Perception d'Emotions Mélangées : Du Corpus Vidéo à l'Agent Expressif

Stéphanie Buisine* — Sarkis Abrilian — Radoslaw Niewiadomski***,**** — Jean-Claude Martin** — Laurence Devillers** — Catherine Pelachaud*****

** LCPI-ENSAM, 151 bd de l'Hôpital, 75013 Paris, France*
stephanie.buisine@paris.ensam.fr
*** LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France*
{sarkis, martin, devil}@limsi.fr
**** LINC, IUT of Montreuil, Univ. Paris 8, 140 rue Nouvelle France, 93100 Montreuil, France c.pelachaud@iut.univ-paris8.fr*
***** Department of Mathematics and Computer Science, University of Perugia, Italy radek@dipmat.unipg.it*

RÉSUMÉ. *Dans la vie de tous les jours, nous faisons fréquemment l'expérience de mélanges d'émotions, notamment lorsque des émotions simultanées se superposent ou lorsque certaines en masquent d'autres. Cet article présente une étude de la perception de comportements émotionnels multimodaux d'Agents Conversationnels Animés. L'objectif de cette expérimentation est de déterminer si on peut détecter correctement les signaux émotionnels portés par différentes modalités (parole, expressions faciales, gestes) lorsqu'ils apparaissent superposés ou masqués. Des comportements émotionnels annotés à partir d'un corpus d'extraits télévisés sont rejoués par un agent expressif à différents niveaux d'abstraction, et nous comparons la perception de ces différents niveaux. Les résultats apportent des éléments de discussion sur l'utilisation de tels protocoles pour l'étude des effets de différents modèles et différentes modalités sur la perception d'émotions complexes.*

ABSTRACT. *Real life emotions are often blended and involve several simultaneous superposed or masked emotions. This paper reports on a study on the perception of multimodal emotional behaviors in Embodied Conversational Agents. This experimental study aims at evaluating if people detect properly the signs of emotions in different modalities (speech, facial expressions, gestures) when they appear to be superposed or masked. We compared the perception of emotional behaviors annotated in a corpus of TV interviews and replayed by an expressive agent at different levels of abstraction. The results provide insights on the use of such protocols for studying the effect of various models and modalities on the perception of complex emotions.*

MOTS-CLÉS : *Agents Conversationnels Animés, Emotions complexes, Génération de comportements multimodaux expressifs, Evaluation.*

KEYWORDS: *Embodied Conversational Agents, Complex emotions, Generation of expressive multimodal behaviors, Evaluation.*

1. Introduction

Les Agents Conversationnels Animés (ACAs) sont particulièrement intéressants dans le contexte d'études expérimentales sur la perception de comportements émotionnels multimodaux, étant donné qu'on peut contrôler leurs signaux comportementaux et même leurs modalités. Dans la vie de tous les jours, les émotions sont souvent complexes et impliquent plusieurs émotions simultanées [11, 12, 23]. Elles surviennent par exemple comme une succession rapide d'émotions, une superposition, une émotion masquée, supprimée ou au contraire exagérée. Nous désignons ces phénomènes sous le terme d'émotions mélangées. Ces mélanges produisent des « expressions faciales multiples simultanées » [20].

Les expressions faciales qui en résultent varient en fonction du type de mélange. Une émotion masquée peut transparaître à travers une émotion montrée [12] ; deux émotions superposées peuvent être exprimées par des éléments faciaux différents (une émotion pouvant être exprimée dans la partie supérieure du visage et une autre dans la partie inférieure) [12]. Un corpus vidéo d'interviews télévisées peut être un moyen d'explorer le comportement de personnes soumises à de telles émotions mélangées, au niveau de leurs expressions faciales, mais également de leurs gestes ou de leurs paroles [8].

Dans le but de comprendre si les zones faciales ou les éléments faciaux jouent des rôles identiques dans la reconnaissance des émotions, des chercheurs ont mené diverses tâches perceptives ou ont étudié la psychologie de l'activité faciale [4, 5, 6, 14]. Ils ont trouvé que les émotions positives sont principalement perçues à partir de l'expression de la partie inférieure du visage (par exemple le sourire) alors que les émotions négatives sont principalement perçues à partir de la partie supérieure du visage (par exemple un froncement de sourcils). A partir de ce type de résultats, nous avons développé un modèle computationnel des expressions faciales des mélanges d'émotions. Il génère les expressions faciales à partir de celles des émotions simples en utilisant les règles de la logique floue [17]. Jusqu'à présent, très peu de modèles d'émotions mélangées ont été développés pour les ACAs. Le calcul de la nouvelle expression se fait plus généralement par interpolation entre les paramètres faciaux d'expressions données [3, 9, 18, 21].

Cet article présente une étude expérimentale dont l'objectif est d'évaluer si on détecte correctement les signaux de différentes émotions dans des modalités multiples (parole, expressions faciales, gestes) quand elles se trouvent superposées ou masquées. Nous comparons la perception de comportements émotionnels dans des corpus d'interviews télévisées avec des comportements similaires rejoués par un agent expressif. Les expressions faciales de l'agent sont définies par deux approches : le modèle computationnel d'émotions mélangées (nommé « reproduction du mélange facial »), ou l'annotation d'expressions faciales à partir des vidéos (« reproduction niveaux multiples »). Nous nous intéressons également à l'évaluation des différences entre la perception visuelle seule et la perception audio-visuelle.

2. Annoter et rejouer des comportements émotionnels multimodaux

Dans le but d'étudier les comportements émotionnels multimodaux de la vie de tous les jours, nous avons collecté un corpus d'interviews télévisées riches émotionnellement [7]. Plusieurs niveaux d'annotation ont été codés manuellement à l'aide d'Anvil [15] par 3 codeurs experts.

Par ailleurs, nous avons créé un ACA, Greta, qui dispose de qualités communicatives conversationnelles et émotionnelles [19]. Notre modèle d'expressivité est fondé sur des études telles que [13, 24, 25].

Comme mentionné précédemment, nous avons défini, à partir d'un corpus, deux approches pour générer les expressions faciales de Greta [17]. Dans l'approche « reproduction niveaux multiples », les paramètres faciaux (mouvements des sourcils, direction du regard, tension de la bouche…), sont spécifiés à partir des annotations manuelles de la vidéo initiale [16].

L'approche « reproduction du mélange facial » (cf. fig. 1) utilise un modèle computationnel pour générer les expressions faciales des mélanges d'émotions [16]. Ce modèle est basé sur une approche de division du visage en zones (sourcils, lèvres, etc.). Différents types d'émotions mélangées (ex : superposition et masquage) sont générés par différentes règles floues (établies à partir de [12]).

(a) déception (b) joie (c) video initiale (d) déception masquée par la joie calculée par "mélange facial"

Fig. 1. Déception masquée par la joie.

3. Protocole expérimental

Les objectifs de notre expérience sont 1) de tester si les sujets perçoivent des émotions mélangées dans les animations de Greta de la même manière que dans les vidéos d'origine et 2) de comparer les deux approches pour la reproduction d'émotions mélangées. Nous avons sélectionné deux extraits d'interviews télévisées, chacune contenant un type de mélange d'émotions : l'un montre un exemple de superposition de colère et de désespoir [2], l'autre une combinaison d'émotions négatives (déception, tristesse, colère) et positives (contentement, sérénité) avec une intention de masquer les émotions négatives derrière un sourire (cf. Fig. 1).

Nous avons demandé à 40 sujets (23 hommes, 17 femmes), âgés de 19 à 36 ans (âge moyen 24 ans), de comparer les vidéos originales et des animations de Greta, et ce dans deux conditions : d'abord sans le son, puis avec le son. Dans chaque condition, les sujets devaient comparer la vidéo originale avec quatre animations différentes : deux animations rejouant chacune des émotion de base (spécifiées avec des données de la littérature [10, 24]) et deux animations des émotions mélangées (générées avec les deux approches citées précédemment). Dans toutes les animations, nous avons utilisé la bande son de la vidéo d'origine. Les sujets devaient classer les 4 animations selon leur degré de similarité avec la vidéo (cf. Fig. 2), puis annoter les émotions qu'ils percevaient dans l'animation la mieux classée (sélection d'un ou plusieurs labels émotionnels dans une liste).

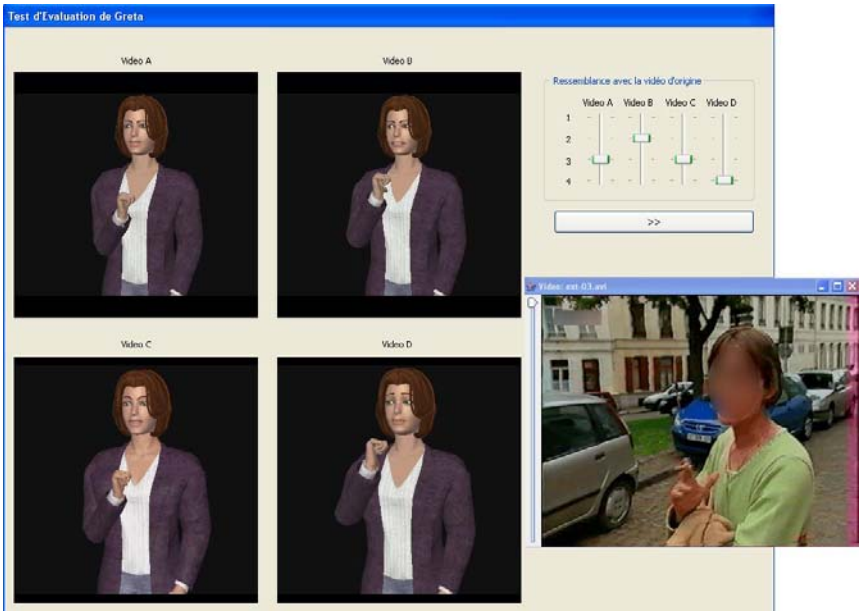

Fig. 2. L'exemple de superposition : 4 animations et 4 sliders pour le classement des animations ; la vidéo originale (non floutée dans le test) est présentée séparément.

4. Résultats

4.1. Superposition colère et désespoir

Nous avons calculé le nombre de fois où chaque animation a été classée comme la plus similaire à la vidéo. Dans la condition sans son, la colère est classée 1ère par 61% des sujets (niveaux multiples 20%, mélange facial 9%, désespoir 9%). Dans la condition avec son, la colère est classée 1ère par 33% des sujets (niveaux multiples 26%, mélange facial 24%, désespoir 17%). Nous avons mené une analyse de variance avec la Condition sonore (avec, sans son) et l'Animation (niveaux

multiples, mélange facial, colère, désespoir) comme facteurs intra-sujets. Le classement des animations a été converti en scores de similarité (le 1er rang est devenu un score de similarité de 3 points, le 4ème rang un score de similarité de 0 point). L'effet principal de l'Animation s'est montré significatif ($F(1/114)=15.86$; $p<0.001$, cf. Fig. 3), ainsi que l'interaction entre Condition sonore et Animation ($F(1/114)=5.98$; $p=0.001$, cf. Fig. 4) : l'effet de l'Animation est hautement significatif dans la condition sans son ($F(3/114)=24.11$; $p<0.001$) alors qu'il est en tendance dans la condition avec son ($F(2/114)=2.42$; $p=0.087$).

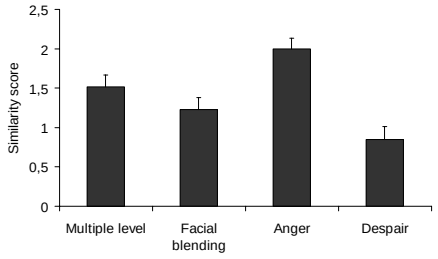

Fig. 3. Scores de similarité des 4 animations avec l'exemple de superposition.

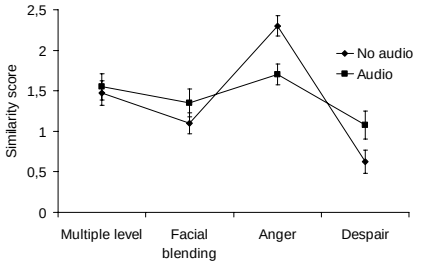

Fig. 4. Scores de similarité des 4 animations en fonction de la condition sonore.

Enfin, le tableau 1 résume les résultats de la tâche d'annotation à l'aide de macro-catégories (colère, tristesse). Ces résultats montrent que, même si l'animation la mieux classée était l'animation « colère », les sujets l'ont perçue comme une combinaison de plusieurs émotions.

Tableau 1. Pourcentages des sujets ayant perçu chaque macro-catégorie d'émotion dans l'animation classée comme la plus similaire à la vidéo témoin.

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

4.2. Déception masquée par la joie

Dans la condition sans son, la joie a été classée 1ère par 40% des sujets (mélange facial 33%, niveaux multiples 20%, déception 7%). Dans la condition avec son, le mélange facial a classé 1er par 38% des sujets (niveaux multiples 27%, joie 24%, déception 11%). Une analyse de variance avec la Condition sonore (avec, sans son) et l'Animation (niveaux multiples, mélange facial, joie, déception) comme facteurs intra-sujets, montre un effet principal de l'Animation ($F(1/114)=18.07$; $p<0.001$, cf. Fig. 6). Par ailleurs, le tableau 2 résume les résultats de la tâche d'annotation.

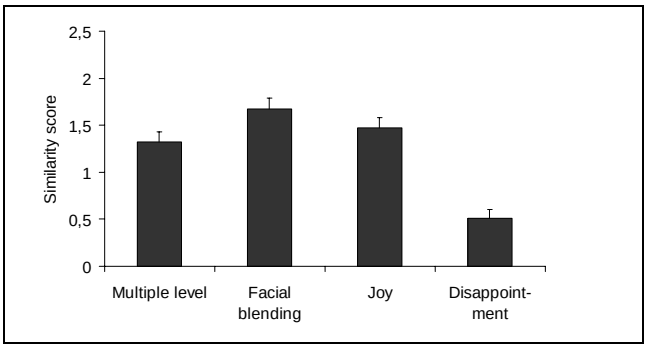

Fig. 6. Scores de similarité des 4 animations avec la vidéo initiale dans l'exemple de masquage.

Tableau 2. Pourcentages de sujets ayant perçu chaque macro-catégorie d'émotion dans l'animation classée comme la plus proche de la vidéo témoin.

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

4.3. Effets des modèles d'émotions complexes

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

5. Discussion

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

Par ailleurs, notre protocole expérimental nous a permis de comparer deux conditions, avec et sans le son. Les 4 animations correspondant à l'exemple de

superposition ont été mieux discriminées dans la condition sans son ; l'ajout d'indices verbaux a eu l'effet de diminuer les différences entre les animations. Dans les animations de superposition, les sujets ont mieux perçu la dimension tristesse / désespoir dans la condition avec le son. Nous pouvons en conclure que, dans cette séquence particulière, la colère a été principalement exprimée par des comportements non verbaux alors que le désespoir a été exprimé par le canal verbal.

Dans l'exemple de la déception masquée par la joie, les sujets ont mieux perçu les émotions négatives dans la condition avec le son. Ceci suggère que la personne dans cette séquence vidéo contrôlait mieux ses comportements non verbaux (émotions positives perçues dans la condition sans son) que ses comportements verbaux (indices négatifs perçus dans la condition avec le son). Cet effet important du comportement verbal peut être dû au contenu sémantique du message, ainsi qu'au fait que nous ayons utilisé la voix originale et non une synthèse vocale. Cette dernière hypothèse est compatible avec de précédentes études suggérant que les indices acoustiques expriment bien la colère et la douleur [1].

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

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

6. Conclusion et perspectives

A la suite de ce travail, nous envisageons d'améliorer notre modèle de génération d'émotions basiques et d'émotions mélangées. Nous souhaitons pouvoir attribuer des poids différents à différentes émotions à combiner. Le modèle computationnel de l'expressivité de l'agent nécessiterait également d'être amélioré pour mieux simuler les mouvements de bras expressifs et pour mieux correspondre aux comportements observés dans les vidéos (en incluant par exemple les mouvements du torse, en permettant de donner des spécifications d'expressivité séparées pour différentes parties du corps, etc.). Enfin, nous pensons que l'influence des canaux visuel et auditif doit être étudiée de manière plus approfondie, par exemple en utilisant un système de synthèse vocale, et en filtrant la sortie audio de sorte à rendre le contenu sémantique inintelligible tout en conservant la prosodie et la qualité de la voix [22].

Remerciements

Ce travail a été partiellement financé par le Réseau d'Excellence FP6 IST HUMAINE (http://emotion-research.net). Nous remercions vivement Maurizio Mancini pour son aide.

Références

1. Abrilian, S., Devillers, L., Buisine, S., Martin, J.-C.: EmoTV1: Annotation of Real-life Emotions for the Specification of Multimodal Affective Interfaces. 11th Int. Conf. Human-Computer Interaction (HCII'2005) (2005) Las Vegas, Nevada, USA

2. Abrilian, S., Devillers, L., Martin, J.-C.: Annotation of Emotions in Real-Life Video Inter-views: Variability between Coders. 5th Int. Conf. Language Resources and Evaluation (LREC'2006) (2006) Genoa, Italy

3. Albrecht, I., Schröder, M., Haber, J., Seidel, H.-P.: Mixed feelings: Expression of non-basic emotions in a muscle-based talking head. Special issue of Journal of Virtual Reality on "Language, Speech & Gesture" 8 4 (2005)

4. Bassili, J. N.: Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. Jour. Pers. Soc. Psychol. 37 11 (1979)

5. Cacioppo, J. T., Petty, R. P., Losch, M. E., Kim, H. S.: Electromyographic activity over facial muscle regions can differentiate the valence and intensity of affective reactions. Jour-nal of Personality and Social Psychology 50 (1986)

6. Constantini, E., Pianesi, F., Prete, M.: Recognizing Emotions in Human and Synthetic Faces: The Role of the Upper and Lower Parts of the Face. Intelligent User Interfaces (IUI'05) (2005) San Diego, CA, USA 20-27

7. Devillers, L., Abrilian, S., Martin, J.-C.: Representing real life emotions in audiovisual data with non basic emotional patterns and context features. 1st Int. Conf. Affective Computing and Intelligent Interaction (ACII'2005) (2005) Beijing, China 519-526

8. Douglas-Cowie, E., Devillers, L., Martin, J.-C., Cowie, R., Savvidou, S., Abrilian, S., Cox, C.: Multimodal Databases of Everyday Emotion: Facing up to Complexity. 9th European Conf. Speech Communication and Technology (Interspeech'2005) (2005) Lisbon, Portugal 813-816

9. Duy Bui, T. Creating Emotions And Facial Expressions For Embodied Agents. PhD Thesis. University of Twente. 2004.

10. Ekman, P.: Emotion in the human face. Cambridge University Press (1982)
11. Ekman, P.: The Face Revealed. Weidenfeld & Nicolson London (2003)
12. Ekman, P., Friesen, W. V.: Unmasking the face. A guide to recognizing emotions from facial clues. Prentice-Hall Inc., Englewood Cliffs, N.J. (1975)
13. Gallaher, P.: Individual differences in nonverbal behavior: Dimensions of style. Journal of Personality and Social Psychology 63 (1992)
14. Gouta, K., Miyamoto, M.: Emotion recognition, facial components associated with various emotions. Shinrigaku Kenkyu 71 3 (2000)
15. Kipp, M.: Gesture Generation by Imitation. From Human Behavior to Computer Character Animation. Boca Raton, Dissertation.com Florida (2004)
16. Martin, J.-C., Abrilian, S., Devillers, L., Lamolle, M., Mancini, M., Pelachaud, C.: Levels of Representation in the Annotation of Emotion for the Specification of Expressivity in ECAs. 5th International Working Conference On Intelligent Virtual Agents (IVA'2005) (2005) Kos, Greece 405-417
17. Martin, J.-C., Niewiadomski, R., Devillers, L., Buisine, S., Pelachaud, C.: Multimodal Complex Emotions: Gesture Expressivity And Blended Facial Expressions. Special issue of the Journal of Humanoid Robotics. Eds: C. Pelachaud, L. Canamero. (to appear)
18. Pandzic, I. S., Forchheimer, R.: MPEG-4 Facial Animation. The Standard, Implementation and Applications. John Wiley & Sons, LTD (2002)
19. Pelachaud, C.: Multimodal expressive embodied conversational agent. ACM Multimedia, Brave New Topics session (2005) Singapore 683 - 689
20. Richmond, V. P., Croskey, J. C.: Non Verbal Behavior in Interpersonal relations. Allyn & Bacon Inc. (1999)
21. Ruttkay, Z., Noot, H., ten Hagen, P.: Emotion Disc and Emotion Squares: tools to explore the facial expression face. Computer Graphics Forum 22 1 (2003)
22. Savvidou, S., Cowie, R., Douglas-Cowie, E.: Contributions of Visual and Auditory Chan-nels to Detection of Emotion. British Psychological Society Annual Conference (NI Branch) (2001) Cavan, Republic of Ireland
23. Scherer, K. R.: Analyzing Emotion Blends. Proceedings of the Xth Conference of the International Society for Research on Emotions (1998) Würzburg, Germany 142-148
24. Wallbott, H. G.: Bodily expression of emotion. European Journal of Social Psychology 28 (1998)
25. Wallbott, H. G., Scherer, K. R.: Cues and Channels in Emotion Recognition. Journal of Personality and Social Psychology 51 4 (1986).