Détection d'émotions dans la voix de patients en interaction avec un agent conversationnel animé

Clément Chastagnol[1,2] *Laurence Devillers*[1,3]
(1) LIMSI-CNRS
(2) Université Paris-Sud Orsay
(3) GEMASS, Université Paris-Sorbonne 4
cchastag@limsi.fr, devil@limsi.fr

RÉSUMÉ

Le projet français ANR ARMEN a pour objectif de construire un robot assistant pour les personnes âgées et handicapées. L'interaction avec le robot est réalisée avec un agent conversationnel animé (ACA), le robot est une plateforme mobile. Ce travail se concentre sur la construction du module de détection d'émotions du système robotique. A cette fin, des données ont été collectées auprès de 77 patients de plusieurs centres médicaux. L'interaction avec les sujets était presque entièrement conduite de manière naturelle en parlant avec l'agent virtuel. La difficulté spécifique de ce projet réside dans la grande variété de voix (âgées, dégradées) et de comportement affectif des utilisateurs. Nos premiers résultats montrent un score de 46% de bonne détection sur quatre classes émotionnelles (Colère, Joie, Tristesse, Neutre). Nous analysons ces scores selon l'âge et la qualité vocale.

ABSTRACT

Emotions detection in the voice of patients interacting with an animated conversational agent

The French ARMEN ANR-funded project aims at building an assistive robot for elderly and disabled people We focus in this paper on the emotion detection module for this robot. The interaction is almost entirely conducted in a natural, spoken fashion with a virtual agent. 77 patients have participated to the data collection. The specific difficulty in this project lies in the large variety of user voices (elderly, damaged) and affective behaviors of the patient. Our first results show 46% of good emotion detection on four classes (Anger, Joy, Neutral and Sadness). We first try to analyze the differences due to age and voice quality.

MOTS-CLÉS : robot assistant, détection d'émotions spontanées, qualité vocale

KEYWORDS : assistive robot, spontaneous emotions detection, vocal quality

1. Introduction

Les machines et les ordinateurs ont vocation à devenir de plus en plus sociales et tournées vers l'utilisateur humain. Les récents développements dans les domaines des robots d'assistance et des interfaces Homme-Machine ont conduit à la prédiction de "robots sociaux" (*social assistive robots*). Le terme a été proposé par Feil-Seifer et Mataric (Feil-Seifer et Mataric, 2005) et définit une machine conçue pour deux objectifs : soutenir et aider physiquement des personnes en situation de handicap moteur et proposer une interaction sociale à l'utilisateur, en général dans le cadre d'une tâche bien

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 137–144,
Grenoble, 4 au 8 juin 2012.

délimitée (rééducation, coaching...). Alors que la manipulation d'objets réels rend nécessaire la présence physique d'un robot, le rôle social peut être assuré par un Agent Conversationnel Animé (ACA) affiché sur un écran. Beaucoup d'efforts ont été mis dans le développement de robots assistants, en particulier à destination des personnes âgées (Graf et al., 2002). Les robots sociaux sont plus récents et ils ont notamment été mis en application pour la thérapie d'enfants autistes (Robins et al., 2005). Enfin, la recherche dans le domaine des interactions sociales avec des ACA s'est concentrée sur le problème des interactions naturelles et multimodales et sur l'évolution de l'engagement de l'utilisateur au cours du temps dans plusieurs tâches comme agent immobilier (Cassel, 2000) ou coach sportif (Bickmore et al., 2005).

L'analyse des états affectifs (émotions, sentiments) et de la personnalité de l'utilisateur est encore très rudimentaire en robotique et se limite souvent à des interactions tactiles (Shibata et Tanie, 1999). C'est néanmoins en comprenant ces facteurs qu'il sera possible d'ajouter des compétences sociales aux robots (Delaborde et Devillers, 2010) ou aux ACA (Schroeder et al., 2008). Les interactions sociales sont caractérisées par un échange continu et dynamique de signaux porteurs d'information. Les humains peuvent communiquer sur plusieurs niveaux simultanément en produisant et en comprenant ces signaux. Parmi les différents canaux de communication utilisés, l'expression vocale communique la plus riche variété d'informations ; c'est aussi la modalité la plus naturelle pour communiquer de la signification, de l'émotion et de la personnalité. L'expression vocale est caractérisée par une composante verbale, porteuse du langage, et par une composante non-verbale ou para-linguistique (prosodie, intonations, hésitations).

Nous présentons ici la conception et la construction d'un module de détection d'émotions pour un robot social d'assistance, interagissant grâce à un agent conversationnel animé (ACA). La section 2 décrit les spécifications du système. Dans la section 3, des détails sur le protocole expérimental de recueil de données émotionnelles spontanées sont présentées. Le corpus collecté est présenté dans la section 4 et des premiers résultats expérimentaux dans la section 5.

2. Spécifications du robot ARMEN

Le projet français ANR ARMEN a pour but de concevoir un robot assistant pour les personnes âgées et handicapées, capable d'aller chercher des objets hors-de-portée ou perdus, les manipuler et d'évoluer dans un environnement réaliste. De plus, il doit pouvoir appeler de l'aide en cas d'urgence et se comporter comme un compagnon de vie en comprenant des discussions simples sur des sujets spécifiques et moduler ses réponses en fonction de l'état émotionnel de l'utilisateur. L'interaction doit se dérouler le plus naturellement possible en parlant à un ACA affiché à un écran. Le système de communication en développement se compose de plusieurs modules : un module de reconnaissance de la parole, un module de détection d'émotions et un module de gestion de dialogue.

La difficulté spécifique à ce projet se situe dans la grande variété de voix des utilisateurs : certains ont subi des interventions chirurgicales (trachéotomie par exemple) qui les empêchent de produire une voix claire et forte. La plupart des personnes handicapées moteur (para- ou tétraplégiques) ont également des voix très faibles car ils

ont perdu le contrôle de leurs muscles abdominaux. Les bruits produits par les canules (valve posée suite à une trachéotomie) et les respirateurs sont également problématiques. Même les voix de personnes âgées en bonne santé peuvent être difficiles à traiter car elles sont parfois complètement dévoisées ou chuchotées.

Il existe peu de corpus disponibles contenant de la parole émotionnelle spontanée (Zeng et al., 2009) et encore moins avec la typologie de voix présente dans ce projet. C'est pourquoi des collectes de données dans les établissements médicalisés partenaires ont été décidées. Dans ces collectes, des émotions ont été provoquées chez des patients en les plongeant de manière la plus proche de l'application finale dans une interaction avec un ACA. L'interaction était structurée autour de scénarios inspirés de la vie quotidienne des patients et conçus avec le personnel des centres médicaux ; ces scénarios étaient choisis pour leur charge émotionnelle potentielle et pour que les patients s'y associent facilement. Le recrutement des patients pour la collecte a été effectué de manière la plus large possible en termes de qualité vocale pour examiner les cas les plus difficiles et pouvoir établir des limites de fonctionnement.

3. Protocole et dispositif expérimentaux

Deux collectes de données ont été organisées à Montpellier en France, en collaboration avec l'association APPROCHE, qui promeut l'utilisation des nouvelles technologies pour aider les personnes dépendantes. Les enregistrements ont eu lieu en juin 2010 et en juin 2011, sur une période de huit jours au total. Trois centres médicaux étaient impliqués : un centre de rééducation fonctionnelle, un EHPAD (Établissement d'Hébergement pour Personnes Âgées Dépendantes) et un centre de vie pour personnes handicapées. La complémentarité de ces trois sites a permis d'enregistrer un large spectre de voix, parfois très marquées.

Les expérimentations se sont déroulées selon la technique du Magicien d'Oz avec un interviewer, un module de dialogue sur un ordinateur portable et un opérateur déclenchant le module à l'insu du sujet, qui pensait réellement avoir une conversation avec le module. Les réactions obtenues sont donc très proches d'une interaction homme-machine en contexte réel. Pour la première collecte, le sujet interagissait uniquement avec une voix synthétique ; un ACA a été ajouté pour la seconde.

Les collectes étaient divisées en trois phases : dans la première, l'interviewer présentait le projet au sujet et expliquait le but de l'expérience. Le sujet était alors invité à jouer des émotions en exagérant le ton de sa voix. Dans la deuxième phase, le sujet interagissait avec le module de dialogue dans le cadre de plusieurs scénarios (8 scénarios courts pour la première collecte, 3 plus développés pour la seconde), conçus pour induire des émotions par projection ; l'interviewer expliquait le scénario courant au sujet et lui demandait de s'imaginer en situation et de faire comprendre au module l'état émotionnel qu'il ressentait. Le sujet interagissait alors avec le module de dialogue piloté par l'opérateur, qui déclenchait des réponses scriptées selon des stratégies pré-établies : comprendre, montrer de l'empathie, de pas comprendre, se tromper... Le dialogue durait en moyenne 4 à 5 tours de paroles par scénario pour la première collecte et jusqu'à 20 pour la seconde. Un extrait du scénario "Colère" de la première collecte est reproduit dans le tableau 1. Le sujet devait expliquer à l'ACA qu'il était énervé car il attendait un

médecin pour examen et qu'il était très en retard. Dans la troisième et dernière phase, le sujet répondait à des questions posées par l'interviewer concernant la qualité de l'interaction, l'acceptabilité de l'ACA et leur propre personnalité.

Patient	Bon bah alors qu'est-ce que c'est ce-ce-ce-ce-cette pagaille là. Je comprends pas hein, il m'avait promis qu'il serait là puis il est pas là, mais... C'est pas possible quoi, y'en a marre hein.
Agent	Tu es en colère ?
Patient	Aaaahlala, ça suffit, je-je-je, ras-le-bol. C'est incroyable quoi, incroyable.
Agent	Oulala, tu as l'air très énervé.
Patient	Ouais ouais ouais. Là maintenant, c'est incroyable.
Agent	C'est vrai que ça fait longtemps, je comprends que ça t'agace.
Patient	Ouais, ça m'agace beaucoup, ouais.

TABLE 1 – Extrait d'un enregistrement d'un patient interagissant avec l'agent dans le scénario « Colère ».

Les scénarios ont été conçus conjointement avec les membres du personnel du centre de rééducation et approuvés par les médecins. Il ont été inspirés de situations de la vie quotidienne et étaient prévus pour se rapprocher de l'expérience réelle que pourrait avoir un utilisateur du robot final. Les deux collectes ont été filmées et enregistrées, les sessions ont duré en moyenne 20 minutes, avec un minimum de 9 minutes et un maximum de 37 minutes. Marie, l'ACA manipulé lors de la seconde collecte, est développé sur la plate-forme MARC développée au LIMSI-CNRS (Courgeon et al., 2008) ; une photo de Marie se trouve en Figure 1. L'ACA était contrôlé par une interface également développée au LIMSI-CNRS pour les besoins de ces collectes et utilisant le langage de balises émotionnelles BML (Vilhjalmsson et al., 2007) pour animer le visage de l'ACA.

Figure 1 - Illustration de Marie, l'ACA interagissant avec l'utilisateur.

4. Présentation du corpus ARMEN

Le corpus ARMEN_1 complet pour la première collecte contient 17,3 heures d'enregistrements audio et vidéo de 52 personnes âgées de 16 à 91 ans. Le corpus ARMEN_2 pour la deuxième collecte contient 8,7 heures d'enregistrements audio et vidéo de 25 personnes de 25 à 91 ans. Les sujets sont atteints de pathologies variées (handicaps physiques et cognitifs) et plus ou moins dépendants selon l'échelle AGGIR utilisée par les médecins français et basée sur la définition de l'US Diagnosed Related Groupe (Fetter et al., 1980).

Pour les deux collectes, les enregistrements audio de la deuxième phase de l'expérience (scénarios) ont été segmentés et étiquetés selon un protocole détaillé par deux annotateurs experts en segments d'au plus 5 secondes, cohérents au niveau du contenu émotionnel. Un schéma d'annotation simple a été utilisé, comprenant 5 étiquettes émotionnelles (Colère, Joie, Neutre, Peur, Tristesse, plus une étiquette "Poubelle" pour éliminer les segments bruités) et une échelle d'Activation à 5 degrés.

Les deux corpus annotés ainsi obtenus (ARMEN_1 et ARMEN_2) sont détaillés dans le tableau 2. Seuls les segments consensuels ont été gardés et utilisés pour les expériences décrites ci-dessous.

5. Premières expériences

Les résultats présentés plus bas montrent des premiers résultats de classification sur les étiquettes émotionnelles uniquement. Elles tentent d'établir une différence de performance selon l'âge et la voix des locteurs et donnent une idée de la complexité des données. Le protocole pour chaque expérience a été le suivant : la classe Neutre a d'abord été sous-échantillonnée pour obtenir une répartition des classes moins déséquilibrée et la classe Peur a été supprimée car elle contenait trop peu d'instances. Des paramètres acoustiques (384 paramètres, utilisés pour le challenge Interspeech 2009 (Eyben et al., 2009)) ont ensuite été extraits des segments audio par la librairie openEAR (Schuller et al., 2009). Une optimisation "grid search" à deux dimensions a été réalisée sur le paramètre de coût C et le paramètre Gamma d'un classifieur SVM avec un noyau à base radiale. Pour chaque couple de paramètres (C, Gamma), une évaluation Leave One Speaker Out a été réalisée, pour s'assurer que le classifieur n'apprenait pas les voix des locuteurs, ce qui est à prendre particulièrement en compte dans le cas de données avec des voix très spécifiques et très différentes. La moyenne non-pondérée des précisions par classe a été utilisée pour quantifier la performance du classifieur, vu le déséquilibre persistant entre les classes.

Un ensemble regroupant les deux corpus ARMEN_1 et ARMEN_2 (ARMEN $1+2$ équilibré) a d'abord été évalué. Puis cet ensemble a été divisé en deux paires de sous-ensembles selon deux critères : l'âge des locuteurs (plus ou moins de 60 ans) et la qualité vocale (normale ou dégradée), selon des informations fournies par des orthophonistes concernant la qualité vocale (volume faible, sauts de volume, timbre de voix altéré, dévoisement...), l'articulation et selon la présence de bruits parasites (respirateurs, valves...).

Nom du corpus	ARMEN_1	ARMEN_2	ARMEN 1+2 équilibré	Voix agées vs jeunes	Voix normales vs dégradées
Nombre de segments consensuels (% du total)	1996 (46%)	1588 (63%)	2080	658 / 997	978 / 677
Score Kappa	0.33	0.37	N/A	N/A	N/A
Nombre de locuteurs	52	25	77	31 / 37	33 / 35
Répartition des classes					
Colère	406 (20%)	92 (6%)	498 (24%)	108 (16%) / 260 (26%)	247 (25%) / 121 (18%)
Joie	427 (21%)	236 (15%)	663 (32%)	231 (35%) / 309 (31%)	383 (39%) / 157 (23%)
Neutre	748 (38%)	1158 (73%)	520 (25%)	164 (25%) / 249 (25%)	244 (25%) / 169 (25%)
Peur	97 (5%)	21 (1%)	0	0	0
Tristesse	318 (16%)	81 (5%)	399 (19%)	155 (24%) / 179 (18%)	104 (11%) / 230 (34%)

TABLE 2 – Détails sur la composition du corpus ARMEN.

Les premiers résultats montrent que le système de détection d'émotions global a une meilleure performance que les systèmes entraînés de manière spécifiques sur une catégorie d'âge ou de qualité vocale donnée.

Quelques remarques peuvent être faites : la Colère est beaucoup mieux reconnue pour les voix jeunes que pour les voix âgées, mais c'est le contraire pour la Joie. Concernant les voix normales, la Tristesse n'est pas reconnue (environ le même niveau que le hasard), mais elle est deux fois mieux reconnue pour les voix dégradées. Une expérience cross-corpus a également été menée (ses résultats ne figurent pas dans le tableau 3) avec les voix normales et dégradées. En entraînant le classifieur sur les voix dégradées et en testant sur les voix normales, la précision pour la classe Tristesse grimpe à 51% alors qu'elle descend à 20% lorsque l'on fait le contraire. Cela suggère que les voix dégradées dans ce corpus expriment la classe Tristesse d'une manière plus séparable des autres classes d'émotion que les voix normales. Il faudrait dependant vérifier d'éventuels effets de différence de taille de données d'apprentissage pour pouvoir conclure.

Sous-ensemble considéré	Score moyen	Colère	Joie	Neutre	Tristesse
Ensemble complet	46,1%	48,0%	53,1%	43,3%	40,1%
Voix jeunes	43,7%	52,7%	46,0%	42,2%	34,1%
Voix âgées	41,0%	19,4%	64,9%	42,1%	37,4%
Voix normales	43,2%	55,5%	51,4%	41,0%	25,0%
Voix dégradées	43,9%	41,3%	39,5%	42,6%	52,2%

TABLE 3 – Premiers résultats.

6. Conclusion

Nos premiers résultats montrent qu'il est difficile de traiter des données spontanées avec une qualité vocale très variable (voix âgées, dégradées...). De prochaines expériences tenteront de déterminer l'empreinte de certaines classes de qualité vocale et d'âge du locuteur avec des ensembles de paramètres acoustiques adaptés (Brendel et al., 2010) et d'améliorer les scores de détection d'émotion en utilisant la sortie du module de reconnaissance de la parole ainsi que des paramètres acoustiques supplémentaires ; les stratégies d'interaction de l'ACA et son niveau d'expressivité seront également l'objet de futures expériences.

Remerciements

Cette étude est financée par le projet français ANR ARMEN (http://projet_armen.byethost4.com). Les auteurs voudraient remercier l'association APPROCHE pour leur assistance durant les collectes de données.

Références

Bickmore, T., Caruso, L., Clough-Gorr, K. et Heeren, T. (2005). It's just like you talk to a friend - Relational agents for older adults. *In Interacting with Computers,* volume 17, numéro 6, pages 711–735.

Brendel, M., Zaccarelli, R., Schuller, B. et Devillers, L. (2010). Towards measuring similarity between emotional corpora. *In Proc. 3rd ELR0A Internat. Workshop on EMOTION*, Valetta, Malte, pages 58–64.

Cassel, J. (2000). More Than Just Another Pretty Face: Embodied Conversational Interface Agents. *In Communications of the ACM*, volume 43, numéro 4, pages 70–78.

Courgeon, M., Martin, J-C. et Jacquemin, C. (2008). MARC: a Multimodal Affective and Reactive Character. *In Proceedings of the 1st Workshop on AFFective Interaction in Natural Environments*, Chania, Crète.

Delaborde, A. et Devillers, L. (2010). Use of non-verbal speech cues in social interaction between human and robot: Emotional and interactional markers. *In Proceedings of the 3rd ACM Workshop on Affective Interaction in Natural Environments*, pages 75–80.

Eyben, F., Wöllmer, M. et Schuller, B. (2009). openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. *in Proc. 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction 2009 (ACII 2009)*, Amsterdam, Pays-Bas, pages 1–6.

Feil-Seifer, D. et Mataric, M.J. (2005). Defining socially assistive robotics. *In Proc. IEEE International Conference on Rehabilitation Robotics (ICORR'05)*, Chicago, Il, USA, pages 465–468.

Fetter, R.B., Shin, Y., Freeman, J.L., Averill, R.F. et Thompson, J.D. (1980). Case-Mix definition by Diagnosis Related Groups. *In Medical Care*, volume 18, numéro 2.

Graf, B., Hans, M., Kubacki, J. et Schraft, R. (2002). Robotic home assistant care-o-bot II. *In Proceedings of the Joint EMBS/BMES Conference*, Houston, TX, USA, volume 3, pages 2343–2344.

Robins, B., Dautenhahn, K., Boekhorst, R. et Billard, A. (2005). Robotic assistants in therapy and education of children with autism: Can a small humanoid robot help encourage social interaction skills?. *In Universal Access in the Information Society (UAIS)*, volume 4, numéro 2, pages 105-120.

Schroeder, M., Cowie, R., Heylen, D., Pantic, M., Pelachaud, C. et Schuller, B. (2008). Towards responsive sensitive artificial listeners. *In Proc. 4th Intern. Workshop on Human-Computer Conversation*, Bellagio, Italie.

Schuller, B., Steidl, S. et Batliner, A. (2009). The Interspeech 2009 Emotion Challenge. *In Proc. of the 10th Annual Conference of the International Speech Communication Association*, Brighton, Royaume-Uni.

Shibata, T. et Tanie, K. (1999). Creation of Subjective Value through Physical Interaction between Human and Machine. *In Proceeding of the 4th International Symposium on Artificial Life and Robotics*, pages 20–23.

Vilhjalmsson, H., Cantelmo, N., Cassell, J., Chafai, N.E., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A.N., Pelachaud, C., Ruttkay, Z., Thórisson, K.R., van Welbergen, H. et van der Werf, R.J. (2007). The Behavior Markup Language: Recent Developments and Challenges. *In Proc. of the 7th International Conference on Intelligent Virtual Agents*, pages 99–111.

Zeng, Z., Pantic, M., Roisman, G.I. et Huang, T.S. (2009). A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 31n numéro 1, pages 39–58.