Who’s next? Integrating Non-Verbal Turn-Taking Cues for Embodied Conversational Agents

Jonathan Ehret∗
Visual Computing Institute,
RWTH Aachen University, Germany

Andrea Bönsch
Visual Computing Institute,
RWTH Aachen University, Germany

Cosima A. Ermert
Institute for Hearing Technology and Acoustics, RWTH Aachen University, Germany

Patrick Nossol
Department of Computer Science, RWTH Aachen University, Germany

Chinthusa Mohanathasan
Work and Engineering Psychology, RWTH Aachen University, Germany

Sabine J. Schlittmeier
Work and Engineering Psychology, RWTH Aachen University, Germany

Janina Fels
Institute for Hearing Technology and Acoustics, RWTH Aachen University, Germany

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

ABSTRACT
Taking turns in a conversation is a delicate interplay of various signals, which we as humans can easily decipher. Embodied conversational agents (ECAs) communicating with humans should leverage this ability for smooth and enjoyable conversations. Extensive research has analyzed human turn-taking cues, and attempts have been made to predict turn-taking based on observed cues. These cues vary from prosodic, semantic, and syntactic modulation over adapted gesture and gaze behavior to actively used respiration. However, when generating such behavior for social robots or ECAs, often only single modalities were considered, e.g., gazing. We strive to design a comprehensive system that produces cues for all non-verbal modalities: gestures, gaze, and breathing. The system provides valuable cues without requiring speech content adaptation. We evaluated our system in a VR-based user study with 32 participants executing two subsequent tasks. First, we asked them to listen to two ECAs taking turns in several conversations. Second, participants engaged in taking turns with one of the ECAs directly. We examined the system’s usability and the perceived social presence of the ECAs’ turn-taking behavior, both with respect to each individual non-verbal modality and their interplay. While we found effects of gesture manipulation in interactions with the ECAs, no effects on social presence were found.

CCS CONCEPTS
• Computing methodologies → Intelligent agents; • Human-centered computing → Natural language interfaces; User studies; Virtual reality.

∗ e-mail: ehret@vr.rwth-aachen.de

1 INTRODUCTION
The ability to take turns in conversations is a fundamental aspect of human communication. However, it remains often unaddressed in exchanges with virtual interlocutors. When designing embodied conversational agents (ECAs) [5] – anthropomorphic, autonomous, virtual agents that use natural language – for interactions in virtual reality (VR), leveraging different modalities of turn-taking can significantly improve the effectiveness and naturalness of such interactions. Skantze [37] describes different modalities used in natural conversations to communicate turn-taking. Those are verbal cues (i.e., syntax, semantics, and pragmatics), prosody, breathing, gaze, and gestures. The goal of this paper, however, is to derive a system producing non-verbal turn-taking cues, so only using the latter three of the aforementioned modalities. The rationale behind that is, that while the speech signal is often predefined (either scripted or for better naturalness even prerecorded [12]), the non-verbal behavior of the conversing ECAs is frequently generated. Furthermore, according to Skantze [37], using gestures and breathing has attracted less attention when designing systems to regulate turn-taking. Since human turn-taking signals are ambiguous and sometimes lack clarity, we decided to use a rule-based system, not a data-driven approach. This system should produce clear and intelligible signals, while still leveraging the subconscious processing skills of humans in conversations. We produce non-verbal turn-taking cues specifically for ECAs in VR since the co-presence of the sender and recipient of such signals appears to play a crucial role in their effectiveness (cf. [37, p. 6]).
In this paper, we will first provide an overview of related work, followed by a description of our derived implementation. Subsequently, we will present a VR study we conducted to evaluate the performance of our system and discuss the insights gained.

2 RELATED WORK

One modality for giving cues about whether an interactor wants to continue speaking (turn-hold) or is willing to pass the turn on to someone else (turn-yield) is gazing behavior, first described by Duncan Jr. (1974) [10]. Extensive research has been conducted on implementing natural gaze patterns for ECAs [34], with a particular emphasis on the execution of gaze shifts, i.e., saccades, (e.g., [1, 26]) and the coordination of eyes, head, and torso movements (e.g., [29, 36]). Furthermore, when generating eye movement for ECAs, eye blinks have to be produced for the gazing to look plausible [38]. Gaze thereby serves two functions at the same time (cf. [18]): observing the world and constituting a behavior that is observed. For the former, a recent approach by Goude et al. [17] uses the saliency of the virtual scene rendered from the perspective of the ECA to automatically generate plausible gaze patterns. However, our focus lies on the latter function. Gaze can, for example, be used to predict intention [3, 20] but emerging gaze patterns can also transport social signals, like turn-taking intent [25]. There is a multitude of observation studies on how gaze is altered by humans during conversations to signal turn-taking (e.g., [25, 28]). These observations are then used to predict who is going to speak up next in a conversation, for example using head orientation only [31] or combining it with eye tracking to enhance accuracy [9]. Jokinen et al. [24] found that eye gaze is especially useful to distinguish whether a speaker is taking a pause to think (turn-hold) or wants to yield the turn. Furthermore, this data is also used to derive gaze models which can then be applied to ECAs or socially-aware robots (e.g., [16, 27]).

Wagner et al. [40] describe gestures as also playing a key role in signaling turn-taking. One important aspect here is that during spontaneous conversations, gestures often terminate before the end of speech when yielding the turn while they extend well beyond the end of the speech when holding the turn [41]. Furthermore, posture shifts occur more frequently at discourse segment boundaries [6].

Several studies compare how combining different modalities improves the clarity of turn-taking. For example, prosody alone is not sufficient to predict turn-taking [32], and combining respiration and gaze yields superior predictions to using gaze alone [21]. Recent approaches using artificial networks combine even more modalities, e.g., acoustic and linguistic [33] combined with visual features automatically extracted from videos [23]. De Coninck et al. [7] chose the opposite way, predicting gesture classes and gaze targets from annotated conversational states. Edlund and Beskow [11] developed the MushyPeek framework, which deliberately manipulated avatar behavior in avatar-mediated communication. Due to these manipulations (e.g., changing gaze behavior or adding raised eyebrows), participants unconsciously changed their communication behavior. Furthermore, ECAs can communicate attitude through their behavior when interrupting [30], which can also be used to shape turn-taking [4]. However, turn-taking behavior can also be manually added to communication with an ECA in a Wizard-of-Oz paradigm to effectively influence the turn-taking during the interaction and create more natural intercourse (e.g., [8]). We refer the interested reader to [37] for further insights into the intricacies of turn-taking.

3 IMPLEMENTATION OF NON-VERBAL TURN-TAKING CUES

Following the findings of Skantze [37] we based our implementation on three non-verbal modalities: gaze, gestures, and breathing. Due to the additive nature of turn-taking cues (cf. [37]) we combined all three to give cues that are as clear as possible. We deliberately excluded syntactic, semantic, or prosodic turn-taking cues since we strove to implement a system that works with any speech material without a need for adaptation.

We structured each conversational act (i.e., a sentence being uttered by one speaker which might be followed by another sentence by the same speaker or a speaker-switch) in three phases, which will be treated differently when generating non-verbal behavior.

1) DuringUtterance: From the start of the sentence up to 1 s before the end.

2) CloseToEnd: The 1 s time frame at the end of the utterance before finishing the sentence. This time frame is chosen in accordance with the evaluation by Ishii et al. [21].
3) Gap between two utterances, which can be uttered by the same speaker (turn-hold) or by different speakers (turn-yield). The gap between two sentences of the same speaker or by different speakers is chosen to last by default 300 ms, which is approximately the median in real-life conversations (c.f. [37]).

For each of the three used modalities, we generated behavior according to these phases. Thereby, we aimed for generating behavior patterns that resemble those observed in real conversations. However, since there are large interpersonal differences in these behaviors we tried to derive simple rules to implement a system that is easy to understand, leveraging our trained skills from human-human interactions. At the same time, we deliberately excluded all nuances and possible ambiguities observed in real conversations reducing some of the complexity.

### 3.1 Gazing

To dynamically adapt the ECAs’ gaze, we first implemented a general gaze controller for the MetaHumans used. We move the eyes, head, and upper body towards the gaze target following the dynamics of the movement in the work by Pejsa et al. [29]. Thereby the eyes start to move earlier and always move all the way to the target, while the head and torso start slightly delayed. Opting for natural gazing, we align the ECA’s head only 80% to the target position, while the remaining 20% are covered by the eyes. Although Sidenmark and Gellersen [36] report that gaze shifts with angles below 25° tend to be performed by eye movement only, this model looked plausible in our scenario. Additionally, the torso aligns 10% with the gaze target. For a more dynamic eye movement, we added optional saccade movement (periodic additional eye rotation, e.g., when listening) based on the findings of Lee et al. [26]. Furthermore, we added blinking following the statistics described by Truotou et al. [38] using cubic ease-in/out and also forcing blinks for larger gaze shifts. To simulate the natural eye contact between humans, we, furthermore, designed the ECAs to periodically alternate the eye they look into while engaging in eye contact with the user.

The gaze behavior is implemented for the use case of two talkers taking turns telling a story to one Addressee. Thereby the talkers always switch roles between Speaker and, while the other one is speaking, Listener. During the phases of the conversational act, we use different gazing patterns for the phases DuringUtterance and CloseToEnd. For the latter, we differentiate between holding the turn and yielding the turn to the next talker During the Gap the behavior of either CloseToEnd realizations is prolonged.

**DuringUtterance**: Following the observations by Rienks et al. [31] the Speaker divides his/her gazes equally between Listener (33%), Addressee(33%), and random gaze targets in the environment (33%, see Figure 3). Also following [31], the Listener gazes twice as much at the Speaker (67%) than at the Addressee (33%). Gaze durations are chosen from a normal distribution ($\mu=2.75\text{ s}, SD = 2.4$), following Ding et al. [9], with a minimal gaze duration clamped at 1.0 s since smaller gaze lengths tended to look very unnatural.

**CloseToEnd** (holding): Following the results of Ishii et al. [22], the Speaker looks at the Listener in 25.1% of the cases and breaks the gaze immediately in case the gaze becomes mutual. In our implementation, each mutual gaze is accordingly broken immediately during this phase by averting the gaze towards a gaze target in the environment. In case the previous gaze ends within this phase (it potentially extends further, see gaze duration distribution above), the Speaker looks at the Listener in 25.1% of the cases and otherwise averts gaze towards an environment gaze target. Headling to the observations of [22], the Listener looks towards the Speaker in 62.5% of the cases (if a new gaze target needs to be chosen) and otherwise simply extends the previous gaze during this phase.

**CloseToEnd** (yielding): To show clear yielding behavior, the Speaker always looks at the Listener, who in this case is the next speaker. In Ishii et al.’s observation, the Speaker looks away in 25% of the cases if the gaze is not mutual [22]. We, however, always keep the gaze at the Listener during this phase for clarity (again only changing the gaze once the previous gaze exceeded the minimal gaze duration of 1 s). Also for clarity, the Listener always looks at the Speaker in this phase (in [22] this was only true in 62.5% of the cases) and averts the gaze immediately into the environment once the gaze is mutual (in [22] this was only observed in 71% of the cases). This is in line with the findings by Oertel et al. [28] in which incoming speakers tended to look away while the previous speaker tried to maintain a mutual gaze. Since the Addressee is never expected to take the turn, he/she is never looked at in CloseToEnd.

In most cases the Addressee is the user him-/herself, so we don’t need to generate gazing behavior. However, to also cover cases in which one ECA takes over the role of the Addressee, we added a simplified model of always looking at the current Speaker,
either virtual or human. Following Wagner et al. [40] listeners predominately engage using head nods when listening. Therefore, we designed the Addressee to produce nods at the end of each sentence of the other ECA, respectively end of turn of the participant, with a chance of 50% to seem more natural and involved.

3.2 Gesturing
Following the observations by Zellers et al. [41], gestures should not terminate in the time frame of 500 ms prior to the speech end if the turn should be held beyond the following gap. Therefore, we manipulated the co-verbal gestures such that in the case of held turns the hands are fixed on the last accent/stroke of the animation before the Gap and prolonged by 300 ms into the gap. This way the hands hold the accent while the rest of the body still performs slight movements in a natural way. This animation is then blended together with the animation played after the Gap. Accordingly, the first accent/stroke is prolonged 300 ms forward, so that the animation does not blend back to an idle pose during the Gap – all co-verbal animations are by default played with 300 ms blend in and out from and to the looped idle animation – and the gesture is held during the Gap (see Figure 2).

3.3 Breathing
As described in [21] and [37], respiration can be a helpful cue for initializing a turn but also for holding a turn. To this end, we extracted inhale audio sequences from the used audio material and replay a randomly selected one during the Gap for the ECA who is going to speak afterward. This is independent of the fact whether this is a turn-hold or whether the turn is passed on in the break, since – as common in natural conversation – the person speaking after the Gap needs to take a breath to have enough air for the following utterance.

4 EVALUATION
To test whether the added non-verbal turn-taking cues are (sub-consciously) perceived as intended, we conducted a VR user study (which constitutes a more realistic setting than [7]). We expected the following hypotheses to be confirmed:

H1 ECAs are rated as more socially present and natural when more modalities of turn-taking cues are shown.
H2 When participants take over an active role in turn-taking, gaps between turns decrease with more modalities of turn-taking cues being shown.
H3 When participants take over an active role in turn-taking, ECAs’ behavior is rated less confusing when turn-taking cues are embedded.

4.1 Material
The study took place in a virtual living room\(^3\) which is populated by two MetaHumans\(^3\). The study was rendered using Unreal Engine 4.27. The ECAs are positioned in front of the participant on both sides of a virtual TV screen, both at 30\(^\circ\) and 1.5 m of the participant, facing him/her (see Figure 3). For the gazing implementation, we defined additional environment gazing targets which were placed on sensible objects/locations in the scene (see Figure 3). As speech content, we utilized family stories from the heard text recall (HTR) task [35], which has speech material from two different voices (female, male) and face tracking data (using iPhone’s ARKit for face tracking) readily available [14]. Each of the 34 texts contains 10 sentences, narrating the stories of different families and providing information about three generations of family members, such as their names, jobs, hobbies, and relationships. Nine questions accompany each text, requiring participants to combine information from different sentences. In the database, suggestions for turn passes between two speakers are given, yielding 4-5 turn changes per text. The number of sentences spoken by one talker in a row is arbitrary while the sum of sentences spoken by each speaker is balanced. These texts were chosen as they originate from a verified paradigm, featuring compatible content complexity throughout the texts, and provide all the necessary information for this evaluation. Additionally, we posed the questions during the first study part, concealing the true purpose of the study, using attentive listening to the stories and recalling their contents as a plausible cover story. Furthermore, this had participants focus carefully on the conversation and thereby also on the non-verbal behavior.

While face tracking data was available and could directly be used to animate the ECAs, full-body movements were missing. Thus, we recorded co-verbal movements for each sentence, using consumer components only (see Figure 4), namely a Vive Pro 2 (head-mounted display (HMD)), two Valve Index Controllers, that support rudimentary finger tracking, and six Vive Trackers, which were attached to the feet, elbows, pelvis, and chest. The recorded rotation and translation data of all nine tracking points was then applied to the MetaHuman skeleton using Unreal Engine’s Full-Body IK\(^3\).

\(^3\)Motion Capture Plugin: https://git-ce.rwth-aachen.de/vr-vis/VR-Group/unreal-development/plugins/MoCapPlugin
4.2 Apparatus
The study was executed on a desktop PC (Intel Core i9-10900X, 32GB RAM, GeForce RTX 3080 Ti). For the presentation a Vive Pro Eye HMD was used, which allowed for eye tracking during the study. Eye tracking was used to identify mutual gaze between the ECAs and the participant and also logged for further analysis. During the study, participants wore the same motion capture setup as described in Section 4.1, so that their movement could be transferred onto a gender-matching full-body avatar and additionally be saved for further analysis. The audio was replayed over Sennheiser HD650 headphones using a Focusrite Scarlett 2i2 3rd Gen audio interface. The scene was auralized with Virtual Acoustics\(^4\) using generic binaural rendering. A static directional filter of human speech was assigned dynamically to the speech sound sources (cf. [13]). For simple study control a Study Framework Plugin\(^5\) for Unreal Engine was utilized.

4.3 Study Design
The study was split into two parts; Listen and Act. In the first part (Listen), participants listened to 10 family stories from the HTR being told by the two ECAs. In the second part (Act), participants took over a part in telling the stories while one of the ECAs represented the addressee. Thereby participants had to directly react to the turn-taking cues given by the ECA.

In both parts, five levels of the Turn-Taking Cues factor (T) are presented:
- \(T_{\text{None}}\): no turn-taking cues are given
- \(T_{\text{Breath}}\): only the breath cues are audible (see Section 3.3)
- \(T_{\text{Gesture}}\): only the gesture cues are shown (see Section 3.2)
- \(T_{\text{Gaze}}\): only the gazing cues are shown (see Section 3.1)
- \(T_{\text{Full}}\): all of the above are combined

When gazing turn-taking cues are not given we tried to generate similar gaze patterns, which, however, do not carry any turn-taking information. To that end we let the ECAs gaze at the other ECA, the participant, and gaze targets in the environment with equal frequencies, using the same gaze length normal distribution we used in the Listening phase (cf. Section 3.1). When gestures are not used as turn-taking cues, we still used gesture holding as described in Section 3.2 but at random gaps. So, if the ECA did not continue after the Gap with a randomly held gesture, that held gesture was interpolated into the idle gesture. The number of held gestures was kept approximately the same as in the conditions using them as turn-taking cues. Inhale sounds were omitted entirely when not used as cues. The different conditions can be seen in the supplemental video\(^6\).

4.4 Study Procedure
After reading a study description for the Listen part and giving their informed consent, participants filled out a demographics questionnaire and were equipped with the tracking hardware (HMD, Valve Index Controllers, six Vive Trackers), used for applying their motions onto their avatar, and headphones. Once immersed, first a calibration of the gender-matched avatar and the eye tracking was performed. After that, the experimenter adjusted the voice detection threshold such that the HMD’s microphone could be used to detect participants starting to speak. Following that, participants undertook one training trial of the Listen part (always using \(T_{\text{Full}}\)). During the Listen part, a male and a female ECA (see Figure 1a) told a family story (see Section 4.1) while using different turn-taking cues to signal turn-taking. Participants were instructed to listen carefully to the stories. Once finished nine questions regarding the stories heard (e.g., "How old is Vincent?") were shown on the virtual TV screen, which participants had to answer orally. The correct answer was presented to the experimenter, who had to log whether the right answer was given by the participants by means of button presses. When all nine questions were answered, a Likert-scale questionnaire assessing Social Presence was presented within the virtual environment. The participants had to point and click on the corresponding answer with the controller. The questionnaire included sub-scales from different questionnaires which we expected, if anything, to change due to the used manipulation. The underlying hypothesis is based on the observations in [39] that higher social presence was found for ECAs exhibiting richer non-verbal behavior For Anthropomorphism the first construct of the Godspeed questionnaire \([2]\) was presented where participants have to pick values on 5-point bipolar scales (e.g., between Fake and Natural). After that the constructs Human-Like Behavior (HLB) and Agent’s Coherence (COH) from the ASA questionnaire \([15]\) were utilized, which had to be answered on a 7-point Likert scale. Once answering those, the actual Listen phase started, repeating the same task as in the training trial 10 times. During these 10 trials, each of the five levels of \(T\) was presented twice. The presentation order of the turn-taking levels and the presented texts is counterbalanced using the Balanced Latin Square method. Participants were asked after each trial whether they wanted to have a break (as an additional field in the last questionnaire) and had to take a break of at least 5 min after completing all 10 trials. At the beginning of the break, a short questionnaire had to be filled out (at a desktop computer) asking for their general experience during the Listen part.

When feeling ready for the next part, participants had to read the study description for the Act part and conduct 10 trials of the

---

\(^4\)https://www.virtualacoustics.org/


\(^6\)Supplemental Video: https://youtu.be/txN9s1ZpMA
Act part which were again foregone by a training trial. During the Act part, the spatial layout remained the same apart from a flip chart being placed between both ECAs. This virtual flip chart was used to present the text that had to be spoken by the participant, since in this study part the participants took over one part in telling the stories (see Figure 1 b)). In this part 10 different stories were used than in the Listen part. While participants told the story with the ECA of opposite gender to their own, the ECA with the same gender took over the role of the Addressee. Participants were shown whether they take the first turn at telling the story and the sentences they have to speak next. However, when the ECA speaks they have no information on when to start and are therefore told to carefully look at the ECA to find out when to speak and then start speaking as quickly as possible. Using the HMD’s microphone and a calibrated speech detection threshold, the start of a participant’s utterance is recognized and the gap length since the end of the ECA’s speech is logged. Once participants are done with their turn (i.e., they read the entire text currently displayed on the flip chart), the experimenter triggers the ECA to continue by means of pressing a button. Additionally, the experimenter logs any attempts to speak during the ECA’s turn. If the participant does not start speaking for 3 s after the ECA is done, the ECA performs a dedicated turn-yielding gesture towards the participant. Once the full story was told we did not ask the related HTR questions but showed a virtual Likert scale questionnaire asking whether it was easy to understand when to speak up, whether the behavior of the partner was confusing or ambivalent, and whether the task was frustrating. All of the above were answered on 7-point Likert scales from −3 (Disagree) to 3 (Agree). Again, the 10 trials were counterbalanced. After finishing this part, participants had to answer a final desktop-based questionnaire and were compensated 15 € for their time. On average the study took 75 min, of which the immersed time for the Listen part was 31.9 min and 11.6 min for the Act part.

4.5 Participants

32 persons (21 male, 11 female) took part in our study. One female participant felt unwell during the execution and had to cancel the study. The remaining participants had a mean age of 25.9 years (SD = 5.0) and all self-reported normal hearing and normal or corrected to normal vision. Four participants (12.5%) were fluent in German while the others had German as their mother tongue (the whole study was conducted in German). Three of the participants (12.5%) never used VR before, seven (21.9%) only once before, 14 (43.7%) less than 10 times, and the rest (21.9%) more frequently.

4.6 Results

Data that is recorded per trial is analyzed using one-way repeated-measure ANOVAs with the single factor T (levels: TNone, TBreath, TGesture, TGaze, TFull). Data is checked before on normality using Shapiro-Wilk tests. Where the assumption of sphericity (evaluated with Mauchly’s test) is violated Greenhouse-Geisser correction is used when interpreting the ANOVA. When applicable paired-sample t-tests with Bonferroni correction are used as post-hoc tests.

Analyzing the questionnaires posed after each trial in the Listen part, we first confirmed the internal validity of the questionnaires by computing their Cronbach’s Alpha, which were $\alpha = .95$ (Godspeed), $\alpha = .93$ (HLB) and $\alpha = .77$ (COH). Averaging the scores per turn-taking level for each participant and computing ANOVAs did not reveal any significant effects (all $F \leq 1.12$ and $p \geq .33$). On average the ratings for anthropomorphism (Godspeed) were $M = 2.7$ (SD = 1.1; from scale $[1, 5]$), for human-like behavior (HLB) $M = 0.6$ (SD = 1.5; from scale $[−3, 3]$), and for coherence (COH) $M = 2.3$ (SD = 0.9; from scale $[−3, 3]$).

Due to the fact that the number of texts used is a multiple of the numbers of levels of $T$, the balanced Latin Square counterbalancing always matched the same text to the same level of $T$. Although the HTR questions were primarily used as a disguise, we still planned to evaluate the performance in the HTR task. However, due to the above-mentioned shortcoming, it is not feasible to evaluate the answers given, since the texts and their questions might vary in difficulty, which might be confounded with experimental variation. In the questionnaire following the Listen part participants rated on a scale from ‘−3 (‘Strongly Disagree’) to 3 (‘Strongly Agree’)’ that the ECAs sounded like humans in the real world ($M = 1.6$, $SD = 1.7$) but did not look as alike to humans in the real world ($M = 0.3$, $SD = 1.7$). Participants on average also stated that they noticed the ECAs signaling to yield or keep the turn ($M = 0.6$, $SD = 1.8$). However, also 19.4% rated this below or equal to $−2$.

A repeated-measures ANOVA (with Greenhouse-Geisser correction) revealed a significant effect of $T$ on the gap participants left before starting to speak once the ECA finished speaking during the Act part, $F(3.04, 91.4) = 4.93, p = .003$. Post-hoc tests revealed a significant difference between $T_{\text{Breath}}$ and $T_{\text{Gesture}} (p = .03)$ and between $T_{\text{Breath}}$ and $T_{\text{Full}} (p = .002)$. There were also two non-significant trends between $T_{\text{None}}$ and $T_{\text{Gesture}} (p = .10)$ and between $T_{\text{None}}$ and $T_{\text{Full}} (p = .10)$, all other $p > .44$ (see Figure 5).

We analyzed the four questions asked after each Act trial (see Section 4.4) for internal consistency. We concluded to analyze the questions for easiness and the inverted answers to the questions whether the ECA’s behavior was ambivalent or confusing together (Cronbach’s $\alpha = .81$). This is called Clarity from here on and is the
mean of the three aforementioned scales (ambiguous and confusing inverted). The question regarding frustration is evaluated separately since it would have reduced the Cronbach’s Alpha score to $\alpha = .79$ and is differently framed. A repeated-measures ANOVA revealed a significant effect of $T$ on Clarity, $F(4, 120) = 5.42, p < .001$. Post-hoc tests showed significant difference between $T_{\text{None}}$ and $T_{\text{Gesture}}$ ($p = .04$) and between $T_{\text{None}}$ and $T_{\text{Full}}$ ($p = .01$), all other $p > .18$ (see Figure 5). For the frustrating questions, no significant effect was found ($F < 1.92, p = .14$), with the means per turn-taking level all between $-2.66$ and $-2.36$. We also checked whether participants tried to speak in a Gap when they should not. In sum this happened 21 times during $T_{\text{None}}$, eight times during $T_{\text{Breath}}$, 13 times while in $T_{\text{Gesture}}$, two times in $T_{\text{Gaze}}$ and six times when all cues are shown in $T_{\text{Full}}$ (of 651 gaps in total). However, a Friedman test (which is the non-parametric equivalent to a repeated-measures ANOVA and had to be used since the assumption of normality was violated), did not show a significant effect of $T$ ($p = .20$). Explicit yield gestures (played after 3 s of silence) were in sum only triggered five times for different participants, so we did not analyze them further.

5.1 Limitations
While the inhalation sound was played at the identical volume as the speech, this modality could still be improved especially for showing the willingness to take over the turn, for example, by a sharp inhalation during another speaker’s turn (we only played inhalation sounds during the gaps). Furthermore, the gaps during the Listen part were static and rather short (all lasting 300 ms) which might have had a negative influence, since the additional modalities might especially play a role in prolonged gaps, e.g., due to thinking. A further aspect we noticed is that the environment gaze targets (cf. Figure 3) were not optimally placed often leading to “averted” gazes which are only slightly off from looking at the participant, which some commented on negatively. Since most of our participants came from the same cultural background (German), the presented results might only be applicable to this cultural group. Another observed behavior we did not consider is that of posture shifts, which, following Cassell et al. [6], appear more frequent at turn shifts.

6 CONCLUSION
In this paper, we presented an approach to generate turn-taking cues for ECAs based on non-verbal behaviors, more specifically: gesture holding, gaze manipulation, and breathing. We conducted a user study to evaluate their efficiency. When only listening to two ECAs jointly telling a family story, no difference in their perceived naturalness and social presence was found. However, when participants joined in taking turns with one ECA, we found an effect of gesture manipulation on the gaps left by participants and also on the perceived clarity of the turn-taking signaling of the ECA. This means that gesture holding seems to be a valuable turn-taking signal.

In future research, we plan to add listening agents to give additional hints to participants by having other bystanders react to the turn-taking cues given by the speakers (cf. [19]). Furthermore, we plan to improve the breathing modality by more variability and potential respiration during the previous speaker’s turn.

ACKNOWLEDGMENTS
This research was funded by the German Research Foundation (DFG) within the project ‘Listening to, and remembering conversations between two talkers: Cognitive research using embodied conversational agents in audiovisual virtual environments’, which is part of the DFG Priority Program “AUDICTIVE” (SPP 2236). The contribution by Sabine J. Schlittmeier was supported by a grant from the Head-Genuit-Foundation (P-16/10-W). Special thanks goes to Malte Kögel for his support in conducting the study.