Le jeu de données Brain-IHM : une nouvelle ressource pour l'analyse des bases cérébrales des conversations humain-humain et humain-agent-virtuel

Magalie Ochs, Roxane Bertrand, Aurélie Goujon, Deirdre Bolger, Anne-Sophie Dubarry, Jean-Marie Pergandi, Philippe Blache

Le jeu de données Brain-IHM : une nouvelle ressource pour l'analyse des bases cérébrales des conversations humain-humain et humain-agent-virtuel

Magalie Ochs[1], Roxane Bertrand[2], Aurélie Goujon[2], Deirdre Bolger[2], Anne-Sophie Dubarry[2], Jean-Marie Pergandi[3], Philippe Blache[2]

[1]Aix Marseille Univ, Université de Toulon, CNRS, LIS , Marseille, France
[2]Aix Marseille Univ, LPL, Aix en Provence, France
[3]Aix Marseille Univ, CRVM, ISM, Marseille, France
{first-name.last-name}@univ-amu.fr

Résumé

Ce papier présente un ensemble de données original d'interactions contrôlées, en se concentrant sur l'étude des feedbacks. Il s'agit d'enregistrements de différentes conversations entre un médecin et un patient, jouées par des acteurs. Dans ce corpus, le patient est principalement un auditeur et produit différents feedbacks, dont certains sont (volontairement) incongrus. De plus, ces conversations ont été resynthétisées dans un contexte de réalité virtuelle, dans lequel le patient est joué par un agent artificiel. Le corpus final est constitué de différents films de conversations humain-humain, ainsi que les mêmes conversations rejouées dans un contexte humain-machine, ce qui donne le *premier corpus parallèle humain-humain/humain-machine*. Le corpus est ensuite enrichi de différentes annotations multimodales aux niveaux verbal et non verbal. De plus, ce corpus a été enrichi de données de perception subjective et de données neurophysiologique. Nous avons en effet conçu une expérience au cours de laquelle différents participants devaient regarder les films du corpus et donner une évaluation de l'interaction. Au cours de cette tâche, nous avons mesuré l'activité cérébrale du participant par électroencéphalographie. L'ensemble de données Brain-IHM est conçu dans un triple but : 1/ étudier les feedbacks en comparant les feedbacks congruents versus incongruents 2/ comparer la perception des feedbacks produit par l'humain versus la machine 3/ étudier les bases cérébrales de la perception des feedbacks.

Mots-clés : Feedbacks, multimodalité, corpus parallèle humain-humain/humain-machine, bases cérébrales du feedback, étude perceptive, EEG, cerveau.

1. Introduction

L'un des principaux enjeux est de comprendre comment l'interaction entre deux humains peut être réussie en se fondant sur le comportement des différents participants au cours d'une conversation. En particulier, parmi les paramètres garantissant la qualité et le succès d'une interaction, l'attitude du participant qui écoute le locuteur principal (hochements de tête, expressions faciales, etc.) est cruciale : une absence de réaction ou une réaction inappropriée entraîne une perte d'engagement de l'interlocuteur et donc un échec de l'interaction (Gratch et al., 2006 ; Bevacqua et al., 2008). Ces attitudes de l'interlocuteur pendant l'écoute sont souvent véhiculées par des feedbacks. Comme le souligne (Allwood, 1992), "*La raison d'être des mécanismes de feedback linguistique est la nécessité de susciter et de donner des informations sur les fonctions communicatives de base, c'est-à-dire le contact continu, la perception, la compréhension et la réaction émotionnelle/attitudinale, d'une manière suffisamment discrète pour permettre à la communication de servir d'instrument pour la poursuite de diverses activités humaines*". Les feedbacks sont très fréquents, généralement courts et multimodaux (verbaux et/ou non verbaux). Ils assurent la cohésion entre les interlocuteurs : ils sont le signe que la conversation est suivie, comprise, acceptée et sont donc essentiels pour une interaction naturelle (Schegloff 1982 ; Alwood et al. 1992).

Cette question est cruciale dans le domaine de l'interaction humain-machine fondée sur les agents conversationnels animés (ACAs) et constitue une condition de notre capacité à développer des environnements dans lesquels le participant humain perçoit le comportement de l'agent artificiel comme crédible et engageant. Cette question est également d'une grande importance lorsque l'on tente de déterminer les différences de perception des autres comportements des participants comparant les situations humain-humain et humain-machine. Au niveau du cerveau, plusieurs travaux ont été réalisés pour analyser l'activité cérébrale des utilisateurs qui interagissent ou observent des agents artificiels (ECA ou robots). Par exemple, dans (Urgen et al., 2018), l'activité cérébrale des utilisateurs mesurée par électroencéphalographie (EEG) est analysée et comparée en fonction du niveau de réalisme du robot. Une corrélation est révélée entre l'activité cérébrale et l'incongruence entre l'apparence et les mouvements. Dans (Rauchbauer et al., 2019), l'activité cérébrale des utilisateurs a été enregistrée en IRMf et comparée selon que le participant interagit avec un robot ou un humain.

Une des problématiques aujourd'hui est qu'il n'existe que peu de ressources pour comparer l'interaction humain-humain et humain-machine en général, et les activités comportementales et neurophysiologiques des utilisateurs en particulier. Une des raisons expliquant ce manque de ressources est la difficulté de créer des corpus comparables (ou mieux encore parallèles) avec l'interaction humain-humain et humain-machine, avec exactement le même contexte et la même dynamique dans l'interaction. En particulier, le comportement de l'agent artificiel doit être comparable à celui d'un humain, en tenant compte des capacités expressives de l'agent virtuel (par exemple, fluidité du geste, aspect caricatural).

Dans cet article, nous décrivons tout d'abord une méthode pour créer un corpus parallèle de conversation humain-humain et humain-machine. Dans une deuxième partie, nous décrivons comment cette méthode a été utilisée pour créer le jeu de données Brain-IHM. Cette ressource est constituée de différentes conversations entre un médecin et un patient. Dans ces conversations, le médecin annonce une mauvaise nouvelle au patient qui l'écoute et produit régulièrement des feedbacks. Le patient est joué par un acteur humain et ensuite rejoué par un agent artificiel. Ce corpus est également accompagné d'une information particulièrement originale : l'activité cérébrale d'un participant qui regarde la conversation.

Le corpus Brain-IHM constitue une ressource unique pour l'étude de deux types de questions :

- La perception du feedback, y compris au niveau du cerveau ;
- La comparaison des interactions humain-humain et humain-machine.

Le jeu de données Brain-IHM présenté dans cet article vise plus spécifiquement à étudier les phénomènes de feedback en général, l'activité cérébrale associée à leur perception, et à comparer les feedbacks produits par un humain ou un agent virtuel. Nous proposons dans la dernière partie de cet article une étude réalisée à partir du jeu de données Brain-IHM, en essayant de comprendre si les feedbacks sont traités automatiquement. Aujourd'hui, la plupart des méthodes existantes pour évaluer un agent artificiel sont fondées sur des évaluations subjectives à travers des questionnaires remplis par les utilisateurs après leur interaction avec l'agent virtuel. Ces travaux permettent de développer une mesure objective de la crédibilité de l'agent virtuel, fondée sur l'électroencéphalographie (EEG). L'activité EEG de l'utilisateur liée à la perception du feedback de l'agent virtuel pourrait constituer un indice objectif de la crédibilité perçue du comportement de l'agent.

2. Les feedbacks

2.1 Les types de feedback

Dans ce projet, nous nous concentrons sur deux types de feedbacks prototypiques (Schegloff, 1982) : les feedbacks incitant la poursuite de la conversation appelés les « continuers » (hochement de tête, oui, mhmh, ok, etc.) et les « assessments » appelés *feedbacks d'évaluation* (accord, surprise, émotion, etc.). Les *continuers* apparaissent généralement dans les 200 ms après l'intervention du locuteur et sont réalisés plus automatiquement sans traduire une évaluation sémantique du discours, contrairement aux *feedbacks d'évaluation* qui apparaissent généralement plus tard (au-delà de 200 ms après l'intervention du locuteur). Nous fournissons pour chacun d'eux une identification précise de leurs paramètres multimodaux (auditifs et visuels).

Au cours d'une conversation naturelle, on s'attend à ce que les participants adoptent des réponses typiques. En tant qu'auditeurs, des feedbacks appropriés sont nécessaires pour une réalisation interactive réussie. La pertinence des réponses de feedback dépend de différents critères tels que leur localisation ou leur valeur sémantique. Par exemple, Bavelas et collaborateurs (2000) ont montré que les *continuers*, qui aident à montrer la construction des connaissances partagées, sont des réponses appropriées lorsqu'ils se produisent au début du récit, tandis que les *feedbacks d'évaluation*, qui ont une fonction évaluative concernant les événements décrits, sont plutôt fournis à la fin. La pertinence d'un feedback peut également dépendre de sa valeur sémantique qui peut être identifiée avec des attributs scalaires liés à la certitude/incertitude, à la compréhension/non-compréhension par exemple (Prévot et al., 2016). Dans le présent document, nous examinons la pertinence des feedbacks produits en fonction de ce critère de valeur sémantique.

Parmi les différents feedbacks produits par les participants dans le corpus collecté, 3 types ont été manipulés. Soulignons que le corpus a été collecté en français, mais la méthodologie ainsi que le type de modélisation appliqué ici peuvent être appliqués à d'autres langues.

Les feedbacks cibles avec les fonctions et les axes sémantiques correspondants :

- *tout à fait ("sure")* tout-à-fait ("sure") correspondant à un feedback de confirmation sur l'axe sémantique attendu / inattendu
- *ah bon ("vraiment", "oh")* reflétant la surprise sur l'axe sémantique incertitude/ certitude
- *oh non ("oh no")* oh non ("oh no") reflétant la déception sur l'axe sémantique attendu/ inattendu

Les feedbacks congrus concernent les éléments exprimant la valeur sémantique appropriée projetée par l'énoncé précédent, tandis que les feedbacks incongrus concernent ceux qui expriment une valeur inappropriée.

Exemples :

Docteur : Je suis votre médecin anesthésiste c'est moi qui vous ai endormi, vous vous en souvenez ?

Patient :

- Feedback congruent : *Tout à fait*
- Feedback incongruent : *oh non*

Docteur : C'est un médicament qui permet de relâcher les muscles

Patient :

- Feedback congruent: *Ah bon*
- Feedback incongruent : *Tout à fait*

Docteur: Ça a dû être un moment très pénible pour vous.

Patient :

- Feedback congruent : *tout à fait*
- Feedback incongruent : *ah bon*

Doctour : Vous éprouvez toujours des difficultés à respirer ?

Patient :

- Feedback congruent : *tout à fait*
- Feedback incongruent : *ah bon*

Docteur : Avez-vous des questions ?

Patient :

- Feedback incongruent : *tout à fait*
- Feedback incongruent : *ah bon*

Les trois types de feedbacks étudiés sont **multimodaux**, c'est-à-dire qu'ils sont exprimés via des signaux verbaux et non verbaux. Par exemple, le feedback *"tout-à-fait"* est associé à un hochement de tête pour renforcer la valeur sémantique de confirmation dans la conversation en face à face. Le feedback *"ah bon"* est produit avec un haussement de sourcils pour souligner la fonction sémantique de la surprise. Le feedback *"oh non"* est produit avec un froncement de sourcils afin de renforcer la déception. Notez que l'incongruité considérée dans cette étude est l'incongruité sémantique, c'est-à-dire l'utilisation d'un feedback avec une

fonction spécifique (par exemple la surprise) dans une situation où un feedback avec une autre fonction est attendu (par exemple la confirmation ou la non-confirmation). Nous ne considérons pas l'incongruité en termes de contradiction entre le signal verbal et non-verbal dans l'expression d'un feedback (par exemple un feedback avec un item verbal *"tout-à-fait"* et avec une expression faciale de surprise).

2.2 Les feedbacks de l'agent virtuel

Les feedbacks décrits ci-dessus ont été reproduits sur l'Agent conversationnel incarné Greta de la plateforme VIB (Pelachaud, 2009). Dans cette plateforme, les signaux de communication, tels que les feedbacks, sont décrits à travers des fichiers FML (Function Markup Language) et BML (Behaviour Markup Language) de SAIBA (Vilhjálmsson et al., 2007). Afin de simuler les feedbacks identifiés, nous avons enrichi la bibliothèque de l'ACA avec un ensemble de fichiers décrivant exactement le comportement multimodal correspondant aux feedbacks.

Figure 1 : Deux feedbacks différents produits par l'ACA

Pour reproduire les feedbacks sur l'ACA, des linguistes, experts de ce phénomène, ont observé les feedbacks exprimés par les humains lors des enregistrements de films (section 3) et ont manipulé la prosodie, les gestes, les mouvements de tête et les expressions faciales de l'ACA pour obtenir des expressions similaires. La figure 1 illustre deux exemples de feedbacks.

Finalement, nous avons créé 6 nouveaux feedbacks de l'ACA correspondant aux besoins. En plus des 3 types de feedbacks étudiés, nous avons créé 3 feedbacks basiques - "D'accord", "Ok", "Oui " - pour assurer une variabilité dans le comportement du patient et un naturel dans la conversation. Il est à noter que l'activité cérébrale de l'observateur n'est pas analysée pour ces trois feedbacks.

3. Créer un corpus *parallèle* (humain-humain et humain-machine)

L'ensemble des données Brain-IHM a été conçu pour explorer la question de la production des feedbacks dans un contexte contrôlé en étudiant leur perception par un participant tiers (ci-après l'observateur). Notre première hypothèse est que la perception d'un feedback par l'observateur est comparable à celle qu'en ont les participants à la conversation. En d'autres termes, lorsqu'une personne entend (et voit) une conversation sans y prêter attention, elle va traiter les différents éléments du dialogue de la même manière que les participants eux-mêmes. Si cette hypothèse est vraie, elle élargit les possibilités d'investigation. En particulier, il devient possible d'étudier les réactions du cerveau dans la perception du feedback.

L'étude du signal cérébral d'un participant à la conversation dans un environnement naturel est une tâche très complexe, le signal étant extrêmement bruité en raison de l'activité musculaire, des mouvements, etc. L'idée de ce jeu de données Brain-IHM a donc été d'enregistrer un dialogue entre deux participants reproduisant une conversation spécifique, dans laquelle l'un d'eux est l'interlocuteur principal tandis que l'autre produit régulièrement des feedbacks.

3.1 Scénarios de dialogue et enregistrement des films

Le thème choisi dans cette étude est un contexte médical dans lequel un médecin doit annoncer une mauvaise nouvelle à un patient. Un tel contexte est particulièrement adapté à l'étude des feedbacks puisque le patient est principalement en position d'écoute. Nous avons d'abord élaboré 3 scénarios, en collaboration avec les médecins partenaires du projet : perforation intestinale suite à une endoscopie, régurgitation de liquide gastrique lors d'une opération, arrêt respiratoire lors d'une opération. Chaque scénario a donné lieu à la création d'un dialogue prototypique dans lequel les prises de tour du médecin sont entièrement spécifiées et suivies de feedbacks produits par le patient. Les dialogues ont été validés par des médecins réels, habitués à être confrontés à de telles situations.

Figure 2 : *Mise en place de l'enregistrement du dialogue médecin-acteur et patient-acteur. Un fond vert a été utilisé pour nous permettre d'intégrer la scène dans un environnement virtuel.*

L'étape suivante a consisté à former deux acteurs, jouant le rôle de chaque participant du dialogue (médecin et patient). Le médecin devait suivre strictement le contenu général de chaque tour, mais avec la possibilité d'adapter la façon dont le contenu a été exprimé. Le patient devait produire exactement le feedback prévu dans le scénario.

Dans chaque dialogue, le patient a produit plusieurs feedbacks (au moins un à chaque fin de tour), parmi lesquels les 18 feedbacks cibles qui seront étudiés. Afin de comparer la perception des différents types de feedbacks, nous avons demandé aux acteurs de jouer deux fois le même scénario (donc le même dialogue) : l'un dans lequel tous les feedbacks sont congrus, le second avec 50% de feedbacks incongrus (avec une incongruité sémantique comme décrit dans la section 2.1).

Exemple de feedbacks incongrus :

Docteur : *Avez-vous des questions ?*
Patient: *Ah bon ?*

Notez que dans la condition d'incongruence, certains feedbacks sont produits de façon congruente afin d'éviter un dialogue trop peu naturel (comme expliqué à la fin de la section 2.2). De plus, dans la perspective de faire varier la production du patient, nous avons également intégré un tour lexicalisé produit par le même patient, en réponse à une question du médecin. Deux médecins-acteurs et deux patient-acteurs ont joué ces scénarios, pendant environ 3 minutes chacun.

Chaque scénario a donné lieu à plusieurs films, faisant varier le couple médecin/patient. La figure 2 illustre les conditions d'enregistrement.

Issus de cette première collection, les meilleurs films ont été sélectionnés par des experts externes en interaction humain-

humain qui devaient évaluer les interactions les plus réalistes. Finalement, deux films ont été choisis pour chaque scénario (6 films au total) : un film dans lequel le patient produit des feedbacks congrus, un second dans lequel il produit des feedbacks congrus et incongrus.

3.2 Vidéo humain-humain et humain-agent virtuel

Chaque film a été monté, ce qui permet au patient d'être vu comme l'illustre la figure 3.b. Le but est de permettre à l'observateur de percevoir le plus clairement possible les feedbacks produits par les acteurs (virtuels ou réels).

Afin de créer le corpus parallèle avec l'acteur virtuel (i.e. répliqué le comportement de l'acteur-patient sur l'agent virtuel), nous avons annoté la vidéo d'interaction humain-humain décrite dans la section précédente en utilisant ELAN pour identifier précisément le moment (début, fin, durée) et le type de réactions exprimées par l'acteur-patient (Figure 3.a).

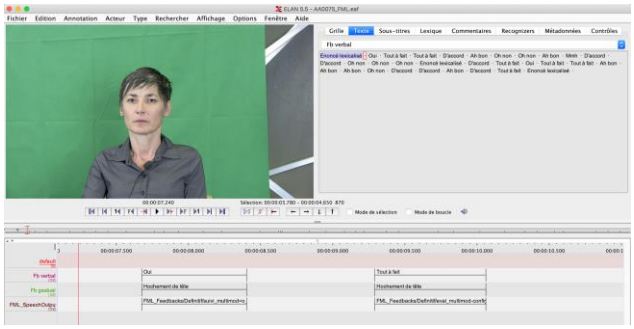


Figure 3.a : Image d'écran de l'annotation des feedbacks du de l'acteur-patient dans ELAN.

L'annotation consiste à spécifier le fichier FML dans la bibliothèque des feedbacks multimodaux produits par l'agent virtuel (voir section 2.2) correspondant aux feedbacks exprimés par le patient dans le film. De plus, l'annotation précise le moment et la durée des feedbacks (Figure 3.a). Un outil a été développé pour générer automatiquement à partir du fichier d'annotation l'animation de l'agent virtuel exprimant les réactions au moment exact. L'outil prend en entrée les fichiers annotés (Figure 3.a) et génère le film correspondant à l'animation du patient virtuel avec les feedbacks appropriés décrits dans le FML (Section 2.2) au moment indiqué et avec la durée correspondante spécifiée dans le fichier ELAN. De cette façon, le comportement de feedbacks de l'agent virtuel est le même que celui de l'acteur-patient.

De plus, les conversations ont été intégrées dans un environnement virtuel conçu dans Unity et validé par des médecins pour représenter une salle de réveil où l'annonce des mauvaises nouvelles se fait généralement dans les hôpitaux. Au total, nous obtenons un corpus de 12 films, 4 par scénario. Pour chaque scénario, nous avons :

- Deux films *humain-humain*, l'un dans lequel l'acteur-patient produit des feedbacks congruents, l'autre avec des feedbacks congruents et incongrus ;
- Deux films humain-agent virtuel, l'un dans lequel le patient artificiel produit des feedbacks congruents (les mêmes feedbacks produits par l'acteur-patient et en même temps) et l'autre avec des feedbacks congruents et incongrus (similaire à l'acteur-patient).

Il est intéressant de noter la triple originalité de l'ensemble de données Brain-IHM : 1/ il s'agit du premier corpus avec une production contrôlée de feedbacks 2/ il contient des conversations parallèles entre humains et entre humains et un agent virtuel 3/ il

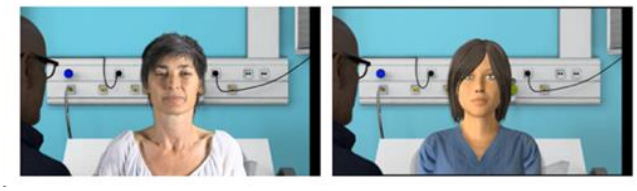


Figure 3.b : Captures d'écran de la vidéo du jeu de données Brain-IHM (à gauche, simulation d'une interaction acteur-docteur / acteur-patient, à droite, re-synthèse du même comportement multimodal de l'acteur-patient sur l'agent virtuel)

contient des informations à la fois verbales, non verbales et électrophysiologiques (Section 4).

3.3 Annotations verbales et non-verbales

Le fait que la production du corpus soit presque entièrement contrôlée facilite les annotations. Au niveau verbal, en particulier, la transcription du dialogue est alignée sur le signal, à l'aide de SPASS (Bigi, 2015). Cela se fait au niveau des phonèmes, offrant une segmentation précise en phonèmes, syllabes et tokens. L'étiquettage morpho-syntaxique est également fourni, ainsi qu'une segmentation syntaxique de niveau intermédiaire dans les tours. La qualité de l'enregistrement du signal offre également la possibilité d'appliquer des analyseurs prosodiques automatiques.

Le niveau non verbal ne concerne que les expressions faciales : il a été demandé aux deux acteurs d'être statiques, l'auditeur n'a produit que des expressions faciales différentes (hochements de tête, sourcils levés, sourires, etc.). Ces informations sont directement annotées sans utiliser d'outil, grâce aux spécifications des scénarios.

Outre ces annotations, la principale originalité du corpus est qu'il est accompagné de deux autres types d'informations : l'une concernant l'évaluation subjective du patient virtuel, la seconde avec l'enregistrement du signal cérébral du participant qui regarde les films. Ces deux informations sont décrites dans la section suivante.

4. Mise en place et conception de l'expérimentation

L'activité cérébrale a été mesurée dans le cadre du paradigme suivant : chaque participant doit regarder des films de dialogues dans 4 conditions différentes : humain/humain, humain/agent virtuel, avec seulement des feedbacks congruents ou congrus et incongrus. Ces conditions permettent de faire différentes comparaisons : communication humain/machine et productions congruentes/incongrues. Plus précisément, l'idée est tout d'abord de confirmer l'hypothèse selon laquelle aucune différence ne peut être observée dans la perception des feedbacks de l'agent humain et de l'agent virtuel. Ceci fait, il devient alors possible d'explorer les effets spécifiques au niveau du cerveau.

Trente-six participants (27 femmes et 9 hommes) ont été recrutés pour l'expérience ; tous les participants ont signé un formulaire de consentement au début de l'expérience et ont été rémunérés pour leur participation. Les sujets recrutés avaient entre 18 et 40 ans.

Les données d'électroencéphalographie (EEG) ont été enregistrées à partir d'un système à 64 électrodes Biosemi Active2, qui amplifie le signal au niveau de la tête du participant via un AD-box, un amplificateur DC. Dans ce système EEG, la masse est remplacée par deux électrodes, la CMS (Common Mode Sense) et la DRL (Driven Right Leg), qui assurent que le potentiel moyen du participant est aussi proche que possible de la tension ADC du Biosemi AD-box. Les signaux ont été enregistrés à une fréquence de 2048 Hz. Des électrodes externes ont été

placées sur le visage pour enregistrer les mouvements oculaires horizontaux et verticaux afin de faciliter leur traitement hors ligne

Pendant l'enregistrement EEG, les participants ont regardé les 4 films différents tout en étant confortablement assis devant un écran d'ordinateur dans une cage de Faraday. Avant de commencer la présentation des films, une très courte vidéo avec du son a été présentée pour régler le niveau sonore des vidéos. Un mode de présentation sonore en champ libre a été utilisé via des haut-parleurs placés à gauche et à droite de l'écran. Les participants ont été invités à regarder passivement les 4 vidéos ; la tâche n'a impliqué aucune réaction de leur part. Chaque film a duré environ 5 minutes et, entre chaque film, il y a eu une courte pause pendant laquelle le participant a répondu à un questionnaire en ligne basé sur le film vu.

Ce questionnaire était composé de 7 questions ou affirmations sur la perception du patient par les participants (virtuelle ou réelle selon la vidéo qu'il venait de visionner) : la *crédibilité* du patient ("Selon vous, dans quelle mesure le patient est-il crédible par rapport aux patients réels ?"), *l'appréciation du patient* par le participant ("Vous aimez le patient"), la *réactivité* du patient ("Avez-vous trouvé le patient réceptif à ce que le médecin a dit ?"), le *naturel* de la conversation ("Avez-vous trouvé la conversation naturelle entre le patient et le médecin ?"), la *compréhension perçue* du patient ("Avez-vous eu l'impression que le patient a compris ce que le médecin a dit ?"), la *performance perçue du médecin* dans la tâche d'annoncer la mauvaise nouvelle ("Pensez-vous que le médecin a bien expliqué le problème et ce qui allait arriver au patient ?" et "Pensez-vous que le médecin a eu des difficultés à annoncer la mauvaise nouvelle au patient ?) Les participants répondaient aux questions ou indiquaient leur niveau d'accord avec l'affirmation au moyen d'une échelle de Likert à 5 points.

L'analyse des réponses aux questions sur la *crédibilité perçue* de l'agent et sur le *caractère naturel de la conversation* nous permet de valider les deux conditions différentes, soit avec seulement des feedbacks congruents (condition congruente), soit avec des feedbacks congruents et incongruents (condition incongrue). Dans les deux conditions, pour la conversation humain-humain et humain-agent virtuel, les participants ont significativement perçu la conversation avec des feedbacks incongrus comme moins naturelle avec un agent moins crédible que la conversation avec seulement des feedbacks congruents. La figure 4 illustre les réponses moyennes des participants à la question sur la crédibilité du patient ("Selon vous, dans quelle mesure le patient est-il crédible par rapport aux patients réels ?") sur une échelle de Likert de 5 points. Les résultats du T-test montrent que l'acteur-patient n'exprimant que des feedbacks congruents (HHC) a été perçu comme significativement plus crédible que lorsqu'il exprimait des feedbacks incongrus (HHI) ($p<0,05$). De la même manière, l'agent virtuel n'exprimant que des feedbacks congruents (condition HAC) a été perçu comme significativement plus crédible que lorsqu'il exprime des feedbacks incongrus (condition HAI) ($p<0,001$).

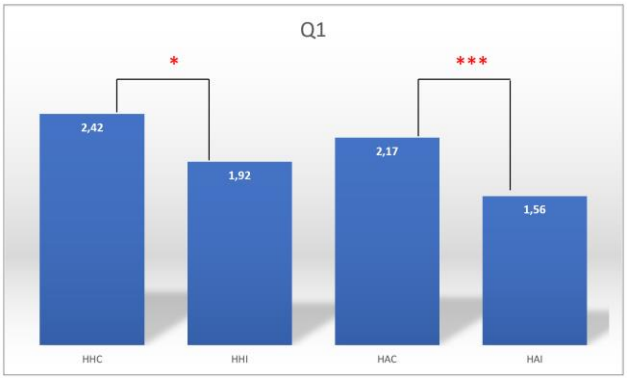


Figure 4 : Moyenne des réponses (entre 0 et 5) à la question "Selon vous, dans quelle mesure le patient est-il crédible par rapport aux vrais patients" pour chaque condition.

La figure 5 illustre la réponse moyenne des participants à la question sur le *caractère naturel* de la conversation ("Avez-vous trouvé la conversation naturelle entre le patient et le médecin ?") sur une échelle de Likert de 5 points. Les résultats du T-test montrent une *différence significative sur la perception du naturel de la conversation* en fonction des feedbacks exprimés : les conversations ont été perçues de manière significativement plus naturelle lorsque les acteurs (virtuels ou réels) ont exprimé des feedbacks congruents que lorsqu'ils en ont exprimé des incongruents.

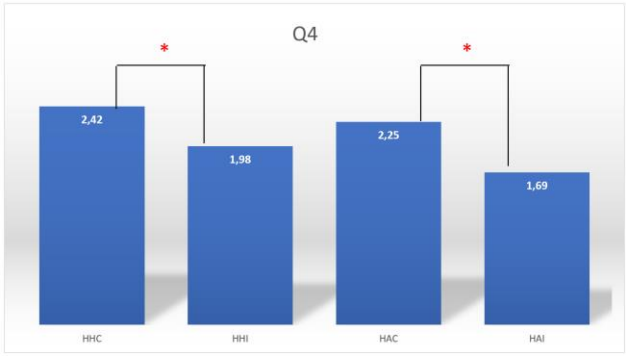


Figure 5 : Moyenne des réponses (entre 0 et 5) à la question "Selon vous, dans quelle mesure le patient est-il crédible par rapport aux vrais patients" pour chaque condition.

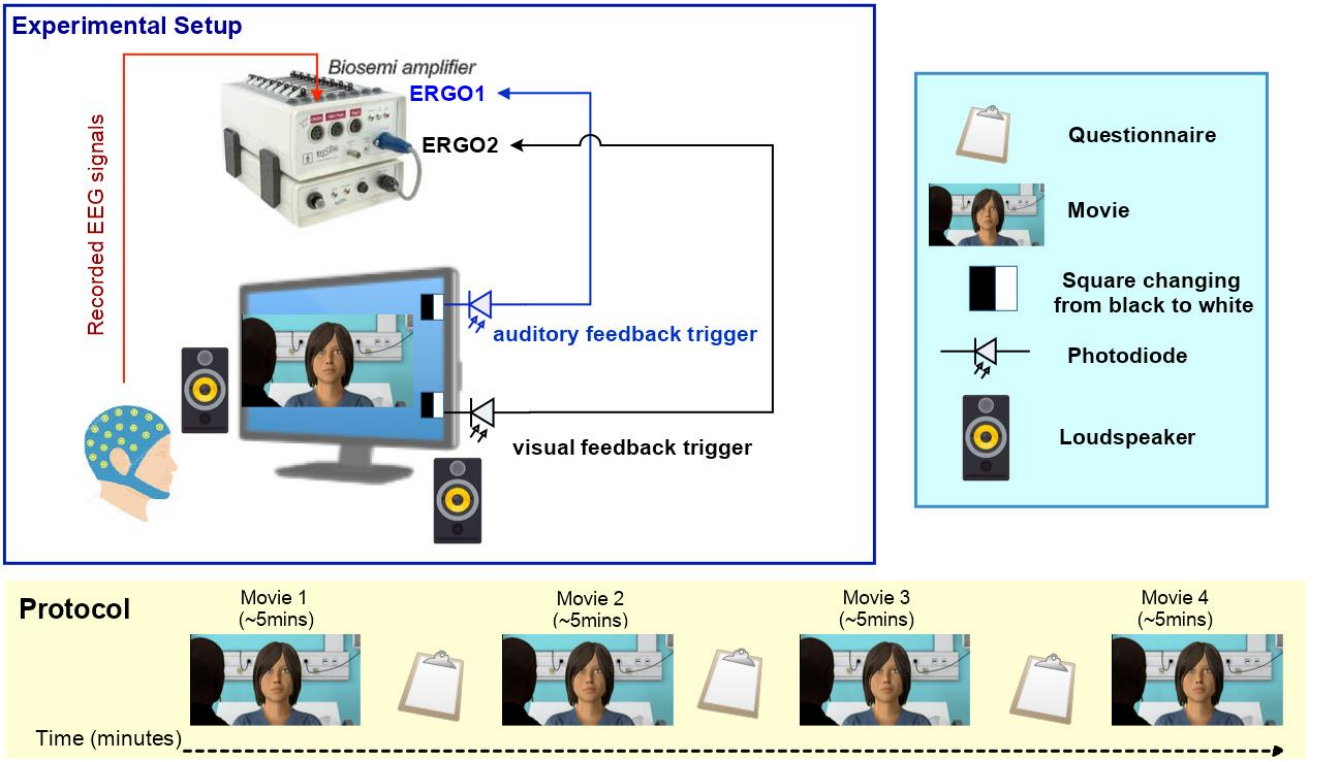
Notez que les résultats du questionnaire sont loin d'être surprenants mais nous permettent de valider que les feedbacks incongrus ont été perçus comme tels.

La présentation des 4 types de vidéos (humain-humain congruent, humain-humain incongruent, humain-agent virtuel congruent et humain-agent virtuel incongruent) a été contrebalancée entre les participants afin d'éviter l'effet d'ordre sur les résultats. Pour synchroniser le début de chaque feedback, auditif et visuel, et le signal EEG, nous avons incrusté deux petits carrés noirs dans chaque vidéo, un pour les feedbacks auditifs et un pour les feedbacks visuels, qui passaient du noir au blanc au début de chaque feedback. Le changement de luminosité de chaque carré coloré a été détecté par des photodiodes dont les signaux étaient enregistrés comme deux canaux EEG auxiliaires via les entrées ERGO1 et ERGO2 du Biosemi AD-box (Figure 6). Ces carrés étaient invisibles pour le participant (cachés par une bande plastique placée sur le côté de l'écran). En outre, pour que les différents types de feedbacks puissent être distingués hors ligne, la durée du changement de luminosité a varié pour les feedbacks visuels et auditifs et pour chacun des 4 types de vidéo, ce qui a permis d'obtenir un signal de photodiode différent pour chaque type de vidéo (humain-humain, humain-agent virtuel), chaque

modalité de feedback (auditif et visuel) et chaque type de feedback (congruent et incongruent). Ainsi, pour déterminer la durée des signaux de photodiode, ces signaux ont été convertis en fonction d'échelon. Donc, chaque feedback correspond à une fonction d'échelon dont t0 est le début du feedback (l'onset) et le type de feedback peut être déterminé en calculant la durée de l'échelon.

5. Perspectives : les bases cérébrales de la perception de feedbacks

Les feedbacks peuvent être considérés au niveau inférieur comme des marqueurs de discours (plus instructifs que référentiels) ou au contraire comme des objets linguistiques à part entière. Dans le premier cas, les feedbacks peuvent être considérés comme des réactions automatiques au discours de l'orateur principal. Certaines études ont montré que de tels feedbacks ne peuvent être prédits qu'à partir du cours du temps (Penteado et al., 2019) : l'apparition d'un feedback semble dépendre principalement de la réalisation du précédent (voir aussi (Ward et Tsukahara, 2000)). Cependant, d'autres travaux ont également montré que les feedbacks dépendent de différentes caractéristiques linguistiques telles que la prosodie (en particulier les pauses) mais aussi de l'information morpho-syntaxique (certains adverbes dans le discours du locuteur principal peuvent augmenter la probabilité d'une réalisation du feedback par l'auditeur (Bertrand et al., 2003 ; Bertrand et al., 2007). Un tel comportement contextuel indique qu'un certain type de traitement linguistique doit être effectué par l'auditeur. Enfin, certains types de feedback (en particulier le feedback d'évaluation) nécessitent une certaine compréhension de la production du locuteur principal, ou en d'autres termes une forme de traitement sémantique, qui est le niveau linguistique le plus élevé. Les feedbacks peuvent alors être considérés de manière très différente, en fonction de leur relation avec le contexte précédent. Dans certains cas, elles semblent se situer au niveau le plus bas, produites presque automatiquement alors que dans d'autres cas, elles semblent nécessiter un traitement linguistique plus profond. En résumé, la principale question de cette étude est de savoir si les feedbacks sont traités automatiquement ou non.



Notre première hypothèse dans cette étude est que la perception des feedbacks repose à la fois sur des mécanismes automatiques et profonds. Nous proposons pour cela d'analyser le signal EEG en réponse à la production des feedbacks. Notre hypothèse est que les feedbacks verbaux et non verbaux suscitent des potentiels evoqués (PE) spécifiques en fonction de la manière dont elles sont traitées. Des travaux antérieurs ont montré qu'une augmentation de la négativité postérieure précoce (Early Posterior Negative-EPN) peut être associée à la perception des expressions faciales, modulant l'amplitude des composantes N170 et P100 (Eimer, 2011 ; Herrmann et al., 2004 ; Sprengelmeyer et al., 2006 ; Krombholz et al., 2007 ; Righart et al., 2006). De la même manière, certains travaux ont mis en évidence des effets similaires dans la perception des mots émotionnels (Wang et al., 2014) et plus précisément la sensibilité de la composante N170 aux émotions et à leur intensité (Sprengelmeyer et al., 2006 ; Krombholz et al., 2007 ; Righart et al., 2006). Ces effets peuvent être associés à un certain type de traitement automatique. L'hypothèse de cette expérience est que la perception du feedback est d'une certaine manière constituée de mécanismes incluant la reconnaissance faciale et émotionnelle. Si c'est le cas, la perception du feedback devrait susciter des composantes précoces comparables telles que N170 et P100. Le deuxième aspect de cette expérience consiste à rechercher des traces de mécanismes profonds, impliquant notamment un certain niveau de traitement sémantique. L'idée est que les feedbacks incongrues, tels que décrits dans la section précédente, provoquent un effet similaire à celui de l'incongruité sémantique. Un tel phénomène est associé à une forte onde négative déclenchée 400 ms après le début du stimulus, connue sous le nom d'effet N400 (Besson et al., 1992 ; Kutas et al., 1980 ; 2011), qui peut également être observée dans le cas de l'incongruité émotionnelle (Leuthold et al., 2011 ; Bartholow et al., 2001).

Notre deuxième hypothèse est qu'il n'y a pas de différence dans la perception des feedbacks produits par un humain ou par un agent virtuel, la signature du signal de l'activité cérébrale devrait être la même. Cette hypothèse est importante pour différentes raisons. Tout d'abord, il est nécessaire d'étudier plus précisément le type de mécanismes qui se produisent au niveau du cerveau lors de la

communication avec un humain ou avec un agent virtuel. Par exemple, nous savons que les expressions faciales sont traitées de la même manière, qu'elles soient produites par un agent virtuel ou un humain (Dyck et al., 2008). L'ensemble de données Brain-IHM ouvrira de nouvelles possibilités pour comparer la communication entre l'humain et la machine. En outre, au niveau méthodologique, ce type de ressources et la manière dont elles sont construites offrent également de nouvelles possibilités d'explorer le rôle de certaines caractéristiques spécifiques dans la production des feedbacks (par exemple, le retard, la prosodie, la synchronisation verbale/non verbale, etc.).

6. Conclusion

Ce papier présente un nouvel ensemble de données originales d'interactions contrôlées, en se concentrant sur l'étude des feedbacks et plus particulièrement sur deux types de feedbacks multimodaux : les feedbacks incitant la poursuite de la conversation appelés les « continuers » (hochement de tête, oui, mhmh, ok, etc.) et les « assessments » appelés feedbacks d'évaluation (accord, surprise, émotion, etc.). Un corpus parallèle de vidéos de conversations entre humains et agents virtuels a été créé. Une méthodologie spécifique a été développée pour créer un tel corpus avec un comportement de l'agent virtuel reproduisant le comportement de l'acteur mais considérant la capacité expressive de l'agent virtuel. Dans cet article, nous nous sommes concentrés particulièrement sur les feedbacks lors d'une conversation en français entre un médecin et un patient dans le contexte de l'annonce d'une mauvaise nouvelle, mais la méthodologie pourrait être reproduite pour analyser d'autres comportements multimodaux dans d'autres langues et d'autres contextes applicatifs. Ce type de corpus permet de comparer la perception de l'interaction humain-humain par rapport à l'interaction humain-machine. De plus, nous avons proposé un dispositif expérimental spécifique et un plan d'expérience pour enregistrer les activités cérébrales des observateurs de l'interaction.

En conclusion, le jeu de données Cerveau-IHM constitue la première ressource de ce type, fournissant une production contrôlée de feedbacks, un corpus parallèle humain-humain/humain-machine, complètement annoté au niveau verbal et non-verbal et un ensemble de données EEG permettant d'étudier le niveau perceptif

7. Remerciements

Ce travail a été financé par le CNRS (PEPS Brain-IHM) et soutenu par les subventions ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI) et ANR-11-IDEX-0001-02 (A*MIDEX).

Notez que ce papier est largement inspiré d'une traduction d'un papier que nous avons soumis à LREC 2020.

8. Réferences

Allwood, J., Nivre, J., & Ahlsén, E. (1992). On the semantics and pragmatics of linguistic feedback. Journal of semantics, 9(1), 1-26.

Bailenson J.N., Swinth, K., Hoyt, C., Persky, S., Dimov, A., Blascovich, J. (2005) The independent and interactive effects of embodied-agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in immersive virtual environments. Presence: Teleoperators and Virtual Environments 14(4)

Bartholow B. D., Fabiani M., Gratton G., & Bettencourt B. A. (2001). A psychophysiological examination of cognitive processing of and affective responses to social expectancy violations. Psychological science, 12(3), 197-204.

Bertrand, Roxane; Espesser, Robert (2003). Prosodic cues of back-channel signals in French conversational speech. International Conference on Prosody and Pragmatics (NWCL)

Bertrand, Roxane; Ferré, Gaëlle; Blache, Philippe; Espesser, Robert; Rauzy, Stéphane (2007). Backchannels revisited from a multimodal perspective. Proceedings of Auditory-visual Speech Processing Besson M., Kutas M., Petten CV. (1992) An Event-Related Potential (ERP) Analysis of Semantic Congruity and Repetition Effects in Sentences, in Journal of Cognitive Neuroscience, 4(2):132-49.

Bevacqua E., M. Mancini, R. Niewiadomski, C. Pelachaud (2007), An expressive ECA showing complex emotions. In proceedings of AISB'07 ``Language, Speech and Gesture for Expressive Characters", 208-216

Bevacqua E., Mancini M., & Pelachaud, C. (2008) A listening agent exhibiting variable behaviour. In International Workshop on Intelligent Virtual Agents (pp. 262-269). SpringerBrigitte Bigi (2015). SPPAS - Multi-lingual Approaches to the Automatic Annotation of Speech. In "the Phonetician" - International Society of Phonetic Sciences,

Chollet M., Ochs M. and Pelachaud, C. (2014) Mining a multimodal corpus for non-verbal signals sequences

Gardner, R. (2001). When listeners talk. Benjamins, Amsterdam.

conveying attitudes, Language Resources and Evaluation Conference (LREC)

Dyck M., Winbeck M., Leiberg S., Chen Y., Gur R. C., & Mathiak K. (2008). Recognition profile of emotions in natural and virtual faces. PloS one, 3(11).

Eimer M. (2011). The Face-Sensitive N170 Component of the Event-Related Brain Potential. In G.Rhodes, A. Calder, M. Johnson & J. V. Haxby (Eds.), Oxford Handbook of Face Perception, Oxford University Press.

Fox Tree J. E. (1999). Listening in on monologues and dialogues. Discourse Processes, 27, 35-53.

Gratch J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R. J., & Morency, L. P. (2006). Virtual rapport. In International Workshop on Intelligent Virtual Agents (pp. 14-27). Springer.

Herrmann M. J., Ehlis A. C., Ellgring H., & Fallgatter A. J. (2004). Early stages (P100) of face perception in humans as measured with event-related potentials (ERPs). Journal of Neural Transmission, 112(8)

Kutas , M. , & Hillyard , S. A . (1980) Reading senseless sentences: brain potentials reflect semantic incongruity. Science, 207, 203 – 204 .

Kutas , M. , & Federmeier , K. D . (2011) Thirty years and counting: fi nding meaning in the N400 component of the event-related brain potential (ERP), Annual Review of Psychology , 62 , 621 – 647 .

Krombholz A., Schaefer F., & Boucsein W. (2007). Modification of N170 by different emotional expression of schematic faces. Biological Psychology, 76(3), 156-162.

Leuthold H., Filik R., Murphy K., & Mackenzie I. G. (2011). The on-line processing of socio-emotional information in prototypical scenarios: inferences from brain potentials. Social Cognitive and Affective Neuroscience, 7(4), 457-466

Morency, L. P., de Kok, I., & Gratch, J. (2010). A probabilistic multimodal approach for predicting listener backchannels. Autonomous Agents and Multi-Agent Systems, 20(1), 70-84.

Ochs M., de Montcheuil G., Pergandi J-M., Saubesty J., Pelachaud C., Mestre, D. and Blache P. (2017) An architecture of virtual patient simulation platform to train doctors to break bad news, Conference on Computer Animation and Social Agents (CASA)

Pelachaud, C. (2009). Studies on gesture expressivity for a virtual agent. Speech Communication, 51(7), 630-639.

Penteado, B. E., Ochs, M., Bertrand, R., & Blache, P. (2019, July). Evaluating Temporal Predictive Features for Virtual Patients Feedbacks. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. ACM.

Poppe, R., Truong, K. P., Reidsma, D., & Heylen, D. (2010). Backchannel strategies for artificial listeners. In International Conference on Intelligent Virtual Agents (pp. 146-158 Springer

Porhet C., Ochs M., Saubesty J., de Montcheuil G., Bertrand R. (2017) Mining a Multimodal Corpus of Doctor's Training for Virtual Patient's Feedbacks, International Conference on Multimodal Interaction (ICMI).

Prévot L., Gorish J., Bertrand R. (2016) A CUP of CoFee: A large collection of feedback utterances provided with communicative function annotations. Language Resources and Evaluation Conference (LREC)

Rauchbauer, B., Nazarian, B., Bourhis, M., Ochs, M., Prévot, L., & Chaminade, T. (2019). Brain activity during reciprocal social interaction investigated using conversational robots as control condition. Philosophical Transactions of the Royal Society B, 374(1771), 20180033.

Righart R. & de Gelder B. (2006). Context influences early perceptual analysis of faces: An electrophysiological study. Cerebral Cortex,16,1249-1257

Schegloff, E.A., 1982. Discourse as an interactional achievement: some uses of ‘‘uhhuh" and other things that come between sentences. In: Tannen (Eds.), Analyzing Discourse: Text and Talk, Georgetown University Press, pp. 71--93.

Sprengelmeyer R. & Jentzsch I. (2006). Event related potentials and the perception of intensity in facial expressions. Neuropsychologia, 44(14), 2899-2906

Urgen, B. A., Kutas, M., & Saygin, A. P. (2018). Uncanny valley as a window into predictive processing in the social brain. Neuropsychologia, 114, 181-185.

Vilhjálmsson, H., Cantelmo, N., Cassell, J., Chafai, N. E., Kipp, M., Kopp, S., ... & Ruttkay, Z. (2007, September). The behavior markup language: Recent developments and challenges. In International Workshop on Intelligent Virtual Agents (pp. 99-111). Springer, Berlin, Heidelberg.

Ward, N. & Tsukahara, W. (2000). Prosodic features which cue back-channel feedback in english and japanese. Journal of Pragmatics, 32: 1177-1207.

Ressources

Le jeu de données Brain-IHM est disponible sur l'ORTOLANG.