Reconnaissance d'Emotions : un Point de Vue Interaction Multimodale

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
Université de Grenoble, CNRS, LIG, Grenoble, France
Laurence.Nigay@imag.fr

RESUME
Le domaine de la reconnaissance d'émotions atteint un stade de maturité où commence à émerger un besoin en termes d'ingénierie et en particulier de modèles de conception. Partant de cette constatation, nous proposons d'exploiter les résultats obtenus en ingénierie de l'interaction multimodale pour les appliquer et les adapter à la reconnaissance d'émotions, une émotion étant intrinséquement multimodale. En particulier, nous adaptons la définition d'une modalité d'interaction au cas des modalités mises en jeu lors de la reconnaissance passive d'émotions. Une modalité définie, nous étudions ensuite les relations entre modalités en nous appuyant sur les propriétés CARE de l'interaction multimodale. Nous soulignons le pouvoir génératif de CARE ainsi que les apports de notre modèle en ingénierie de la reconnaissance d'émotions.

MOTS CLES : reconnaissance d'émotions, multimodalité.

ABSTRACT
Analysis of emotion recognition is a young but maturing research field, for which there is an emerging need for engineering models and in particular design models. Addressing these engineering challenges of emotion recognition, we reuse and adapt results from the research field of multimodal interaction, since the expression of an emotion is intrinsically multimodal. In this paper, we refine the definition of an interaction modality for the case of passive emotion recognition. We also study the combination of modalities by applying the CARE properties. We highlight the benefits of our design model for emotion recognition.

CATEGORIES AND SUBJECT DESCRIPTORS: H.5.2 User Interfaces: Theory and methods.

IHM'10 © ACM, 2010. This is the author's version of the work. It is posted here by permission of ACM for your personal use. Not for redistribution. The definitive version was published in Actes de la 22ème Conférence Francophone sur l'Interaction Homme-Machine, IHM'10.

GENERAL TERMS: Human factors, Theory, Algorithms.

KEYWORDS: Emotion Recognition, Multimodality

INTRODUCTION

La maturité croissante du domaine de la reconnaissance d'émotions en informatique fait émerger de nouveaux besoins en termes d'ingénierie et en particulier en termes de modèles de conception. Après une phase de réplication, durant laquelle nombreux ont été les travaux proposant des systèmes de reconnaissance [6] [11] [14], nous entrons progressivement dans une phase d'empirisme [5], où des modèles pour la conception sont mis au point [7].

La plupart des systèmes conçus permettent une reconnaissance passive des émotions : l'utilisateur ne cherche pas à communiquer consciemment ses émotions à la machine, il se laisse observer. Pour définir l'émotion, nous nous basons sur la théorie de Scherer [12][13]. Une émotion y est caractérisée par une expression hautement synchronisée : le corps entier (visage, membres, réactions physiologiques) réagit à l'unisson et l'expression émotionnelle humaine est clairement multimodale. Partant de ces deux constats, le besoin de modèle de conception et le caractère multimodal d'une expression émotionnelle, nous proposons un modèle de conception de systèmes de reconnaissance passive des émotions dont l'originalité est de considérer la reconnaissance d'émotions comme une forme d'interaction multimodale avec la machine. Se faisant, nous pouvons exploiter des résultats issus de l'interaction multimodale à la reconnaissance d'émotions passives. Dans une première partie, nous rappelons la définition d'une modalité d'interaction et les espaces de conception de la multimodalité. Nous spécialisons ensuite ces résultats au cas de la reconnaissance d'émotions. Nous concluons enfin sur les bénéfices de notre modèle de conception inspiré de l'interaction multimodale pour la reconnaissance d'émotions.

INTERACTION MULTIMODALE

Le paradigme de la multimodalité en Interaction Homme-Machine se caractérise par la mise à disposition de plusieurs moyens de communication entre l'utilisateur et le système, comme l'illustre l'exemple canonique du "mets çà là" combinant parole et geste [1]. Focalisant sur

l'interaction multimodale en entrée (de l'utilisateur vers le système), nous nous basons sur la définition d'une modalité d'interaction donnée par Nigay dans [9]. Une modalité y est définie par la relation suivante :

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

• d est un dispositif physique d'interaction : souris, caméra, microphone, capteurs de mouvement, GPS, écran, etc.
• 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 considérant 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 correspond aux actions humaines à bas niveau d'abstraction : l'humain agit sur le dispositif. Le système représentationnel se situe 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) ? Du point de vue système, le couple renseigne sur le dispositif mis en œuvre ainsi que sur le domaine et le format des données échangées entre l'homme et la machine. Enfin, 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.

Une modalité d'interaction définie, nous retenons deux modèles de conception de la multimodalité : l'espace TYCOON [8] et les propriétés CARE [9]. Ces deux espaces définissent des relations entre modalités d'interaction similaires. Nous avons privilégié les propriétés CARE, car elles ont donné lieu à la conception d'outils basés composants ICARE [2] et plus récemment OpenInterface [10], pour concevoir des applications interactives multimodales. Les propriétés CARE [9] permettent de caractériser l'interaction multimodale ; l'acronyme CARE signifie Complémentarité, Assignation, Redondance, et Equivalence. L'assignation signifie l'absence de choix : une modalité est assignée à la réalisation d'une tâche. Deux modalités sont équivalentes si elles peuvent être utilisées alternativement pour accomplir une tâche. Le choix de la modalité peut alors être effectué par l'utilisateur ou le système. La redondance dénote l'utilisation séquentielle ou parallèle de plusieurs modalités d'interaction équivalentes : la redondance d'information en provenance de l'utilisateur implique la prise en compte d'une seule des modalités d'interaction par le système, l'autre pouvant éventuellement contribuer à fiabiliser l'expression obtenue. Enfin, la complémentarité entre modalités d'interaction exprime qu'il faut utiliser toutes les modalités pour obtenir une commande complète. Cela signifie qu'aucune des modalités ne suffit seule.

REDEFINITION POUR LA RECONNAISSANCE D'EMOTIONS

Pour spécialiser la définition d'une modalité d'interaction au cas de la reconnaissance d'émotions, nous nous limitons dans nos travaux au cas de la reconnaissance passive des émotions, qui regroupe néanmoins la grande majorité des systèmes de reconnaissance existants [4]. En reconnaissance passive, l'utilisateur est scruté par le système. Il ne cherche pas à communiquer consciemment ses émotions à la machine, mais se laisse observer.

En nous référant à Scherer [12], nous notons qu'une émotion implique des réactions physiques et physiologiques hautement synchronisées (c'est-à-dire exprimées à l'unisson). Un système de reconnaissance d'émotions est implicitement caractérisé comme multimodal s'il est capable d'effectuer une reconnaissance selon au moins deux canaux de communication émotionnelle à la fois, ces canaux étant le visage, la voix, le mouvement et la gestuelle, et l'ensemble des réactions physiologiques.

D'une part, ces différents canaux de communication émotionnelle ne constituent pas une taxonomie unanime : ainsi Scherer [12] considère les aspects moteurs (mouvement, visage) et physiologiques (y incluant la cause des intonations de la voix).

D'autre part, en nous basant sur la définition (1) d'une modalité d'interaction, une modalité serait alors définie par un couple <dispositif, canal de communication émotionnelle>. Cette définition permet, comme la définition (1), de prendre en compte les points de vue système et utilisateur d'une modalité. Selon le point de vue utilisateur, cette définition 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 le système. 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 le système. En conclusion, le point de vue utilisateur pour la définition d'une modalité dans notre cadre d'étude est inutile.

Nous ne considérons donc plus que l'aspect système de la modalité. Or, cette définition en canaux de communication émotionnelle est trop grossière pour représenter de façon efficace les différents flux de données dans la séquence de traitements menant à une reconnaissance. Elle ne permet pas de rendre compte des multiples formats de données et des combinaisons possibles. Nous étendons donc la définition d'une modalité en assimilant chaque flux de données présent dans le système comme un système représentationnel. On réécrit alors en développant (1) dans le contexte de la reconnaisssance d'émotion :

$$modalité = <\ldots << d, sr_1^C >, sr_2^C > \ldots sr_n^C >, sr_1^A > \ldots >, sr_m^A >, sr_1^I > \ldots sr_p^I > \quad (2)$$

où la séquence des systèmes représentationnels explicite les transferts subis par une donnée depuis le dispositif jusqu'à son interprétation finale. Le dispositif est assimilé à une unité de capture et chaque système représentationnel correspond à un format de données. Nous distinguons trois niveaux d'abstraction dans la séquence de systèmes représentationnels : le niveau Capture (pour aquérir l'information du monde réel), le niveau Analyse (pour extraire les caractéristiques émotionnellement pertinentes), et enfin le niveau Interprétation (pour inférer une émotion à partir des caractéristiques extraites). Nous définissons alors 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. Nous schématisons ces trois niveaux d'abstraction à la Figure 1 en considérant qu'un système représentationnel correspond à un composant logiciel comme dans notre modèle d'architecture noté « Branche émotion » [3] pour la reconnaissance d'émotions. Le modèle d'architecture « Branche émotion » est inspiré des outils basés composants ICARE [2] et OpenInterface [10] pour l'interaction multimodale.

Cette redéfinition de la modalité permet de plus de réutiliser directement les propriétés CARE de la multimodalité. En considérant la définition (2), 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'une combinaison 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.

APPORTS DU POINT DE VUE INTERACTION MULTIMODALE POUR LA RECONNAISANCE D'EMOTIONS

Plusieurs bénéfices découlent de la spécialisation de la définition d'une modalité et des propriétés CARE de la multimodalité.

Un premier apport de notre modèle concerne l'exploration de l'espace des possibilités en nous reposant sur CARE. Ainsi nous constatons que dans les travaux existants l'accent a principalement été mis sur la robustesse de la reconnaissance. Un recensement des systèmes existants montre que l'équivalence (choix de modalités) est une combinaison qui n'est jamais utilisée, à aucun des niveaux que ce soit le niveau Capture (choix du dispositif), le niveau Analyse (choix des caractéristiques à extraire) ou le niveau Interprétation (choix de la méthode d'interprétation ou du modèle d'émotion à utiliser). Ce constat met en avant l'aspect génératif des propriétés CARE. L'équivalence de modalités est une combinaison intéressante en reconnaissance d'émotions à explorer : elle permet par exemple à un système de laisser l'utilisateur préférer une reconnaissance par caméra des expressions faciales au lieu d'une reconnaissance par signaux physiologiques (plus fiables et difficiles à feindre), lui laissant plus de marge de manœuvre pour contrôler ce qu'il veut laisser transparaître ; ou au contraire de préférer une reconnaissance par signaux physiologiques si la possibilité de feindre ou cacher des émotions est indésirable.

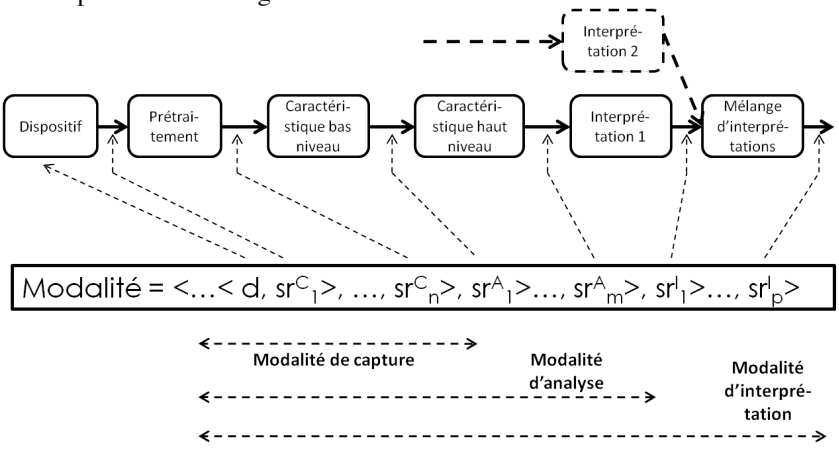

Figure 1 : Une modalité (définition (2)) comme une chaîne de composants à trois niveaux d'abstraction Capture, Analyse et Interprétation.

Nous avons ainsi développé un système de reconnaissance des émotions par le mouvement intitulé eMotion qui intègre l'équivalence de deux dispositifs du niveau Capture et de deux systèmes représentationnels du niveau Analyse. Dans eMotion, deux capteurs de mouvement sont laissés au choix : des capteurs donnant la position des poignets (capteurs Polhemus Liberty) ou des capteurs donnant les coordonnées du corps complet (combinaison de capture de mouvement Moven - XSens). De plus il est possible de choisir le système représentationnel du niveau Analyse pour le calcul de la direction sagittale du mouvement (avant-arrière). Par exemple, dans un cas applicatif de eMotion qu'est le ballet et la reconnaissance des émotions des danseurs, il est possible de choisir une composante sagittale relative (le danseur avance ou recule, l'orientation de son tronc définissant l'avant), ou absolue (le danseur avance en se rapprochant du public et recule en s'en éloignant, quelle que soit son orientation).

Un autre apport de notre modèle est lié à la conception logicielle. En considérant qu'un système représentationnel de la définition (2) est un composant logiciel, nous construisons un système de reconnaissance d'émotions fortement modulaire en assemblant des composants définis à trois niveaux, Capture, Analyse et Interprétation (Figure 1). Se faisant nous appliquons notre modèle d'architecture noté « Branche émotion » [3]. Cette structuration très modulaire implique une forte modifiabilité du code du système de reconnaissance d'émotions et permet ainsi une exploration efficace de plusieurs solutions de reconnaissance. Par exemple, dans notre système de reconnaissance eMotion basé sur les mouvements du corps, il est possible de passer d'une capture du corps reposant sur la combinaison Moven, à une autre capture du corps basé sur la combinaison Moven mais où les positions des poignets sont mesurées non plus par la combinaison Moven mais par les capteurs Polhemus (Moven+polhemus). Ces deux possibilités (l'une reposant sur une unique solution de capture, l'autre sur deux solutions complémentaires), sont équivalentes au niveau Capture et les composants des niveaux Analyse et Interprétation ne sont pas modifiés. Au contraire si nous considérons maintenant la capture des positions des poignets, cette modalité de capture n'est pas équivalente à celle basée sur tout le corps. Une modalité fournissant les coordonnées du corps entier et une modalité où seules les positions des poignets sont connues ne sont pas équivalentes. Le passage de l'une à l'autre entraîne alors des changements qui sont propagés dans les composants du niveau supérieur, le niveau Analyse : ainsi le calcul de la vitesse du mouvement est différent selon que l'on dispose de la position du bassin ou non.

CONCLUSION

Nous avons présenté la spécialisation des définitions et concepts de l'interaction multimodale au domaine de la reconnaissance passive des émotions. Nous avons défini une modalité dans le cadre de la reconnaissance d'émotion. Cette définition permet l'application des propriétés CARE entre modalités. Le modèle conceptuel résultant a des apports à la fois en conception du système de reconnaissance d'émotions, par exemple en identifiant des combinaisons de modalités encore peu explorées, mais aussi en conception logicielle en définissant une décomposition fortement modulaire du code du système de reconnaissance d'émotions. Une perspective à ces travaux serait de définir un environnement graphique de développement où le développeur assemblerait des composants (de type Capture, Analyse et Interprétation) pour définir le système de reconnaissance d'émotions.

REMERCIEMENTS

Les résultats présentés dans cet article sont partiellement financés par l'ANR - projet CARE (Cultural Expérience: Augmented Reality and Emotion) - www.careproject.fr

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*. Ph.D. thesis, Université Joseph Fourier, Grenoble 1, 2006.

[3] 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.

[4] Clay, A. *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*. Ph.D. thesis, Université Bordeaux 1, Bordeaux, 2009.

[5] Gaines, B.R., Modeling and Forecasting the Information Sciences. Information Sciences 57-58,1991, p. 3-22.

[6] Jaimes, A. and Sebe, N. Multimodal human-computer interaction: A survey. *Comput. Vis. Image Underst.*, 108(1-2):116–134, 2007.

[7] 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.

[8] Martin, J.C. TYCOON: Theoretical framework and software tools for multimodal interfaces. *Intelligence and Multimodality in Multimedia interfaces*, 1998.

[9] 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.

[10] OpenInterface European project. IST Framework 6 STREP funded by the European Commission (FP6-35182). www.oi-project.org.

[11] 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.

[12] 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.

[13] 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.

[14] 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