

A2MIMO : Agent Conversationnel Animé et Analyse Multimodale pour la Santé Mentale

Soumaya Sabry
Université Caen Normandie,
ENSICAEN, CNRS, Normandie Univ,
GREYC UMR 6072
Caen, France

Antoine Cochard
Université Caen Normandie,
ENSICAEN, CNRS, Normandie Univ,
GREYC UMR 6072
Caen, France

Gaël Dias
Université Caen Normandie,
ENSICAEN, CNRS, Normandie Univ,
GREYC UMR 6072
Caen, France

Résumé

A2MIMO (*Artificial Agent for Mind Monitoring*) est une plateforme intégrée d'entretien psychiatrique assisté par un agent conversationnel animé. Le système couple l'agent Greta, hébergé dans une interface patient immersive sous Unity3D, à un serveur central asynchrone qui orchestre la session, gère le dialogue conduit autour d'un questionnaire clinique, et exécute en parallèle un moteur d'analyse multimodale en temps réel (visage, posture, texte; voix et dynamique de frappe en cours d'intégration). Une interface clinicien dédiée permet la supervision en direct ou en différé de l'entretien ainsi que la génération automatique de comptes-rendus. A2MIMO ne pose pas de diagnostic : il est conçu comme un outil d'aide au dépistage et au suivi, en complément de la consultation, et comme une infrastructure de recherche pour la constitution d'un corpus multimodal annoté automatiquement. La plateforme est accessible via la salle psychosensorielle du Laboratoire de Recherche GREYC, et un déploiement clinique avec le Centre Hospitalier Universitaire de Caen est en préparation.

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**; • **Applied computing** → **Health informatics**.

Mots-clés

Agent conversationnel animé, Santé mentale, Analyse multimodale

1 Motivation

L'évaluation clinique en santé mentale repose largement sur l'auto-révélation du patient [18], sur la qualité de l'interaction et sur l'observation de signaux comportementaux subtils, plutôt que sur des marqueurs biologiques objectifs [1]. Cette dépendance au jugement clinique, combinée à des contraintes structurelles fortes (p. ex. ressources limitées, temps d'écoute restreint, stigmatisation persistante) motive le développement d'outils numériques capables d'assister les praticiens sans s'y substituer. Dans ce contexte, les agents conversationnels animés (ACA) constituent une voie prometteuse. En couplant dialogue parlé et comportements multimodaux expressifs, ils favorisent l'engagement et la révélation de soi dans un cadre non-jugeant [12, 16].

Nous présentons **A2MIMO** (Figure 1), une plateforme articulant deux interfaces complémentaires autour d'un serveur central. Du côté patient, un environnement immersif héberge l'agent Greta et collecte simultanément les signaux multimodaux émis durant l'entretien (visage, posture, voix, texte, dynamique de frappe), sans capteur porté par le patient; des capteurs d'ambiance (température,

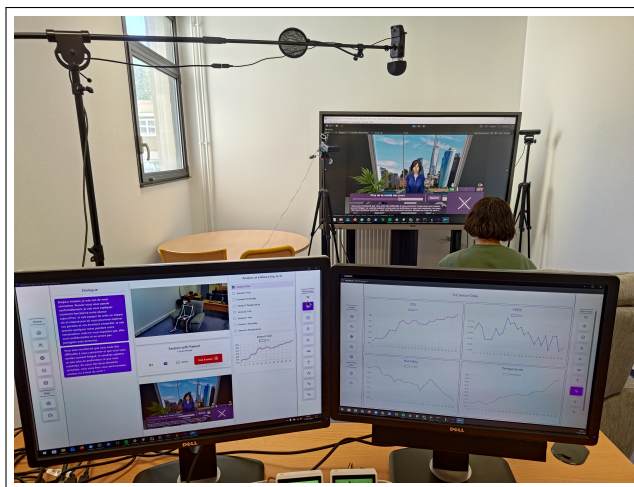


FIGURE 1 : Plateforme A2MIMO hébergée en salle psychosensorielle du Laboratoire GREYC.

humidité, qualité de l'air, CO₂) documentent les conditions d'enregistrement. Du côté clinicien, une interface dédiée (PsyUI) restitue ces signaux et les indicateurs comportementaux calculés en parallèle, offrant au praticien une vue d'ensemble de la séance. L'intérêt clinique est double : une passation standardisée que le clinicien peut revoir en différé, et des indicateurs comportementaux utiles. Toutes les données brutes et les sorties d'analyse sont conservées, ouvrant la voie à des analyses fines hors-ligne et à la constitution progressive d'un corpus multimodal.

Par rapport aux plateformes existantes (Virtual Human Toolkit [8], SimSensei [7]), A2MIMO n'apporte pas un nouveau modèle d'agent mais une intégration pensée pour la clinique : clinicien dans la boucle, dialogue ancré sur des questionnaires validés, modèles exécutés en local et moteur d'analyse extensible par *plug-ins*.

2 Architecture de la Plateforme

A2MIMO adopte une architecture distribuée et événementielle, organisée autour de trois modules fonctionnels (Figure 2).

Le **Module patient** (en bleu) regroupe l'interface patient sous Unity3D et l'agent Greta. L'interface fournit un environnement 3D où l'ACA est placé au centre de l'écran, accompagné d'un panneau textuel pour la question courante et d'un curseur de sévérité pour coter chaque item; les questions ouvertes reçoivent une réponse orale ou, au choix, écrite. Le poste de capture intègre deux caméras (visage et buste), un microphone et la dynamique de frappe. Greta,

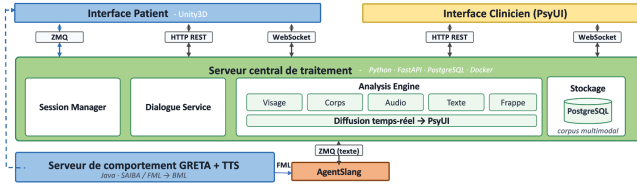


FIGURE 2 : Vue d'ensemble de l'architecture A2MIMO. Le code couleur distingue les trois modules fonctionnels : interface patient et ACA Greta (bleu), serveur central de traitement (vert) et interface clinicien PsyUI (jaune).

implémenté en Java selon le cadre SAIBA et le langage FML [5], synthétise les comportements verbaux et non-verbaux de l'agent (parole, regard, gestes, expressions faciales) en sortie MPEG-4.

Le **Module serveur central** (en vert), écrit en Python avec FastAPI, SQLAlchemy et PostgreSQL et conteneurisé avec Docker, joue le rôle d'orchestrateur asynchrone. Il héberge un *Session Manager* qui contrôle le cycle de vie de la session, un *Dialogue Service* qui gère la progression du dialogue, et un *Analysis Engine* modulaire dédié à l'analyse multimodale. Un modèle Whisper-tiny local transcrit les réponses orales en texte, et un modèle Llama 3.1 (8B), aussi en local, reformule les questions, demande d'élaborer et converse avec le patient dans un cadre de *chitchat* dyadique.

Le **Module clinicien** (en jaune), PsyUI, est une application bureau multiplateforme (Tauri, React). Elle propose un tableau de bord en temps réel agrégeant les flux bruts (vidéo, audio, transcription, capteurs d'ambiance) et les indicateurs comportementaux calculés.

La communication entre composants est assurée par WebSocket pour les flux multimodaux haut-débit, par REST/HTTP pour la gestion de session et des rapports, et par ZMQ via le *middleware* AgentSlang pour la transmission texte → FML vers Greta.

3 Dialogue Conversationnel Adapté

Le dialogue de A2MIMO s'appuie sur les questionnaires de dépistage de la dépression (PHQ-9 [10]) et de l'anxiété (GAD-7 [19]). Les items et leur cotation ne sont jamais modifiés : l'agent les énonce tels quels, précédés de phrases de contexte qui donnent à l'échange la forme d'une conversation. Le *Dialogue Service* gère les tours de parole, enregistre l'historique du dialogue et pilote Greta via AgentSlang jusqu'à la formule de clôture. L'adaptation repose sur un mécanisme d'élaboration optionnel, activé sous quatre conditions cumulatives : la session est configurée en mode IA, la question autorise explicitement l'élaboration, la réponse du patient est positive, et aucune élaboration n'a déjà été générée pour ce tour. Le contexte (question, réponse, historique) est alors transmis au modèle de langue, qui produit une question de suivi, prononcée par Greta sans incrémenter l'index du questionnaire. Le risque de réponse inadaptée est ainsi borné : le modèle ne modifie ni les items ni leur cotation, n'intervient qu'une fois par tour, est désactivable, et le clinicien qui supervise en direct peut interrompre la session.

4 Analyse Multimodale en Temps Réel

Le moteur d'analyse repose sur l'idée de *queue-and-worker* : chaque modalité dispose d'une file d'attente dans laquelle les signaux acquis sont déposés au fil de l'eau, et un travailleur dédié les consomme à son rythme avant de diffuser les résultats. Acquisition, inférence et diffusion sont ainsi totalement découplées. Les points d'entrée WebSocket (/face, /audio, /dialogue, ...) diffusent les flux bruts à la PsyUI afin de les visualiser, et les déposent en parallèle dans des files dédiées par modalité. Un *throttling* temporel limite la fréquence d'inférence sur les flux vidéo pour préserver la réactivité. Les modalités traitées sont :

- **Visage** : détection par cascade de Haar [21], suivie d'un CNN d'émotion (7 classes, images 48×48 en niveaux de gris),
- **Posture** : extraction de points clés via MediaPipe [14],
- **Texte** : adaptateurs LoRA [9] attachés à un *backbone* XLM-RoBERTa partagé [6], pour détecter émotions et sentiments,
- **Voix** (prototype) : MFCC, énergie et hauteur (prosodie).

Le choix d'un *backbone* unique par modalité, complété par des têtes ou adaptateurs légers, permet d'éviter le coût mémoire d'un modèle indépendant par tâche tout en gardant l'extensibilité. Un emplacement de *plug-in* est explicitement réservé pour intégrer de futurs modules (p. ex. empathie, adaptation stylistique, fatigue, stress). Les résultats sont rediffusés vers la PsyUI sur des points d'entrée d'analyse dédiés (/face_analysis, /text_analysis, etc.), permettant au clinicien de suivre la dynamique comportementale du patient en parallèle du contenu verbal. Ces indicateurs restent unimodaux, donc interprétables par le clinicien ; les flux étant horodatés et synchronisés, leur fusion [17] et l'évaluation des modèles seront menées sur le corpus clinique à venir.

5 Perspectives

À court terme, Whisper-tiny sera remplacé par un modèle de transcription plus robuste et le module vocal étendu à l'analyse de la prosodie. La dynamique de frappe sera intégrée comme modalité à part entière du moteur d'analyse, celle-ci étant de plus en plus reconnue comme un biomarqueur numérique d'humeur et de fatigue mentale [20]. Deux pistes complémentaires, suggérées par nos partenaires cliniciens, exploitent la plateforme comme dispositif d'observation active. La première consiste à introduire, sous condition et de manière contextuelle, un trait d'humour dans le dialogue afin de capter la réaction du patient (p. ex. variations de la fréquence cardiaque [11], entropie du rire [15]). La seconde vise à mesurer l'évitement du regard ou l'altération de la frappe face à des stimuli subjectivement anxiogènes : cet évitement constitue un biomarqueur reconnu des biais attentionnels associés à l'anxiété et du traitement émotionnel [2, 3]. À plus long terme, nous prévoyons de comparer deux configurations de l'agent Greta : une version standard et une version dotée de dialogues empathiques adaptatifs [4]. L'objectif est de créer un cadre interactionnel non-jugeant dans lequel les patients se sentent suffisamment en sécurité pour se livrer pleinement, plutôt que de se sentir évalués [13], et de mesurer l'effet de cette empathie sur l'alliance thérapeutique, la richesse des auto-révélation et l'engagement. Le déploiement au CHU est soumis à approbation éthique (consentement éclairé, données hébergées à l'hôpital).

Références

- [1] Anissa Abi-Dargham, Scott J. Moeller, Farzana Ali, Christine DeLorenzo, Katharina Domschke, Guillermo Horga, Amandeep Jutla, Roman Kotov, Martin P. Paulus, Jose M. Rubio, et al. 2023. Candidate biomarkers in psychiatric disorders: state of the field. *World Psychiatry* 22, 2 (2023), 236–262.
- [2] Thomas Armstrong and Bunmi O. Olatunji. 2012. Eye tracking of attention in the affective disorders: A meta-analytic review and synthesis. *Clinical Psychology Review* 32, 8 (2012), 704–723.
- [3] Yair Bar-Haim, Dominique Lamy, Lee Pergamin, Marian J. Bakermans-Kranenburg, and Marinus H. van IJzendoorn. 2007. Threat-related attentional bias in anxious and nonanxious individuals: a meta-analytic study. *Psychological Bulletin* 133, 1 (2007), 1–24.
- [4] Mukesh Barange, Sandratra Rasendrasoa, Maël Bouabdelli, Julien Saunier, and Alexandre Pauchet. 2022. Impact of adaptive multimodal empathic behavior on the user interaction. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents (IVA)*. 1–8.
- [5] Elisabetta Bevacqua, Ken Prepin, Radoslaw Niewiadomski, Etienne de Sevin, and Catherine Pelachaud. 2010. GRETA: Towards an interactive conversational virtual companion. In *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*. John Benjamins Publishing, 143–156.
- [6] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. 8440–8451.
- [7] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhomme, et al. 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. 1061–1068.
- [8] Arno Hartholt, David Traum, Stacy C. Marsella, Ari Shapiro, Giota Stratou, Anton Leuski, Louis-Philippe Morency, and Jonathan Gratch. 2013. All Together Now: Introducing the Virtual Human Toolkit. In *Intelligent Virtual Agents (IVA 2013) (Lecture Notes in Computer Science, Vol. 8108)*. Springer, 368–381.
- [9] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.
- [10] Kurt Kroenke, Robert L. Spitzer, and Janet B. W. Williams. 2001. The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine* 16, 9 (2001), 606–613.
- [11] Helmut K. Lackner, Elisabeth M. Weiss, Günter Schulter, Helmut Hinghofer-Szalkay, Andrea C. Samson, and Ilona Papousek. 2013. I got it! Transient cardiovascular response to the perception of humor. *Biological Psychology* 93, 1 (2013), 33–40.
- [12] Liliana Laranjo, Adam G. Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie Y. S. Lau, and Enrico Coiera. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association* 25, 9 (2018), 1248–1258.
- [13] Gale M. Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior* 37 (2014), 94–100.
- [14] Camillo Lugesia, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019).
- [15] Jorge Navarro, Raquel Del Moral, Pedro Cuesta-Alvaro, Rafael Lahoz-Beltra, and Pedro C. Marijuán. 2016. The entropy of laughter: Discriminative power of laughter's entropy in the diagnosis of depression. *Entropy* 18, 1 (2016), 36.
- [16] Simon Provoost, Ho Ming Lau, Jeroen Ruwaard, and Heleen Riper. 2017. Embodied conversational agents in clinical psychology: a scoping review. *Journal of Medical Internet Research* 19, 5 (2017), e151.
- [17] Syed Arbaaz Qureshi, Sriparna Saha, Mohammed Hasanuzzaman, and Gaël Dias. 2019. Multitask representation learning for multimodal estimation of depression level. *IEEE Intelligent Systems* 34, 5 (2019), 45–52.
- [18] Enikő Éva Savander, Jukka Hintikka, Mariel Wuolio, and Anssi Peräkylä. 2021. The patients' practises disclosing subjective experiences in the psychiatric intake interview. *Frontiers in Psychiatry* 12 (2021), 605760.
- [19] Robert L. Spitzer, Kurt Kroenke, Janet B. W. Williams, and Bernd Löwe. 2006. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of Internal Medicine* 166, 10 (2006), 1092–1097.
- [20] Claudia Vesel, Homa Rashidisabet, John Zulueta, Jonathan P. Stange, Jennifer Duffecy, Faraz Hussain, Andrea Piscitello, John Bark, Scott A. Langenecker, Shannon Young, Erin Mounts, Larsson Omberg, Peter C. Nelson, Raeanne C. Moore, Dave Koziol, Keith Bourne, Casey C. Bennett, Olusola Ajilore, Alexander P. Demos, and Alex Leow. 2020. Effects of mood and aging on keystone dynamics metadata and their diurnal patterns in a large open-science sample: A BiAffect iOS study. *Journal of the American Medical Informatics Association* 27, 7 (2020), 1007–1018.
- [21] Paul Viola and Michael Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 1. I–511–I–518.