Natural Turn-Taking with Embodied Conversational Agents

Jonathan Ehret*
Visual Computing Institute
RWTH Aachen University, Germany

Andrea Bönsch
Visual Computing Institute
RWTH Aachen University, Germany

Torsten W. Kuhlen
Visual Computing Institute
RWTH Aachen University, Germany

1 Introduction

Adding embodied conversational agents (ECAs) to immersive virtual environments (IVEs) becomes relevant in various application scenarios, for example conversational systems. For successful interactions with these ECAs, they have to behave naturally, i.e. in the way a user would expect a real human to behave. Teaming up with acousticians and psychologists, we strive to explore turn-taking in VR-based interactions between either two ECAs or an ECA and a human user. Although the underlying project has a larger scope, we primarily focus on the ECAs' turn-taking in this report. We want to explore which gestures are suitable to facilitate turn-taking, and how these turn-taking cues impact (i) the perceived social presence of the ECA, (ii) the user's listening effort as well as (iii) remembering what has been said.

The heard text recall (HTR) tasks developed by Fintor et al. [1] is well suited to objectively measure listening effort in combination with testing the memory performance. In [1], two speakers alternatingly explained family stories from which participants had to remember different facts and even combine various information pieces to answer questions (e.g., *How old is Vincent?* for the story in Table 1) after listening to the speakers. However, so far this technique was only used in very artificial setups, e.g., the speech was only presented acoustically. We plan to use this paradigm in IVEs enhancing it such that the stories are presented by two ECAs, showing no, misleading, or fitting turn-taking behavior. Furthermore, we want to examine the influence of fitting or non-fitting co-verbal gesturing of the ECAs during presentation. We will thereby try to relate the objective performance in the HTR task with the subjectively perceived social presence. In case a correlation can be detected, we will introduce listening effort as a proxy to shed further light onto the nuances of human interactions with ECAs.

In an initial step we therefore need to find fitting and non-fitting co-verbal gestures and also model appropriate turn-taking signals. Skantze describes in a recent review of turn-taking for conversational systems [2] that there are various turn-yielding cues that could be of interest here. The most important aspects are linguistic features, prosody, breathing, gaze, and gestures. Since we plan to use pre-recorded speech, we will not alter the content (linguistic features) or prosody of those. Therefore, we want to focus on gazing at the next speaker in the gap between two inter-pausal units (IPUs) to signal turn-yielding, audibly breathing between two IPUs to signal turn-holding, and gestures being prolonged into the gaps to signal turn-holding. This kind of co-verbal behavior is especially interesting to evaluate in IVEs since co-presence of the interlocutors is important for efficient signaling of turn-taking (cf. [2]).

2 User Study

To that end, we want to conduct a user study using a head-mounted display (HMD). In the first study part (P_{Listen}), participants rate conversations between two ECAs presenting these family stories to the participant to find good candidates for fitting and non-fitting co-verbal gestures for further studies using motion-captured movements. In the second part (P_{Act}), participants then take over the role of one of the interlocutors. They are presented with all sentences (e.g., on a virtual blackboard), knowing only where the ECA should take over, so when they have to signal turn-yielding. Particularly, participants do not know when the turn is passed back by the ECA (only seeing labels marked with 2* in Table 1). The ECA is controlled by the experimenter to react on the turn-taking cues of the participant. As a between-subjects factor we vary whether the ECAs perform turn-taking signals or not at all, while using recorded gestures and facial movement for both. The gaps between turns (i.e. the time between the ECA ending an IPU and the participant starting) and perceived social presence are then measured as dependent variables. Furthermore, we record the turn-taking signals of the participants by means of eye tracking and hand controller movement and document any occurence of unintended speech overlap.

1	Rosa lives alone in a small 2-room apartment in Hamburg since her retirement as a postal worker
2*	Her twin brother Vincent is also still alive.
2	He is 80 years old and fortunately lives only two streets away.
2	Vincent thinks it's shameful how little Rosa's three children care about her.
1*	Her daughter Frida doesn't visit her at all, and Rosa's eldest son Gustav only drops by when a business trip takes him nearby.
2*	Rosa's second daughter Clarissa already has two children.
...	...

Table 1: Excerpt from a family story developed by Fintor et al. [1]. Texts are presented by two talkers (labeled 1 & 2 in the first column). For P_{Act} the participants (1) would only see labels marked with 2*, before which they need to pass the turn on to the ECA. They are told that the ECA counterpart only sees labels for sentences where to yield the turn back to the participant (1*), so they need to watch out when to continue. All other speaker labels, excluding who starts, would be hidden, while all texts are visible.

3 Discussion Points

The main questions we want to discuss during the workshop are:

- Is our approach promising to research turn-taking of ECAs in IVEs?
- Does reading out the text in P_{Act} conflict with the turn-yielding signals of the participant?
- What would work best as non-fitting gestures during P_{Listen}?

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SPP2236 AUDICTIVE – 444724862, Listening to, and remembering conversations between two talkers: Cognitive research using embodied conversational agents in audiovisual virtual environments.

References

[1] E. Fintor, L. Aspöck, J. Fels, and S. J. Schlittmeier. The role of spatial separation of two talkers' auditory stimuli in the listener's memory of running speech: listening effort in a non-noisy conversational setting. *Int. J. Audiol.*, pp. 1–9, 2021. doi: 10.1080/14992027.2021.1922765

[2] G. Skantze. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech and Language*, 67(101178), 2021. doi: 10.1016/J.CSL.2020.101178

*e-mail: ehret@vr.rwth-aachen.de