SPASCA: Social Presence and Support with Conversational Agent for Persons Living with Dementia

Ali Köksal[1], Jingjing Gu[2], Kotaro Hara[2], Jing Jiang[2], Joo-Hwee Lim[1], Qianli Xu[1]

[1]Institute for Infocomm Research (I²R), Agency for Science, Technology and Research (A*STAR), Singapore
[2]Singapore Management University, Singapore
qxu@i2r.a-star.edu.sg

Abstract

We present SPASCA - a conversational AI system that promotes psychological and cognitive well-being of persons living with dementia (PLWD). This system features an AI agent that provides social presence and support to PLWD through verbal communications, without physical presence of human caregivers. The system integrates (1) a novel dialogue model that generates dialogue items relevant to the user's experiences and lifestyle, (2) a digital avatar in the form of a talking head with the identity of a caregiver who is familiar to the demented user. We develop prototypes that adopt various interaction modalities and conversational styles and report the pros and cons of different system configurations through expert review. Our system shows the potential of conversational AI for personalized and affordable healthcare services.

Introduction

For persons living with dementia (PLWD), constant intervention from humans is necessary due to both physical and psychological needs of the individuals. However, such intervention is often taxing to the caregivers. Computer-assisted interventions have the potential to alleviate such a problem by augmenting human caregivers. For example, simulated presence therapy (SPT) is an intervention program that engages a PLWD via pre-recorded videos (Abraha et al. 2020), which nevertheless does not support true dialogues. Latest generative AI technologies (e.g. OpenAI-o1, Claude3.5, etc.) show impressive conversational capabilities, whereas they typically use single modality (i.e. speech) and lack knowledge of personal experience and preference. A computer-assisted intervention program should, as much as possible, simulate intelligence and interactivity of true human caregivers, such as realistic appearance, multimodal interaction, and empathy.

We develop an AI-driven interactive avatar to provide users engaging conversational experience with two main technical components: (1) a dialogue model (DM) to generate speech content that is suitable to PLWD, (2) a video synthesis model that generates lip-synchronized talking head with proper head motions and facial expressions. In this paper, we present the design of prototype systems (Fig. 1) and demonstrate their functionalities.

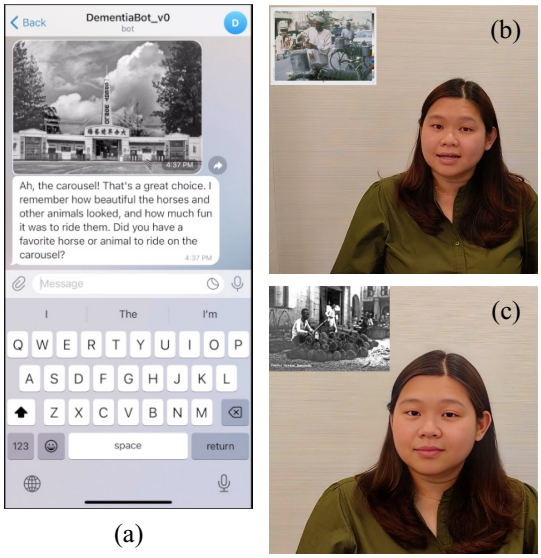

Figure 1: User interfaces of three prototypes. (a) Text-based chatbot using customized dialogue model, (b) Avatar with pre-recorded video following fixed scripts, (c) Avatar with talking face synthesis.

System Design

We develop the system in three stages: (1) data collection and analysis, (2) model training, and (3) system integration. Fig. 2 shows an overview of the process.

Data Collection and Analysis We collected data from conversations between a PLWD and a trained physiotherapist following a style of reminiscence therapy. Ten sessions of conversations were recorded from which we extracted the speech data and applied analytical methods to gain insights (Hara et al. 2024). The outcome was used to design the conversational style of the AI agent. In a separate session, we recorded videos of two practitioners who acted out the care-giving process in a studio setting. Video clips of dialogue items were extracted from the recording to implement the semi-interactive agent (Prototype 2). We also extracted a few face images for person identity in the talking head synthesis model.

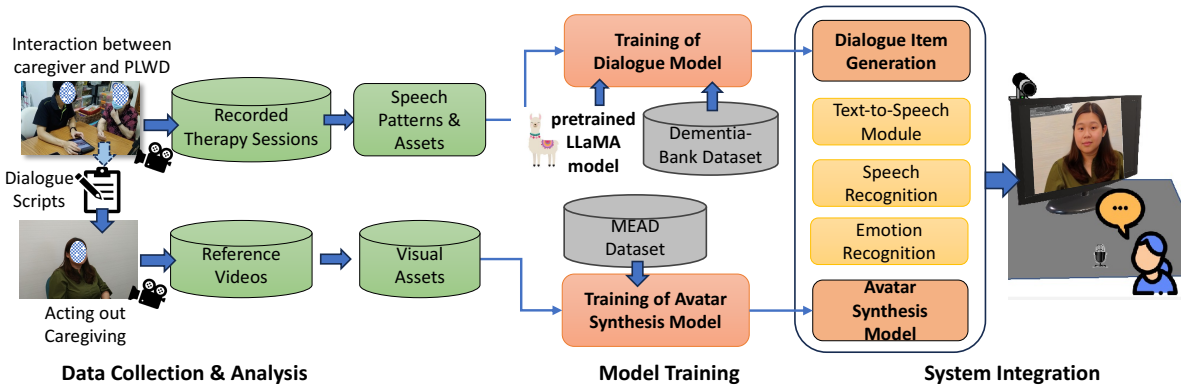

Figure 2: Overview of the SPASCA system.

Model Training : To train the dialogue Model, we conducted customization by prompting LLama2 (Touvron et al. 2023) based on conversations from real-world data. The customized DM generates dialogue contents that simulate the conversational style of a physiotherapist (Yang et al. 2024). Next, we train a video generation model to synthesise a realistic talking head avatar with accurate lip-synchronization (refer to Appendix) (Köksal, Xu, and Lim 2024). The model is capable of controlling facial features (i.e. emotional expression patterns) without requiring driving videos, unlike existing solutions such as Deep-Live-Cam (https://deeplive.cam/) and LivePortrait (Guo et al. 2024).

System Integration

We integrate the DM and avatar synthesis models into a prototype system that supports voice communications. As auxiliary modules, open-source or off-the-shelf tools/APIs are adopted, including a text-to-speech model using Google Cloud API, speech recognition using Whisper (Radford et al. 2022), and emotion recognition using DeepFace (Serengil and Ozpinar 2021).

In a typical setup, a digital avatar is shown on the computer screen and a webcam oversees the user. The user voice input is captured via a 'press-to-talk' mode. The voice is processed by the speech recognition module, and the text is sent to the DM. A text response is generated and converted into an audio clip, which is sent to the video synthesis module. Another input to the video synthesis module is the user's emotional state, which is recognized by the emotion recognition module. The video synthesis model generates a video clip of a talking face that makes the utterance with the matching facial expression. A web browser displays the video clip and hosts the interaction. We also design a few baseline prototypes according to the functionalities of the DM and video synthesis model.

1. **Dialogue model:** Other than the aforementioned customized DM (i.e. `DM-full`), an ablated DM is designed to follow the structured scripts (i.e. `DM-scripted`). The dialogue logic is implemented using simple decision trees. The resultant DM will generate (retrieve) responses based on pre-defined options, thus restricting the interaction to a confined set. This ablated DM is intended to verify if a simple semi-interactive system is sufficient to PLWD, whose speech patterns are considered simpler.
2. **Talking face:** A pseudo-synthesis model is developed that retrieves (instead of generates) the video clips corresponding to the utterance (denoted as `TF-recorded`). This is only applicable when the system adopts the `DM-scripted`. The advantage of this model is that it is faster and the visual fidelity is higher than generated videos (denoted `TF-generated`). To further study the effect of visualization, an alternative UI is developed, which does not show the talking head avatar. In other words, it only supports text- and audio-based interaction. For implementation, we adopt the Telegram platform, that enforces a text messaging style conversation.

Based on the above configurable functions, we design three prototypes. **Prototype-1** (Fig. 1(a)) is a chatbot that supports text and audio interaction, without showing the virtual avatar. Through prototype-1, we want to explore if the self-paced dialogue interaction is acceptable to PLWD. **Prototype-2** (Fig. 1(b)) adopts `DM-scripted` and `TF-recorded`. **Prototype-3** (Fig. 1(c)) adopts `DM-full` and `TF-generated`. By comparing user reactions in prototypes 2 and 3, we want to gain insights into the need for richness and variability of dialogue content, and the fidelity of visualization. As future work, we will conduct evaluation on the prototypes and report the results.

Summary

We present SPASCA - a conversational AI system that promotes psychological and cognitive well-being of persons living with dementia. The system integrates a novel dialogue model that generates dialogue items relevant to the user's experiences and lifestyle, and a pose and emotion controllable digital avatar with arbitrary identity (e.g. a caregiver). We develop prototypes that adopt various interaction modalities and conversational styles. This work paves the way toward customizable and controllable interactive agent with social intelligence to support personalized healthcare services. We are in process of evaluating the performance and efficacy of the system with target users.

Ethical Statement

The study was approved by the Institutional Review Board of the Agency for Science, Technology and Research (A*STAR), Singapore (Reference Number: 2022-127), and Singapore Management University IRB (Approval Number: IRB-22-171-A098(1122).

Acknowledgments

This research is supported by the SMU-A*STAR Joint Lab in Social and Human-Centred Computing (Grant No. SAJL-2022-HAS002 & C232918002). We would like to thank *Dementia Singapore* for the invaluable support to data collection, conceptualization, and system development.

References

Abraha, I.; Rimland, J. M.; Lozano-Montoya, I.; Dell'Aquila, G.; Vélez-Díaz-Pallarés, M.; Trotta, F. M.; Cruz-Jentoft, A. J.; and Cherubini, A. 2020. Simulated presence therapy for dementia. *The Cochrane database of systematic reviews*, 4: CD011882.

Guo, J.; Zhang, D.; Liu, X.; Zhong, Z.; Zhang, Y.; Wan, P.; and Zhang, D. 2024. LivePortrait: Efficient Portrait Animation with Stitching and Retargeting Control. *ArXiv*, abs/2407.03168.

Hara, K.; Natalie, R.; Cheong, W. S.; Gu, J.; and Xu, Q. 2024. Exploring Conversations between a Practitioner and a Person with Dementia. In *ACM ASSETS'24*.

Köksal, A.; Xu, Q.; and Lim, J.-H. 2024. Talking Face Generation via Face Mesh - Controllability without Reference Videos. *2024 IEEE Conference on Artificial Intelligence (CAI)*, 1380–1386.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *ArXiv*, abs/2212.04356.

Serengil, S. I.; and Ozpinar, A. 2021. HyperExtended LightFace: A Facial Attribute Analysis Framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, 1–4. IEEE.

Touvron, H.; Martin, L.; Stone, K. R.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D. M.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A. S.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I. M.; Korenev, A. V.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *ArXiv*, abs/2307.09288.

Yang, Y.; Huang, H.; Achananuparp, P.; Jiang, J.; and Lim, E.-P. 2024. Speaker Verification in Agent-Generated Conversations. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 Long Papers)*.