Combinaisons d'expressions vocales, faciales et posturales des émotions chez un agent animé : ce que perçoivent les utilisateurs

Céline Clavel celine.clavel@limsi.fr
Laurence Devillers devil@limsi.fr
Jean-Claude Martin martin@limsi.fr

Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur
B.P. 133
91403 ORSAY CEDEX – FRANCE

Résumé :
L'objectif de ce papier est de présenter l'évaluation d'un système de communication à distance médiatisée par un avatar capable d'exprimer les émotions détectées dans la voix de l'utilisateur. Un module de reconnaissance audio des émotions exprimées dans la voix et une librairie d'expressions non verbales émotionnelles ont ainsi été développés. Le focus de ce papier porte plus particulièrement sur l'impact de l'intégration de ces deux composants sur la perception émotionnelle et sur la qualité perçue de la synchronisation audio/vidéo.
Mots-clés : Agent conversationnel, émotion, détection, voix, comportement non verbal

Abstract:
The aim of this paper is to present the evaluation of a mediated communication system by a virtual agent who can express emotions which are detected in the voice of the user. An audio recognition module of expressed emotions in the voice and a library of nonverbal emotional expressions have been developed. The interest of this paper focuses on the impact of integration of these two components on the emotional perception and the perceived quality of the audio / video.
Keywords: Virtual agent, emotion, detection, voice, non-verbal behavior

1 Introduction

Communiquer revient à échanger des signaux verbaux et non verbaux comme les expressions faciales ou posturales, la gestualité, l'intonation de la voix… Une partie de ces signaux se rapporte à la dimension émotionnelle. Comprendre les émotions véhiculées par autrui est un enjeu majeur de la communication. En effet, l'émotion est considérée comme une construction sociale dans la mesure où elle est constituée d'un ensemble de réponses prescrites qui se réfèrent à des normes ou à des attentes partagées [3 ; 15]. Ainsi, les expressions émotionnelles permettent d'établir, maintenir, cesser une relation sociale ou de donner du sens sur la nature de la relation établie [12]. Elles favorisent ou dissuadent également l'expression de certains comportements chez autrui.

Aujourd'hui, une nouvelle forme d'interaction se développe. L'homme est amené de plus en plus fréquemment à interagir avec des agents conversationnels animés. Ce sont des personnages animés de telle sorte qu'ils sont capables d'aider un utilisateur du web dans la résolution de diverses tâches : environnements virtuels de formation, commerce électronique, réservation de billets…

Alors que plusieurs études ont déjà évalué la perception par des utilisateurs d'expressions d'émotions dans des modalités isolées (parole, expressions faciales ou postures), nous connaissons encore peu de choses sur la manière dont ces expressions sont perçues lorsqu'elles sont combinées. Le système que nous avons utilisé pour effectuer cette étude est un prototype développé dans le cadre du projet Affective Avatars[1]. Il permet d'animer en temps réel des agents animés capables d'exprimer les émotions détectées dans la voix de l'utilisateur. La voix de l'utilisateur sert donc d'interface de contrôle de l'expressivité de l'avatar.

Pour se faire, nous décrivons dans la section 2 les modules du prototype Affective Avatar qui ont été utilisées. La section 3 présente l'évaluation du système développé et plus particulièrement l'examen de la cohérence perçue entre l'expressivité vocale de l'utilisateur et l'expressivité corporelle de

[1] Cette étude a été soutenue par le projet "ANR-07-TLOG-Affective Avatars – Cap Digital"

l'avatar qui le représente. Enfin nous terminons par une conclusion et les perspectives qu'ouvre cette étude.

2. Description du système

Le système proposé permet d'animer un agent virtuel à partir des émotions qui sont détectées dans la voix de l'utilisateur. Deux modules interviennent dans le traitement des émotions.

2.1. Module de détection des émotions

En parole, on distingue en général deux types d'information : les informations linguistiques et paralinguistiques. Concernant ces dernières, les différents indices peuvent être regroupés en trois catégories : les indices suprasegmentaux (fréquence fondamentale, énergie, durée, contour intonatif), les indices segmentaux (articulation, durée) et les indices intra-segmentaux (qualité vocale). Parmi les propriétés considérées comme pertinentes pour caractériser les manifestations émotionnelles dans la voix, les indices prosodiques sont les plus communément utilisés. De plus, les modèles les plus souvent employés pour la détection des émotions sont les SVM (Support Vector Machine), les GMMs (Mélange de Gaussiennes), les HMMs (Modèles de Markov cachés) les kNN, (K plus proches voisins), et les arbres de décision. Les résultats obtenus par ces différents classificateurs sont souvent comparables [4].

Toutefois, à l'heure actuelle, les études sur la détection des émotions dans la voix présentent un certain nombre de limites. Tout d'abord, la plupart d'entre elles se sont focalisées sur un ensemble minimal d'émotions : positives et négatives [14], émotion vs. état neutre [5], frustration/colère vs. neutre/amusé [2]. Il existe très peu de systèmes de détection des émotions qui vont au-delà de la classification de 4 à 5 émotions de base (très différentes). Les taux de détection deviennent très bas sur des émotions nuancées. Peu d'études sont menées avec des données issues de corpus enregistrés dans des contextes réels. Or l'étude de voix spontanées montre que les émotions sont souvent mélangées ou masquées dans les interactions naturelles et que les émotions spontanées sont produites à travers des indices de différents niveaux : acoustiques, prosodiques, des « affect bursts » (rire, toux, souffle), disfluences, indices lexicaux et dialogiques. Ainsi le challenge des études sur les comportements émotionnels à partir de données naturelles réside dans la multiplicité d'indices à la fois linguistiques et paralinguistiques témoignant des émotions. Il s'agit donc de construire des modules de détection des émotions en temps réel robustes aux variabilités d'expressivité prosodiques et linguistiques des locuteurs.

Pour cela, il faut utiliser un grand nombre de paramètres pour caractériser les émotions et développer des corpus de données naturelles où l'apprentissage se réalise à partir d'un grand nombre de voix (une centaine dans le cadre de cette étude). Ainsi, deux corpus ont été collectés et annotés pour servir de base d'apprentissage (3 heures de parole) et un corpus de test de 15 locuteurs a été collecté. Le schéma d'annotation utilisé est composé de catégories verbales et de dimensions. Ce schéma permet d'annoter des émotions complexes. Les modèles (arbres de décisions, SVM) ont été construits soit à partir des dimensions (comme l'activation), soit à partir de catégories émotionnelles (3, 4 ou 5 émotions).

Les résultats montrent peu de différences entre la librairie OpenEar (plateforme de référence pour la détection des émotions). Un protocole 10-folds cross-validation «Speaker – Indépendant» a été mis en place. L'évaluation finale a été menée avec un corpus de test collecté auprès de 15 personnes [6 ; 16 ; 17].

2.2. Module de l'expressivité non verbale de l'agent

De nombreuses études ont considéré les expressions faciales d'émotion [10] mais aussi posturales [9 ; 18]. Plusieurs travaux se sont appuyés sur ces données afin de nuancer les comportements expressifs de l'avatar en fonction des émotions exprimées [1 ; 11 ; 7]

Cet état de l'art nous a servi à définir des spécifications des expressions non-verbales (expressions faciales, posture, mouvement) de

quelques émotions de base (colère, peur, tristesse, joie, surprise). Ces spécifications informelles ont permis la conception des expressions faciales et animations posturales pour un personnage virtuel. Ces animations combinant expressions faciales et posturales ont ensuite été validées par des tests perceptifs [8].

2.3. Interfaçage des deux modules

Les postures et expressions du visage qui ont été créées sont rassemblées dans une bibliothèque de gestes expressifs. Les transitions entre animations sont gérées à travers un graphe d'états découpé en "zones émotionnelles". On est ainsi en mesure de piloter automatiquement un personnage à partir de la voix en synchronisant à la fois des gestes associés aux périodes de parole et de silence et les gestes expressifs associés aux émotions.

3. Evaluation

Les différents composants du système ayant été évalués indépendamment [16 ; 6 ; 17 ; 8], nous connaissons leurs points forts et leurs limites. Ainsi concernant l'expressivité non verbale, les expressions de colère, joie, et tristesse avec deux niveaux d'intensité sont bien reconnues aussi bien quand les modalités sont présentées de manière indépendante, expression du visage seule ou expression posturale seule que lorsque les expressions faciale et posturale sont combinées. La limite principale concernant l'expressivité non verbale est que la dimension temporelle n'a pas été prise en compte. L'objectif de l'étude que nous présentons dans cette section vise à étudier comment l'agent est perçu quand les deux principaux composants sont présents (expressivité audio et expressivité visuelle). Nous avons donc réalisé un test perceptif à partir d'enregistrements audio-vidéo du système pour mesurer les émotions perçues au cours de ces clips vidéo et la qualité perçue de l'interaction audio/vidéo.

3.1. Méthode

3.1.1. Participants

20 participants dont 8 femmes et 12 hommes ont passé l'expérimentation. La moyenne d'âge de l'échantillon est de 27 ans. Les participants ont un niveau d'étude compris entre bac +4 et doctorat. Ils sont tous issus d'un milieu informatique.

3.1.2. Matériel

Pour évaluer l'intégration des deux composants, il nous a semblé intéressant de manipuler le *contenu sémantique du message verbal* de l'agent et le *niveau de concordance des différentes modalités expressives*. Ainsi nous nous intéressons à l'impact de la présence d'un indice sémantique dans le message verbal de l'agent sur la perception du sujet. La perception est-elle favorisée lorsque l'agent dit « je suis un agent expressif en colère » (présence d'un indice sémantique) par rapport à la condition où l'agent dit « je suis un agent virtuel expressif » (absence d'un indice sémantique). La seconde variable manipulée correspond à la concordance entre les modalités expressives et nous nous interrogeons sur l'impact de la non concordance sur la perception. Lorsque la voix, le message et l'expression non verbale ne concordent pas, quel impact cela a-t-il sur la perception ? Le tableau 1 présente l'ensemble des conditions expérimentales proposées au cours de cette étude.

Type de message linguistique	Concordance entre l'audio et la vidéo	Audio	Vidéo
Présence d'un indice sémantique (triste, en colère, joyeux)	concordants	colère	colère
		joie	joie
		triste	triste
	non concordants	colère	joie
		triste	neutre
		colère	neutre
		joyeux	triste
		joyeux	colère 1
		joyeux	colère 2
Absence d'indice sémantique (je suis un agent virtuel expressif)	concordants	colère	colère
		joie	joie 1
		joie	joie 2
		triste	triste
		neutre	neutre
	non concordants	joie	colère
		triste	colère

TAB. 1 Ensemble des conditions expérimentales

Ainsi le matériel expérimental se compose de 16 enregistrements vidéo dont 7 enregistrements où l'agent exprime un message linguistique sémantiquement neutre (*absence d'indice sémantique sur l'émotion exprimée*) et 9 enregistrements où l'agent véhicule un

message linguistique sémantiquement émotionnel (*présence d'un indice sémantique sur l'émotion exprimée*).

Parmi les clips vidéo *où il n'y pas d'indice sémantique dans le message linguistique de l'agent*, 5 animations sont *concordantes* : Le locuteur avait l'intention d'exprimer de la joie dans sa voix, le système a bien détecté cette émotion, l'agent exprime donc cette émotion par son visage et sa posture. Deux autres animations sont *non concordantes*. Le locuteur avait l'intention d'exprimer une émotion mais c'est une autre émotion qui a été détectée par le système et c'est donc cette autre émotion que l'agent exprime par ses expressions faciales et ses postures.

Parmi les clips vidéo où il y a *un indice sémantique dans le message linguistique* de l'agent, 3 animations sont *concordantes*. 6 autres animations sont *non concordantes*.

3.2. Hypothèses et mesures

3.2.1. Les émotions perçues

La première série de mesures concernait les émotions perçues par les sujets au cours de chaque séquence vidéo. Les participants devaient indiquer dans quelle mesure ils percevaient de la colère, de la peur, de la surprise, de la tristesse, de la joie, du dégout et de la neutralité pour chaque clip vidéo via des échelles de Likert de 0 « pas du tout » à 4 « énormément ».

Notre but était ici de tester si la perception émotionnelle des utilisateurs pour les animations où la concordance audio/vidéo est vérifiée, est moins confuse que pour les animations où la concordance audio/vidéo n'est pas respectée. On s'attendait à ce que ce résultat soit d'autant plus vrai que les sujets étaient confrontés à un message émotionnel.

Pour les animations non concordantes, il s'agissait de voir l'impact des modalités expressives sur la perception. Le message linguistique sémantique est-il prévalent sur les messages facial et corporel ? Que se passe-t-il en l'absence d'indice sémantique dans le message linguistique ? Les sujets perçoivent-ils le décalage audio vidéo ou s'appuient-ils exclusivement sur les messages facial et corporel pour élaborer leur jugement ?

3.2.2. La qualité de l'interaction audio/vidéo

Le second groupe de mesure visait à recueillir des informations sur le ressenti subjectif des sujets quant à la qualité de la synchronisation audio/vidéo. 8 items ont été élaborés à cette fin. Il s'agissait de savoir si les sujets percevaient et étaient gênés par le léger décalage entre l'audio et la vidéo induit par la détection automatique des émotions.

Nous nous attendions à ce que les sujets lorsqu'ils étaient confrontés aux animations *non concordantes*, jugent l'interaction audio vidéo de moins bonne qualité que lorsqu'ils étaient confrontés à des animations où la *concordance* audio vidéo était respectée. Nous faisions l'hypothèse que ce résultat serait d'autant plus vrai que les sujets étaient dans la condition « présence d'un indice sémantique ».

3.3. Protocole

Les participants ont passé le test et répondu individuellement au questionnaire. La moitié des participants a été confrontée à la version «*présence d'un indice sémantique*» concordante ou non, le personnage indique dans son message l'émotion qu'il est censé exprimer. L'autre moitié a été confrontée à la condition ne donnant *pas d'indice dans le message verbal* de l'agent sur son état émotionnel. Le premier groupe de sujets a donc visionné 9 vidéos et le second 7.

Les passations se sont réalisées de manière individuelle sur un écran d'ordinateur. A l'issue de chaque vidéo, les participants devaient répondre à un questionnaire.

3.4. Résultats

3.4.1. Résultats concernant la perception émotionnelle

Perception émotionnelle des animations dans la condition : <u>présence d'un indice sémantique</u>

Le tableau 2 présente une vue synthétique de la perception émotionnelle moyenne de chaque

animation *cohérente* (la même information émotionnelle est censée être véhiculée par les différentes modalités expressives : voix, contenu linguistique et expression non verbale). Les résultats ont ensuite été analysés via des T de Student.

Animations			Emotions Perçues						
émotion détectée dans la voix	émotion indiquée dans le message linguistique	émotion exprimée par l'agent via les expressions faciales et posturales	joie	peur	colère	tristesse	dégout	surprise	neutralité
colère	colère	colère	0,7	0	**3**	0	0,9	0	0,5
joie	joie	joie	**2,8**	0	0,1	0	0	0,1	0,9
tristesse	tristesse	tristesse	0,3	0	0	**3,1**	0,9	0	0

TAB. 2 Perception émotionnelle moyenne des animations cohérentes en «présence d'un indice sémantique»

Les scores de perception pouvaient varier de 0 à 4. On observe que la perception émotionnelle de ces trois animations est conforme à ce qui était attendu. Les animations qui ont une certaine cohérence multimodale sont donc relativement bien perçues par les sujets.

Concernant *les animations non cohérentes* (des informations différentes sont véhiculées dans la voix, le message linguistique et les expressions non verbales de l'agent), le tableau 3 présente les résultats concernant la perception émotionnelle moyenne.

Animations				Emotions Perçues						
Numéro Animation	émotion détectée dans la voix	émotion indiquée dans le message linguistique	émotion exprimée par l'agent via les expressions faciales et posturales	joie	peur	colère	tristesse	dégout	surprise	neutralité
1	joie	colère	joie	2,70	0,30	0,20	0,00	0,10	0,60	0,50
2	neutre	colère	neutre	0,60	0,20	0,40	2,00	0,90	0,00	0,90
3	colère 1	joie	colère 1	1,70	0,00	1,70	0,00	0,50	0,10	0,30
4	colère 2	joie	colère 2	1,00	0,00	2,30	0,00	0,70	0,20	0,40
5	triste	joie	triste	0,30	0,00	0,00	3,10	1,60	0,00	0,30
6	neutre	triste	neutre	0,70	0,20	0,50	2,10	1,10	0,10	0,80

TAB. 3 : Perception émotionnelle moyenne des animations non cohérentes en «présence d'un indice sémantique»

Lorsque la correspondance entre les différentes modalités expressives n'est pas respectée, c'est-à-dire que l'émotion indiquée dans le message linguistique de l'agent n'est pas détectée dans la voix et donc n'est pas exprimée par le comportement non verbal de l'agent, cela ne favorise pas la perception émotionnelle.

Ces résultats suggèrent que les sujets ne s'appuient pas principalement sur le message linguistique comme on aurait pu s'y attendre mais élaborent plutôt leur perception à partir de l'expression non verbale de l'agent. Ces résultats semblent conforter nos spécifications concernant l'expression non verbale des agents qui est bien reconnue.

Enfin concernant les animations 3 et 4, où l'agent se déclare « joyeux » et exprime deux intensités différentes de colère (colère 1 et colère 2), les sujets perçoivent à la fois de la colère et de la joie. Cependant les sujets attribuent plus de colère et moins de joie à l'animation 4 qui exprime d'un point de vue non verbal une colère d'intensité plus forte. Ce résultat confirme nos spécifications concernant l'intensité expressive de la colère.

Perception émotionnelle des animations dans la condition : absence d'un indice sémantique

Les sujets sont confrontés ici à un agent qui n'indique pas dans son message linguistique l'émotion qu'il exprime, mais son expression est concordante à l'émotion que le locuteur avait l'intention d'exprimer. Le tableau 4 présente ces résultats. L'émotion la plus perçue au cours de chaque clip vidéo est l'émotion cible, c'est-à-dire l'émotion traitée dans la voix et exprimée par le comportement non verbal de l'agent. Les sujets, même en l'absence d'un indice sémantique, sont capables de percevoir les émotions traitées par le système et véhiculées par l'agent via son expressivité non verbale.

Cependant, les résultats soulignent également le fait que pour toutes les animations, les sujets perçoivent un peu de joie. Les deux animations censées véhiculer des émotions négatives (les animations 1et 4 du tableau 4) sont perçues comme véhiculant un peu de joie. Ce résultat peut être compris par le fait que l'agent au cours de chaque clip vidéo exprime une émotion via une séquence d'animations composées de l'animation neutre puis de l'animation expressive pour terminer par une expression neutre. L'apex de l'animation globale correspond à l'animation expressive. Toutefois, l'animation neutre n'est pas tout à fait neutre, en effet l'agent utilisé est censé être un agent assistant donc plutôt convivial et son expression

neutre est donc empreinte d'un léger sourire. Aussi lorsqu'il n'y pas d'indice sémantique dans le message linguistique de l'agent, les sujets semblent porter beaucoup d'attention à l'ensemble de son expressivité non verbale, traiter plus d'informations et de ce fait percevoir plus d'émotions.

Animations			Emotions Perçues						
Numéro Animation	émotion détectée dans la voix	émotion exprimée par l'agent via les expressions faciales et posturales	joie	peur	colère	tristesse	dégout	surprise	neutralité
1	Colère	Colère	**1,00**	0,10	**2,50**	0,00	0,10	0,40	0,20
2	Joie	Joie 1	**3,50**	0,00	0,00	0,00	0,00	0,40	0,60
3	Joie	Joie 2	**3,20**	0,00	0,20	0,10	0,10	0,90	0,10
4	Triste	Triste	**1,30**	0,20	1,20	**2,40**	0,70	0,10	0,60
5	Neutre	Neutre	**1,20**	0,00	0,00	0,10	0,00	0,00	**2,60**

TAB. 4 : Perception émotionnelle moyenne des animations cohérentes pour la condition «absence d'un indice sémantique»

Lorsque le système n'a pas détecté l'émotion que le locuteur avait l'intention d'exprimer, la perception émotionnelle en est affectée (tableau 4). La mauvaise détection du système induit de la confusion et le message émotionnel véhiculé par l'agent n'est pas clairement interprété par les utilisateurs qui perçoivent un mélange d'émotions.

Animations				Emotions Perçues						
Numéro Animation	émotion intentionnée dans la voix	émotion détectée dans la voix	émotion exprimée par l'agent via les expressions faciales et posturales	joie	peur	colère	tristesse	dégout	surprise	neutralité
1	joie	colère	colère	**1,80**	0,00	**2,50**	0,20	0,30	0,00	0,40
2	triste	colère	colère	0,00	0,20	**2,30**	**1,40**	0,90	0,00	0,80

TAB. 5 Perception émotionnelle moyenne des animations **incohérentes** en «l'absence d'un indice sémantique»

Comparaison de la perception émotionnelle moyenne des animations en condition <u>présence</u> d'un indice sémantique vs en condition <u>absence</u> d'un indice sémantique

L'analyse des résultats via des T de Student concernant la comparaison des deux conditions révèle que la perception émotionnelle ne varie pas en fonction des conditions expérimentales pour les animations cohérentes. Quelle que soit la nature du message linguistique (absence vs présence d'un indice sémantique), les sujets rapportent une perception émotionnelle équivalente.

Concernant les animations non cohérentes, la comparaison ne peut porter que sur les animations combinant de la joie dans la voix et de la colère dans l'expression faciale et posturale de l'agent. C'est la seule combinaison qui se retrouve dans les deux conditions expérimentales. A nouveau quelle que soit la condition expérimentale, la perception de ces animations incongruentes ne varie pas. La présence de l'indice sémantique ne favorise pas la perception du message véhiculé dans la voix.

3.4.2. Résultats concernant la perception de la qualité de la synchronisation audio/vidéo

La qualité de la synchronisation est calculée à partir des scores obtenus aux 8 items évaluant la qualité de l'interaction audio/vidéo. Cet indice peut prendre des valeurs comprises entre 0 et 32. Un score proche de 32 rapporte une évaluation très satisfaisante de la synchronisation audio/vidéo et un score proche de 0 au contraire montre que les sujets sont très critiques à l'égard de la synchronisation audio/vidéo.

Perception de la qualité de la synchronisation audio-vidéo pour les animations cohérentes en <u>présence</u> d'un indice sémantique

Le tableau 6 présente ces résultats pour les trois animations qui véhiculent la même émotion via les différentes modalités expressives. Seule l'animation de la joie à un niveau de satisfaction relativement satisfaisant. La qualité moyenne est de 18,80 ce qui est supérieure à la moyenne (16). Pour les deux autres animations, les sujets jugent la qualité de la synchronisation comme peu satisfaisante.

Animations			Qualité Globale
Emotion détectée dans la voix	Emotion indiquée dans le message linguistique	Emotion exprimée par l'agent via les expressions faciales et posturales	
colère	colère	colère	11,00
joie	joie	joie	18,80
tristesse	tristesse	tristesse	13,90

TAB.6 Qualité moyenne de la synchronisation audio/vidéo pour les animations cohérentes en «présence d'un indice sémantique»

Perception de la synchronisation audio-vidéo

pour les animations cohérentes en l'absence d'un indice sémantique

Le tableau 7 présente ces résultats pour les cinq animations qui véhiculent la même émotion via les différentes modalités expressives. Trois animations sur cinq obtiennent un niveau de satisfaction relativement satisfaisant (supérieur à la moyenne).

Animations			Qualité Globale
Numéro Animation	Emotion détectée dans la voix	Emotion exprimée par l'agent via les expressions faciales et posturales	
1	Colère	Colère	17,50
2	Joie	Joie 1	18,30
3	Joie	Joie 2	14,60
4	Triste	Triste	10,80
5	Neutre	Neutre	18,90

TAB.7 Qualité moyenne de la synchronisation audio/vidéo pour des animations cohérentes en «l'absence d'un indice sémantique»

Comparaison de la qualité moyenne perçue de la synchronisation audio-vidéo des animations en condition présence d'un indice sémantique vs en condition absence d'un indice sémantique

Les deux sections précédentes semblent suggérer que les sujets ont jugé le système comme plus performant (c'est-à-dire comme proposant des animations correspondant mieux aux expressions verbales) quand l'agent ne proposait pas d'indice sémantique dans son message linguistique que l'inverse. Les différentes animations cohérentes en condition «absence d'un indice sémantique» obtiennent de meilleurs scores en qualité de synchronisation que les animations cohérentes en condition «présence d'un indice sémantique». Ces résultats ont été analysés via des T de Student afin de vérifier que ces différences sont significatives. Concernant les animations relatives à la joie et à la tristesse, aucune différence ne s'observe. La présence ou l'absence d'un indice émotionnel dans le message verbal de l'agent n'influe pas sur la qualité perçue de la synchronisation audio-vidéo pour ces animations. Ce résultat ne se vérifie pas pour les animations relatives à la colère. Les sujets jugent la qualité meilleure dans la condition « absence d'indice » que dans la condition « présence d'indice ».

Ces résultats nous questionnent quant à la temporalité du comportement non verbal associé à l'expression multimodale de l'émotion Il est possible que selon la nature de l'émotion, la temporalité de l'expression varie. Le système de détection/animation développé, nécessite une certaine latence entre la détection des émotions dans la voix de l'utilisateur et l'expression non verbale associée. Cette latence devient moins frappante quand le discours du locuteur ne se limite pas à une seule phrase ce qui n'est pas le cas dans le protocole expérimental réalisé. Ces résultats suggèrent qu'on attendrait une expression rapide pour un évènement relatif à de la colère alors qu'à des évènements reliés à la tristesse ou la joie, l'expression pourrait arriver plus tardivement. Ceci expliquerait pourquoi les sujets jugent la qualité de l'interaction audio/video plus mauvaise lorsqu'ils sont confrontés à un agent affirmant exprimer de colère et mettant relativement du temps à l'exprimer que lorsqu'ils sont face à un agent n'affirmant pas quelle émotion il exprime. Le message linguistique empreint d'un indice sémantique concernant l'expression de l'agent induirait plus d'attente chez les utilisateurs concernant l'expressivité de l'agent et cela aurait un impact différent sur la qualité perçue selon les émotions traitées.

4. Discussion et perspectives

Cette évaluation perceptive concernant l'intégration du système de détection émotionnelle et d'expression émotionnelle non verbale suggère que les sujets perçoivent mieux les animations émotionnelles quand un indice sémantique est présent dans le message verbal de l'agent. Cependant, cet indice dégrade la perception de la synchronisation audio vidéo en créant probablement chez l'utilisateur des attentes quant à l'expressivité de l'agent.

Cette étude a été réalisée auprès d'un petit échantillon et devrait être reproduite afin de confirmer ces premiers résultats. Enfin, elle questionne au sujet de la temporalité, de la

dynamique de l'expression non verbale et de la qualité de mouvement (Wallbott 1998) et les relations entre la forme des gestes et les émotions [13].

Références

[1] André, E., Klesen, M., Gebhard, P., Allen, S., Rist, T. (2000). *Exploiting Models of Personality and Emotions to Control the Behavior of Animated Interface Agents* In J. Rickel (Ed.), Workshop on "Achieving Human-Like Behavior in Interactive Animated Agents". 3-7.

[2] Ang, J., Dhillon, R., Krupski, A., Shriberg, E. & Stolcke, A. (2002.). *Prosody-Based Automatic Detection of Annoyance and Frustration in Human-Computer Dialog*. Proceedings of International Conference on Spoken Language Processing, 2037-2040.

[3] Averill, J. R. (1980). *A contructivist view of emotion*. In R. Plutchick & H. Kellerman (Eds), Emotion, theory, research, and experience: Theories of emotions. New York: Academic Press.

[4] Batliner, A. Hacker, Ch. Steidl, S., Nöth, E., S. D'Arcy, Russell, M & Wong, M. (2004). *"You stupid ting box"- children interacting with the AIBO robot: A cross-linguistic emotional speech corpus*. 4th international Conference on Language Resources and Evaluation, 171-174.

[5] Batliner, A., Fisher, K., Huber, R., Spilker, J. & Noth, E. (2003). *How to Find Trouble in Communication*. Speech Communication, 40, 117-143.

[6] Brendel, M., Zaccarelli, R., Devillers, L. *"Building a system for emotions detection from speech to control an affective avatar"*, LREC 2010

[7] Buisine, S., Abrilian, S., Niewiadomski, R., Martin, J.-C., Devillers, L. and Pelachaud, C. (2006). *Perception of Blended Emotions: from Video Corpus to Expressive Agent.* IVA'2006, 93-10

[8] Clavel, C., Plessier, J., Martin, J.C., Ach, L., Morel, B. (2009). *Combining Facial and Postural Expressions of Emotions in a Virtual Character*. IVA'09, 287 - 300.

[9] Coulson, M. (2004). *Attributing emotion to static body postures Recognition accuracy Confusions and viewpoint dependence*. Journal of nonverbal behavior, 28 (2), 117-139

[10] Ekman, P. and Friesen, W. (1975). *Unmasking the face. A guide to recognizing emotions from facial clues*, Prentice-Hall Inc., Englewood Cliffs, N.J.

[11] Gillies, M., Crabtree, I. B. and Ballin, D. (2006). *Individuality and Contextual Variation of Character Behaviour for Interactive Narrative*. AISB Workshop

[12] Kaiser,S. Wehrle, T. & Schenkel, K. (2009) *Expression faciale des émotions*. In D. Sander, & K. Scherer, (Eds). Traité de psychologie des émotions. Paris : Dunod, 77-108

[13] Kipp, M., Martin, J.-C. (2009). *Gesture and Emotion: Can basic gestural form features discriminate emotions?* ACII-09

[14] Lee, C.M., Narayanan, S. & Pieraccini, R. (2002). *Combining acoustic and language information for emotion recognition.* Proc. International Conference on Spoken Language Processing, 873-876.

[15] Parkinson, B. (1996). *Emotions are social*. British Journal of Psychology

[16] Rollet, N., Delaborde, A., Devillers, L., (2009). "Protocol CINEMO: The use of fiction for collecting emotional data in naturalistic controlled oriented context", ACII 2009

[17] Schuller, B., Rollet, N., Zaccarelli, R., Devillers, L. (2010). "*CINEMO: A French Spoken Language Resource for Complex Emotions: Facts and Baselines*" LREC

[18] Wallbott, H. G. (1998). *Bodily expression of emotion.* European Journal of Social Psychology (28), 879-896.