Preuve de la compatibilité des modèles de la reconnaissance d'émotions et de la multimodalité par plongement

Alexis Clay[*]
[*]ESTIA-Technopole Izarbel, 64210 Bidart, France
a.clay@estia.fr

Nadine Couture[*,†]
[†]LaBRI, UMR 5800, Université de Bordeaux, France
n.couture@estia.fr

Laurence Nigay
LIG, UMR 5217, Université Joseph Fourier, France
Laurence.Nigay[redacted]

RESUME
Le domaine de la reconnaissance d'émotions atteint un stade de maturité où commence à émerger un besoin en termes de modèles pour la conception. Partant de cette constatation, nous proposons la branche émotion, un modèle d'architecture basé composants pour la reconnaissance d'émotions, s'intégrant dans les modèles classiques de systèmes interactifs. Dans cet article, nous plongeons notre modèle dans les théories établies de l'interaction multimodale. Pour cela, nous redéfinissons une modalité dans le cadre de la reconnaissance passive des émotions. Nous montrons que cette définition permet à notre modèle d'hériter de concepts et propriétés de l'interaction multimodale, permettant une réutilisation de ses résultats et ouvrant ainsi des espaces de conceptions encore inexplorés.

MOTS CLES : reconnaissance d'émotions, modèle d'architecture, multi modalité, plongement.

ABSTRACT
The emotion recognition field is a young but maturing field, for which needs for designing models begin to emerge.We consider emotion recognition to be a multimodal interaction with the machine. From this point of view, we designed a component-based architecture model for emotionnaly-wise interactive systems : the emotion branch. In this paper, we immerse our architecture model in the field of multimodal interaction. We redefine a modality within the frame of passive emotion recognition and show that doing so enables us to benefit from concepts from multimodal interaction (CARE properties and software specifications) and to highlight unexplored design spaces.

CATEGORIES AND SUBJECT DESCRIPTORS : H.5.2 User Interfaces : theory and method.

GENERAL TERMS : Human factors, theory.

KEYWORDS: emotion recognition, architecturural model, multimodality, embedding.

INTRODUCTION

La reconnaissance d'émotions en informatique est un domaine jeune mais dont la maturité croissante fait émerger de nouveaux besoins en terme de modèles pour la conception. Après une phase de réplication, durant laquelle nombreux ont été les travaux proposant des systèmes de reconnaissance [12] [6] [17], nous entrons progressivement dans une phase d'empirisme, où des modèles pour la conception sont mis au point [8]. La plupart des systèmes conçus permettent une reconnaissance passive des émotions : l'utilisateur n'est pas cognitivement impliqué dans le processus de reconnaissance. Nous proposons un modèle pour la conception de systèmes de reconnaissance passive des émotions. Notre approche a pour originalité de considérer la reconnaissance d'émotions comme une forme d'interaction avec la machine. Nous avons donc conçu la branche émotion, un modèle d'architecture basé composants s'intégrant dans des architectures classiques pour des applications interactives [5].

Nous nous basons sur la théorie de Scherer et son modèle à composants pour définir l'émotion [14][15]. Une émotion y est caractérisée par une expression hautement synchronisée : le corps entier (visage, membres, réactions physiologiques) réagit à l'unisson. L'expression émotionnelle humaine est clairement multimodale.

Dans cet article, nous traduisons notre point de vue "Interaction homme machine" de la reconnaissance d'émotions en plongeant notre modèle dans les concepts connus de l'interaction multimodale. Un plongement au sens mathématique d'un ensemble A vers un ensemble B se traduit par l'existence d'une fonction injective de A sur B. Dans cet article, nous montrons qu'en assimilant les chaînes de composants de notre modèle d'architecture à des modalités, on redéfinit la reconnaissance passive des émotions comme une interaction potentiellement multimodale avec la machine, effectuant par là-même notre plongement. Il est alors possible d'appliquer directement à notre modèle des propriétés bien connues de l'interaction multimodale [11]. Cette application directe fait

émerger plusieurs avantages. Notamment, nous faisons hériter (au sens programmation objet) nos types de composants pour la reconnaissance d'émotion de types de composants existants pour l'interaction multimodale. Nous faisons également apparaître de nouveaux espaces de conceptions pour la reconnaissance passive des émotions.

Dans une première partie, nous présenterons la branche émotion, un modèle d'architecture basé composants pour la conception de systèmes sensibles aux émotions. Dans une deuxième partie, nous étendons la définition classique d'une modalité dans le cadre de notre modèle, le plongeant ainsi dans le domaine de l'interaction multimodale. Dans la troisième partie nous appliquons à notre modèle les relations entre modalités connues en interaction multimodale. Enfin, nous voyons en dernière partie les bénéfices de ce plongement.

LA BRANCHE EMOTION

Nous présentons dans cette section notre modèle d'architecture basé composants pour la reconnaissance des émotions par ordinateur, appelé "la branche émotion". Le modèle et notamment ses différents types de composants sont présentés plus en détails dans [5].

L'analyse des systèmes de reconnaissance d'émotions existants font tout d'abord émerger une décomposition en trois niveaux remplissant chacun une fonction déterminée : les niveaux Capture, Analyse et Interprétation. Au niveau Capture, l'information est captée du monde réel et en particulier de l'utilisateur grâce à des dispositifs, par exemple caméra ou microphone. Cette information est ensuite analysée dans le niveau Analyse, où des caractéristiques émotionnellement pertinentes sont extraites des données capturées. Enfin, les caractéristiques extraites sont interprétées pour obtenir une émotion. Ce découpage en trois niveaux - Capture, Analyse et Interprétation - est classique en reconnaissance d'émotions et forme un motif fonctionnel sur lequel nous nous appuyons pour développer notre modèle.

Notre modèle d'architecture propose cinq types de composants (figure 1). Chaque composant s'abonne et

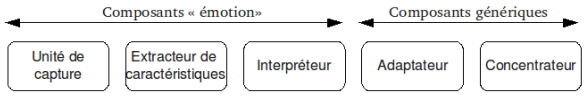

Figure 1 : Les cinq types de composants de la branche émotion.

émet un ou plusieurs flux de données. L'unité de capture a pour rôle de faire l'interface avec un dispositif physique de capture des données. L'extracteur de caractéristiques analyse les données qui lui sont fournies en entrée pour en extraire une ou plusieurs caractéristiques émotionnellement pertinentes. Un interpréteur reçoit les valeurs de plusieurs caractéristiques. Son rôle est d'en interpréter des émotions. Cette interprétation est soumise au modèle des émotions considéré (modèles discrets, continus, ou componentiel), ainsi qu'à l'algorithme informatique utilisé (e.g., réseau de neurones, modèles de Markov cachés). Le modèle dispose également de deux types de composants non reliés à une logique "reconnaissance d'émotions". L'adaptateur a pour rôle de modifier un flux de données. Il peut s'agir d'une simple modification de format comme d'un traitement lourd sans rapport avec la reconnaissance (suivi 3D par caméra par exemple). Le concentrateur a pour rôle d'amalgamer plusieurs flux de données selon une stratégie *ad hoc*.

Il est à noter que le style à composants sur lequel s'appuie notre modèle permet d'avoir un modèle ouvert, extensible, modifiable et réutilisable. Construire une application revient à assembler les composants en les faisant s'abonner aux flux de données d'autres composants (la figure 2 représente un exemple d'assemblage). Cette approche permet également d'extraire des caractéristiques de haut niveau se reposant sur de caractéristiques bas niveau, en enchaînant les composants de type "extracteur de caractéristique". Par exemple, un extracteur E_1 peut calculer les points d'intérêts du visage et fournir ces points extraits à un extracteur E_2, qui grâce à une mémoire tampon calculera les déplacement de ces points d'intérêt.

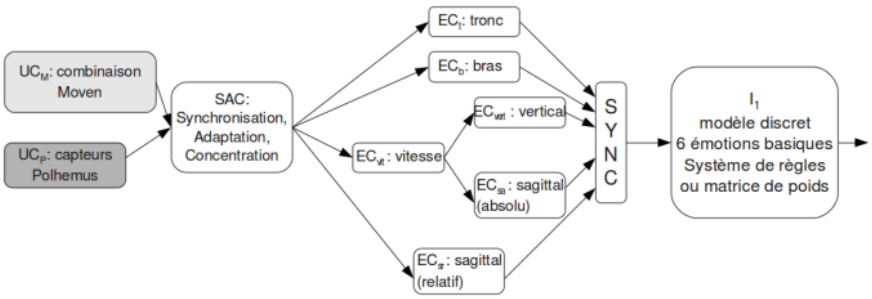

Figure 2 : Exemple d'assemblage des composants.

Il permet enfin d'effectuer une reconnaissance multimodale des émotions.

L'expression émotionnelle est intrinsèquement multimodale. Selon Scherer [14], une émotion est caractérisée par une forte synchronisation : le corps réagit à l'unisson, *via* les expressions faciales, la gestuelle et le mouvement, les intonations vocales, et les réactions physiologiques. Dans un souci de cohérence avec la définition de modalité que nous utilisons dans le reste de cet article, nous appelons ces différents canaux "*canaux de communication émotionnelle*", faisant ainsi écho à la notion d'*affective channels* de Picard [13]. Nous pouvons contrôler nos expressions émotionnelles, certains canaux (le visage, les mouvements des mains) étant plus faciles à contrôler que d'autres (les réactions physiologiques) [10].

La majorité des systèmes existants n'effectuent leur reconnaissance que sur un seul canal de communication. On trouve une majorité écrasante de systèmes de reconnaissance se basant sur les expressions faciales [17], mais il existe des systèmes de reconnaissance par la voix (par exemple [7]), les mouvement [4] ou les réactions physiologiques (par exemple [3]). Il est difficile de classer les différents canaux dans leur capacité à exprimer les émotions. La grande variété des protocoles expérimentaux et des méthodes de reconnaissance fait qu'il est impossible de comparer les résultats des différents systèmes existants. Ces dernières années ont vu l'émergence de quelques systèmes proposants une reconnaissance multicanaux [17] et ont permis de montrer qu'une reconnaissance multicanaux donne de meilleurs résultats que les reconnaissance sur les canaux pris séparéments.

Nous avons donc placé la multimodalité au cœur de notre modèle dès sa conception. Pour cela, nous plongeons ce modèle dans les théories établies de la multimodalité. Nous redéfinissons donc le terme "modalité" dans notre cas de reconnaissance passive des émotions.

DEFINITION DE LA MULTIMODALITE POUR LA BRANCHE EMOTION

Classiquement dans la littérature, les systèmes de reconnaissance d'émotion sont considérés comme "multimodaux" s'ils permettent une reconnaissance basée sur plusieurs canaux de communication émotionnelles en même temps : les expressions faciales, la voix, la gestuelle, et les réactions physiologiques. Dans cet article, nous proposons de considérer la reconnaissance d'émotions comme une interaction entre l'homme et la machine.

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 [1]. Nous nous basons sur la définition d'une modalité donnée par Nigay dans [11]. 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 : quel canal de communication utiliser (e.g. la voix), comment mettre en forme l'information pour être compris de la machine (e.g. langage pseudo-naturel)? 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.

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 :

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

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 une tâche complète. 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.

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. En adaptant littéralement cette définition à celle donnée dans l'équation (1), on obtient comme définition d'une modalité :

$$modalité = < dispositif, \quad canal\ de\ communication\ émotionnelle > \quad (3)$$

Cette définition correspond à la définition implicitement utilisée dans le domaine de la reconnaissance d'émotions. Un système est communément admis comme étant multimodal s'il effectue simultanément une reconnaissance sur plusieurs canaux de communication émotionnelle. Cette définition présente cependant plusieurs limitations. La principale limitation vient du fait que comme la définition (1), elle met en œuvre un point de vue système et un point de vue utilisateur. Dans ce dernier point de vue, elle fait ressortir le choix du canal de communication à utiliser et la mise en forme de l'information selon ce canal de façon à être compris par la machine. Or, comme nous l'avons explicité précédemment, l'émotion est un phénomène hautement synchronisé et il n'y a pas de choix du canal de communication émotionnelle. De plus nous nous plaçons dans le cas d'une reconnaissance passive, où l'utilisateur est scruté et n'est pas activement sollicité pour communiquer son état émotionnel. L'utilisateur ne cherche donc pas à mettre en forme l'information émotionnelle de son expression pour être compris par la machine. Dans cette définition, le point de vue utilisateur de la définition n'est donc pas utile. En se restreignant à des considérations techniques, la définition (3) présente des limitations au niveau de sa granularité, trop grossière. La définition ne permet pas de rendre compte des multiples formats de données et des combinaisons possibles. Enfin, la séparation en cinq canaux de communication émotionnelle n'est pas forcément soutenue. Ainsi Scherer distingue les composants "expression motrice" et "processus neurophysiologiques" [14] et ne distingue pas les différents canaux.

En conclusion, la définition (3) est une définition trop naïve pour permettre d'appliquer les notions d'interaction multimodale en reconnaissance passive des émotions. Nous avons donc choisi 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 [13]. Un système de reconnaissance s'appuyant sur plusieurs de ces canaux est un système "multicanaux". Nous proposons, dans le prochain paragraphe, une définition de la modalité, dans le contexte de la reconnaissance d'émotion, permettant la réutilisation des concepts de l'interaction multimodale.

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 [16] pour la multimodalité en sortie de systèmes interactifs. Les systèmes représentationnels peuvent appartenir aux niveaux Capture, Analyse ou Interprétation. Contrairement à la définition implicite de la littérature en reconnaissance d'émotions (cf (3)), nous considérons donc qu'une application de reconnaissance d'émotions est multimodale si elle met en œuvre plusieurs modalités telles que définies par l'équation (1).

Nous adoptons un point de vue système uniquement. Et 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 :

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

On réécrit alors en développant (2) dans le contexte de la reconnaisssance d'émotion :

$$modalité = \begin{array}{l} <\ldots<< d, sr_1^C >, sr_2^C >\ldots sr_n^C >, \\ sr_1^A >\ldots>, sr_m^A >, \\ sr_1^I >\ldots sr_p^I > \end{array} \quad (4)$$

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. Dans notre modèle, le dispositif est assimilé à une unité de capture et chaque système représentationnel est un format de données dans une chaîne de composants (voir figure 3).

Nous rappelons que 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.

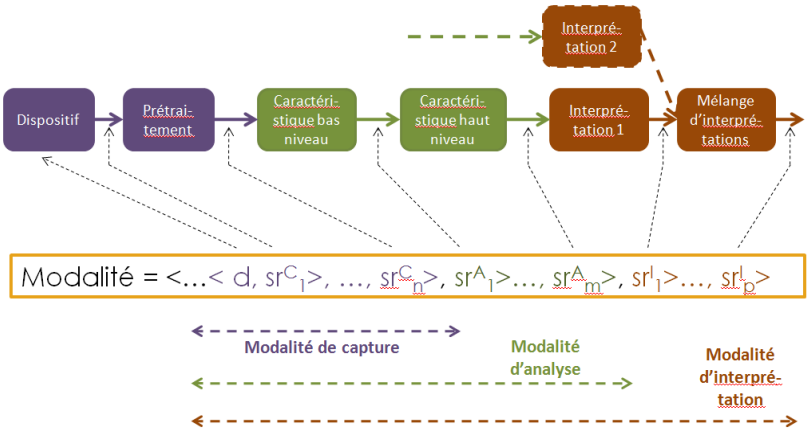

Figure 3 : Traduction d'une chaîne de composants en une modalité**.**

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.

Dans cette section nous avons redéfini une modalité dans le cas de la reconnaissance passive des émotions. Plus spécifiquement et dans le cadre de notre modèle d'architecture, nous avons considéré une chaîne de composants (c'est-à-dire une séquence de composants connectés entre eux) comme une modalité d'interaction. Nous avons donc effectué un plongement de notre modèle vers le domaine de l'interaction multimodale. Ce plongement nous permet d'appliquer directement à notre modèle un ensemble de propriétés sur les relations entre modalités, et de faire découler des bénéfices de cette application.

APPLICATION DES PROPRIETES CARE

L'identification d'une modalité dans le cadre de notre modèle permet d'appliquer les principes de l'interaction multimodale.

Tout d'abord, pour caractériser la multimodalité, nous retenons deux espaces de conception qui caractérisent les relations entre modalités d'interaction : l'espace TYCOON [9] et les propriétés CARE de la multimodalité [11]. Ces deux espaces définissent des relations similaires. Par cohérence avec notre choix de définition d'une modalité, nous avons privilégié les propriétés CARE, qui ont de plus donné lieu à la conception de l'outil basé composants ICARE pour concevoir des applications interactives multimodales, et dont nous réutilisons les spécifications objet.

Les propriétés CARE [11] permettent de caractériser l'interaction multimodale ; l'acronyme CARE signifie 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. Dans notre modèle, une modalité est une chaîne de composants commençant à une unité de capture (voir figure 3). L'assignation signifie l'absence de choix. Dans notre modèle, une modalité est assignée à un composant si l'usage de cette modalité est obligatoire pour faire fonctionner le composant. Deux modalités sont équivalentes pour un composant si elles fournissent des flux de données selon un même système représentationnel et peuvent être utilisées alternativement sur ce composant. Deux modalités sont redondantes pour un composant si elles sont équivalentes et utilisées simultanément. Enfin, des modalités sont complémentaires pour un composant si elles ne sont pas équivalentes et sont toutes nécessaires au fonctionnement du composant.

En considérant la définition étendue de la multimodalité (4), 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 :

• $\forall X \in \{C, A, I\}, \forall Y \in \{C, A, I\}, X \leq Y, \forall i \in \mathbb{N}, sr_i^X$ 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.

• $\forall X \in \{C, A, I\}, \quad \forall Y \in \{C, A, I\}, \quad \forall Z \in \{C, A, I\}, X \leq Y \leq Z, \forall i \in \mathbb{N}, \forall j \in \mathbb{N}, sr_i^X$ et sr_j^Y sont équivalents pour sr_{i+1}^Z si l'un ou l'autre permettent de former sr_{i+1}^Z.

• $\forall X \in \{C, A, I\}, \quad \forall Y \in \{C, A, I\}, \quad \forall Z \in \{C, A, I\}, X \leq Y \leq Z, \quad \forall i \in \mathbb{N}, \quad \forall j \in \mathbb{N}, \quad sr_i^X$ 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.

• $\forall X \in \{C, A, I\}, \quad \forall Y \in \{C, A, I\}, \quad \forall Z \in \{C, A, I\}, X \leq Y \leq Z, \forall i \in \mathbb{N}, \forall j \in \mathbb{N}, sr_i^X$ 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.

BENEFICES DU PLONGEMENT

L'application des propriétés CARE apporte deux avantages pour notre modèle. Tout d'abord cela permet la réutilisation de spécifications logicielles existantes de composants pour la conception d'applications multimodales. Les types de composants de la branche émotions héritent ainsi de types de composants pour l'interaction multimodale. En particulier, l'outil ICARE [2] est un système basé composants permettant la création d'applications interactives multimodales, et nous faisons hériter les composants de notre modèle des types de composants définis pour ICARE. Ensuite, l'analyse des système de reconnaissance existant à la lumière des propriétés CARE permet de réutiliser les espaces de caractérisation classiques de la multimodalité pour faire émerger des espaces de conceptions non encore explorés pour les systèmes de reconnaissance d'émotions.

Héritage de types de composants existants pour l'interaction multimodale

ICARE se base sur la définition 1 et sur les propriétés CARE pour définir trois types de composants : le type *dispositif*, le type *système représentationnel*, et le type *combinaison*, qui permet l'implémentation de la redondance et de la complémentarité de modalités. Ces types de composants ont été spécifiés dans [16] et prennent en compte de nombreux travaux en interaction multimodale. Notre redéfinition de la multimodalité, ainsi que l'identification des propriétés CARE dans le cadre de notre modèle nous permet d'identifier nos types de composants comme héritant des types de composants spécifiés pour ICARE.

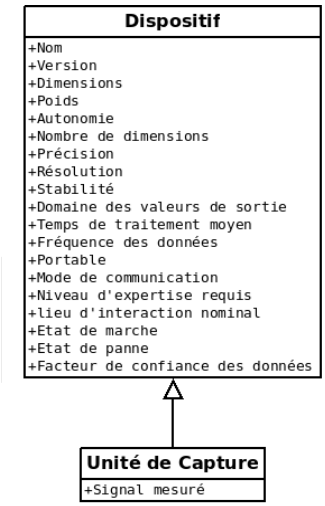

Figure 4 : L'unité de capture hérite du type de composant Dispositif.

Nous identifions tout d'abord le type de composant *unité de capture* comme héritant du type de composant ICARE *dispositif* (figure 4). Une chaîne de caractère nommant le signal mesuré (e.g. "coordonnées poignet gauche") vient en sus des attributs hérités.

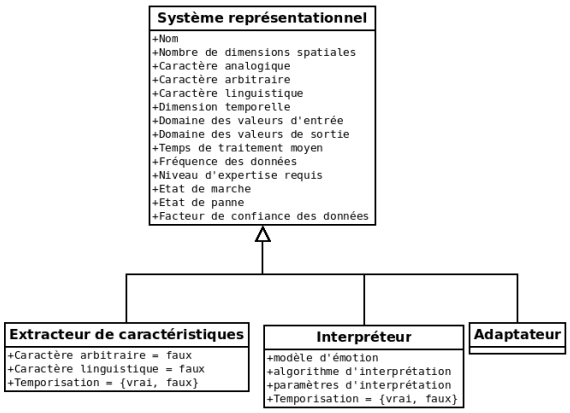

Figure 5 : Les extracteurs de caractéristiques, interpréteurs et adaptateurs héritent du type de composant système représentationnel**.**

Nous identifions ensuite les types de composant *extracteur de caractéristique* et *interpréteur* comme héritant tous deux du type de composant ICARE *système représentationnel* (figure 5). L'extracteur de caractéristique présente une caractéristique supplémentaire : la temporisation (cas des caractéristiques bloquantes [5]. Les caractères arbitraire

et linguistique d'un extracteur sont tous deux à faux : les caractéristiques ne sont pas choisies de façon arbitraire mais nous ne considérons pas qu'elles forment un langage de l'émotion. L'interpréteur présente quant à lui quatre attributs supplémentaires : la temporisation, le modèle d'émotion choisi, l'algorithme d'interprétation, et les paramètres d'interprétation. Nous identifions ensuite le type de composant *adaptateur* comme étant un type de composant ICARE *système représentationnel*.

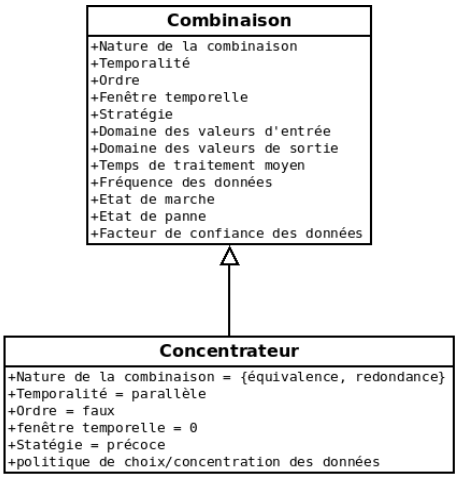

Figure 6 : Le concentrateur hérite du type de composant Combinaison.

Enfin, nous identifions notre type de composant *concentrateur* comme héritant du type de composant ICARE combinaison (figure 6). Cette spécialisation entraîne l'affectation de valeurs à certains attributs. Tout d'abord, l'attribut "Nature de la combinaison" est restreint aux valeurs {équivalence, redondance}. En effet, le système d'abonnement entre composants permet de prendre en charge les relations d'assignation et de complémentarité. 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*.

Cadre génératif des propriétés CARE

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 7 illustre les combinaisons de modalités proposées par les propriétés CARE utilisées dans les travaux existants en reconnaissance d'émotions.

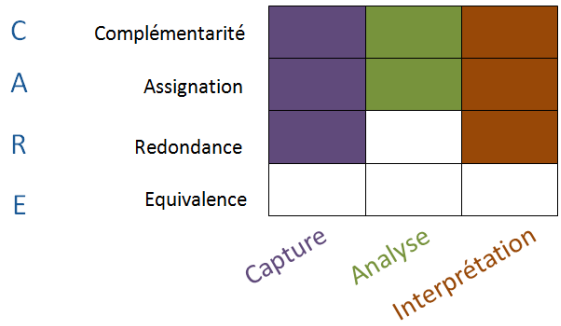

Figure 7 : En blanc, les cas non explorés dans les systèmes existants de reconnaissance d'émotions.

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 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.

CONCLUSION

Nous avons présenté dans cet article comment nous lions notre modèle d'architecture pour la reconnaissance passive des émotions et les concepts de l'interaction multimodale. L'expression émotionnelle étant hautement synchronisé, nous avons montré qu'il était possible de se limiter aux aspects système des réflexions sur la multimodalité dans le cadre d'une reconnaissance passive. Nous avons affiné la définition de la modalité dans le cadre de la reconnaissance d'émotion et de notre modèle « la branche émotion ». Cette définition permet l'application des relations CARE entre modalités. Ceci apporte plusieurs bénéfices : notamment, nous reprenons les spécifications de composants existants pour l'interaction multimodale. De plus, les propriétés CARE présentent un pouvoir génératif dans le cadre de la reconnaissance passive des émotions, en faisant apparaître des espaces de conception encore inexplorés.
Partant de ce constat, une perspective de ce travail, démarrée dans le cadre du développement du logiciel eMotion, est d'explorer les possibilités offertes par l'équivalence des modalités en reconnaissance d'émotions, l'équivalence étant une relation entre modalités ignorée par les systèmes actuels.

BIBLIOGRAPHIE

1. Bolt, R.A. "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.
2. Bouchet, J. *Ingénierie de l'interaction multimodale en entrée Approche à composants ICARE*. PhD thesis, Université Joseph Fourier, Grenoble I, 2006.
3. Chanel, G., Kronegg, J., Grandjean, D. and Pun, T. Emotion assessment: Arousal evaluation using EEG's and peripheral physiological signals. *Lecture Notes in Computer Science*, 4105:530, 2006.
4. Clay, A., Couture, N. and Nigay, L., Towards Emotion Recognition in Interactive Systems: Application to a Ballet Dance Show. In *WinVR'09, Proceeding of the World Confe-rence on Innovative Virtual Reality*, ASME-AFM, pages 19–24, 02 2009.
5. Clay, A., Couture, N. and Nigay, L. 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.
6. Jaimes, A. and Sebe, N. Multimodal human-computer interaction: A survey. *Comput. Vis. Image Underst.*, 108(1-2):116–134, 2007.
7. Jin, X. and Wang, Z. 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.
8. Lisetti, C.L. Le paradigme MAUI pour des agents multimodaux d'interface homme machine socialement intelligents. *Revue d'Intelligence Artificielle, Numéro Spécial sur les Interactions Emotionnelles*, 20(4-5):583–606, 2006.
9. Martin, J.C. TYCOON: Theoretical framework and software tools for multimodal interfaces. *Intelligence and Multimodality in Multimedia interfaces*, 1998.
10. Morris, D., Friedhoff, H. and Dubois, Y. *La clé des gestes*. B. Grasset, 1978.
11. Nigay, L. and Coutaz, J. 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.
12. Pantic, M, Sebe, N., Cohn, J. F. and Huang T. 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.
13. Picard, R.W. *Affective computing*. MIT press, 1997.
14. Scherer, K.R. 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.
15. Scherer, K.R. Feelings integrate the central representation of appraisal-driven response organization in emotion. In *Feelings and emotions: The Amsterdam symposium*, pages 136–157, 2004.
16. Vernier, F. *La multi modalité en sortie et son application à la visualisation de grandes quantités d'information*. PhD thesis, Université Joseph Fourier, Grenoble I, 2001.
17. Zeng, Z., Pantic, M., Roisman, G.I. and Huang, T.S. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.