Génération par diffusion conditionnelle de comportements non-verbaux pour les entretiens motivationnels

Nezih Younsi, Catherine Pelachaud, Laurence Chaby

Génération par diffusion conditionnelle de comportements non-verbaux pour les entretiens motivationnels

Nezih Younsi
ISIR - Sorbonne université
Paris, France
younsi@isir.upmc.fr

Catherine Pelachaud
CNRS - ISIR - Sorbonne université
Paris, France
Catherine.pelachaud@isir.upmc.fr

Laurence Chaby
ISIR - Paris Cité
Paris, France
Laurence.chaby@isir.upmc.fr

ABSTRACT

L'entretien motivationnel (EM) est une approche de communication centrée sur le client, visant à encourager le changement de comportement. Cet article présente un modèle de génération de comportement conçu pour simuler les comportements d'un thérapeute virtuel dans le cadre de l'EM, basé sur une analyse détaillée des échanges entre thérapeute et client. Les actes de dialogue dans l'EM, désignent les diverses formes de communication employées par le thérapeute et le client. En exploitant le corpus AnnoMI, qui documente des interactions d'EM humain-humain, nous avons identifié des co-occurrences entre les expressions faciales et les actes de dialogue du thérapeute et du client. Notre modèle s'appuie sur ces interactions pour simuler de manière réaliste les comportements d'un thérapeute virtuel. Une attention particulière a été accordée aux expressions faciales afin d'enrichir la dynamique interactionnelle entre le thérapeute et le client. Notre modèle repose sur une structure Observation–Action, intégrant une approche de diffusion conditionnelle et entraînée sur le corpus AnnoMI, permet de générer des comportements de thérapeute virtuel réactifs et adaptés aux signaux non verbaux du client.

KEYWORDS

Modèle de diffusion conditionnelle, Adaptation non-verbal, Entretien Motivationnel

1 INTRODUCTION

L'adaptation dans les interactions humain-humain est un processus multimodal qui se manifeste à travers différents niveaux de communication [9]. Les interlocuteurs ajustent leurs comportements l'un à l'autre, créant un échange fluide qui peut être observé en termes de comportements verbaux [10], mais aussi non verbaux [2, 22, 25]. Ces ajustements comportementaux se produisent à la fois consciemment et inconsciemment, servant à améliorer la qualité de l'interaction et à atteindre des objectifs communs [9]. Cette interaction complexe de signaux verbaux et non verbaux est essentielle pour faciliter une communication efficace et de qualité. Dans le domaine de l'interaction humain-agent, l'objectif est d'émuler ces comportements adaptatifs dans les interactions entre les humains et les Agents Conversationnels Animés (ACA), qu'ils soient physiques (comme les robots sociaux) ou virtuels. L'ajustement des signaux verbaux et non verbaux en temps réel, similaire à celui observé chez les humains, est essentiel pour développer des ACA capables de faciliter des échanges naturels et de qualité. [1]. Cette adaptation va au-delà de la simple réplication des comportements humains ; [6]. Notre étude se concentre spécifiquement sur la génération d'expressions faciales adaptatives pour les ACA, en développant un modèle génératif basé sur l'apprentissage automatique pour générer des expressions faciales pertinentes de thérapeutes virtuels lors de séances d'entretiens motivationnels avec des clients humains.

L'entretien motivationnel (EM) est une approche de communication centrée sur le client conçue pour faciliter sa motivation à changer de comportement [21]. En établissant une relation interpersonnelle, les thérapeutes cherchent à optimiser la qualité des séances [26][4] [11]. Pour mesurer l'adhérence aux changements de comportement des clients, le schéma du Motivational Interviewing Skill Code (MISC) [20] est fréquemment utilisé. Il catégorise différents aspects des interactions thérapeute-client, mettant l'accent sur les stratégies permettant aux thérapeutes de soutenir efficacement les clients dans leur processus de changement.

Nous voulons baser notre modèle du thérapeute virtuel d'EM sur le comportement des thérapeutes humains. Pour ce faire, nous avons d'abord analysé le corpus AnnoMI [28], une collection de vidéos d'entretiens thérapeute-client avec des actes de dialogue annotés selon le code MISC spécifique aux EM [21]. Nous avons procédé à l'analyse de la dynamique des expressions faciales entre le thérapeute et le client, selon les comportements annotés par l'EM. Suite à cela, nous avons construit une architecture basé sur un modèle de diffusion conditionnelle dans le but d'apprendre à reproduire la dynamique des expressions faciales du thérapeute virtuel, en prenant les actes de dialogue d'EM et les expressions faciales du client humain comme condition.

2 ARCHITECTURE ET TRAVAUX EN COURS

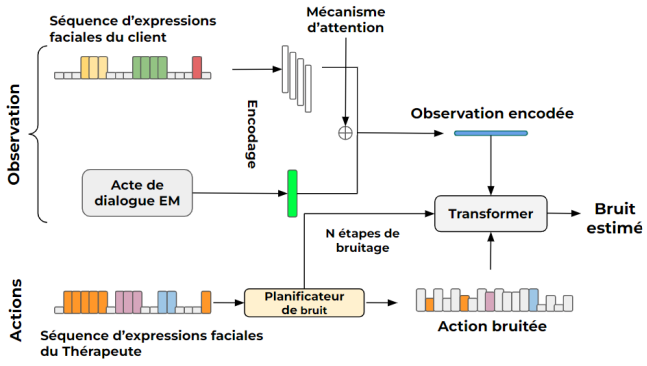

Figure 1: Architecture de l'estimateur de bruit du modèle de diffusion

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM

De nombreuses approches et modèles ont été conçus pour adapter les comportements des ACA à ceux des interlocuteurs humains. Des approches d'apprentissage supervisé telles que les réseaux de neurones, les Transformers, et les BI-LSTM ont également été utilisées [16] [24] [12] [27], ainsi que des modèles génératifs comme les Réseaux antagonistes génératifs (GANs) ou les auto-encodeur variationnels (VAE) pour produire des comportements dans de nouvelles situations, souvent conditionnant une modalité de comportement sur d'autres [15] [7] [14]. Les modèles de diffusion qui ont récemment montré des performance surpassant celle des précédant dans bien des domaines (Génération d'image [17], Vision par ordinateur [8], Modélisation multimodale [3]) ont aussi été appliqués à l'interaction humain-machine [23] [30] [19], pour cloner le comportement humain ou générer des gestes communicatifs conditionnés multimodalement. Les résultats obtenues par ces architectures ont motivé l'utilisation d'un modèle de diffusion conditionnel pour simuler l'adaptation interpersonnelle durant les entretiens motivationnels, utilisant une approche Observation/Action pour générer les expressions faciales de l'agent en réponse aux comportements de l'utilisateur.

Le corpus AnnoMI comprend 133 vidéos d'interactions thérapeute-client [28], annotées selon le Motivational Interviewing Skill Code (MISC) [20] qui décrit des actes de dialogue propres au EM, notamment *Change talk, sustain talk, neutral* pour le client et *information, question, reflection, advice* pour le thérapeute. Après extraction et synchronisation des Unités d'Action faciales (AUs) [13] via OpenFace [5] avec les annotations EM, une analyse initiale a révélé des co-occurrences entre certaines AUs propre à chaque interlocuteur. Cela nous a conduit à regrouper les AUs en catégories d'expression faciales : *Mouth up* (AU12, AU06 et AU25) , *Mouth down* (AU14 et AU15), *Nose wrinkle* (AU10 et AU09) ainsi que *Neutral* correspondant à une activation nulle ou de faible intensité [29]. Puis, nous avons effectué une analyse par extraction de séquences qui nous a permis d'identifier des co-occurrences entre ces catégories d'expressions faciales des thérapeutes et clients et les actes de dialogue propres aux EM. Les thérapeutes montrent notamment une propension à exprimer des expressions mouth up, soulignant leur rôle actif dans le soutien positif aux clients durant l'entretien. Ces résultats ont motivé le développement d'un modèle génératif d'expressions faciales du thérapeute, conditionné sur les expressions faciales du client et les actes de dialogue d'EM. Pour cela, la base de données a été restructurée en un format Observation/Action représentant chaque tour de parole; l'Observation (qui sera la condition du modèle génératif) comprend les expressions faciales du client et ses actes de dialogue, et l'Action inclut les catégories d'expressions faciales du thérapeute. Toutes les séquences d'expressions faciales sont standardisées à 25 frames par seconde, chaque frame reflétant l'activation ou non de l'une des trois catégories d'expressions faciales. L'autre composante de l'observation inclut un vecteur spécifiant le type d'acte de dialogue d'EM exprimé pendant le tour de parole, offrant ainsi un cadre structuré et normalisé regroupant les informations nécessaires pour l'entraînement du modèle de diffusion.

L'architecture du modèle génératif comprend trois composants : le planificateur de bruit DDPM (Denoising Diffusion Probabilistic Model), le modèle d'estimation du bruit, et le modèle de diffusion conditionnelle. Ce dernier intègre les deux premiers composants pour réaliser le processus de diffusion (phase de bruitage) et la phase d'échantillonnage (phase de débruitage) responsable de la génération de l'action correspondante étant donné une observation comme condition. Ces trois composants sont assemblés suivant un pipeline qui transforme les données en un espace latent, puis inverse le processus pour générer de nouveaux échantillons de données étant donné des conditions inédites.

Le modèle d'estimation de bruit est le coeur du processus de diffusion inverse et donc en majeur partie responsable de la génération de données. Au centre de l'apprentissage du modèle de diffusion, il apprend à estimer le bruit ajouté par le planificateur DDPM à la cible. Dans le cas de la génération d'images, l'estimateur de bruit utilise une architecture U-net. Cependant, pour les séquences temporelles, cette approche n'est pas adaptée. Notre modèle utilise des transformeurs pour traiter les séquences d'observation, capturant efficacement les relations et dépendances temporelles complexes grâce à des mécanismes d'attention multi-têtes et un codage positionnel. Cela permet de prédire le bruit ajouté à l'action cible par le DDPM à après un certain nombre de pas de bruitage qui est lui aussi pris en compte lors de l'estimation par le transformeurs.

La phase d'échantillonnage (*sampling*) représentant le processus inverse de diffusion, permet de générer graduellement une action cible à partir d'un bruit gaussien, en se basant sur une observation conditionnelle. Inspirée de l'algorithme de DDPM [17], cette phase débute avec un bruit gaussien de même dimension que l'action cible, puis appelle itérativement l'estimateur de bruit pré-entrainé pour estimer ce dernier à chaque étape de débruitage, conditionné par l'observation associée. Après un certain nombre d'itérations, correspondant aux nombres d'étapes de bruitage définies durant l'entraînement, on obtient la prédiction finale de la cible.

Suite à l'entraînement du modèle d'estimation de bruit avec ajustement des divers hyper-paramètres (taille de batch = 128, taux d'apprentissage cyclique [0.0001, 0.01], nombre de pas de bruitage = 500), nous avons enregistré une fonction de perte à 0.016 sur le jeu de données de validation après 1000 cycles d'entraînement. Concernant la phase de sampling, nous avons exploré plusieurs architectures pour ajuster la diversité ainsi que la qualité des séquences d'actions générées par le modèle. L'application de la méthode d'estimation par noyaux *Kernel density estimation* (KDE) qui, pour chaque condition d'observation donnée, génère plusieurs prédictions avant de choisir, via une estimation de densité par noyau, l'action avec la plus grande vraisemblance, a permis d'améliorer significativement la fidélité des actions générées au regard des conditions observées. Nous avons observé que la précision des séquences d'actions générées s'améliorait avec l'augmentation du nombre K d'échantillons générés par observation, jusqu'à atteindre un seuil optimal à $K = 5$, au-delà duquel les performances se stabilisaient. Étant donné que les interactions sociales peuvent être assimilées à un environnement stochastique, une seule observation peut très bien correspondre à plusieurs actions appropriées. Ainsi, la diversité des séquences d'actions, en plus de leur qualité, devient cruciale dans le développement de modèles génératifs pour simuler les comportements verbaux et non verbaux. Afin d'ajuster et améliorer la diversité du modèle nous adoptant la méthode *Classifier free guidance* (CFG) [18]. Cette technique, qui consiste à entraîner le modèle d'estimation de bruit de manière à la fois conditionnelle et non conditionnelle, a démontré une nette amélioration, tant en termes

de qualité qu'en diversité des résultats générés, surtout avec un ratio de 10% pour le non conditionnel et 90% pour le conditionnel.

Après l'implémentation de notre modèle, nous envisageons maintenant de mener des évaluations objectives et subjectives. Pour les premières nous prévoyons des analyses d'ablation et pour les dernières des tests de perception avec des vidéos d'EM avec un thérapeute virtuel.

3 REMERCIEMENT

Ce travail a été effectué dans le cadre du Projet ANR-JST-DFG PANORAMA.

REFERENCES

[1] Sean Andrist, Bilge Mutlu, and Adriana Tapus. 2015. Look like me: matching robot personality via gaze to increase motivation. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 3603–3612.

[2] Michael Argyle and Mark Cook. 1976. Gaze and mutual gaze. (1976).

[3] Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18208–18218.

[4] Zachary G Baker, Emily M Watlington, and C Raymond Knee. 2020. The role of rapport in satisfying one's basic psychological needs. *Motivation and emotion* 44 (2020), 329–343.

[5] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 59–66.

[6] Atef Ben Youssef, Mathieu Chollet, Hazaël Jones, Nicolas Sabouret, Catherine Pelachaud, and Magalie Ochs. 2015. Towards a socially adaptive virtual agent. In *Intelligent Virtual Agents*. Springer, Delft, The Netherlands, 3–16.

[7] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. 2021. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE virtual reality and 3D user interfaces (VR)*. IEEE, 1–10.

[8] Emmanuel Asiedu Brempong, Simon Kornblith, Ting Chen, Niki Parmar, Matthias Minderer, and Mohammad Norouzi. 2022. Denoising pretraining for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4175–4186.

[9] Judee K Burgoon, Lesa A Stern, and Leesa Dillman. 1995. *Interpersonal adaptation: Dyadic interaction patterns*. Cambridge University Press.

[10] Chris Cherpas. 1992. Natural language processing, pragmatics, and verbal behavior. *The Analysis of verbal behavior* 10 (1992), 135–147.

[11] Edward L Deci and Richard M Ryan. 2012. Self-determination theory. *Handbook of theories of social psychology* 1, 20 (2012), 416–436.

[12] Soumia Dermouche and Catherine Pelachaud. 2019. Engagement modeling in dyadic interaction. In *2019 international conference on multimodal interaction*. 440–445.

[13] Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior* (1978).

[14] Mireille Fares. 2020. Towards multimodal human-like characteristics and expressive visual prosody in virtual agents. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 743–747.

[15] David Greenwood, Stephen Laycock, and Iain Matthews. 2017. Predicting head pose from speech with a conditional variational autoencoder. ISCA.

[16] Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. 2018. Evaluation of speech-to-gesture generation using bi-directional LSTM network. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 79–86.

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[18] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

[19] Shivam Mehta, Siyang Wang, Simon Alexanderson, Jonas Beskow, Éva Székely, and Gustav Eje Henter. 2023. Diff-TTSG: Denoising probabilistic integrated speech and gesture synthesis. *arXiv preprint arXiv:2306.09417* (2023).

[20] William R Miller, Theresa B Moyers, Denise Ernst, and Paul Amrhein. 2003. Manual for the motivational interviewing skill code (MISC). *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico* (2003).

[21] William R Miller and Stephen Rollnick. 2012. *Motivational interviewing: Helping people change*. Guilford press.

[22] Lisette Mol, Emiel Krahmer, Alfons Maes, and Marc Swerts. 2012. Adaptation in gesture: Converging hands or converging minds? *Journal of Memory and Language* 66, 1 (2012), 249–264.

[23] Tim Pearce, Tabish Rashid, Anssi Kanervisto, Dave Bignell, Mingfei Sun, Raluca Georgescu, Sergio Valcarcel Macua, Shan Zheng Tan, Ida Momennejad, Katja Hofmann, et al. 2023. Imitating human behaviour with diffusion models. *arXiv preprint arXiv:2301.10677* (2023).

[24] Najmeh Sadoughi and Carlos Busso. 2017. Joint learning of speech-driven facial motion with bidirectional long-short term memory. In *Intelligent Virtual Agents: 17th International Conference, IVA 2017, Stockholm, Sweden, August 27-30, 2017, Proceedings 17*. Springer, 389–402.

[25] Karen L Schmidt and Jeffrey F Cohn. 2001. Human facial expressions as adaptations: Evolutionary questions in facial expression research. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists* 116, S33 (2001), 3–24.

[26] Peggy Van Minkelen, Carmen Gruson, Pleun Van Hees, Mirle Willems, Jan De Wit, Rian Aarts, Jaap Denissen, and Paul Vogt. 2020. Using self-determination theory in social robots to increase motivation in L2 word learning. In *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*. 369–377.

[27] Jieyeon Woo, Catherine Pelachaud, and Catherine Achard. 2023. ASAP: Endowing Adaptation Capability to Agent in Human-Agent Interaction. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 464–475.

[28] Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023. Creation, Analysis and Evaluation of AnnoMI, a Dataset of Expert-Annotated Counselling Dialogues. *Future Internet* 15, 3 (2023), 110.

[29] Nezih Younsi, Catherine Pelachaud, and Laurence Chaby. 2024. Beyond Words: Decoding Facial Expression Dynamics in Motivational Interviewing. In *LREC-COLING*. Turin, Italy.

[30] Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. 2023. Taming Diffusion Models for Audio-Driven Co-Speech Gesture Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10544–10553.