Nous examinons tout d'abord l'ajout d'un nouveau composant dans le système.

Le premier cas est celui où la modalité définie par le composant à ajouter est totalement équivalente à au moins une modalité de même niveau dans le système, c'est-à-dire qu'il existe au moins un composant de même niveau Capture, Analyse ou Interprétation dont le système représentationnel de sortie soit équivalent à celui du composant à ajouter. Les modifications requises sont alors inscrites au niveau fonctionnel considéré et à son composant correspondant. Dans eMotion par exemple, nous pourrions ajouter un système de marqueurs optiques sur la combinaison que nous utilisons actuellement. Ces marqueurs optiques fournissent également les coordonnées du corps. Les systèmes représentationnels de la combinaison et des marqueurs optiques sont équivalents. Il appartient alors au concepteur de décider de l'utilisation du nouveau dispositif. Etant équivalents, ces dispositifs sont interchangeables : le concepteur peut préférer utiliser la combinaison dans de mauvaises conditions d'éclairage.

Ces dispositifs peuvent être également mis en redondance, afin de pallier aux défauts des deux systèmes. Dans ces deux cas, il est nécessaire de rajouter un composant concentrateur pour gérer l'équivalence/redondance, et de modifier le composant Capture afin qu'il prenne en compte la nouvelle unité de capture et le concentrateur, et les nouvelles connexions dues à ces nouveaux composants. Le reste du système n'est pas impacté. Si un extracteur de caractéristiques est ajouté, alors seul le composant Analyse est impacté. Si un interpréteur est ajouté, seul le composant Interprétation est impacté.

Le deuxième cas est celui où il n'existe pas de modalité de capture (respectivement d'analyse, interprétation) équivalente à celle définie par la nouvelle unité de capture (respectivement extracteur de caractéristiques, interpréteur), ou lorsque le concepteur décide de ne pas concentrer deux modalités équivalentes. Dans ce cas, le nouveau flux de données est agrégé aux autres en sortie du composant Capture (respectivement Analyse, Interprétation). Le nouvel agrégat de flux ainsi défini doit pouvoir être pris en charge par le composant Analyse (respectivement le composant Interprétation, le système interactif) qui est également modifié. L'agent racine est alors modifié pour assurer la liaison entre les composants médians (Capture, Analyse, Interprétation). Par exemple, si nous ajoutons une caméra vidéo à eMotion, le composant Capture sera modifié afin d'envoyé le flux vidéo au composant Analyse ; le composant Analyse sera modifié pour envoyer ce flux vidéo aux extracteurs travaillant sur la vidéo. Comme il n'en existe pas dans la version actuelle, il est nécessaire d'ajouter des extracteurs de caractéristiques travaillant sur le flux vidéo. Si nous ajoutons un calcul de l'expansion des bras, équivalent à l'extracteur déjà présent travaillant sur des coordonnées, alors il est possible de concentrer les flux d'expansion des bras (obtenus séparément par la vidéo et par les coordonnées) : le composant Interprétation n'est pas impacté. Si nous ajoutons une nouvelle caractéristique comme l'Index de Contraction (voir paragraphe 3.4.2, page 57), alors le flux de sortie est modifié : le composant Interprétation est impacté. Un nouvel interpréteur est alors nécessaire pour prendre en compte l'index de contraction.

La présence d'un niveau médian de composants Capture, Analyse et Interprétation ne limite pas la modifiabilité de notre système. Elle permet de structurer les modifications à apporter en une hiérarchie. En effet, les modifications apportées à ces composants médians ne concernent que l'agencement des connexions entre composants feuilles. Ce niveau médian est un choix implémentationnel. Une architecture PAC ne comportant qu'un agent racine et des agents feuilles est possible : la gestion de toutes les connexions serait alors déportée vers l'agent racine. L'architecture PAC du système interactif et la hiérarchie en racine, composants médians et feuille permet également de structurer l'interaction. De ce point de vue, ajouter un composant revient à ajouter un agent PAC, rattaché à un composant médian.

eMotion achève donc la validation de notre modèle en l'instanciant à un cas réel. Nous avons présenté dans cette section l'application des concepts architecturaux vus aux chapitres 4 et 5. Une des principales contributions de ce système est que son architecture - selon la branche émotion pour le côté fonctionnel, couplé à l'architecture PAC pour le contrôle des connexions - offre un canevas logiciel permettant de créer d'autres systèmes de reconnaissance d'émotions. En effet, dans son architecture, eMotion ne limite pas la reconnaissance à la gestuelle. Son

implémentation actuelle, ne gérant que le mouvement, est le résultat d'un choix implémentationnel lié à notre cas applicatif présenté au chapitre suivant. Des extensions d'eMotion pourraient être implémentées afin de permettre une reconnaissance par les expressions faciales, la voix, ou les réactions du système nerveux autonome. Contrairement à un outil comme ICARE ou EyesWeb, eMotion ne permet pas à un concepteur d'assembler directement des composants entre eux mais offre un squelette d'interface qu'un programmeur peut reprendre et étendre.

Dans la prochaine section, nous mettons l'accent sur la composition de modalités et illustrons en particulier le cas de l'équivalence. Cette composition permet au concepteur ou au système d'effectuer des choix de la modalité à utiliser et n'est pas présente dans les systèmes multimodaux existants de reconnaissance d'émotions.

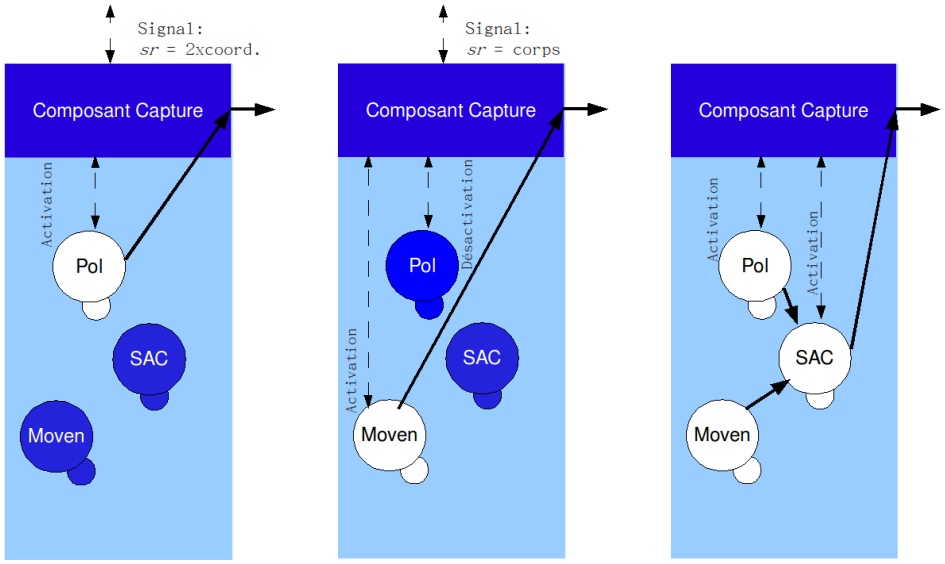
6.4 Illustration des propriétés CARE pour la reconnaissance d'émotions : cas de l'équivalence

Dans cette section, nous reprenons les propriétés CARE appliquées à la reconnaissance d'émotions et illustrons leur pouvoir génératif vu au paragraphe 4.4.3, page 87. La figure 25 (page 88) montre en particulier que l'équivalence des modalités, impliquant un choix de la part du système ou du concepteur, ne se retrouve pas dans les systèmes existants de reconnaissance d'émotions. eMotion permet de tels choix en proposant plusieurs modalités équivalentes dans le système. Nous présentons également un exemple de complémentarité des dispositifs.

Grâce à la hiérarchie du système en agents PAC et aux communications de contrôle entre facettes Contrôle des différents agents, chaque composant est capable de contrôler et réorganiser les connexions de ses agents fils. Les composants peuvent être ainsi activés ou désactivés et les flux de données redirigés par le composant père. Dans cette section nous présentons trois choix possibles pour le système. Le premier est le choix du dispositif par le concepteur. Le second est un choix système, directement dépendant du choix du dispositif par le concepteur : le système réorganise son fonctionnement interne en fonction du dispositif utilisé. Enfin, le dernier est un choix du concepteur au niveau Analyse. La figure 52 illustre la séquence de choix du dispositif sur laquelle nous nous appuyons pour décrire les changements internes au composant Capture.

6.4.1 Choix du dispositif par le concepteur

Au lancement du système, le concepteur se voit proposer le choix *via* un *widget* graphique du dispositif à utiliser : la combinaison de capture du mouvement Moven, les capteurs Polhemus, ou une combinaison des deux dispositifs, où les capteurs Polhemus sont affectés aux poignets du sujet et remplacent les mesures donnée par la combinaison. Ce *widget* graphique est la facette Présentation du composant Capture. La facette Contrôle de ce composant Capture est donc directement prévenue de la sélection ou désélection de l'un des dispositifs.



(a) Choix des capteurs Polhemus seuls.

(b) Déconnexion des capteurs Polhemus et choix de la combinaison Moven.

(c) Rétablissement des capteurs Polhemus pour une combinaison de capteurs.

Fig. 52: Séquence de changement de dispositif par le concepteur : impact sur le composant Capture.

Capteurs Polhemus seuls

A l'initialisation, les deux unités de capture sont désactivées et aucune connexion de flux de données n'est établie. Lorsque le concepteur sélectionne l'utilisation des capteurs Polhemus (figure 52a) en cochant la case correspondante, le composant Capture active tout d'abord l'unité de capture correspondant aux capteurs. Le flux de sortie de l'unité de capture est connecté au flux de sortie du composant Capture. Ce dernier signale alors à l'agent racine que le système représentationnel utilisé est celui d'un couple de coordonnées des poignets du sujet.

Si le concepteur désélectionne les capteurs Polhemus, le composant Capture désactive à nouveaux toutes les unités de capture et déconnecte tous les flux de données afin de restaurer la configuration initiale. Il signale également à l'agent racine qu'aucun système représentationnel n'est utilisé.

Combinaison Moven seule

Lorsque le concepteur sélectionne l'utilisation de la combinaison Moven (après être revenu à l'état initial du système en décochant l'utilisation des capteurs Polhemus), le composant Capture agit de façon similaire au paragraphe précédent (figure 52b) : l'unité de capture correspondant à la combinaison est activée, son flux de sortie connecté à la sortie du composant Capture, et un signal est envoyé à l'agent racine indiquant que le système représentationnel

utilisé colporte les coordonnées du corps complet.

Utilisation simultanée : complémentarité des dispositifs

Lorsque le concepteur, après avoir sélectionné l'utilisation de la combinaison Moven, sélectionne également l'utilisation des capteurs Polhemus, la facette Contrôle du composant Capture entame le processus suivant, illustré à la figure 52c :

1. Le flux issu de la combinaison est déconnecté du flux de sortie du composant Capture.
2. L'unité de capture correspondant aux capteurs Polhemus est activée.
3. Les flux des deux unités de capture sont ensuite connectés au composant SAC (Synchronisation-Adaptation-Concentration). Ce composant synchronise tout d'abord les deux flux, traduit ensuite les coordonnées des capteurs Polhemus dans le repère de la combinaison Moven, et enfin remplace la coordonnées des poignets données par la combinaison par celles données par les capteurs. Le flux de sortie est donc un flux de coordonnées du corps complet.
4. Le flux de sortie du composant SAC est connecté à la sortie du composant Capture.

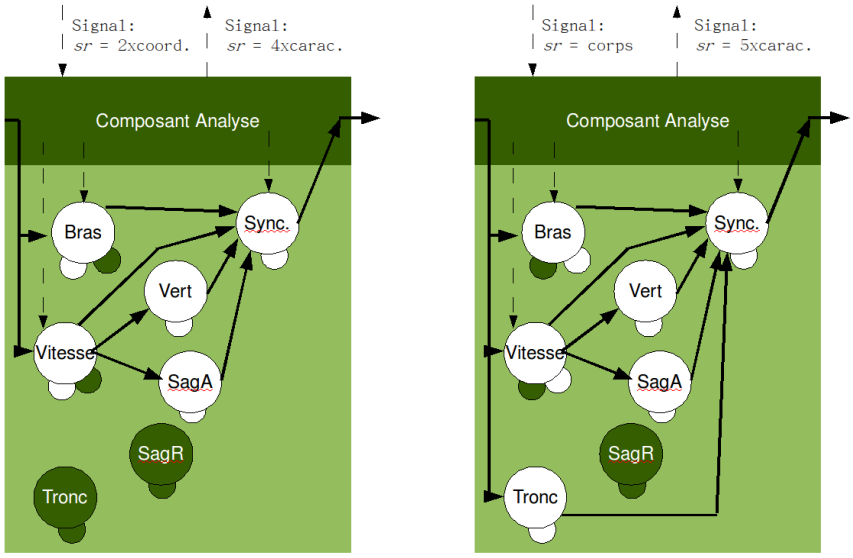
Contrairement aux paragraphes précédents, l'agent racine n'est ici pas notifié d'un changement de système représentationnel ; en effet, en connectant d'abord la Moven puis en rajoutant les capteurs Polhemus, il n'y a pas de changement. Le composant Capture peut ainsi procéder à sa réorganisation interne sans en avertir son père.

Dans la version actuelle d'eMotion, l'utilisation simultanée de la combinaison et des capteurs Polhemus permet au concepteur de contrôler les poignets du sujet portant la combinaison, comme s'il tenait ses mains en face à face. Cette possibilité illustre la possibilité d'intégration de composants spécifiques à une expérimentation de type "Magicien d'Oz" : le sujet en combinaison se meut de façon naturelle, tandis qu'un compère utilise les capteurs Polhemus pour obtenir des valeurs franches de l'écartement des bras (bras complètement écartés ou complètement repliés). D'un point de vue artistique et considérant notre travail avec les danseurs, l'utilisation simultanée des capteurs Polhemus et de la combinaison Moven permet à une personne tenant les Polhemus de "mener" les bras du danseur.

Enfin, le choix du dispositif par le concepteur n'est perçu que par le composant Capture. Le changement de dispositif est abstrait en un changement de système représentationnel de sortie. Ainsi, des trois modes de capture possibles (capteurs Polhemus, combinaison de capture du mouvement, ou combinaison des deux dispositifs), seuls deux systèmes représentationnels sont utilisés : le flux de couple de coordonnées (capteurs Polhemus) et le flux des coordonnées du corps complet (combinaison Moven et combinaison des capteurs Polhemus et de la combinaison Moven). Le reste du système n'est averti que du changement de système représentationnel. Ce changement de système représentationnel est considéré comme un changement de contexte d'interaction, qui amène le système à faire des choix explicités dans la partie suivante.

6.4.2 Choix de caractéristiques par le système selon la modalité de capture choisie

Le changement de système représentationnel au niveau Capture implique des choix d'organisation des composants dans les niveaux Capture et Analyse. A l'annonce d'un changement de système représentationnel de capture par l'agent racine, les composants Analyse et Interprétation modifient leur organisation interne pour s'adapter au nouveau contexte d'interaction. La figure 53 illustre les deux cas de changement de système représentationnel et l'impact sur le composant Analyse.



(a) Réorganisation du composant Analyse dans le cas d'un *sr* <couple de coordonnées>.

(b) Réorganisation du composant Analyse dans le cas d'un *sr* <coordonnées du corps>.

Fig. 53: Réorganisation du composant Analyse après changement de système représentationnel de capture.

Système représentationnel <*couple de coordonnées*>

Lorsque le concepteur choisit de n'utiliser que les capteurs Polhemus, l'agent racine est informé que le système représentationnel utilisé est un couple de coordonnées des poignets (figure 53a). Il signale alors au composant Analyse ce changement. Ce dernier va procéder à une réorganisation interne en suivant le processus suivant :

1. Le composant Analyse active les extracteurs de caractéristique "bras" et "vitesse" et leur signale le système représentationnel utilisé. Chacun de ces composants, grâce à cette information, active alors la facette Abstraction correspondante. Au niveau de l'implémentation, le composant "bras" calcule l'expansion des bras en seuillant la distance entre les deux capteurs ; le composant "vitesse" calcule la vitesse moyenne des deux capteurs.
2. Le composant Analyse active les composants "Vert." et "SagA". Il connecte leurs entrées

à la sortie du composant "vitesse" délivrant le vecteur vitesse courant. Le composant "Vert" ne considère que la composante en z, le composant "SagA" la composante en x. La composante sagittale est ici forcément calculée dans le repère absolu.

3. Le composant de synchronisation est activé. Les sorties de chaque extracteur y sont connectées. La sortie du composant de synchronisation est connectée à la sortie du composant Analyse.
4. Le composant Analyse signale à l'agent racine que le système représentationnel utilisé en sortie du composant Analyse est l'agrégat des quatre caractéristiques activées.

De la même façon, le composant Interprétation est notifié du système représentation en sortie du composant Analyse. L'interpréteur est activé et il active sa facette Abstraction lui permettant d'interpréter ces quatre caractéristiques.

Système représentationnel <*coordonnées du corps*>

Lorsque le concepteur choisit d'utiliser la combinaison en plus des capteurs, l'agent racine est alors informé du changement vers un système représentationnel <*coordonnées du corps*> (figure 53b). Le message est relayé au composant Analyse qui entame le processus suivant :

1. Le composant Analyse signale aux composants "Bras" et "vitesse" du changement de système représentationnel. Ceux-ci réorganisent également leur agencement interne et changent de facettes Abstraction. Le composant "Bras" vérifie maintenant la distance entre poignets et tronc en plus de la distance entre poignets. Le composant "vitesse" calcule le vecteur vitesse du bassin. Les composants "Vert." et "SagA", ne sont pas notifiés ; en effet, de leur point de vue, aucun changement ne survient.
2. Le composant Analyse active le composant "Tronc" de calcul de l'expansion du tronc et connecte sa sortie au composant de synchronisation.
3. Le composant de synchronisation est notifié qu'il doit maintenant synchroniser cinq caractéristiques.
4. Le composant Analyse notifie l'agent racine que son flux de sortie est désormais un agrégat de cinq caractéristiques.

Le composant Interprétation est ensuite notifié du changement des caractéristiques extraites. L'interpréteur change alors de facette Abstraction pour prendre en compte ces cinq caractéristiques.

Le changement du dispositif choisi par le concepteur ne déclenche donc pas forcément une modification du système. Seul lorsque le système représentationnel de capture est changé peut on voir une réorganisation des composants au sein du système. Cette réorganisation est automatique et effectué par le système ; elle correspond à des choix du système lui permettant de s'adapter à deux contextes d'interaction.

6.4.3 Choix de composants équivalents par le concepteur

Le dernier choix possible est à effectuer par le concepteur : il s'agit du choix du calcul de la composante sagittale du mouvement. Deux composants permettent ce calcul dans notre système. Le composant "SagA" ("Sagittal Absolu") calcule la composante sagittale du mouvement dans un repère fixe. Dans le cadre d'un spectacle de danse par exemple, cette caractéristique permet de savoir si le danseur se déplace vers l'avant de la scène (donc vers le public) ou vers l'arrière. Le composant "SagR" (pour "Sagittal Relatif") calcule le déplacement sagittal par rapport à l'orientation du tronc et est donc relatif au sujet. Cette caractéristique traduit un déplacement en avançant ou en reculant. Ce dernier composant nécessite les informations concernant le tronc du sujet et n'est donc disponible que lorsque le système représentationnel des coordonnées du corps est utilisé.

Le concepteur peut sélectionner l'un ou l'autre de ces composants grâce à un *widget* graphique, qui est la facette Présentation du composant Analyse. Initialement, le composant "SagA" est actif. Le choix du calcul d'une composante sagittale relative par le concepteur est directement signifié au composant Analyse. Celui-ci déconnecte alors les flux d'entrée et de sortie du composant "SagA" et connecte l'entrée du composant "SagR" à son propre flux d'entrée. Il connecte ensuite le flux de sortie du composant "SagR" à son propre flux de sortie.

Cette réorganisation est interne au composant Analyse et ne modifie pas son flux de sortie. Le reste du système n'est donc pas notifié de cette réorganisation.

6.5 Expérimentation

Dans cette expérimentation, la capture de mouvement se fait par la combinaison Moven. Au niveau Analyse, l'utilisation de la combinaison permet de calculer la caractéristique d'expansion du tronc, qui joue un rôle principal dans la détermination de l'émotion exprimée. Enfin, nous choisissons de calculer la direction sagittale relativement au mouvement du danseur : un mouvement est considéré vers l'avant si le tronc est orienté dans le sens du mouvement, vers l'arrière si le tronc est orienté dans le sens opposé au mouvement. Cette caractéristique a été choisie car *a priori* plus proche de l'intention de De Meijer dans sa description de la caractéristique sagittale dans [46].

Le but de l'expérimentation était de valider eMotion en tant que système de reconnaissance d'émotions en utilisant la configuration décrite ci-dessus. Cette validation nous était nécessaire pour utiliser eMotion dans notre cas d'application décrit au chapitre suivant : utiliser les émotions reconnues pour paramétrer les éléments virtuels dans le cadre de l'augmentation d'un spectacle de danse.

6.5.1 Déroulement de l'expérimentation

Dans cette expérimentation, nous avons procédé à l'enregistrement de séquences dansées expressives grâce à la combinaison Moven et l'évaluation de ces séquences par un groupes

d'évaluateurs humains. Nous avons donc mis au point un protocole expérimental permettant de remplir ces deux objectifs. Cette expérimentation a été réalisée dans le cadre du projet CARE[6], financé par l'Agence Nationale de la Recherche.

Cadre de l'expérimentation

Le sujet enregistré était le danseur et chorégraphe professionnel Gaël Domenger, affilié au Ballet Biarritz, spécialisé dans l'improvisation. Les préconisations sur l'évaluation d'un système de reconnaissance vues à la partie 2.5 (page 46) ne s'appliquaient pas dans notre cas : en effet, notre but était d'effectuer une reconnaissance sur des séquences dansées et non sur des expressions spontanées de l'émotion. Le matériel utilisé a été la combinaison Moven et le logiciel Moven studio. Une caméra FireWire a également été utilisée afin de permettre une analyse vidéo du mouvement par le laboratoire GIPSA[7], produisant ainsi, pour chaque séquence dansée, la séquence vidéo et la séquence de coordonnées à des fins de comparaison. Chaque enregistrement commençait par un *clap* des mains afin de synchroniser coordonnées et vidéo.

L'expérimentation s'est déroulée dans les locaux du Ballet Biarritz. La salle fournie présentait l'avantage d'une bonne luminosité et d'un fond uni, contrastant fortement avec la combinaison, assurant ainsi de bonnes conditions pour une analyse vidéo.

Le groupe d'évaluateurs était limité à sept personnes, dont cinq provenaient du monde la recherche (doctorant ou enseignants chercheurs). Une personne était liée au domaine de la culture, et une autre était étudiant en informatique. Etant donné le faible nombre d'évaluateurs, nous n'avons pas pris en compte la distribution selon le sexe et l'âge.

Enregistrement des séquences dansées

L'expérimentation a été découpée en quatre sessions d'enregistrement/évaluation. La première partie visait à tester les six émotions d'Ekman où la surprise et la colère étaient divisées respectivement en surprise positive et négative et en colère chaude et froide. Les huit émotions résultantes (joie, dégoût, tristesse, peur, colère chaude, colère froide, surprise positive, surprise négative) ont été exprimées trois fois chacune. Le danseur avait donc pour rôle de jouer 24 séquences dans un ordre tiré aléatoirement. Les consignes spécifiaient une position de départ droite, bras le long du corps. Un *clap* devait suivre ; puis le danseur improvisait une séquence d'une minute environ en faisant monter progressivement l'intensité de l'émotion exprimée. Dans cette partie, le danseur était directement informé de l'émotion à jouer par le label correspondant ("la joie", "la colère",... etc.). Dans cette session, une pause de quelques minutes a été introduite après la quatorzième séquence. Un problème technique a entraîné une deuxième pause à la vingtième séquence.

La deuxième partie présentait un protocole similaire ; le danseur ne devait ici pas exprimer des émotions mais les tempéraments extrêmes situés aux extrémités du cube PAD [95]

[6]Cultural experience : Augmented Reality and Emotion, http ://www.careproject.fr/
[7]http ://www.gipsa-lab.inpg.fr/

(voir partie 1.3.2, page 24). Le danseur devait donc tour à tour exprimer un tempérament exubérant, ennuyé, dépendant, dédaigneux, relâché, anxieux, docile et hostile. La troisième session étendait le protocole aux relations interpersonnelles : ami avec quelqu'un, ennemi avec quelqu'un, dominant et soumis. Enfin, la dernière session avait pour but de tester l'expression de mélange d'émotions. Pour cette session, nous avons écarté les variations de colère et de surprise pour ne tester que les émotions primaires d'Ekman : la joie, la colère, la surprise, la tristesse, la peur et le dégoût. Contrairement aux autres sessions, l'approche utilisée ici a été une approche scénario. Au lieu de signifier au danseur l'émotion à exprimer, un scénario court lui était dit, devant lui permettre de mieux appréhender l'improvisation. Sept mélanges ont été ainsi enregistrés par les scénarii suivants :

- Peur + colère : "Quelqu'un te menace et tu as à la fois peur d'une attaque et en colère à cause de la provocation."
- Tristesse + peur : "Tu as perdu ta maison dans un désastre et apprends qu'il y a toujours du danger."
- Dégoût + surprise : "Tu es dégoûté par quelque chose d'inattendu."
- Tristesse +colère : "Un automobiliste a écrasé ton chien."
- Joie + tristesse : "Tu te sens nostalgique d'une expérience douce-amère."
- Dégoût + colère : "Tu dis à quelqu'un : "Comment oses-tu me montrer quelque chose d'aussi dégoûtant ?""
- Joie + colère : "Ton enfant a fait une amusante bêtise et tu veux montrer ta colère alors que tu as envie de rire."

Evaluation des séquences dansées

Les évaluations de chaque séquence expressive ont été faites sur le principe du choix forcé. Les observateurs se sont vu remettre, avant chaque session, un feuillet comportant un tableau constitué du numéro de la séquence vue sur les lignes et des états affectifs à choisir en colonnes. Les observateurs avaient pour consigne d'identifier de une à trois émotions dans une séquence. Dans aucune session les observateurs n'étaient prévenus si l'expression dansée se faisait sur des états affectifs uniques ou sur des mélanges. Les observateurs pouvaient donc observer une émotion unique, un mélange d'émotions ou une succession d'expressions dans la même séquence. Pour annoter cela, il a été demandé aux observateurs de noter d'un "1" la première émotion reconnue, "2" la deuxième, et "3" la troisième. Des émotions mélangées étaient annotées d'un même numéro.

L'expérimentation décrite dans cette partie nous a permis de rassembler des évaluations humaines, en direct, aux, mêmes séquences dansées que pour le test d'eMotion, afin de servir de point de référence. Nous avons de plus constitué un corpus de séquences dansées expressives de divers état affectifs en utilisant deux mesures : l'enregistrement de coordonnées du corps par la Moven et des séquences vidéo de chaque séquence dansées. La présence d'un *clap* en début et en fin de séquence permet de synchroniser, pour une même séquence, enregistrement vidéo et coordonnées. Ce corpus ouvre certaines perspectives et peut trouver diverses utilités, même hors du domaine de la reconnaissance d'émotions ; par exemple, pour un système de suivi 3D du corps humain, les enregistrements de coordonnées pourraient permettre de quantifier facilement la précision du suivi basé vidéo.

Emotion	Pourcentage de reconnaissance
joie	71,43%
dégoût	47,62%
tristesse	61,90%
peur	28,57%
colère chaude	85,71%
colère froide	85,71%
surprise positive	19,05%
surprise négative	9,52%

TAB. 5: Pourcentages de reconnaissance des séquences dansées expressives par les observateurs humains.

6.5.2 Résultats

Les feuillets de chaque participant ont été relevés après chaque session. Nous avons calculé une moyenne de reconnaissance en prenant en compte les différentes possibilités. Ainsi, la reconnaissance exacte de l'intention d'expression était déterminée comme une reconnaissance parfaite ; une bonne réponse au milieu d'autres était pondérée pour affaiblir son impact sur le pourcentage de reconnaissance. Par exemple, pour une séquence exprimant la joie, une évaluation faisant apparaître la joie en deuxième position après une autre émotion, ou en mélange avec une autre émotion, était pondérée par un facteur 0.5. Dans ce mémoire, nous ne relatons que les résultats de la première session, qui ont été utilisés comme références pour valider eMotion. Le tableau 5 montre ces résultats selon les huit émotions testées.

Dans ces résultats d'évaluation par des observateurs humains, nous constatons qu'il existe des disparités entre émotions quant à leurs taux de reconnaissance. En particulier, la surprise négative présente un pourcentage de reconnaissance inférieur à 10% ; un choix d'une émotion au hasard dans la liste présenterait une reconnaissance de 12,5%.

Nous avons fourni à eMotion les mêmes séquences, enregistrées grâce à la combinaison de capture du mouvement. eMotion effectuant une reconnaissance à chaque mesure, les pourcentages donnés représentent le nombre de mesures sur lesquelles la bonne émotion a été reconnue par rapport au nombre total de mesures dans une séquence. De plus, eMotion ne reconnaissant que les six émotions de base d'Ekman, nous avons considéré comme bonnes réponses la surprise pour les séquences exprimant la surprise positive et négative, et la colère pour les séquences exprimant la colère froide ou chaude. La table 6 reporte les résultats obtenus par eMotion sur les huit émotions exprimées par le danseur, en regard des évaluations humaines.

Analyse des résultats

La tristesse et la joie exceptées, le pourcentage de reconnaissance présente des résultats faibles. La surprise positive et négative et cependant mieux reconnue par le système que par les humains. L'absence de résultats, dans le cadre de cette expérience, peut s'expliquer par les fenêtres temporelles utilisées par les humains et par le système, et par le jeu du danseur. Un

Emotion	Pourcentage de reconnaissance par eMotion	Pourcentage de reconnaissance par l'humain
joie	51.26%	71,43%
dégoût	0.7% (tristesse, 43.2%)	47,62%
tristesse	55.95%	61,90%
peur	7.95% (tristesse 51.73%)	28,57%
colère chaude	4.22% (tristesse 36.69%)	85,71%
colère froide	0.6% (tristesse 43.86%)	85,71%
surprise positive	20.54% (joie 47.52%)	19,05%
surprise négative	32.4% (joie 39%)	9,52%

Tab. 6: Pourcentages de reconnaissance des séquences dansées expressives par eMotion (entre parenthèses, 'émotion la mieux reconnue et son pourcentage) en regard des observateurs humains.

humain juge l'expression émotionnelle d'une séquence sur l'ensemble de celle-ci, soit sur une durée d'environ une minute. eMotion ne considère qu'un laps de temps très court, de l'ordre du dixième de seconde, durant laquelle elle analyse le mouvement et infère une émotion. Pour une émotion comme le dégoût, le danseur exécute quelques mouvements évoquant des spasmes. Ces mouvements sont typiques et durant ces mouvements, eMotion reconnait effectivement le dégoût en majorité. La durée totale de leur exécution ne constitue par contre qu'une petite fraction de la durée totale d'une séquence. Pendant le reste du temps, le danseur marche ou effectue des mouvements non reconnus comme le dégoût. D'un point de vue humain par contre, ces quelques mouvements suffisent à évaluer l'ensemble de la séquence. Il en est de même pour la peur : le danseur effectue quelques retraits rapides mais joue aussi, dans ses mouvements, une sorte de curiosité et un intérêt pour l'objet imaginaire provoquant cette peur. eMotion étant incapable de discerner cette subtilité, elle ne reconnait la peur que dans les phases où le danseur recule violemment, comme pour fuir.

La tristesse, à l'inverse, est bien reconnue. Ceci est dû au fait que les caractéristiques principales de la tristesse (dos voûté, vitesse lente) sont conservées pendant toute la séquence : le danseur ne saute pas soudainement et redresse rarement le tronc. La joie est tout aussi bien reconnue, pour les mêmes raisons : tout au long de la séquence, le danseur est actif, rapide, érige le tronc, tend les bras...

D'une manière générale et après visionnage des enregistrements vidéo et animation d'un squelette 3D grâce aux coordonnées enregistrées, il apparaît que notre calcul de reconnaissance sur une séquence, se basant sur le pourcentage de mesures bien reconnues sur le nombre total de mesure, n'est pas un indicateur valide des performances de notre système. Une autre validation est donc à prévoir. Cette expérimentation montre par contre qu'un calcul de l'émotion à une telle fréquence ne permet également pas une reconnaissance précise de l'émotion, dont l'expression s'étale sur un laps de temps de l'ordre de quelques secondes. L'interpréteur doit donc être modifié afin de prendre en compte les dernières secondes de mouvement dans son ensemble.

eMotion contribue à évaluer notre modèle conceptuel présenté au chapitre 5. L'évaluation du système a permis de cerner les points forts et les points faibles de notre reconnaissance. Les erreurs de reconnaissance, principalement sur le dégoût et la peur, sont dues principalement à des considérations temporelles et peuvent être corrigées grâce à l'extraction de nouvelles caractéristiques et l'implémentation d'une nouvelle interprétation. Notre modèle permet la modification du système : les nouvelles caractéristiques nécessitent l'implémentation de nouveaux extracteurs qui seront ensuite branchés dans le système. De même, établir une nouvelle interprétation nécessite de créer un nouveau composant interpréteur à brancher dans le système. Les composants déjà présents ne requièrent aucune modification dans cette évolution du système.

Par cette évaluation, nous avons validé notre système pour la reconnaissance de deux émotions : la joie et la tristesse. Ce prototype est utilisé dans nos travaux sur l'augmentation d'un spectacle de danse par l'émotion exprimée par le danseur, décrits au chapitre 7.

6.6 Conclusion

Nous avons présenté, dans ce chapitre, le système interactif eMotion de reconnaissance d'émotions par la gestuelle. D'un point de vue architecture, eMotion repose sur le découpage fonctionnel en trois niveaux Capture, Analyse et Interprétation qui se transposent dans son implémentation. eMotion instancie le modèle de la branche émotion vu au chapitre 5. Le système s'appuie sur le modèle PAC pour l'interaction graphique avec le concepteur. De plus, il illustre certaines combinaisons de modalités héritées de l'approche "interaction multimodale" qui n'ont pas encore été développées dans les systèmes actuels de reconnaissance d'émotions, notamment le choix d'une modalité par le concepteur *via* le choix du dispositif de capture, et le choix du système de l'analyse à effectuer selon le système représentationnel de capture utilisé.

D'un point de vue du domaine de la reconnaissance d'émotions, eMotion permet la reconnaissance d'émotions basiques en se basant entièrement sur un travail issu de la littérature en psychologie. Le système n'a donc pas subi d'entraînement préalable, les règles d'interprétation étant déjà établies dans [46]. L'expérimentation mise en œuvre a permis de révéler l'importance de l'aspect temporel de la reconnaissance d'émotions : en particulier, une reconnaissance rapide associée à un mécanisme de vote sur une séquence ne permet pas une reconnaissance d'émotions exprimées de façons ponctuelles comme le dégoût ou la surprise. Cette expérimentation servait un double but d'exploration et de validation de notre système. Les résultats obtenus permettent de valider partiellement notre système pour des émotions exprimables dans la durée, en particulier la joie et la tristesse.

eMotion définit un canevas logiciel extensible basé sur notre cadre architectural conceptuel. Il fournit un ensemble de services paramétrables pour reconnaître une émotion basée sur les mouvements. Pour illustrer l'utilisation de ce canevas logiciel, nous présentons au chapitre suivant une application exploitant ce canevas pour reconnaître l'émotion d'un danseur.

Chapitre 7

eMotion appliqué à la danse : reconnaissance d'émotions et réalité augmentée

Nous avons présenté, au fil de ce mémoire, les différentes étapes de la construction d'un système de reconnaissance d'émotions, de l'intégration de ses concepts dans ceux de l'interaction multimodale vers la conception d'un modèle d'architecture et jusqu'à la réalisation d'une application de reconnaissance basée sur la gestuelle. Dans ce chapitre, nous présentons une application pratique de nos contributions au cas de la danse et du ballet augmenté. La danse est une forme d'art qui, dans sa plus pure expression, se réduit au mouvement d'un danseur. Comme toute forme d'art, elle déclenche chez le public des réactions émotionnelles variées (admiration, extase ou ennui). Dès le début de nos travaux, nous avons collaboré avec les danseurs et chorégraphes de l'institution Malandain Ballet Biarritz[1] [36].

L'objectif de notre collaboration avec des danseurs était de définir comment les émotions reconnues à partir des mouvements d'un danseur pouvaient être utilisées pour moduler les éléments virtuels mis en jeu sur la scène du ballet. Notre objectif était d'augmenter la scène du ballet à partir d'émotions reconnues et d'effectuer cette modulation en temps réel, afin de garder à chaque représentation son caractère unique [35].

Dans le cadre de nos travaux, cette collaboration nous a permis d'appliquer le canevas logiciel eMotion à un cas concret. L'implémentation d'eMotion n'est pas dépendante d'un cadre applicatif : les caractéristiques de mouvement utilisées sont des caractéristiques générales. L'implémentation actuelle d'eMotion est donc une solution à la fois tournée vers notre cadre applicatif (seule une reconnaissance par le mouvement a été actuellement implémentée) et générique dans le choix d'une reconnaissance sur des caractéristiques générales de mouvement.

Nous présentons donc, dans la première section, notre domaine d'application : la danse. Cette section permet de présenter les principes fondamentaux de la danse et d'introduire le

[1] http ://www.malandainballet.com/

travail collaboratif que nous avons mené avec les danseurs. Nous abordons ensuite le domaine de la réalité augmentée, et en tirons des concepts et des choix pour la conception de notre propre système. Nous décrivons ensuite nos travaux concernant la tâche d'augmenter une scène de ballet. Nous décrivons ensuite nos travaux concernant l'augmentation de la scène conçue et développée, puis des applications concrètes de ces travaux et de cette collaboration sous la forme d'évènements scientifiques et artistiques. Enfin, nous concluons sur une expérimentation exploratoire ayant pour but de déterminer l'impact des augmentations sur la perception par le public des émotions exprimées par un danseur.

7.1 Cadre d'application : réalité augmentée pour la danse

Nous présentons dans cette section notre cadre d'application : la réalité augmentée pour un spectacle de danse. Nous y décrivons donc les deux domaines de la danse et de la réalité augmentée.

7.1.1 Principes fondamentaux de la danse

Lorsque l'on parle de danse sans rentrer dans les problématiques liées au style et à la technique, il faut considérer quel est le bien commun entre tous ces styles et ces différentes techniques. Il est également nécessaire d'être d'accord sur des principes qui peuvent sembler de prime abord des évidences car ces évidences servent d'outils à la danse pour s'exprimer et peuvent donner, à ceux qui ont la volonté de l'analyser, le moyen de détacher leur observation d'un contexte particulier défini par un consensus ou un goût individuel. Cette démarche de détachement ne peut donc s'effectuer sans que soit défini ce que sont les fondamentaux de la danse, au sens le plus général possible du terme et hors des définitions que pourrait lui donner l'histoire de son évolution. La première évidence à laquelle nous sommes confrontés pour parler des fondamentaux de la danse, c'est que la danse est une pratique physique et que par conséquent parler de danse revient forcément à parler du corps.

Le corps

Le corps est à la fois un outil mécanique et la porte de nos relations avec le reste du monde. Le danseur explore cette dualité à travers son travail et fait du corps un sujet artistique.

Le corps est un outil. Il s'agit d'un assemblage mécanique comportant des limitations. Pour un danseur, une grande partie du travail au cours des ans est d'arriver à appréhender cette mécanique et ses limitations afin de constamment améliorer la perception de son propre corps. Le danseur se détache ainsi de son corps jusqu'à parfois le considérer comme une machine et explorer la façon dont il peut bouger et s'articuler, et utiliser cette mécanique complexe dans un contexte de danse.

Le corps définit un intérieur et un extérieur. L'intérieur du corps est l'outil et l'assemblage d'organes que le danseur apprend à connaître. L'extérieur est le reste du monde : le corps permet l'interaction avec l'extérieur. Tout comme le danseur explore les possibilités de son

propre corps, il développe également les relations qu'a ce corps avec l'extérieur. La danse est ainsi une réflexion dans un contexte global, culturel, politique ou social. Ainsi, plusieurs sortes de danses existent dans le monde, impliquant des focalisations sur des membres différents.

Son rôle d'outil et son rôle de lien avec l'extérieur placent le corps comme sujet d'étude. Le corps est un outil artistique comme le pinceau, mais dont les liens avec l'intérieur (esprit) et l'extérieur (les autres) font qu'il pense et s'interroge. Qu'implique de dévoiler un mouvement ? Un corps ? Le corps comme sujet prenant en compte l'intérieur et l'extérieur passe aussi par les mouvements classiques des différentes traditions voire des costumes traditionnels (par exemples, les danses traditionnelles vietnamiennes)

Le corps est un outil ; le mouvement est la création. La notion de mouvement fait intervenir les notions de déplacement dans l'espace et le temps. Le mouvement comme lien avec l'extérieur renvoie à considérer une fonction de communication avec l'autre.

Le temps

Le temps est un des deux paramètres intrinsèques du mouvement : un mouvement ne peut être que dynamique.

Le temps renvoie tout d'abord au vieillissement. Le corps vieillit et se modifie au cours de la vie du danseur, qui voit ses possibilités se réduire : les techniques doivent donc changer. Cette notion de temps, métaphore d'une vie, renvoie au cycle de la naissance, de l'évolution, et de la mort. Ce cycle se retrouve dans un mouvement, ou dans une danse : le mouvement naît, évolue, puis meurt.

A l'échelle d'une danse, le temps donne le tempo et le rythme. La musique n'est pas nécessaire pour donner ce rythme. Le corps lui-même, par le rythme cardiaque ou la respiration, permet de structurer le mouvement. Certaines danses vont cependant s'effectuer dans le silence et vont donc ainsi livrer le danseur au temps sans lui donner d'autres partenaires que son propre corps. La pause est également un élément à part entière du mouvement, permettant d'exprimer aux autres le rythme du corps du danseur.

La danse est souvent associée à la musique. La musique se base également sur le rythme et le tempo, par des jeux de vitesses d'exécution, des silences, etc. Musique et danse sont complémentaires sans empiètement : la danse est visuelle, la musique sonore.

L'espace

L'espace est le deuxième paramètre intrinsèque du mouvement. Il est divisé en deux : l'espace proche et l'espace lointain, reflétant les concepts de kinésphère et d'espace général posés par Laban et décrits au paragraphe 3.2 (page 53).

L'espace proche englobe le corps dans une sphère délimitée par l'extension de ses membres (bras et jambes). Cette sphère est divisée verticalement en six niveaux : le sol, les genoux, le bassin, les épaules, la tête et un niveau correspondant à la main lorsque les bras sont tendus vers le haut. Cette structuration de l'espace offre au danseur une carte d'exploration de son espace proche. L'espace lointain englobe le reste de l'espace (la scène, par exemple). Il n'est accessible que par un déplacement du danseur. L'espace lointain est lui-même divisé en niveaux horizontaux et verticaux, quadrillant l'espace volumique de la scène. Tout comme pour l'espace proche, ce quadrillage offre une carte de l'espace au danseur afin de permettre l'exploration. Par défaut, l'espace lointain est dirigé de la scène vers le public. Ce repère est cependant parfois modifié par les chorégraphes pour offrir un autre angle de vue au public.

C'est dans l'espace lointain que se trouvent les espaces proches des autres. Les différents espaces proches peuvent être organisés dans l'espace lointain, à la fois dans leurs positions et leurs mouvements les uns par rapport aux autres. En effet, le rapport à l'autre est le dernier fondement de la danse.

Relation à l'autre

La danse est un art expressif et établit une communication avec le public. Le corps en tant que sujet reflète les liens extérieurs, avec les autres. La relation avec le public et les autres danseurs est une composante fondamentale de la danse. Les autres danseurs permettent en particulier de donner une nouvelle référence dans l'espace général en y déterminant un repère. Le public peut ainsi se situer dos à la scène qui se joue. L'orientation du danseur dans l'espace général, et l'orientation de l'espace général par rapport au public ont un rôle communicatif exploré par les chorégraphes.

Le mouvement expressif en danse

La danse, grâce aux fondamentaux décrits ci-dessus, permet l'expression d'émotions ou de concepts abstraits. Par exemple, certains spectacles ne mettent pas en œuvre une musique mais un texte : les mouvements du danseur viennent alors appuyer aussi bien la musicalité de la lecture (rythme, tempo, intonations) que le sens du texte et de la phrase. Pour cela, les danseurs analysent leurs mouvements afin de dévier les habitudes sociales de position et les mouvements communicatifs vers la danse. Extraire les paramètres communicatifs des mouvements quotidiens leur permet de les réinjecter pour créer un mouvement de danse et ainsi adapter une interaction entre deux danseurs.

7.1.2 Définition de la réalité augmentée

Mallem et Roussel présentent dans [92] un résumé, datant de 2008, des principes, des technologies et des applications de la réalité augmentée. Nous présentons ici quelques principes de la réalité augmentée, utiles pour nos travaux et issus de cet état de l'art.

Mallem et Roussel définissent la réalité augmentée (RA) de la façon suivante :

> "La réalité augmentée (RA) est un concept rendu possible par un système capable de faire coexister spatialement et temporellement un monde virtuel avec l'environnement réel. Cette coexistence a pour objectif l'enrichissement de la perception de l'utilisateur de son environnement réel par des augmentations visuelles, sonores ou haptiques. L'environnement peut être présent dans l'environnement réel (réalité augmentée en vision directe sur le site) ou peut être perçu à distance (réalité augmentée en vision indirecte généralement hors du site)."

Milgram et Kishino définissent dans [99] un continuum entre environnement réel et environnement virtuel (voir figure 54). Les deux extrêmes sont le monde physique et le monde numérique (réalité virtuelle). Entre les deux se situe la réalité mixte : un mélange de réel et de virtuel. L'environnement dominant (réel ou virtuel) permet de distinguer entre réalité augmentée et virtualité augmentée. Dubois distingue réalité et virtualité augmentées [48] en considérant l'objet de la tâche. Si l'objet de la tâche est réel, il s'agit alors de réalité augmentée : des éléments virtuels sont ajoutés au réel (figures 55a et 55b). Si l'objet de la tâche est virtuel, il s'agit de virtualité augmentée : des éléments réels sont ajoutés au virtuel (par exemple, le système GeoTUI [124], figure 55c).

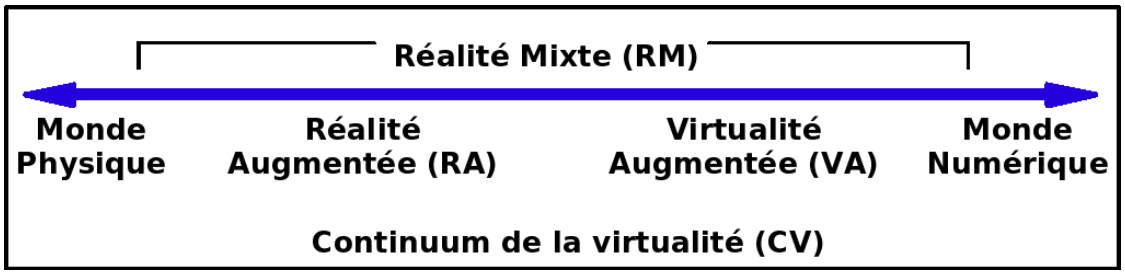
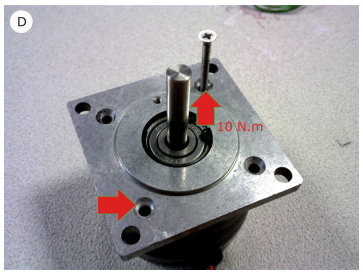


Fig. 54: Le continuum de Milgram et Kishino [99].

En réalité augmentée, l'environnement réel se voit ajouter des éléments virtuels. Notre cas applicatif entre donc dans le cadre de la réalité augmentée : la scène réelle est augmentée par des éléments virtuels. Les éléments virtuels ajoutés au réel peuvent faire appel à plusieurs modalités, comme le son [91] ou le toucher [41] ; l'humain se reposant principalement sur la vue, cette modalité est cependant la plus explorée. En se limitant à la vue, le but de la réalité augmentée est donc de permettre à l'humain de voir du virtuel intégré au monde réel, en y ajoutant du contenu graphique en relation avec la scène réelle.

Un système de réalité augmentée se doit donc de posséder plusieurs fonctionnalités. La figure 56 [92] représente les différentes briques d'un système de réalité augmentée. Au plus bas niveau, le système repose sur trois piliers. Tout d'abord, le système doit être capable d'assurer un suivi de l'utilisateur et de recaler les données visuelles afin de les intégrer le mieux possibles dans l'image réelle. En effet, l'humain détecte particulièrement bien les décalages créant un inconfort visuel. Le deuxième pilier est le rendu graphique des données. Il peut s'agir d'indications textuelles, de flèches ou de graphismes plus complexes. Enfin, le système repose sur une technologie d'affichage permettant de faire percevoir à l'humain les objets

(a) Chirurgie augmentée. Figure tirée de [63].



(b) Vue de l'opérateur dans le système T.A.C. : Pointage d'une pièce d'un assemblage mécanique [13].

(c) Le système GeoTUI [124].

FIG. 55: Exemples de réalité (55a et 55b) et de virtualité (55c) augmentées.

virtuels. Le système s'appuie sur ces piliers et propose ainsi un mélange entre virtuel et réel. Le système doit donc proposer une présentation du monde virtuel, une scénarisation de celui-ci (*authoring*) et des techniques et interfaces permettant à l'utilisateur d'interagir avec lui. Ces différentes briques sont englobées dans l'application manipulée par l'utilisateur.

Ces briques permettent à un système de réalité augmentée d'ajouter au réel des informations virtuelles, et de rendre ces informations perceptibles par l'utilisateur. De nombreux systèmes de réalité augmentée sont appliqués aux domaines de la chirurgie et de la mécanique. Par exemple, Grimson *et al.* [63] projettent directement sur le crâne du patient pendant une chirurgie les informations de scanner collectées antérieurement (figure 55a, page 158). Bottecchia *et al.* proposent dans [13] le système T.A.C (TéléAssistance Collaborative), permettant à un opérateur de maintenance mécanique de recevoir des instructions sur la réparation d'un moteur de la part d'un expert distant grâce au paradigme POA (*Picking Outlining Adding*). L'expert à distance peut voir la scène vue par les yeux de l'opérateur, sélectionner et surligner des éléments du moteur et d'ajouter des éléments virtuels (flèches par exemple) permettant de pointer les éléments lors d'une phrase ambigüe ("Mets cette pièce ici"). L'opérateur voit donc la scène réelle augmentée par les indications de l'expert (figure 55b, page 158), simulant ainsi la coprésence de l'expert auprès de l'opérateur.

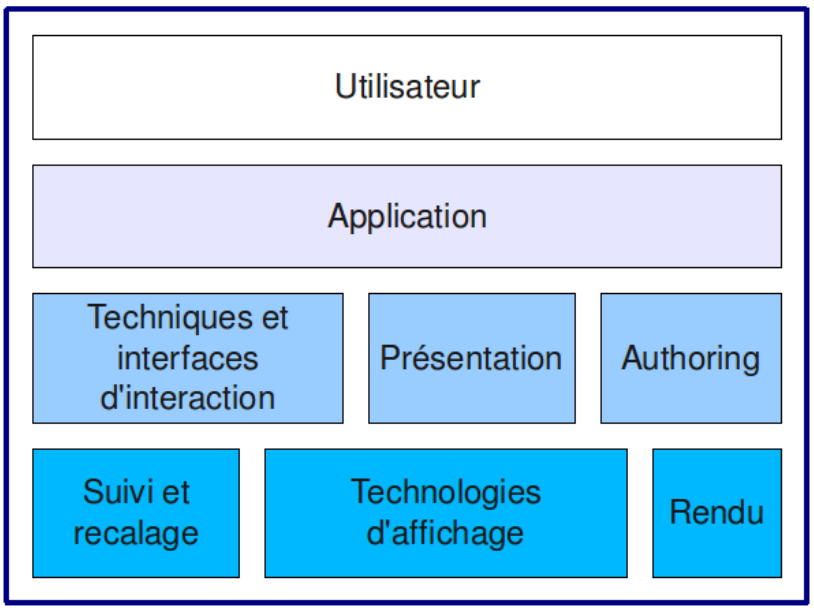


Fig. 56: Les différentes briques composant un système de réalité augmentée. Figure tirée de [92]

7.1.3 Suivi de l'utilisateur

La plupart des systèmes de réalité augmentée cherchent à suivre le point de vue de l'utilisateur, afin d'y insérer du contenu virtuel. Un suivi vidéo est généralement effectué, selon des méthodes basées image (détection et suivi de points d'intérêts dans l'image) ou objet (suivi de la position par projection du modèle 3D connu de l'objet à suivre). Des capteurs de suivi du mouvement peuvent être utilisés seuls ou combinés avec les techniques de vision par ordinateur. Le suivi de l'utilisateur est souvent un point critique en réalité augmentée mais ne s'applique pas au cas du ballet augmenté. En effet, dans notre cas, les utilisateurs percevant les augmentations sont le public, que nous considérons comme définissant un point de vue unique et statique sur la scène augmentée.

7.1.4 Affichage des données virtuelles et intégration dans le réel

Il existe trois types de dispositifs de visualisation en réalité augmentée [92] : les affichages *headworn* sont portés directement sur la tête. Il s'agit de lunettes ou de casques capables d'afficher une image devant les yeux de l'utilisateur. Ces dispositifs se divisent en deux catégories : les dispositifs *optical see-through* apportent une vision directe de la réalité (comme à travers des lunettes classiques). Le dispositif est alors capable d'afficher des éléments virtuels dans une zone du champ de vision. Les dispositifs de type *video see-through* capturent la réalité grâce à une caméra ; le flux vidéo est alors augmenté. L'utilisateur ne voit plus la réalité de façon directe mais à travers le flux vidéo augmenté. Les affichages *handheld* se font grâce à un dispositif mobile tenu à la main comme un téléphone portable ou un PDA. Une caméra fixée sur le dispositif capture un flux vidéo. Ce flux est ensuite augmenté et affiché sur l'écran du

dispositif. Enfin, les affichages de types projectifs (*projective*) utilisent des projecteurs permettant d'afficher directement des éléments virtuels sur les objets réels. Cette dernière technique est celle qui s'adapte le mieux à notre cas. En effet, elle ne nécessite pas des membres du public qu'ils disposent d'un affichage personnel (*headworn* ou *handheld*). De plus, considérer le public comme un point de vue unique et statique permet de s'affranchir des contraintes de mobilité que remplissent de tels types d'affichage : l'affichage projectif ne pénalise pas le public.

7.2 Reconnaissance d'émotions pour augmenter une scène de ballet

Dans cette section, nous décrivons le travail que nous avons fait en collaboration avec Malandain Ballet Biarritz, avec en vue un objectif à long terme : la création d'un ballet augmenté se basant sur la reconnaissance des émotions d'un danseur. Nous décrivons ici nos motivations scientifiques et les motivations artistiques de nos collaborateurs ; nous décrivons ensuite les technologies choisies pour progresser vers la mise en place d'un ballet augmenté. Nous présentons ensuite le système existant utilisé, se basant sur les émotions reconnues par eMotion pour modifier des éléments virtuels. Nous décrivons ensuite deux évènements mis en place en collaboration avec Malandain Ballet Biarritz. Enfin, nous décrivons une expérimentation ayant pour but de déterminer les augmentations permettant le mieux de convoyer l'émotion exprimée par le danseur.

7.2.1 Motivations

Notre collaboration avec Malandain Ballet Biarritz est le fruit d'une volonté tant du côté scientifique que du côté artistique de rapprocher ces deux domaines extrêmement différents. L'établissement d'un dialogue commun a d'ailleurs été une phase primordiale de cette collaboration.

Motivation scientifique

La danse définit un cadre d'étude idéal pour étudier les émotions basées sur les mouvements. De plus la réalité augmentée est souvent utilisée dans le cadre de ballets.

Un premier objectif est l'application du canevas logiciel eMotion au cadre de la danse. eMotion n'est pas soumis à un domaine particulier ; les caractéristiques qui y sont considérées sont des caractéristiques génériques de mouvement. Le concepteur peut cependant paramétrer l'extraction des caractéristiques afin d'utiliser eMotion dans un cadre particulier. Ici, le canevas logiciel dispose d'une machine dédiée, qui fournit les émotions reconnues au système d'augmentation.

Notre deuxième objectif est l'exploration du potentiel de la reconnaissance d'émotions par ordinateur mêlée à la réalité augmentée dans le contexte d'un spectacle de ballet afin de mieux convoyer le message du chorégraphe et suggérer de nouvelles situations artistiques. Plusieurs

spectacles augmentés ont été montés et joués dans les quinze dernières années. L'évolution des technologies et des systèmes de réalité augmentée a permis aux artistes de les utiliser comme outils dans leurs créations. Le spectacle "The Plane[2]" a ainsi unifié la danse, le théâtre et éléments virtuels dans un duo entre un danseur et sa propre image. Le spectacle "Hand-Drawn Spaces" [75] présentait une chorégraphie en trois dimensions de personnages dessinés à la main, où les mouvements du danseur réel étaient appliqués au personnage virtuel. Une interaction en temps réel a été introduite par "The Jew of Malta[3]" où des coupes architecturales virtuelles de bâtiments et les costumes virtuels des danseurs étaient générés en temps réel, en fonction de la musique et de la position du chanteur sur la scène. Notre motivation est également de pousser l'interaction entre danseur et éléments virtuels en y incluant l'émotion exprimée.

Motivation artistique

La reconnaissance d'émotions par ordinateur appliquée à la danse est un contexte de recherche profitant aussi bien à l'artiste qu'au scientifique. Communiquer une émotion à un ordinateur, système si différent de l'humain, implique de revenir aux fondamentaux de la danse : l'espace, le temps et l'autre, et de reconstruire la danse à partir de ce nouveau paramètre de départ : le danseur n'exprime plus pour des humains. Le processus de recherche induit implique un requestionnement de "l'interprétation" d'une émotion et par là même un requestionnement de la relation entre son corps et son esprit que le danseur a pu établir au cours de sa carrière. La reconnaissance d'émotions force à distinguer l'expérience d'une émotion et son expression. Toutes ces nuances impliquent de la part du danseur une analyse profonde du mouvement et de ses qualités. Si le scientifique a pour rôle d'observer le mouvement pour permettre une reconnaissance toujours plus précise, le danseur quant à lui doit être capable d'analyser suffisamment ses propres mouvements pour pouvoir en isoler les paramètres lui permettant de mieux se faire comprendre par la machine ; le but final étant une convergence des capacités de reconnaissance de la machine vers les capacités de reconnaissance de l'humain.

Les différentes propositions de la réalité augmentée permettent à la danse d'avancer dans ses principes en partageant le processus de mise en scène avec de nouvelles formes d'art n'ayant jamais établi de contact avec la danse et son développement. Pour les danseurs, la recherche collaborative entre art et science sur les relations entre l'homme et la machine et ses interactions, a généré une esthétique propre et une forme de raisonnement intéressantes à inclure dans le processus de mise en scène. Un objectif est ainsi de développer un processus de mise en scène basé sur les principes de la recherche scientifique pour expliciter la recherche effectuée en reconnaissance d'émotions par des augmentations, des lumières, du son, et donner un sens à la présentation de cette recherche en tant que spectacle. Les outils qu'apporte la réalité augmentée dans ce cadre permettent de relier les différents éléments de notre recherche entre eux en offrant aux chorégraphes les moyens de partager avec le public les émotions échangées entre le danseur et la machine.

[2]Site web du Troika Ranch : www.troikaranch.org/

[3]Page web "The Jew of Malta" :http ://www.joachimsauter.com/en/projects/vro.html

Enfin, la finalité de cette collaboration est également la mise en place d'un ballet augmenté. Nous avons donc créé des "briques d'augmentation" - des éléments virtuels qu'il est possible de composer et mettre en œuvre dans un spectacle. Nous avons en particulier développé une ombre virtuelle, dont la taille et la couleur change selon les émotions exprimées par le danseur (voir le paragraphe 7.2.4).

7.2.2 La réalité augmentée appliquée à un spectacle de danse

L'augmentation d'une scène de ballet en vue d'un spectacle augmenté est un cas très particulier de réalité augmentée. Le but est ici d'augmenter la scène par des éléments virtuels. Ces éléments virtuels étant générés par les mouvements du danseur et devant interagir avec lui, il est nécessaire de pouvoir capter les mouvements de ce dernier. Le danseur n'est par contre pas celui qui perçoit les éléments virtuels ; le public est ici prioritaire. L'interaction est donc, dans notre cas, tripartite et fait intervenir le danseur, le système et le public. Contrairement au danseur, le public est passif.

Technologies de capture du mouvement

Certains prototypes issus de la recherche permettent un suivi 3D basé caméra d'un corps humain. Nous avons d'emblée rejeté cette solution, qui nécessite un éclairage constant et optimal, à l'opposé d'un éclairage de scène (spots lumineux de différentes couleurs pouvant se mouvoir). Nous nous sommes orientés vers des solutions commerciales de capture du mouvement par combinaison. Welch et Foxlin présentent en 2002 un état de l'art des technologies de capture du mouvement [150]. Bien que des améliorations technologiques soient apparues depuis, les auteurs décrivent les forces et faiblesses des différentes technologies et fournit une base solide à notre choix de technologie pour la capture de mouvement. Welch et Foxlin décrivent le capteur idéal comme étant petit (de la taille d'une puce), indépendant (ne nécessitant pas d'autres éléments dans l'environnement ou sur l'utilisateur), complet (pouvant mesurer les six degrés de liberté), précis, rapide, résistant aux occlusions, robuste, tenace (pouvant capter la cible le plus loin possible), sans fil et à un prix abordable. Les différentes technologies existantes en 2002 ne respectaient qu'au maximum quatre critères parmi ces dix. Les progrès en miniaturisation et en couplage des capteurs fait qu'en 2009 de nombreux systèmes utilisent plusieurs technologies en même temps afin de pallier aux inconvénients de chacune. Nous décrivons au paragraphe 7.2.3 la combinaison que nous utilisons, remplissant de façon acceptable les dix critères cités.

Les technologies mécaniques consistent à relier l'objet à suivre à un socle grâce à un assemblage mécanique permettant le mouvement. L'assemblage mécanique dispose de senseurs capables de mesurer les angles de rotation ou de translation de ses différentes parties et, ainsi, de recalculer précisément la position et l'orientation de l'objet suivi. Cette technologie permet une grande précision et permet de plus l'ajout d'un retour d'effort, comme le dispositif Phantom[4]. Cependant, une telle technologie limite le mouvement à un faible volume, la rendant totalement inadéquate à une utilisation sur une scène. Les technologies optiques se basent

[4]http ://www.sensable.com/haptic-phantom-desktop.htm

sur l'utilisation de capteurs lumineux actifs (LEDs) ou passif (réfléchissant une lumière extérieure). Les ondes lumineuses utilisées peuvent se trouver dans le spectre visible ou infrarouge. L'avantage de cette technologie est sa précision et sa rapidité. Elle n'est cependant pas résistante aux occlusions et nécessite un environnement lumineux contrôlé. Ces deux faiblesses invalident l'utilisation d'une technologie optique sur scène. Les technologies acoustiques se basent sur la différence de phase entre les signaux sonores issus de sources différentes. Grandement sensibles au bruit, elles ne peuvent être retenues dans un environnement où de la musique est constamment jouée.

Les technologies inertielles reposent sur l'utilisation de MEMS (*MicroElectroMechanical Systems*). Elles permettent des capteurs réduits, insensibles aux occlusions, complets et rapides. La précision peut être meilleure en regard d'autres technologies mais les mesures devant subir une (rotations) ou deux (accélérations) intégrations pour en déduire l'orientation et la position, ces capteurs sont sujet au *drift*, c'est-à-dire à une propagation et à une amplification de l'erreur au cours du temps. Enfin, les technologies magnétiques s'appuient sur la génération d'un champ magnétique par une source. Les capteurs sont également petits et insensibles aux occlusions. La précision décroît cependant rapidement avec l'éloignement par rapport à la source. De plus, la présence d'objets métalliques affecte le champ magnétique et donc la précision des capteurs. Le champ magnétique terrestre est utilisable pour calculer l'orientation du capteur par rapport au nord magnétique terrestre. La combinaison que nous utilisons pour capturer les mouvements du danseur utilise ces deux technologies en combinaison.

Affichage des éléments virtuels

L'augmentation d'une scène de ballet a pour objectif d'augmenter l'expérience du public lors du spectacle. Les éléments virtuels ont donc pour vocation primaire d'être perçus par le public et non par les danseurs. Dans ce contexte, la projection du contenu virtuel sur la scène est clairement mieux adaptée que les affichages de type *headworn* ou *handheld* (paragraphe 7.1.4, page 159). Nous faisons ici l'hypothèse que le public est situé en face de la scène et partage donc à peu près le même point de vue ; dans une salle classique, cette hypothèse est fausse, le public pouvant se placer sur les côtés (balcons) ou largement en hauteur (paradis) par rapport à la scène.

Les éléments virtuels doivent permettre une interaction avec le danseur. Il n'est pas forcément nécessaire que ce dernier puisse les percevoir ; les danseurs sont entraînés à interagir, dans leurs chorégraphies, avec des objets imaginaires. L'affichage doit par contre pouvoir s'adapter aux mouvements et à la position du danseur sur la scène.

Techniques d'interaction, Présentation et Authoring

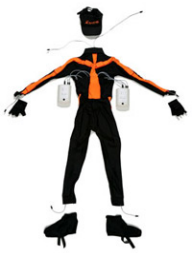
Le projet d'augmentation d'un ballet en est à sa phase de conception des moyens techniques pour l'augmentation. L'interaction proposée dans ce cadre est la reconnaissance des émotions exprimées par le danseur, impactant le comportement des objets virtuels. Nous n'avons pas encore abordé, dans ce cadre, les concepts de présentation et d'*authoring* (scénarisation de l'espace virtuel), qui fournissent des perspectives intéressantes pour la suite de nos travaux.

7.2.3 Choix des technologies

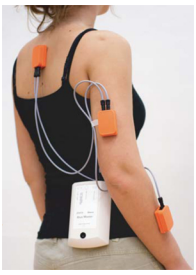
Notre cadre d'application entraîne certaines contraintes limitant le choix des technologies à utiliser pour la capture du mouvement du danseur et l'affichage d'éléments graphiques sur scène.

Choix du système de capture

Nous avons choisi, pour la capture du mouvement en situation de spectacle de ballet, la combinaison de capture du mouvement Moven illustrée à la figure 57. Cette combinaison s'appuie sur des capteurs combinant technologies inertielle et magnétique. Les senseurs inertiels permettent de déterminer l'axe vertical et de mesurer l'orientation du capteur. Le capteur magnétique permet de corriger les dérives des capteurs inertiels en mesurant l'orientation par rapport au nord magnétique. La combinaison est munie de 16 capteurs délivrant leur orientation par rapport à une repère absolu direct (ou l'axe x est celui pointant vers le nord magnétique, et l'axe z l'axe vertical orienté vers le haut). Les capteurs sont insensibles aux occlusions. Les 16 capteurs sont reliés par des fils à deux centrales XBus Master se portant dans le dos, au niveau du bassin (voir figure 57b). Chaque centrale fournit l'énergie requise par les capteurs grâce à des batteries, fusionne les signaux de chaque capteur et envoie les données à un ordinateur par Bluetooth. La combinaison permet donc une capture sans fil du mouvement, laissant au danseur la possibilité de se mouvoir dans un espace de quelques dizaines de mètres de rayon. Cet espace étant soumis à la bonne réception du signal Bluetooth, l'installation de nouveaux récepteurs standards permet de l'élargir. Enfin, la combinaison fournit une précision ($<0.5^{\circ}$), une résolution (0.05°) et une fréquence de mesure (jusqu'à 120Hz) suffisantes pour nos besoins[5].



(a) La combinaison Moven.



(b) Centrale XBus Master et capteurs MTx.

FIG. 57: La combinaison Moven de chez XSens.

La combinaison Moven est donc techniquement adaptée aux besoins d'une utilisation sur scène. L'enveloppe rigide des capteurs permet des mouvements doux au sol. De plus son esthétique reste discrète en comparaison d'autres solutions et permet en particulier au danseur capturé de porter des costumes. Enfin, la combinaison est considérée dans notre cas comme

[5]Données constructeur.

non-intrusive : en effet, les phases d'habillage et de calibration font partie de la préparation du danseur avant d'entrer en scène (elles s'assimilent au revêtement d'un costume).

La combinaison Moven est vendue avec le logiciel propriétaire Moven studio. Ce logiciel propose une interface pour la calibration, le suivi et la capture des mouvements par la combinaison. Un modèle biomécanique permet de calculer les positions et orientations de 23 segments du corps depuis les orientations des 16 capteurs physiques. Le corps évolue alors dans un repère absolu dont l'orientation est déterminée par le nord magnétique et la verticale. L'origine du repère est définie par la position du talon droit lors de la calibration. Les fonctions avancées du logiciel permettent de retravailler le mouvement calculé pour obtenir de meilleurs résultats en vue de l'animation d'un personnage virtuel. Le logiciel Moven studio possède également une option lui permettant de délivrer *via* le réseau les coordonnées capturées sous forme de paquets UDP en temps réel.

Choix de la technologie d'affichage

Dans le cadre du ballet augmenté, nous nous sommes tournés vers des technologies projectives : l'affichage du virtuel sur le réel permet à tout le public de percevoir l'augmentation.

Le système Eyeliner de l'entreprise Musion[6] repose sur l'utilisation d'un film semi-réfléchissant sur lequel peut être projeté n'importe quel contenu graphique (vidéo ou 3D). Le film semi-réfléchissant est tendu entre l'avant de la scène et le plafond. Les danseurs se trouvant derrières sont vus par le public. Un projecteur situé dans la salle projette sur le film le contenu virtuel. Dans une salle sombre, le public a ainsi l'impression que le réel (acteurs situés derrière le film) et le virtuel (images projetées sur le film) sont mêlés. Cette immersion est renforcée par la possibilité de découper le film, créant ainsi des passages pour les danseurs ; ces derniers peuvent alors passer devant les images affichées. Un aperçu de cette technologie est disponible en vidéo sur le site de Musion, où la chanteuse Madonna effectue un court duo avec le groupe virtuel Gorillaz. Compte tenu du coût d'un tel système, nos choix s'orientent vers une projection sur le mur du fond de la scène ainsi que sur son sol. La collaboration avec les danseurs et chorégraphes du ballet Malandain Ballet Biarritz nous a également amené à considérer d'autres dispositifs pour la projection des données, notamment un système de pointage permettant de n'augmenter qu'une faible surface mobile ou encore un système de projection sur cylindre.

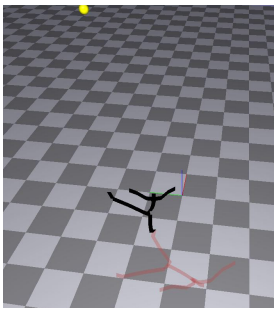
7.2.4 Architecture générale

Le système permettant de générer des augmentations prenant en compte les mouvements et émotions exprimées du danseur s'architecture autour de trois applications connectées entre elles en réseau : le logiciel Moven permet la capture du mouvement sur une machine ; notre canevas logiciel eMotion est lancé sur une deuxième. La troisième application se base sur les mouvements et l'émotion reconnue pour générer du contenu virtuel.

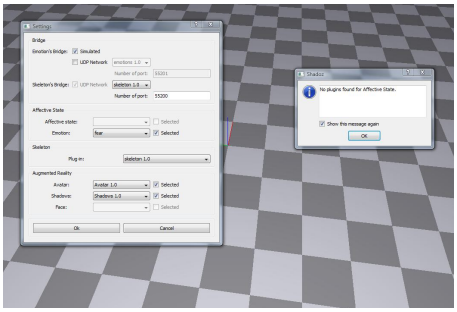
[6] http ://www.musion.co.uk/

L'application ShadoZ

L'application ShadoZ a été développée par Elric Delord dans notre laboratoire dans le cadre du projet CARE[7] (figure 58). ShadoZ est la contraction des mots "Shadow" et de l'axe Z d'un espace à trois dimensions. Son rôle est d'utiliser à la fois les informations de mouvement fournies par Moven studio et l'émotion reconnue par eMotion pour créer une ombre suivant les mouvements du danseur mais dont la taille et la couleur diffèrent selon l'émotion exprimée (voir figure 58a). Nous nous sommes basés sur les travaux de Birren [8] et Valdez [141] pour le choix de la couleur et de la taille à affecter à l'ombre pour une émotion donnée. L'application ShadoZ permet un lissage des émotions fournies par eMotion ainsi que l'attribution d'émotions de l'ombre différentes de celle exprimée par le danseur. Cette attribution est dirigée par une surjection de l'ensemble des émotions exprimées par le danseur dans l'ensemble des émotions exprimées par l'ombre. L'application ShadoZ offre une interface graphique permettant à l'utilisateur de définir cette surjection (voir figure 58b).



(a) L'avatar fil de fer et son ombre expressive.



(b) Interface graphique permettant à l'utilisateur de définir une surjection entre les émotions exprimées par le danseur et celles exprimées par son ombre.

FIG. 58: L'application ShadoZ.

Fonctionnement général du système

Le système général est distribué sur trois machines, communiquant par paquets UDP (voir figure 59). La première machine héberge l'application Moven Studio permettant de capturer les coordonnées du corps à chaque mesure. L'application eMotion est lancée sur une deuxième machine et prend en entrée le flux UDP des coordonnées du corps. Elle délivre un flux UDP des émotions reconnues. L'application ShadoZ prend ces deux flux en entrée et les synchronise pour créer l'ombre virtuelle du danseur.

7.3 Applications de notre système

La collaboration de Malandain Ballet Biarritz à nos travaux nous a permis d'appliquer le système eMotion à des situations concrètes. Nous décrivons ici trois évènements : deux

[7] http ://www.careproject.fr

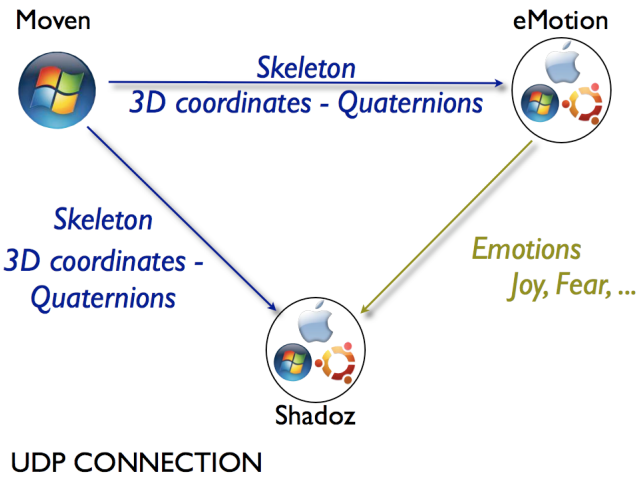


Fig. 59: Le système complet est distribué sur trois machines hétérogènes fonctionnant sous Microsoft Windows Vista, Apple MacOSX et Linux Ubuntu.

conférences dansées et un spectacle de réalité augmentée en appartement.

7.3.1 Conférences dansées

La conférence dansée est une forme rarement utilisée de présentation de travaux scientifiques. Dans cette conférence dansée, nous avons mélangé communication scientifique et improvisation dansée pour s'extraire des formes classiques de la présentation scientifique et de la représentation artistique [31]. Le danseur portait la combinaison de capture du mouvement Moven et improvisait sur un texte lu, accompagné de musique, de jeux de lumières, de diapositives illustrant le discours et de son avatar 3D reproduisant ses mouvements. Cette forme de dialogue a permis au public de la danse d'intégrer une problématique de recherche, et au public scientifique de s'extraire d'une approche purement technique et de mieux appréhender les tenants et aboutissants de la recherche en reconnaissance d'émotions. Le dialogue établi avec le public à la fin de la présentation dansée a permis d'expliciter l'interaction entre chercheurs et danseurs et de favoriser l'interaction entre chercheurs, danseurs, et le public.

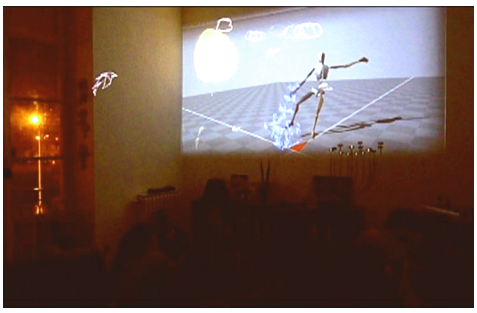
Une autre conférence a été l'occasion pour nous de présenter nos travaux de façon interactive en faisant une démonstration de notre système d'émotion, devant un public exclusivement constitué de scientifiques [32]. Cette conférence a été l'occasion pour nous de faire participer le public à l'expérimentation d'eMotion qui reconnaissait les émotions du danseur dont les mouvements étaient capturés pendant la présentation (figure 60a).

7.3.2 Spectacle augmenté en appartement

Le festival des Ethiopiques, à Bayonne, regroupe divers artistes volontaires qui se produisent gratuitement en différents endroits de la ville. Profitant de ce festival, nous avons réalisé, en collaboration avec Gaël Domenger du Malandain Ballet Biarritz, un spectacle augmenté en

appartement [33]. L'évènement a pris la forme d'un spectacle improvisé dans l'appartement de l'un des artistes et était ouvert au public, offrant ainsi une atmosphère confortable bien que légèrement étrange. Deux joueurs de Saz (un luth kurde) et d'accordéon ont improvisé une atmosphère musicale sur laquelle de la poésie était lue. Gaël Domenger portait la combinaison de capture du mouvement et improvisait une danse sur la musique et le texte. L'avatar du danseur (un squelette en 3D mimant les gestes du danseur) était projeté sur le mur. Pendant ce temps, des animations (développée sur Adobe Flash) choisies, placées et paramétrées en temps réel par Aymeric Reumaux[8] étaient apposées en surimpression sur l'avatar 3D. Au début du spectacle, le danseur était dans une pièce séparée ; le public ne pouvait voir que son *alter ego* virtuel projeté au mur et les animations (figure 60b). A la fin du spectacle, le danseur est venu danser au milieu du public pour interagir avec lui. Le danseur virtuel, le danseur réel et son ombre sur le mur ont alors formé un trio rassemblant le monde réel et le monde virtuel par leurs interactions avec le public, les animations et entre eux.

(a) Démonstration interactive de l'application eMotion.



(b) Scène des Ethiopiques montrant le danseur, son avatar virtuel, et les animations en surimpression.

FIG. 60: Evènements conjoints entre danse et informatique.

7.4 Impact de diverses augmentations

Dans le cadre de l'augmentation d'un spectacle de ballet, nous avons cherché à déterminer quelles augmentations pouvaient le mieux véhiculer les émotions exprimées par le danseur [30]. Nous avons donc mis au point une expérience exploratoire permettant de tester trois augmentations différentes : un avatar 3D de type "bonhomme fil de fer" seul, ce même avatar 3D accompagné de son ombre réagissant aux émotions exprimées par le danseur, et le système MARC, un visage réaliste expressif développé au Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI) de l'Université Paris-6, avec qui nous avons collaboré dans le déroulement de l'expérimentation. Pour l'expérimentation nous avons donc mis en relation les applications Moven Studio, eMotion, ShadoZ, et MARC (voir figure 61a).

[8] http ://orangers.online.fr/

7.4.1 Choix des séquences dansées

Pour les besoins de l'expérience, nous avons tout d'abord sélectionné les séquences les mieux reconnues par l'application eMotion (obtenant des résultats supérieurs à 70%), avec une reconnaissance humaine supérieure à 50%. Nous avons complété l'ensemble de ces séquences avec des séquences bien reconnues par le système même si la reconnaissance humaine était faible. Nous avons ainsi obtenu cinq séquences pour nos tests, représentant la tristesse (deux séquences), la joie, la surprise négative, et une combinaison de joie+tristesse. La seconde séquence représentant la tristesse montrait la particularité de n'être reconnue humainement qu'à 42% mais d'être reconnue par eMotion sur 56% de sa durée. Nous avons différencié ces séquences en nommant la première "tristesse rH" (reconnue humainement) et la seconde "tristesse nrH" (non reconnue humainement). Ces cinq danses expressives ont constitué la base de notre expérimentation.

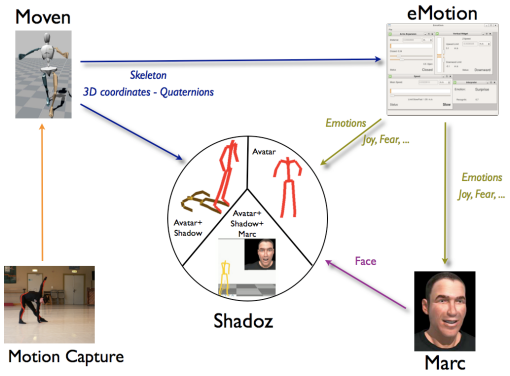
7.4.2 Modalités étudiées

Cinq modalités ont été explorées dans ce test perceptif. Nous avons utilisé les cinq danses affectives pour créer pour chacune cinq vidéos. Chaque vidéo correspondait à une modalité. La première modalité était la séquence vidéo originale, utilisée comme stimulus de référence durant le test. La seconde modalité était une vidéo du même mouvement et du même point de vue, mais reproduit par un avatar 3D de type "bonhomme fil de fer" de couleur noire. Cette modalité a été choisie pour mesurer l'impact d'une représentation minimaliste du danseur sur les capacités de perception émotionnelle des sujets de test. La troisième modalité était une vidéo du visage expressif réaliste MARC. L'application eMotion a été utilisée pour reconnaître les émotions d'une danse expressive. La séquence d'émotions correspondante a été ensuite exprimée par MARC. La quatrième modalité utilisait le logiciel ShadoZ : l'avatar fil de fer était augmenté d'une ombre à la couleur et à la taille dépendant de l'émotion reconnue. L'avatar était de la même couleur que son ombre. La cinquième et dernière modalité combinait l'avatar, son ombre expressive, et le visage expressif de MARC dans une seule vidéo (voir figure 61b).

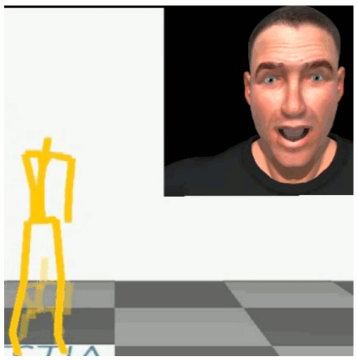
7.4.3 Méthode

Nous avons donc obtenu un total de 25 séquences vidéos (5 émotions $\times$ 5 modalités). 25 séries de 5 vidéos chacune ont été générées de façon semi-aléatoire : chaque série comportait l'expression de chacune des émotions et de chaque modalité. 25 sujets (10 femmes et 15 hommes) se sont portés volontaires pour participer à l'évaluation des séquences vidéo. Les sujets étaient des étudiants ou des employés de l'université, dont les âges variaient entre 23 et 62 ans.

Les sujets devaient tout d'abord mentionner leur âge, sexe et profession avant de répondre à trois questions à choix multiples : "Combien de spectacles de danse avez vous vu ?", "Diriez vous que vous êtes sensibles au comportement des autres d'une manière générale ?", et "Comment noteriez-vous votre empathie, c'est-à-dire votre capacité à reconnaître et comprendre son état émotionnel chez une personne ?". Une série de vidéos a ensuite été présentée à chaque sujet sur un ordinateur portable. Un discours prédéfini leur expliquait la marche à suivre. Pour chaque vidéo de la série, les sujets pouvaient reconnaître jusqu'à trois émotions différentes



(a) Le système complet constitué des applications Moven Studio, eMotion, ShadoZ et MARC



(b) Exemple de la cinquième modalité utilisée, combinant l'avatar fil de fer, son ombre expressive, et le visage expressif MARC

FIG. 61: Tests perceptifs : système mis en place et modalité combinant les différentes augmentations.

sur un tableau à choix forcé. L'émotion reconnue comme la plus intense devait être notée "1", "2" pour l'émotion d'intensité médiane, et "3" pour la moins intense. Les sujets pouvaient noter plusieurs émotions *ex-æquo* dans le cas d'intensités similaires. Dans ce cas, les sujets ne pouvaient reconnaître qu'une ou deux émotions. Les sujets étaient autorisés à regarder chaque vidéo plusieurs fois et n'avaient aucune contrainte de temps pour finir l'évaluation.

7.4.4 Résultats et discussion

Une analyse du χ^2 a été menée sur les résultats des évaluations par le LIMSI. La première analyse a été effectuée sur l'impact de l'émotion et de la modalité considérée dans l'attribution de l'émotion par l'utilisateur. Les résultats obtenus n'ont pas permis de confirmer nos hypothèses ; en effet, la reconnaissance des émotions simples ont été bonnes pour chacune des modalités présentées, avec des écarts insuffisants pour en tirer des conclusions. La séquence représentant une émotion complexe (joie+tristesse) a été mal reconnue dans la modalité de référence (vidéo) et la modalité réduisant l'information (l'avatar fil de fer seul). L'ajout d'augmentations (avatar et son ombre, MARC, et combinaison de l'avatar, son ombre et MARC) présente une meilleure perception de la part des sujets ; les trois modalités sont alors équivalentes. D'autres analyses ont été effectuées selon les différentes variables caractérisant les sujets, mais aucune variable réellement discriminante n'est apparue entre sexes, groupes d'âge et expérience des spectacles de danse.

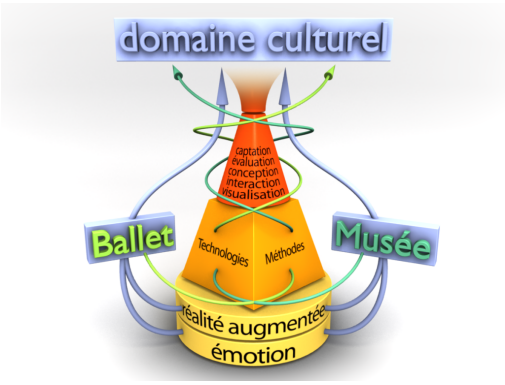
L'expérimentation décrite dans cette partie a été menée dans un but exploratoire. Ses résultats nous orientent sur la voie de l'augmentation d'émotions complexes. Les augmentations semblent en effet aider un observateur à percevoir les émotions complexes. Or, les émotions exprimées au cours d'un ballet sont bien souvent des émotions complexes, plus difficiles à reconnaître pour le public. Cette expérimentation et ses résultats nous engagent

donc à approfondir notre recherche et à développer de nouvelles augmentations en vue d'une expérimentation à plus grande échelle.

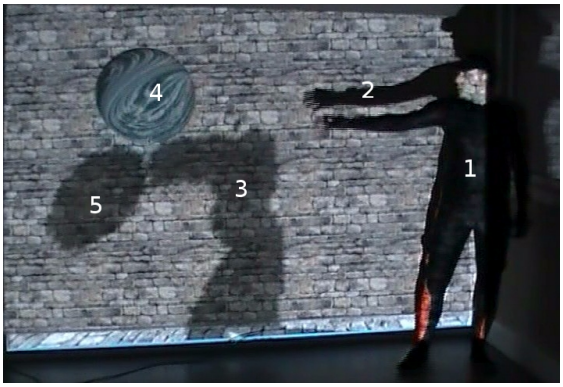
7.5 Conclusion

Nous avons décrit dans ce chapitre le cas applicatif de nos travaux en reconnaissance d'émotions : l'augmentation d'un spectacle de ballet, où les augmentations prennent comme paramètre les émotions exprimées par le danseur. Ce cas applicatif est un exemple d'utilisation du canevas logiciel eMotion à un domaine d'application particulier.

La collaboration avec les danseurs a été entamée dès le début de nos travaux, ce qui nous a permis d'établir un dialogue commun facilitant la collaboration. Ainsi, la danse a été, dans le cadre de nos recherches, un outil de recherche, de validation, d'ouverture à de nouvelles perspectives, et un domaine d'application pour nos travaux. Nous avons montré comment notre canevas logiciel eMotion a été intégré à une application pour augmenter un ballet. Notre canevas a été utilisé pour reconnaître les émotions que traduisent les mouvements d'un danseur.



(a) Schéma représentant les axes de recherche du projet CARE.



(b) Augmentations du logiciel ShadoZ. (1) danseur, (2) ombre réelle, (3) ombre virtuelle, (4) globe virtuel, (5) ombre du globe virtuel.

Fig. 62: Perspectives : le projet CARE et l'application ShadoZ

Nos premiers travaux et la collaboration avec d'autres laboratoires français ont permis le lancement du projet ANR-RIAM CARE (Cultural experience : Augmented Reality and Emotion), dont le but est de contribuer tant au niveau théorique qu'applicatif à l'intégration de l'interaction émotionnelle et de la réalité augmentée dans l'expérience culturelle (voir figure 62a). L'application ShadoZ a été développée dans le cadre de ce projet et permet, en s'appuyant sur les émotions reconnues par eMotion, de générer des augmentations visuelles à projeter sur scène. Les premières versions du logiciel ShadoZ ont visé à répondre aux attentes des chorégraphes de Malandain Ballet Biarritz et implémenter un ombre virtuelle expressive.

La dernière version du logiciel permet d'intégrer diverses augmentations et de les projeter sur la scène (voir figure 62b).

La collaboration avec les danseurs nous a permis de mettre en place des cas réels, hors laboratoire, de spectacles augmentés. Nous avons également expérimenté, de façon exploratoire, l'impact des augmentations sur la perception des émotions par le public.

L'utilisation de la reconnaissance d'émotions et de la réalité augmentée pour un spectacle de danse ouvre de nombreuses perspectives. Du point de vue de la reconnaissance d'émotions, la première perspective est l'extension de l'application à des caractéristiques d'expression émotionnelles plus proches de la danse que les caractéristiques générales utilisées actuellement dans notre système. Il convient d'approfondir les techniques de réalité augmentée pour définir les interactions entre le danseur et les éléments virtuels, la scénarisation du monde virtuel et sa présentation. Enfin, si la reconnaissance d'émotions offre au danseur une piste de recherche sur son propre corps et son mouvement dans l'espace, la réalité augmentée génère de nombreuses perspectives artistiques par la diversité des technologies d'affichage et l'absence de limitations dans la création de contenu graphique.

Conclusion

Les travaux présentés contribuent au domaine de l'ingénierie logicielle des systèmes de reconnaissance d'émotions au sein d'un système interactif sous deux formes complémentaires. En effet nos contributions à la fois conceptuelles et pratiques incluent un motif architectural conceptuel pour la reconnaissance d'émotions, la branche émotion, ainsi qu'une réalisation logicielle de cette architecture sous la forme d'un canevas générique et extensible pour la reconnaissance d'émotions basée sur les mouvements, eMotion.

Pour conclure ce manuscrit, nous rappelons nos contributions pour la conception logicielle des systèmes sensibles aux émotions En accord avec nos objectifs initiaux, ces contributions concernent les aspects conceptuels et pratiques de la conception logicielle. Nous identifions ensuite des perspectives à court terme basées sur les limites identifiées de notre travail, et enfin trois ouvertures précises à long terme que nous donnons à nos travaux.

Résumé des contributions

La branche émotion part des requis identifiés dans notre état de l'art et s'inspire des modèles existants pour les systèmes interactifs multimodaux, ce domaine bénéficiant déjà d'une solide assise en ingénierie logicielle. Elle est donc découpée en trois niveaux capture, analyse et interprétation reflétant les niveaux signal, caractéristique et décision classiquement rencontrés dans les systèmes existants. Nous avons montré son intégration dans des modèles de référence en IHM. En particulier, nous nous sommes attachés à l'instanciation des concepts existants en interaction multimodale (notamment les propriétés CARE de la multimodalité) au domaine de la reconnaissance d'émotions et traitons leur pouvoir génératif. Après avoir exploré le cadre conceptuel de la branche émotion, nous avons détaillé sa structure interne. Nous avons donc décrit l'ensemble des composants constituant la branche émotion ainsi que leurs spécifications. Nous nous sommes basés sur la définition retenue d'une émotion pour ramener le problème de la fusion des modalités à un problème de synchronisation et avons exposé notre solution. Le moteur de synchronisation que nous avons proposé travaille en arrière-plan des composants de la branche émotion et a pour rôle d'assurer la communication entre ces composants en synchronisant les flux de données au sein du système de reconnaissance. Enfin, nous avons en partie validé la branche émotion par la modélisation de travaux existants selon notre modèle.

Nous avons complété la validation de notre cadre conceptuel par la mise en œuvre logicielle d'un système de reconnaissance d'émotions basée sur les mouvements. L'architecture d'eMotion, faisant intervenir orthogonalement le modèle de la branche émotion pour le côté fonctionnel et le modèle PAC pour la hiérarchie de contrôle des agents, se traduit par une architecture implémentationnelle extensible, modifiable, non limitée à la reconnaissance des émotions par le mouvement. Le système eMotion complète de plus la validation de notre modèle conceptuel en explicitant le pouvoir génératif des propriétés CARE appliquées à la reconnaissance d'émotion. En effet, le canevas eMotion propose plusieurs combinaisons de modalités, en particulier l'équivalence, peu étudiée dans les systèmes de reconnaissance d'émotions. Le canevas eMotion implémente des caractéristiques de mouvements de travaux existants et s'appuie sur les analyses existantes pour son interprétation.

Enfin, eMotion a permis l'ouverture de notre sujet au domaine artistique par la danse et, en informatique, à celui de la réalité augmentée. En effet, le canevas eMotion est appliqué à la reconnaissance des émotions exprimées par un danseur. Ces émotions reconnues sont utilisées par un système de réalité augmentée afin de moduler les éléments virtuels projetés sur la scène du ballet. Dans la perception du public, les émotions exprimées par le danseur sont ainsi surlignées par le comportement des objets virtuels sur la scène.

Perspectives de développement

Nous entrevoyons pour nos travaux de multiples perspectives. Là encore, la dualité, conception et réalisation d'un système, apparaît dans nos propositions. Ces dernières s'organisent en deux parties : les extensions à court terme, et les prolongements ouvrant le sujet à plus long terme.

Extensions

Les perspectives à court terme que nous souhaitons développer impactent en priorité l'aspect pratique de nos travaux. Plusieurs perspectives sont liées au développement du canevas eMotion.

La première est le développement du moteur de synchronisation décrit au chapitre 5 (section 5.5, page 113), afin de mettre en œuvre des tests nous permettant de valider notre proposition ou à défaut, d'évaluer l'impact des limitations que nous lui avons identifiées. La deuxième est l'extension de la reconnaissance à de nouvelles caractéristiques et l'implémentation d'une interprétation des caractéristiques sur une fenêtre temporelle plus large. Cette extension de notre système fera l'objet d'une nouvelle expérimentation à concevoir et à mettre en œuvre. Une telle expérimentation est envisageable sans mettre à contribution le danseur : une nouvelle évaluation sur les vidéos enregistrées peut être mise en place et servir de nouvelle référence à la validation. Enfin, notre quatrième perspective est d'éprouver le caractère extensible du canevas logiciel eMotion aux trois niveaux d'abstraction. Dans le cadre du projet ANR CARE, nous envisageons de rajouter un suivi optique (vidéo ou infrarouge) afin d'améliorer le suivi de la position du danseur sur la scène. Au niveau analyse, nous avons entamé l'implémentation de nouvelles caractéristiques tirées de [143]. Enfin, au niveau Interprétation, nous envisageons

l'intégration d'un système tiers de reconnaissance au sein de l'application. Nous envisageons notamment l'intégration d'un système basé vidéo développé par le GIPSA.

Outre ces travaux concernant eMotion, nous souhaitons aussi enrichir notre application sur le ballet, en focalisant sur les présentations de l'émotion reconnue. En particulier, nous souhaitons concevoir d'autres augmentations que celles proposées actuellement afin d'enrichir le potentiel au niveau de l'interaction du danseur avec les éléments virtuels et d'approfondir le couplage entre cette interaction active et l'action des émotions reconnues sur ces éléments virtuels.

Prolongements

Tout au long de ce mémoire, nous avons identifié les enjeux liés aux systèmes de reconnaissance d'émotions. Tous n'ont pas été traités dans nos travaux et pourraient constituer des pistes de recherche. Parmi ces enjeux et pour étendre nos travaux sur le long terme, nous considérons trois pistes de recherche précises, l'une conceptuelle, l'autre pratique et enfin une ouverture vers un autre domaine applicatif : la première est l'extension de notre cadre architectural conceptuel au cas du bouclage analyse-interprétation, la seconde la définition d'un atelier logiciel pour la branche émotion et la troisième l'exploitation de l'émotion pour la sculpture dansée.

Une première perspective vise l'extension de notre cadre conceptuel, la branche émotion. Au sein de la branche émotion, la possibilité et l'impact d'un bouclage entre les niveaux interprétation et analyse reste à étudier. En effet, un a priori basé sur des interprétations passées peut influer sur la perception des caractéristiques (par exemple un rictus pris pour un sourire). Dans la branche émotion, cette influence n'est pas représentée explicitement, bien qu'elle puisse être implémentée dans l'interprétation. Il s'agit ici de donner une dimension temporelle à la reconnaissance d'émotions qui dans la branche émotion n'est pas explicite.

Une deuxième perspective vise à définir un outil d'assemblage basé sur notre modèle conceptuel pour définir un système de reconnaissance d'émotions. Générateur de systèmes de reconnaissances d'émotions, cet outil à concevoir serait une contribution plus générique que notre canevas eMotion puisqu'il permettrait aux concepteurs de définir un système de reconnaissance par assemblage de composants logiciels. Nous voyons dans cette contribution la différence faite en IHM entre un squelette d'interface qui nécessite de programmer et un générateur d'interfaces. Pour démarrer ce travail, nous envisageons d'étudier la plateforme open source OpenInterface qui permet de spécifier par assemblage de composants une interaction multimodale. Basé sur cette plateforme, les enjeux seraient alors de peupler la plateforme de composants organisés selon les niveaux fonctionnels de notre cadre conceptuel et d'intégrer notre moteur de synchronisation au sein du noyau sous-jacent d'OpenInterface.

Enfin nous envisageons d'étendre nos travaux applicatifs par un approfondissement du lien que nous avons mis en place entre danse, reconnaissance d'émotions et réalité augmentée. Nous nous inspirons des travaux de Schkolne *et al.* [138] sur le dessin de surface à main levée pour imaginer une forme de sculpture dansée, où le danseur pourrait faire apparaître et sculpter

de la matière virtuelle au gré de ses mouvements. La reconnaissance des émotions du danseur peut ainsi être utilisée comme forme d'interaction avec le monde virtuel, l'émotion modifiant les propriétés (forme, couleur, etc.) du "pinceau virtuel". Une fois un objet sculpté, nous envisageons également d'utiliser l'émotion exprimée pour modifier les propriétés physiques qui lui sont attribués : par exemple, sa densité ou son élasticité. L'émotion peut impacter de la même manière les propriétés du monde virtuel comme la direction et la force de la gravité. Cette ouverture de nos travaux à son application au cas des sculptures dansées s'appuierait sur une collaboration initiée par nos travaux avec Malandain Ballet Biarritz.

Nous finissons ce mémoire par une question que nous n'avons pas abordée dans nos travaux mais qui est importante et intimement liée au domaine de la reconnaissance d'émotions. Qu'en est-il de l'éthique ? Cette question est déjà importante en l'état actuel de la recherche, notamment pour l'évaluation de systèmes de reconnaissance. Jusqu'à quel point peut-on manipuler un sujet pour le faire vivre les expériences émotionnelles prévues par le protocole de test ? Il paraît évident que mettre en contact un phobique à l'objet de sa phobie pour provoquer la peur dépasse les limites éthiques. En France, la CNIL (Commission Nationale de l'Informatique et des Libertés[9]) peut valider ou refuser les protocoles expérimentaux visant à déclencher des émotions chez les sujets. Nos émotions sont parmi les expériences les plus intimes que nous puissions avoir. Durant nos présentations, nous avons souvent observés un malaise chez les non-informaticiens testant notre système. L'aspect "magique" de toute nouvelle technologie aidant, ces utilisateurs, peu familiers du domaine et en particulier de ses limitations, croyaient être confrontés à un système capable de révéler aux personnes présentes leur état émotionnel. Ce malaise est *a priori* dû à la confrontation sans préparation avec le système ; Reynolds et Picard montrent que prévenir les utilisateurs via un contrat préalable quelles émotions seront reconnues par la machine permet de lever en grande partie le sentiment de violation de vie privée [122]. La plupart des personnes ayant testé notre système de reconnaissance ont également immédiatement relevé le potentiel de surveillance d'un tel système, soulignant l'appréhension que peut susciter la reconnaissance automatique des émotions.

[9] http ://www.cnil.fr

Bibliographie

[1] S. Abrilian, L. Devillers, S. Buisine, and J.C. Martin. Emotv1 : Annotation of real-life emotions for the specification of multimodal affective interfaces. In *11th Int. Conf. on Human-Computer Interaction (HCII'05)*. Lawrence Erlbaum Associates, Inc, 2005.

[2] H. Ahn and R.W. Picard. Affective-cognitive learning and decision making : A motivational reward framework for affective agents. In *Proceedings of the 1st International Conference on Affective Computing and Intelligent Interaction (ACII'05)*, pages 863–866. Springer, 2005.

[3] N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences : A meta-analysis. *Psychological Bulletin*, 111(2) :256–274, 1992.

[4] F. Aznar, M. Sempere, M. Pujol, and R. Rizo. Bayesian Emotions : Developing an Interface for Robot/Human Communication. *Lecture Notes in Computer Science : Advances in Artificial Intelligence*, 3673 :507–517, 2005.

[5] T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, and S. Kollias. Emotion analysis in man-machine interaction systems. *Lecture Notes in Computer Science : Machine Learning for Multimodal Interaction*, 3361 :318–328, 2004.

[6] L. Bass. A metamodel for the runtime architecture of an interactive system : the UIMS tool developers workshop. *SIGCHI Bull.*, 24(1) :32–37, 1992.

[7] L. Bass, C. Buhman, S. Comella-Dorda, F. Long, and J. Robert. Volume 1 : Market Assessment of Component-Based Software Engineering. Technical report, Carnegie Mellon University, Software Engineering institue, Mai 2000. Technical note CMU/SEI-2001-TN-007.

[8] F. Birren. *Color psychology and color therapy : a factual study of the influence of color on human life*. Kessinger Publishing, 2006.

[9] H. Bloch, R. Chemama, A. Gallo, P. Leconte, J.F. Le Ny, J. Postel, S. Moscovici, M. Reuchlin, and E. Vurpillot. *Grand dictionnaire de la psychologie*. Larousse, 1973.

[10] R.A. Bolt. "put-that-there" : Voice and gesture at the graphics interface. In *SIGGRAPH '80 : Proceedings of the 7th annual conference on Computer graphics and interactive techniques*, pages 262–270, New York, NY, USA, 1980. ACM.

[11] R.T. Boone and J.G. Cunningham. Children's decoding of emotion in expressive body movement : The development of cue attunement. *Developmental Psychology*, 34 :1007–1016, 1998.

[12] R.T. Boone and J.G. Cunningham. Children's expression of emotional meaning in music through expressive body movement. *Journal of Nonverbal Behavior*, 25(1) :21–41, 2001.

[13] S. Bottecchia, J.M. Cieutat, C. Merlo, and J.P. Jessel. A new AR interaction paradigm for collaborative teleassistance system : the POA. *International Journal on Interactive Design and Manufacturing*, 3(1) :35–40, 2009.

[14] J. Bouchet. *Ingénierie de l'interaction multimodale en entrée Approche à composants ICARE*. PhD thesis, Université Joseph Fourier, Grenoble I, 2006.

[15] J. Bouchet, L. Nigay, and D. Balzagette. ICARE : Approche à composants pour l'interaction multimodale. *Actes des Premières Journées Francophones : Mobilité et Ubiquité 2004*, pages 36–43, 2004.

[16] W. Burleson, RW Picard, K. Perlin, and J. Lippincott. A platform for affective agent research. In *Workshop on Empathetic Agents, International Conference on Autonomous Agents and Multiagent Systems*, 2004.

[17] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces (ICMI'04)*, pages 205–211, New York, NY, USA, 2004. ACM.

[18] A. Camurri, R. Chiarvetto, A. Coglio, M. Di Stefano, C. Liconte, A. Massari, C. Massucco, D. Murta, S. Nervi, G. Palmieri, et al. Toward Kansei Information Processing in music/dance interactive multimodal environments. In *Proceedings of the Italian Association for Musical Informatics (AIMI) International Workshop - Kansei : The Technology of Emotion*, pages 74–78, 1997.

[19] A. Camurri, P. Coletta, A. Massari, B. Mazzarino, M. Peri, M. Ricchetti, A. Ricci, and G. Volpe. Toward real-time multimodal processing : EyesWeb 4.0. In *Proceedings of Artificial Intelligence and the Simulation of Behaviour (AISB) convention : Motion, Emotion and Cognition*, 2004.

[20] A. Camurri, B. Mazzarino, R. Trocca, and G. Volpe. Real-time analysis of expressive cues in human movement. *Proc. CAST01, GMD, St. Augustin-Bonn*, 2001.

[21] A. Camurri, B. Mazzarino, and G. Volpe. Analysis of Expressive Gesture : The Eyes Web Expressive Gesture Processing Library. *Lecture notes in computer science*, pages 460–467, 2004.

[22] A. Camurri, M. Ricchetti, and R. Trocca. Eyesweb - toward gesture and affect recognition in dance/music interactive systems. In *Proceedings of the 1999 IEEE International Conference on Multimedia Computing and Systems (ICMCS'99)*, volume 1, page 9643, Los Alamitos, CA, USA, 1999. IEEE Computer Society.

[23] A. Camurri and R. Trocca. Analysis of expressivity in movement and dance. *Proceedings of CIM-2000, L'Aquila, AIMI*, 2000.

[24] A. Cardon, J.C. Campagne, and M. Camus. A self-adapting system generating intentional behavior and emotions. *Lecture notes in computer science*, 3825 :33, 2006.

[25] G. Castellano. *Movement expressivity analysis in affective computers : from recognition to expression of emotion*. PhD thesis, University of Genova, 2008.

[26] G. Castellano, L. Kessous, and G. Caridakis. Multimodal emotion recognition from expressive faces, body gestures and speech. In *Proc. of the Doctoral Consortium of 2nd International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2007. Lisbonne, Portugal.

[27] G. Castellano, M. Mortillaro, A. Camurri, G. Volpe, and K. Scherer. Automated analysis of body movement in emotionally expressive piano performances. *Music Perception*, 26(2) :103–119, 2008.

[28] G. Chanel, J. Kronegg, D. Grandjean, and T. Pun. Emotion assessment : Arousal evaluation using EEG's and peripheral physiological signals. *Lecture Notes in Computer Science*, 4105 :530, 2006.

[29] D. Chi, M. Costa, L. Zhao, and N. Badler. The emote model for effort and shape. In *SIGGRAPH '00 : Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 173–182, New York, NY, USA, 2000. ACM Press/Addison-Wesley Publishing Co.

[30] A. Clay, M. Courgeon, N. Couture, E. Delord, C. Clavel, and J.-C. Martin. Expressive virtual modalities for augmenting the perception of affective movements. In *AFFINE '09 : Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots*, pages 1–7, New York, NY, USA, 2009. ACM.

[31] A. Clay, N. Couture, and G. Domenger. Reconnaissance d'émotions : application à la danse. conférence dansée à Anglet, France, 2008.

[32] A. Clay, N. Couture, and G. Domenger. Capture d'émotions et reconnaissance par la gestuelle. conférence dansée interactive, Workshop ERGOIA 2009 à Paris, France, 2009.

[33] A. Clay, N. Couture, G. Domenger, E Delord, and A. Reumaux. Improvisation dansée augmentée en appartement. Festival des Ethiopiques 2009 à Bayonne, France, 2009.

[34] Alexis Clay, Nadine Couture, and Laurence Nigay. Engineering affective computing : a unifying software architecture. In *Proceedings of the 3rd IEEE International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009. (ACII'09)*, pages 1–6, 2009.

[35] Alexis Clay, Nadine Couture, and Laurence Nigay. Towards Emotion Recognition in Interactive Systems : Application to a Ballet Dance Show. In ASME, editor, *Proceedings of the ASME/AFM 2009 World Conference on Innovative Virtual Reality (WinVR2009) World Conference on Innovative Virtual Reality (WinVR'09)*, volume 2009, pages #704, pp. 19–24, Châlon-sur-Saône France, 2009.

[36] Alexis Clay, Elric Delord, Nadine Couture, and Gaël Domenger. Augmenting a Ballet Dance Show Using the Dancer's Emotion : Conducting Joint Research in Dance and Computer Science. In *Arts and Technology*, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, pages 148–156. Springer Berlin Heidelberg, 2010. Arts and Technology First International Conference, ArtsIT 2009, Yi-Lan, Taiwan, September 24-25, 2009, Revised Selected Papers.

[37] M. Coulson. Attributing emotion to static body postures : Recognition accuracy, confusions, and viewpoint dependence. *Journal of nonverbal behavior*, 28(2) :117–139, 2004.

[38] J. Coutaz. PAC, an object oriented model for dialog design. In H.J. Bullinger and B. (eds.) Shackel, editors, *Proceedings of the 2nd IFIP International Conference on Human-Computer Interaction (INTERACT 87)*, volume 87, pages 431–436, 1987.

[39] J. Coutaz and L. Nigay. *Architecture logicielle conceptuelle des systèmes interactifs*, pages 207–246. Hermes Publ., 2001. Chapitre 7 : Analyse et Conception de l'Interaction Homme-Machine dans les systèmes d'information, Kolski (Ed.).

[40] J. Coutaz, L. Nigay, D. Salber, A. Blandford, J. May, and R. Young. Four easy pieces for assessing the usability of multimodal interaction : the CARE properties. In K. (Ed.) Nordby, editor, *Proceedings of IFIP TC13 Fifth International Conference on Human-Computer Interaction (INTERACT'95)*, volume 95, pages 115–120, 1995.

[41] N. Couture and S. Minel. Tactimod dirige et oriente un piéton. In *Proceedings of Ubimob'06*, pages 9–16. IEEE France / ACM France, 09 2006.

[42] N. Couture, G. Rivière, and P. Reuter. *The Engineering of Mixed Reality Systems*, chapter Tangible Interaction in Mixed Reality Systems (chapitre 5). Dubois E., Gray P. and Nigay L. (Eds.), 2009. à paraître.

[43] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and JG Taylor. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1) :32–80, 2001.

[44] A.R. Damasio. *Descartes'error*. Avon Books New York, 1995.

[45] C. Darwin, P. Ekman, and P. Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 2002. écrit par Darwin, C. première publication en 1872.

[46] M. DeMeijer. The contribution of general features of body movement to the attribution of emotions. *Journal of Nonverbal Behavior*, 13(4) :247–268, 1989.

[47] R. Descartes and G. Rodis-Lewis. *Les passions de l'âme*. Vrin, 1988.

[48] E. Dubois. *Chirurgie Augmentée, un cas de Réalité Augmentée. Conception et réalisation centrées sur l'utilisateur*. PhD thesis, Laboratoire de Communication Langagière et Interaction Personne-Système (IMAG), Université Joseph Fourier, 2001. Thèse de doctorat Informatique, 275 pages.

[49] P. Ekman. Basic emotions. *Handbook of cognition and emotion*, pages 45–60, 1999.

[50] P. Ekman, W. Friesen, and J. Hager. *Facial Action Coding System : the manual*. A Human Face, 2002. HTML demonstration version, http ://www.face-and-emotion.com/dataface/facs/manual/TitlePage.html.

[51] P. Ekman and W.V. Friesen. *The repertoire of nonverbal behavior*. Mouton de Gruyter, 1969.

[52] P. Ekman and W.V. Friesen. Facial Action Coding System (FACS) : A technique for the measurement of facial action. *Palo Alto, CA : Consulting*, 1978.

[53] P. Ekman and W.V. Friesen. A new pan-cultural facial expression of emotion. *Motivation and Emotion*, 10(2) :159–168, 1986.

[54] P. Ekman and W.V. Friesen. Hand movements. *Communication Theory*, 22(4) :273, 2007.

[55] P. Ekman, W.V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W.A. LeCompte, T. Pitcairn, P.E. Ricci-Bitti, et al. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, 53(4) :712–717, 1987.

[56] R. El Kaliouby and P. Robinson. Generalization of a vision-based computational model of mind-reading. In *Proceedings of the First International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 582–589. Springer, 2005. Beijing, Chine.

[57] M.S. El-Nasr, J. Yen, and T.R. Ioerger. Flame—fuzzy logic adaptive model of emotions. *Autonomous Agents and Multi-Agent Systems*, 3(3) :219–257, 2000.

[58] F. Février, E. Jamet, G. Rouxel, V. Dardier, and G. Breton. Induction d'émotions pour la motion capture dans une situation de vidéo-conversation. In *Proceedings du Workshop sur les Agents Conversationnels Animés (WACA)*, 2006. Toulouse, France.

[59] B.R. Gaines. Modeling and forecasting the information sciences. *Inf. Sci.*, 57-58 :3–22, 1991.

[60] P. Gebhard. Alma : a layered model of affect. In *AAMAS '05 : Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pages 29–36, New York, NY, USA, 2005. ACM.

[61] K. Ghamen and A. Caplier. Estimation of Anger, Sadness and Fear Expression Intensity based on the Belief Theory. In *Proceedings ACIT*, pages –, Hammamet Tunisie, 2008.

[62] J. Gratch and S. Marsella. Tears and fears : modeling emotions and emotional behaviors in synthetic agents. In *AGENTS '01 : Proceedings of the fifth international conference on Autonomous agents*, pages 278–285, New York, NY, USA, 2001. ACM.

[63] W.E.L. Grimson, G.J. Ettinger, S.J. White, T. Lozano-Perez, W.M. Wells III, and R. Kikinis. An automatic registration method for frameless stereotaxy, imageguided surgery, and enhanced reality visualization. *IEEE Transactions on medical imaging*, 15(2) :129–140, 1996.

[64] H. Gunes, M. Piccardi, and T. Jan. Face and body gesture recognition for a vision-based multimodal analyzer. In *VIP '05 : Proceedings of the Pan-Sydney area workshop on Visual information processing*, pages 19–28, Darlinghurst, Australia, Australia, 2004. Australian Computer Society, Inc.

[65] A. Guye-Vuilleme and D. Thalmann. A high-level architecture for believable social agents. *Virtual Reality*, 5(2) :95–106, 2000.

[66] Z. Hammal, L. Couvreur, A. Caplier, and M. Rombaut. Facial expression classification : An approach based on the fusion of facial deformations using the transferable belief model. *Int. J. Approx. Reasoning*, 46(3) :542–567, 2007.

[67] J. Healey and R. Picard. Digital processing of affective signals. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 6, 1998.

[68] J. Healey, J. Seger, and R. Picard. Quantifying driver stress : developing a system for collecting and processing bio-metric signals in natural situations. *Biomedical sciences instrumentation*, 35 :193–198, 1999.

[69] J. Hodgson. *Mastering movement : the life and work of Rudolf Laban.* Theatre Arts Books, 2001.

[70] S. Ioannou, L. Kessous, G. Caridakis, K. Karpouzis, V. Aharonson, and S. Kollias. Adaptive on-line neural network retraining for real life multimodal emotion recognition. *Lecture Notes in Computer Science*, 4131 :81, 2006.

[71] H. Ishii, S. Ren, and P. Frei. Pinwheels : visualizing information flow in an architectural space. In *Proceedings of the 2001 Conference on Human Factors in Computing Systems*, pages 111–112. ACM New York, NY, USA, 2001.

[72] C.E. Izard. *The face of emotion.* Appleton-Century-Crofts New York, 1971.

[73] A. Jaimes and N. Sebe. Multimodal human-computer interaction : A survey. *Comput. Vis. Image Underst.*, 108(1-2) :116–134, 2007.

[74] X. Jin and Z. Wang. An Emotion Space Model for Recognition of Emotions in Spoken Chinese. In *Proceedings of the First International Conference on Affective Computing and Intelligent Interaction (ACII)*, page 397. Springer, 2005. Beijing, Chine.

[75] P. Kaiser. Hand-drawn spaces. In *SIGGRAPH '98 : ACM SIGGRAPH 98 Electronic art and animation catalog*, page 134, New York, NY, USA, 1998. ACM.

[76] A. Kapoor, R.W. Picard, and Y. Ivanov. Probabilistic combination of multiple modalities to detect interest. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)*, volume 3, pages 969–972, Washington, DC, USA, 2004. IEEE Computer Society.

[77] K. Karpouzis, A. Raouzaiou, and S. Kollias. 'Moving'avatars : emotion synthesis in virtual worlds. *Human-Computer Interaction : Theory and Practice*, 2 :503–507, 2003.

[78] A. Kendon. *Gesture : Visible action as utterance*. Cambridge Univ Pr, 2004.

[79] A. Kendon and W.S. Language. How gestures can become like words. *Crosscultural perspectives in nonverbal communication*, pages 131–141, 1988.

[80] E. Keogh and C.A. Ratanamahatana. Exact indexing of dynamic time warping. *Knowledge and Information Systems*, 7(3) :358–386, 2005.

[81] K.H. Kim, S.W. Bang, and S.R. Kim. Emotion recognition system using short-term monitoring of physiological signals. *Medical and biological engineering and computing*, 42(3) :419–427, 2006.

[82] D. Kirsch. The Sentic Mouse : Developing a tool for measuring emotional valence. Technical report, MIT Media Laboratory Perceptual Computing Section, 1997.

[83] J. Klein, Y. Moon, and RW Picard. This computer responds to user frustration : Theory, design, and results. *Interacting with computers*, 14(2) :119–140, 2002.

[84] B. Kort, R. Reilly, and RW Picard. An affective model of interplay between emotions and learning : Reengineering educational pedagogy-building a learning companion. In *Proceedings of the 2001 IEEE International Conference on Advanced Learning Technologies*, pages 43–46, 2001.

[85] G. E. Krasner and S. T. Pope. A cookbook for using the model-view controller user interface paradigm in Smalltalk-80. *J. Object Oriented Program.*, 1(3) :26–49, 1988.

[86] D. Lalanne, L. Nigay, P. Palanque, P. Robinson, J. Vanderdonckt, and J.-F. Ladry. Fusion engines for multimodal input : a survey. In *Proceedings of the 2009 international conference on Multimodal interfaces (ICMI-MLMI'09)*, pages 153–160, New York, NY, USA, 2009. ACM.

[87] M. Lamolle, M. Mancini, C. Pelachaud, S. Abrilian, J.C. Martin, and L. Devillers. Contextual factors and adaptative multimodal human-computer interaction : multi-level specification of emotion and expressivity in embodied conversational agents. *Lecture Notes in Computer Science*, 3554 :225–239, 2005.

[88] C. Lisetti and G. Bastard. MAUI : a Multimodal Affective User Interface Sensing User's Emotions based on Appraisal Theory-Questions about Facial Expressions, 2004.

[89] C.L. Lisetti. Le paradigme MAUI pour des agents multimodaux d'interface homme-machine socialement intelligents. *Revue d'Intelligence Artificielle, Numero Special sur les Interactions Emotionnelles*, 20(4-5) :583–606, 2006.

[90] O. Losson. *Modélisation du geste communicatif et réalisation d'un signeur virtuel de phrases en langue des signes française*. PhD thesis, Université de Lille 1, 2000.

[91] K. Lyons, M. Gandy, and T. Starner. Guided by voices : An audio augmented reality system. In *International Conference on Auditory Display*. Citeseer, 2000.

[92] M. Mallem and D. Roussel. Réalitée augmentée : Principes, technologies et applications. Site Web : Techniques de l'ingénieur (http ://www.techniques-ingenieur.fr/), 2008. disp. à l'adresse : http ://www.techniques-ingenieur.fr/book/te5920/realite-augmentee.html.

[93] B. Mansoux, L. Nigay, and J. Troccaz. Output multimodal interaction : The case of augmented surgery. In *Proceedings of HCI 2006, Human Computer Interaction, People and Computers XX, The 20th BCS HCI Group conference in co-operation with ACM (London, UK)*, pages 177–192, 2006.

[94] J.C. Martin. TYCOON : Theoretical framework and software tools for multimodal interfaces. *Intelligence and Multimodality in Multimedia interfaces*, 1998.

[95] A. Mehrabian. Outline of a general emotion-based theory of temperament. *Explorations in temperament : International perspectives on theory and measurement*, pages 75–86, 1991.

[96] A. Mehrabian. Framework for a comprehensive description and measurement of emotional states. *Genetic, Social, and General Psychology Monographs*, 121(3) :339, 1995.

[97] A. Mehrabian. Communication without words. *Communication Theory*, page 193, 2007.

[98] A. Mehrabian and JA Russell. The basic emotional impact of environments. *Perceptual and motor skills*, 38(1) :283–301, 1974.

[99] P Milgram and F Kishino. A Taxonomy of Mixed Reality Visual Displays. *IEICE Transactions on Information Systems*, E77-D(12) :1321–1329, 1994.

[100] J. Montepare, E. Koff, D. Zaitchik, and M. Albert. The use of body movements and gestures as cues to emotions in younger and older adults. *Journal of Nonverbal Behavior*, 23(2) :133–152, 1999.

[101] D. Morris, H. Friedhoff, Y. Dubois, and Y. Dubois. *La clé des gestes*. B. Grasset, 1978.

[102] S. Mota and R.W. Picard. Automated posture analysis for detecting learner's interest level. In *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03. Conference on*, volume 5, page 49, 2003.

[103] L. Nigay. *Conception et modélisation logicielles des systèmes interactifs : application aux interfaces multimodales*. PhD thesis, Laboratoire de Génie Informatique (IMAG), Université Joseph Fourier, 1994. Thèse de doctorat Informatique, 315 pages.

[104] L. Nigay and J. Coutaz. A design space for multimodal systems : concurrent processing and data fusion. In *CHI '93 : Proceedings of the INTERACT '93 and CHI '93 conference on Human factors in computing systems*, pages 172–178, New York, NY, USA, 1993. ACM.

[105] L. Nigay and J. Coutaz. A generic platform for addressing the multimodal challenge. In *CHI '95 : Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 98–105, New York, NY, USA, 1995. ACM Press/Addison-Wesley Publishing Co.

[106] L. Nigay and J. Coutaz. Espaces conceptuels pour l'interaction multimédia et multimodale. *TSI, spécial Multimédia et Collecticiel, AFCET & Hermes Publ.*, 15(9) :1195–1225, 1996.

[107] C.S. O'Bryne, L. Cañamero, and J.C Murray. The importance of the body in affect-modulated action selection : A case study comparing proximal versus distal perception in a prey-predator scenario. In *Proceedings of the 2009 IEEE International Conference on Affective Computing and Intelligent Interaction*, 2009. à paraître.

[108] M. Paleari and C. L. Lisetti. Toward multimodal fusion of affective cues. In *HCM '06 : Proceedings of the 1st ACM international workshop on Human-centered multimedia*, pages 99–108, New York, NY, USA, 2006. ACM.

[109] M. Pantic and I. Patras. Dynamics of facial expression : Recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 36(2) :433–449, 2006.

[110] M. Pantic, N. Sebe, J. F. Cohn, and T. Huang. Affective multimodal human-computer interaction. In *MULTIMEDIA '05 : Proceedings of the 13th annual ACM international conference on Multimedia*, pages 669–676, New York, NY, USA, 2005. ACM.

[111] S. Pasquariello and C. Pelachaud. Greta : A simple facial animation engine. In *6th Online World Conference on Soft Computing in Industrial Appications, Session on Soft Computing for Intelligent 3D Agents*, 2001.

[112] C. Peter, E. Ebert, and H. Beikirch. A wearable multi-sensor system for mobile acquisition of emotion-related physiological data. In *Proceedings of the 1st International Conference on Affective Computing and Intelligent Interaction, Beijing*, pages 691–698. Springer, 2005.

[113] G. E. Pfaff, editor. *User Interface Management Systems*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1985.

[114] R.W. Picard. *Affective computing*. MIT press, 1997.

[115] R.W. Picard. Human-computer coupling. *Proceedings of the IEEE*, 86(8) :1803–1807, 1998.

[116] R.W. Picard. Building HAL : Computers that sense, recognize, and respond to human emotion. In Rogowitz B.E. and Pappas T.N. (Eds), editors, *Proceedings of the 2001 SPIE Conference on Human Vision and Electronic Imaging*, volume 4299, pages 518–523, 2001.

[117] R.W. Picard and J. Healey. Affective wearables. *Personal Technologies*, 1(4) :231–240, 1997.

[118] R.W. Picard, S. Papert, W. Bender, B. Blumberg, C. Breazeal, D. Cavallo, T. Machover, M. Resnick, D. Roy, and C. Strohecker. Affective learning—a manifesto. *BT Technology Journal*, 22(4) :253–269, 2004.

[119] R.W. Picard and W. Rosalind. Toward agents that recognize emotion. *VIVEK-BOMBAY-*, 13(1) :3–13, 2000.

[120] R. Plutchik and H.R. Conte. *Circumplex models of personality and emotions*. American Psychological Association Washington, DC, 1997.

[121] F.E. Pollick, H.M. Paterson, A. Bruderlin, and A.J. Sanford. Perceiving affect from arm movement. *Cognition*, 82(2) :51–61, 2001.

[122] C. Reynolds and R. Picard. Affective sensors, privacy, and ethical contracts. In *CHI '04 : CHI '04 extended abstracts on Human factors in computing systems*, pages 1103–1106, New York, NY, USA, 2004. ACM.

[123] C. Reynolds and R.W. Picard. Designing for affective interactions. In *Proceedings of the 9th International Conference on Human-Computer Interaction*, 2001.

[124] G. Rivière, N. Couture, and Jurado F. Tangible user interfaces for geosciences. *Society of Exploration Geophysicists, Technical Program Expanded Abstracts*, 28(1), 2009. à paraître.

[125] J.A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6) :1161–1178, 1980.

[126] J.A. Russell and A. Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3) :273–294, 1977.

[127] J. Scheirer, R. Fernandez, and R. W. Picard. Expression glasses : a wearable device for facial expression recognition. In *CHI '99 : CHI '99 extended abstracts on Human factors in computing systems*, pages 262–263, New York, NY, USA, 1999. ACM.

[128] J. Scheirer and R. Picard. Affective objects. Technical report, MIT Media Laboratory Perceptual Computing Section, 2000.

[129] K.R. Scherer. *On the nature and function of emotion : a component process approach.* Approaches to emotion. NJ : Erlbaum, Hillsdale, k.r. scherer and p. ekman (eds.) edition, 1984.

[130] K.R. Scherer. Adding the affective dimension : A new look in speech analysis and synthesis. In *Fourth International Conference on Spoken Language Processing*, 1996.

[131] K.R. Scherer. Emotions as episodes of subsystem synchronization driven by nonlinear appraisal processes. *Emotion, development, and self-organization : Dynamic systems approaches to emotional development*, pages 70–99, 2000.

[132] K.R. Scherer. Psychological models of emotion. *The neuropsychology of emotion*, pages 137–162, 2000.

[133] K.R. Scherer. Feelings integrate the central representation of appraisal-driven response organization in emotion. In *Feelings and emotions : The Amsterdam symposium*, pages 136–157, 2004.

[134] K.R. Scherer. Which emotions can be induced by music ? what are the underlying mechanisms ? and how can we measure them ? *Journal of New Music Research*, 33(3) :239–251, 2004.

[135] K.R. Scherer, R. Banse, H.G. Wallbott, and T. Goldbeck. Vocal cues in emotion encoding and decoding. *Motivation and Emotion*, 15(2) :123–148, 1991.

[136] K.R. Scherer et al. HUMAINE deliverable D3c : Preliminary plans for exemplars : Theory, 2004.

[137] K.R. Scherer and H.G. Wallbott. Analysis of nonverbal behavior. *Handbook of discourse analysis*, 2 :199–230, 1985.

[138] S. Schkolne, M. Pruett, and P. Schröder. Surface drawing : creating organic 3d shapes with the hand and tangible tools. In *Proceedings of the 2001 SIGCHI conference on Human factors in computing systems (CHI'01)*, pages 261–268, New York, NY, USA, 2001. ACM.

[139] N. Sebe, I. Cohen, and T.S. Huang. Multimodal emotion recognition. *Handbook of Pattern Recognition and Computer Vision*, pages 981–256, 2005.

[140] M. Serrano and L. Nigay. Temporal aspects of care-based multimodal fusion : from a fusion mechanism to composition components and woz components. In *Proceedings of the 2009 international conference on Multimodal interfaces (ICMI-MLMI '09)*, pages 177–184, New York, NY, USA, 2009. ACM.

[141] P. Valdez and A. Mehrabian. Effects of color on emotions. *Journal of experimental psychology. General*, 123(4) :394–409, 1994.

[142] F. Vernier. *La multimodalité en sortie et son application à la visualisation de grandes quantités d'information*. PhD thesis, Université Joseph Fourier, Grenoble I, 2001.

[143] G. Volpe. *Computational models of expressive gesture in multimedia systems*. PhD thesis, University of Genova, 2003.

[144] E. Vyzas and R. W. Picard. Offline and Online Recognition of Emotion Expression from Physiological Data. In *Emotion-Based Agent Architectures Workshop Notes, Internationall Conference on Autonomous Agents*, pages 135–142, 1999.

[145] E. Vyzas and R.W. Picard. Affective pattern classification. *Emotional and Intelligent : The Tangled Knot of Cognition*, pages 176–182, 1998.

[146] J. H. Walker, L. Sproull, and R. Subramani. Using a human face in an interface. In *CHI '94 : Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 85–91, New York, NY, USA, 1994. ACM.

[147] H.G. Wallbott. Bodily expression of emotion. *European Journal of Social Psychology*, 28(6) :879–896, 1998.

[148] H.G. Wallbott and K.R. Scherer. Cues and channels in emotion recognition. *Journal of Personality and Social Psychology*, 51(4) :690–699, 1986.

[149] N. Wang and SC Marsella. Introducing EVG : An Emotion Evoking Game. *Lecture Notes in Computer Science*, 4133 :282, 2006.

[150] G. Welch and E. Foxlin. Motion tracking : No silver bullet, but a respectable arsenal. *IEEE Computer Graphics and Applications*, 22(6) :24–38, 2002.

[151] J.-J. Wong and S.-Y. Cho. Facial emotion recognition by adaptive processing of tree structures. In *SAC '06 : Proceedings of the 2006 ACM symposium on Applied computing*, pages 23–30, New York, NY, USA, 2006. ACM.

[152] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang. A survey of affect recognition methods : Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1) :39–58, 2009.

[153] Z. Zeng, J. Tu, M. Liu, T. Zhang, N. Rizzolo, Z. Zhang, T. S. Huang, D. Roth, and S. Levinson. Bimodal hci-related affect recognition. In *ICMI '04 : Proceedings of the 6th international conference on Multimodal interfaces*, pages 137–143, New York, NY, USA, 2004. ACM.

[154] Zhihong Zeng, Yuxiao Hu, Ming Liu, Yun Fu, and Thomas S. Huang. Training combination strategy of multi-stream fused hidden markov model for audio-visual affect recognition. In *MULTIMEDIA '06 : Proceedings of the 14th annual ACM international conference on Multimedia*, pages 65–68, New York, NY, USA, 2006. ACM.