An Extensible Framework for Real-Time Conversational Avatars

Rahil (Somayeh) Jafaritazehjani, Johanna Björklund
rahil@cs.umu.se , johanna@cs.umu.se

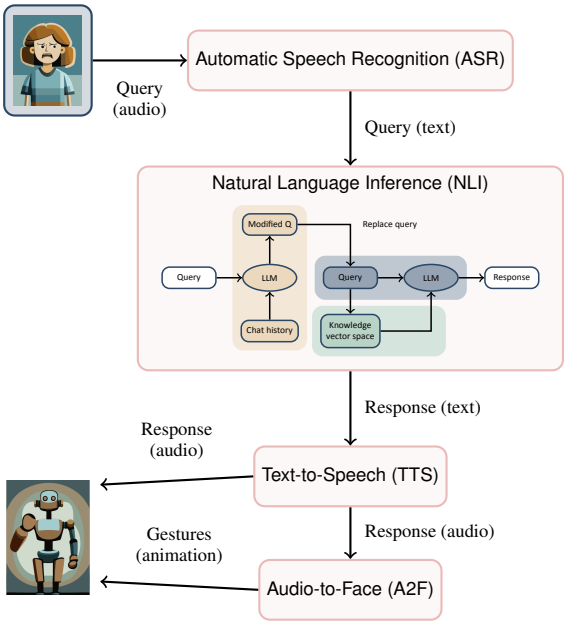
Abstract

The avatar framework is introduced to serve as a tool in simulation studies, or as a means of evaluating solutions for, e.g., automatic speech recognition or information processing in a larger context. It helps with prototyping and give a clearer understanding of the risks and opportunities associated with virtual agents.

A central goal of AI is to enable humans to interact with computers as naturally as they do with other people. The accomplishment of this aim requires comprehensive solutions to several challenging problems, including the automated processing and generation of speech, gestures, semantics, and pragmatics. As technology advances in these areas, avatars—essentially digital embodiments of chatbots—will become more integrated into our daily lives. It is believed that the additional modalities can enhance user experience, increase engagement, and lead to clearer communication. We are already seeing avatars being used in areas such as research,entertainment, gaming, education, healthcare, and marketing (Trepte and Reinecke, 2010; Kyrlitsias and Michael-Grigoriou, 2022).

Continued progress in the field requires a combination of competences, ranging form behavioural psychology to information retrieval and cloud computing. To lower the threshold for collaboration between research groups focusing on individual aspects of the problem, researchers linked to the Wallenberg Research Arena for Media and Language collaborate with the Ukrainan IT-company SoftServe to provide a transparent and extensible framework for real-time conversational avatars.

The schema on the right outlines the architecture of the avatar framework which is an end-to-end system enabling natural interaction between human and computer. The blocks labelled as ASR, TTS and A2F represent components which translate between modalities as follows. ASR is responsible for recognizing the audio input and transforming it to textual representation. TTS and A2F blocks both transform the textual response to an input query to audio and animation respectively. The block labelled as NLI denotes the LLM-driven processing units responsible for generating a textual response relevant to the input of the block (query in the architecture of the figure).



The framework consists of a base model (in blue) which together with a domain-specific knowledge component (in green) form the knowledge-based model. The can also be extended (in gold) to consider conversation history and contextual knowledge to provides more accurate and factual responses to the queries. The knowledge-based framework allows the avatar to be adapted practical use cases as it essentially implements retrieval augmented generation (Lewis et al., 2020; Karpukhin et al., 2020; Gao et al., 2024). Table 1 illustrates the result of providing the avatar with contextual knowledge in the case it is to be applied as a digital assistant to provide people with the information they need. The sample responses of the models indicates that the knowledge-based model generates

Query	Base Model	K.B Model
Who is the project manager of the arena?	The project is the person responsible for managing construction of the arena.	*Name* Project Manager, *Name* University, *Name* company.

Table 1: Comparing the responses of Base and knowledge-based (K.B) Models to the same input query, where proper names have been replaced by the label *Name*.

more accurate and factual response to the query as compared to the base model.

The value of the framework lies in the ease with which the existing components can be replaced, and new components added. For instance, to test the impact of new models on users for gesture generation, one simply changes between API calls of the models, leaving the remaining parts of the avatar untouched. Similarly, to understand the value of emotion-detection for naturalistic human-computer interaction, one can add a pre-processing step to generate a textual description of their emotion to be passed on to the NLI. The framework thus accelerates progress in the field by facilitating the evaluation of key technologies in a holistic setting.

References

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Christos Kyrlitsias and Despina Michael-Grigoriou. 2022. Social interaction with agents and avatars in immersive virtual environments: A survey. *Frontiers in Virtual Reality*, 2:786665.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, abs/2005.11401.

Sabine Trepte and Leonard Reinecke. 2010. Avatar creation and video game enjoyment. *Journal of Media Psychology*.