4th International Symposium on New Frontiers in Human-Robot Interaction
21–22 April 2015, Canterbury, UK
M. Salem, A. Weiss, P. Baxter, K. Dautenhahn (Eds.)

Turn-yielding cues in robot-human conversation

Jef A. van Schendel and **Raymond H. Cuijpers**[1]

Abstract. If robots are to communicate with humans in a successful manner, they will need to be able to take and give turns during conversations. Effective and appropriate turn-taking and turn-yielding actions are crucial in doing so. The present study investigates the objective and subjective performance of four different turn-yielding cues performed by a NAO robot. The results show that an artificial cue, flashing eye-LEDs, lead to significantly shorter response times by the conversational partner than not giving any cue and was experienced as an improvement to the conversation. However, stopping arm movement or head turning cues showed, respectively, no significant difference or even longer response times compared to the baseline condition. Conclusions are that turn-yielding cues can lead to improved conversations, though it depends on the type of cue, and that copying human turn-yielding cues is not necessarily the best option for robots.

1 INTRODUCTION

"Beep boop!" Will our future robot partners communicate with us like Star Wars' R2D2? A more desirable future would be one where we can interact with robots in a fluent and pleasant manner, using the same natural language we use to talk to other people.

As robots grow more advanced, they are able to help us out in more areas of our lives. An area of interest is for instance elderly care, since healthcare costs in European countries are on the rise [6], and the 80+ population in Europe is expected to more than double from 2013 to 2050 [23]. Robots could increase cost-efficiency and have shown positive effects in this area [5].

But no matter what type of work, socially assistive robots as they are called [22], should be not just able to successfully perform their tasks, but deal with human beings in an appropriate, respectful and productive manner. This requires a way to naturally communicate with them, which involves taking and giving turns. This is also called managing the conversational floor.

1.1 Turn-taking

To manage the conversational floor, humans make use of turn-taking and turn-yielding cues [8]. One way to give such cues is through speech itself: the intention to yield a turn can be made clear through syntax (for instance, ending with a direct question) but also changes in intonation or speaking rate [10, 13]. Using these cues requires understanding what is being said, which is difficult for robots. Another way is through non-verbal cues, given through body movement or gaze direction [16]. The major advantage of non-verbal cues is that they do not require speech to be intelligible.

[1] Human Technology Interaction group, Eindhoven University of Technology, the Netherlands, email: r.h.cuijpers@tue.nl

Existing research has investigated ways for robots and other agents to shape and guide a conversation. Positive results have been found when robots have been used to implement conversational gaze behavior [2, 18, 21] and gestures [14, 17], likewise with agents who make use of eye gaze [1, 7, 19], especially when it is appropriate in context [9, 12]. Other researchers investigated both gestures and eye gazing by robots and, in certain combinations, found positive effects on message retention [24] and persuasion [11]. Others still moved on from dyadic sessions to conversations where a robot speaks with multiple people, so-called multiparty settings [3, 4, 15, 18, 25].

Since non-verbal cues have shown promising results in studies such as these, and can be implemented relatively easily for robots, they are of interest for the present study.

While turn-taking has been investigated in many studies, most of them evaluate a combination of turn-yielding cues as a whole and do not compare the effectiveness of isolated turn-yielding cues. Some authors, such as [4], have built interaction models for agents that include turn-yielding. In their study, the assessment of turn-yielding behavior is mixed with other types of interaction. Additionally, the subjective assessment is based on a single condition and is not compared to other models, which makes it difficult to understand the relative contribution of different turn-yielding cues. Therefore, we designed a study in which we can compare the effectiveness and user evaluation of a number of non-verbal turn-yielding cues. The response time of the conversation partner is used as an objective measure, because a shorter response time could mean better and more fluent conversational flow. Shiwa and colleagues [20] already showed that this does not necessarily signify a more pleasant interaction, which is why a questionnaire is used to evaluate the participants' opinion on the value of the different cues. This study will give us further insights in how to employ non-verbal turn-yielding and turn-taking cues during human-robot interaction.

1.2 Turn-yielding cues

Four different turn-yielding cues were selected, based on existing literature.

The first two were based on common human cues and labelled *turn head* and *stop arms*. The former means that the speaker directs its gaze away from the conversational partner during speaking, then returns to the partner when yielding the turn [16]. For the latter, the speaker uses co-speech gestures while talking, but stops doing so when finished. It is based on the idea that interlocutors make certain continuous movements during speaking, but stop moving as a sign that their turn is over [16].

For the third cue, an artificial action was chosen, namely *flash eyes*, where the robot briefly increases the brightness of its eye-LEDs. This condition was added to investigate whether cues have to be based on existing human behavior or not. This cue is not natural in the sense

that it is humanlike, but it is a very common way to communicate non-verbally for robots (and many other technical devices).

The last cue was called *stay silent* and served as the baseline condition. Here, the robot simply stopped speaking with no further action.

These four cues were performed by a robot in dyadic sessions with human conversational partners. In order to generate a large number of turn-yielding events we developed a new task where the participant and the robot took turns to verbally cite the letters of the alphabet. As soon as the robot stopped citing, the participant continued citing letters. After a few letters, the robot continued again. The turn-yielding cues employed by the robot were manipulated.

2 METHOD

2.1 Participants

A total of 20 participants took part in the experiment. One was unable to complete the task and therefore the data in question was not used in the analysis. Roughly half of the participants were recruited from the J.F. Schouten participant database, while the others were recruited through word-of-mouth and invitations via social networks. The only requirement set beforehand was that the participants were able of hearing. Of the 19 participants, 13 were female. All participants were offered monetary compensation or course credits for their time.

2.2 Design

The performed experiment had a within-subjects, repeated measures design with four conditions.

The independent variable in this study was the turn-yielding cue used by the robot. The four conditions, as described under 1.2, were labelled *stay silent*, *stop arms*, *turn head* and *flash eyes*. These were randomly selected by the robot during the experiment.

The dependent variable was the response time of the participant. Specifically, this time was defined as the length in milliseconds between the start of the robot's turn-yielding cue and the beginning of the participant's speech.

Additionally, the participants filled out a questionnaire after the experiment. The questionnaire began by asking the participants which of the four cues they remember noticing. Then, a number of questions asked about their opinion on the four conditions, using a five-point Likert scale. The order of the questions was randomized for each participant in order to minimize ordering bias.

2.3 Setup

This study used a 58-centimeter tall humanoid robot called NAO, developed by Aldebaran Robotics. It has 25 degrees of freedom for movement and various sensors. Of particular interest for this study was its microphone, however, due to unsatisfactory performance during pre-tests, an external microphone was used for the experiment. Both the NAO and the microphone were connected to a laptop, used for controlling the experiment and saving the data.

The experiment took place in the GameXPLab, a laboratory modelled after a living room at Eindhoven University of Technology. Participants were seated in front of a small desk, with the NAO on top of the desk and a small wireless microphone placed between them.

2.4 Procedure

During a short introduction, the participants were given their task: together with the NAO, they were to repeatedly cite the letters of the alphabet. The NAO would start and after a randomly chosen amount of letters it would stop speaking and perform one of the turn-yielding cues. Then, the participant would continue until the NAO started speaking again. The robot autonomously decided when to speak by listening for 2, 3 or 4 utterances after which it waited for a silence to start speaking. The number of utterances determines which letter should be used next. Occasionally, the robot made a mistake (e.g. when it mistook another sound for an letter) or interrupted a person, but this was never a problem from the user's point of view. A small timing delay (0.5s) was added to make the flow as natural as possible. This cycle continued for roughly 15 minutes with each participant.

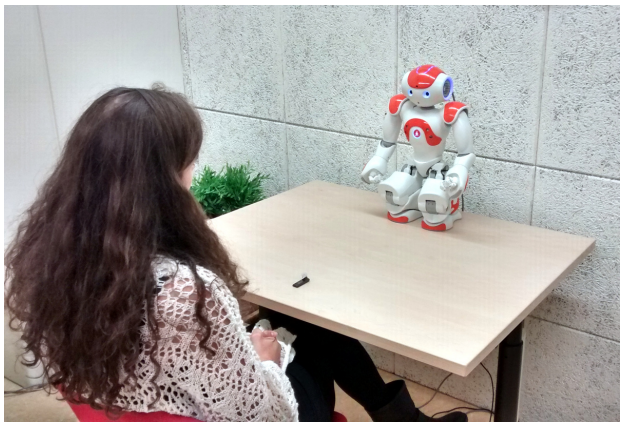

Figure 1. Experiment setup

This particular task was chosen for several reasons. First, the answers by the participants would mostly be single-syllable words, which would make them easier to accurately detect with the microphone and enable the robot to count them, so it would know where to continue the series. The second reason was the assumption that the participants would be able to recall the letters of the alphabet with minimal effort, thereby minimizing the influence of recollection time. Thirdly, the advantage of using a fixed sequence would be to avoid the need for the participant to decide on what to say. In other words, the aim was to control for possibly confounding variables such as recollection time or deliberation time.

Afterwards, the participants filled out a questionnaire (further described under 2.2).

3 RESULTS

3.1 Experiment results

The experiment data was edited and analyzed using SPSS. A number of false positives were recorded as notes during the experiment. After these were removed, a total of 1310 valid data points were left, or about 68.9 recorded measurements per participant.

The distribution of the response time data was found to be skewed right ($\text{skewness} = 1.520 \pm 0.068$) and peaked ($\text{kurtosis} = 5.370 \pm 0.135$). To increase normality it was logarithmically transformed. Histograms of the original (a) and log-transformed (b) data can be found in Figure 3.1. As can be seen, the normality was much improved: the distribution of the transformed data is approximately symmetric ($\text{skewness} = -0.079 \pm 0.068$) and less peaked

(kurtosis $= 0.421 \pm 0.135$).

Table 1 shows the reaction times of the four conditions. Since the distribution of reaction times is skewed we transformed the data using the natural logarithm (ln) before computing the means and standard errors (middle two columns). The last two columns show the reaction times transformed back to the normal time domain.

A one-way ANOVA showed that there was a significant difference between groups ($F(3, 1306) = 15.407$, $p < 0.001$). Levene's test indicated equal variances ($p = 0.644$).

A Tukey HSD post-hoc test revealed that the response time was significantly lower for the *flash eyes* action ($M = 854$ ms, $p = 0.006$) yet significantly higher for the *turn head* action ($M = 1033$ ms, $p = 0.003$) when compared to the *stay silent* condition (M = 944 ms). There was no significant difference between the *stay silent* condition and the *stop arms* action ($M = 916$ ms, $p = .829$).

Additionally, the mean response time for the *turn head* condition was significantly higher than both the *stop arms* ($p < 0.001$) and *flash eyes* ($p < 0.001$) conditions. There was, however, no significant difference between the *flash eyes* and *stop arms* conditions ($p = 0.071$). Post-hoc results are shown in Table 2. A bar chart visualising the means of the four conditions can be found in Figure 4.

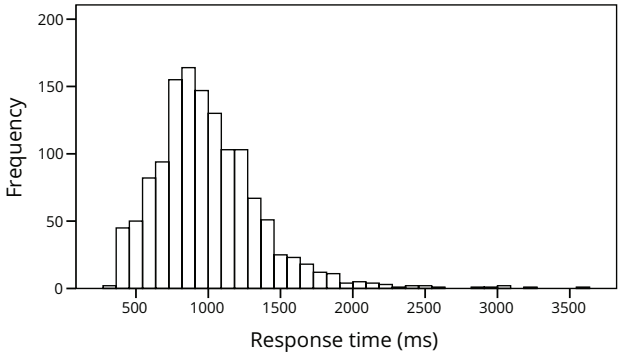

Figure 2. Original data

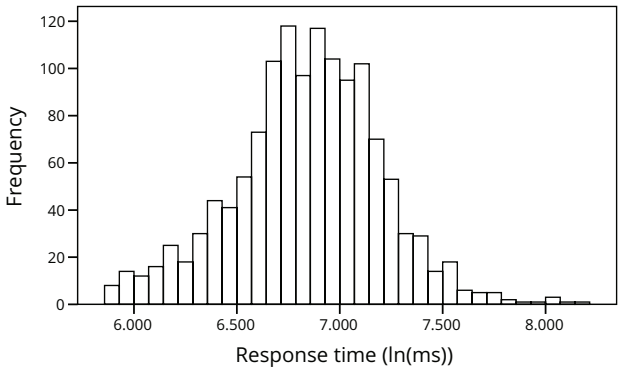

Figure 3. Histograms showing the distribution of response times

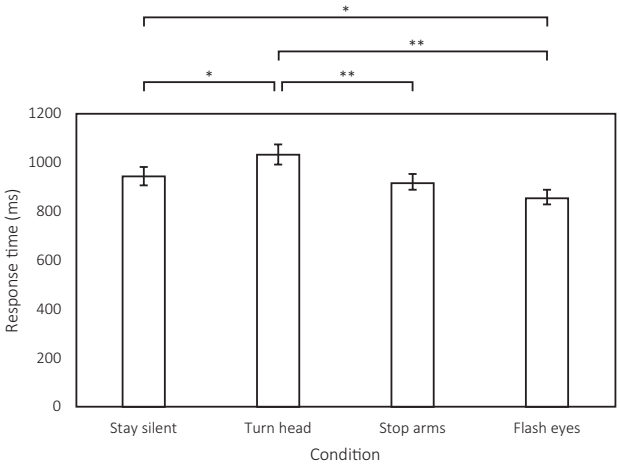

Figure 4. Means of the four conditions. Error bars represent 95% CI. Bars denoted with * differ at significance level < 0.01, bars with ** at significance level < 0.001.

Linear regression on the response times with trial number as the independent variable showed that these times did not decrease after sequential trials (*stay silent* $p = 0.759$; *turn head* $p = 0.224$; *flash eyes* $p = 0.368$), except for the *stop arms* condition ($p = 0.001$). For this last condition, response times decreased by 207 ms after 115 trials, as shown in Figure 5.

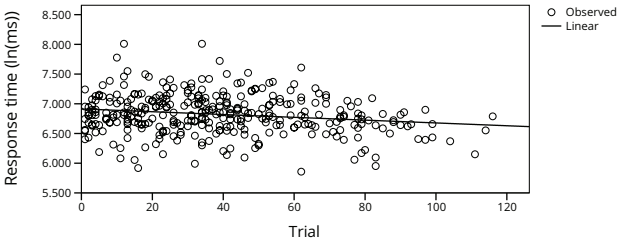

Figure 5. Scatter plot and fitted line of all response times in the *stop arms* condition.

3.2 Questionnaire results

The data gathered with the questionnaire (N = 19) was edited and analyzed using SPSS, in several steps.

The first part of the questionnaire was used as a confirmation of which cues were noticed by the participants. Cues that went unnoticed were excluded from the data.

Furthermore, the questionnaire included pairs of opposite questions, phrased positively and negatively, to avoid acquiescent bias. An example of such a pair is "...improved the flow of the conversation" and "...did not improve the conversation". Before analysis, negatively phrased questions had their answers mirrored.

Principle component analysis was used to identify the underlying factors and group the variables. After applying varimax rotation,

Table 1. Reaction times of the four conditions in the log-transformed and normal domain. SE is the standard error of sample mean. N is the number of turn yields (1310 in total).

Condition	N	Mean (ln(ms))	SE (ln(ms))	Mean (ms)	SE (ms)
Stay silent	331	6.85	.020	944	±19
Turn head	337	6.94	.019	1033	20/-19
Stop arms	334	6.82	.018	916	17/-16
Flash eyes	308	6.75	.020	854	±17

Table 2. Post-hoc test results of the response times

(I) condition	(J) condition	Mean difference (I-J, ln(ms))	SE (ln(ms))	Sig.
Stay silent	Turn head	-0.95	.027	.003
	Stop arms	.023	.027	.829
	Flash eyes	.091	.028	.006
Turn head	Stay silent	.095	.027	.003
	Stop arms	.118	.027	.000
	Flash eyes	.186	.028	.000
Stop arms	Stay silent	-.023	.027	.829
	Turn head	-.118	.027	.000
	Flash eyes	.068	.028	.071
Flash eyes	Stay silent	-.091	.028	.006
	Turn head	-.186	.028	.000
	Stop arms	-.068	.028	.071

three components were found with an Eigenvalue over 1, accounting for 35.1, 28.2 and 13.2 percent, respectively, of the total variance.

The rotated component matrix, shown in Table 3, shows which questions load on which components after rotation. Based on this data, the three components were named *Pleasant*, *Improvement* and *Noticeable*. Table 4 shows which questions make up which components.

After identifying the components, a one-way ANOVA on the combined questions showed that there was a significant difference between groups for the *Improvement* ($F(3, 292) = 8.998$, $p < 0.001$) and *Noticeable* ($F(3, 70) = 3.081$, $p = 0.033$) components, but not for the *Pleasant* component ($F(3, 218) = 0.602$, $p = 0.614$).

A Tukey HSD post-hoc test performed on the *Improvement* and *Noticeable* components showed that there were several significant differences between the means of the questionnaire responses. *Flash eyes* scored significantly higher on *Improvement* than both *stop arms* ($p < 0.001$) and *stay silent* ($p = 0.001$). Also, *stop arms* scored higher than *stay silent* on *Noticeable* ($p = 0.040$).

The post-hoc test results for the *Improvement* and *Noticeable* components can be found in Table 5 and 6, respectively. A graphical summary of all the components can be found in Figure 6.

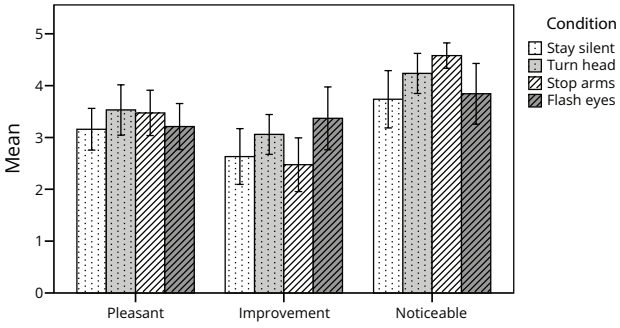

Figure 6. Means of the four conditions for every component. Error bars represent 95% CI.

4 DISCUSSION

The present study investigated different turn-yielding cues to be used by a robot in robot-human conversation. An experiment and questionnaire measured the performance and rating of the different cues. The results show that using a turn-yielding cue can lead to faster response times by the conversational partner compared to the baseline condition. One of the cues, namely *flash eyes*, produced the lowest response times and was rated higher on Improvement than the baseline condition and any other cue. The results, therefore, partially confirm the hypothesis that turn-yielding cues by a robot can improve robot-human conversation.

4.1 Different types of cues

The *flash eyes* cue lead to faster response times and had the highest *Improvement* rating by the participants. However, other cues showed different results. The *turn head* cue showed significantly longer response times compared to *staying silent*. Moreover, while the *stop arms* condition was rated as more noticeable than *staying silent*, there was no significant difference between the mean response times of these two cues.

There was a difference of 179 ms between the means of the response times for the *flash eyes* and *turn head* cues. A conclusion could be that while turn-yielding cues have the potential to lead to decreased response times, the type of cue matters a great deal.

Table 3. Rotated component matrix. Questions marked with * were mirrored.

Question	Component 1	Component 2	Component 3
...made it obvious it was my turn	.913	.094	.035
...had no clear meaning*	.868	.023	-.110
...did not improve the conversation*	.723	.459	.067
...improved the flow of the conversation	.703	.465	-.143
...was uncomfortable*	.074	.871	-.101
...was friendly	.142	.863	.155
...felt natural	.415	.560	-.096
...was hard to notice*	-.060	-.007	.986

Table 4. Components and related questions. Questions marked with * were mirrored.

Component 1, *Pleasant*	Component 2, *Improvement*	Component 3, *Noticeable*
...was uncomfortable*	...made it obvious it was my turn	...was hard to notice*
...was friendly	...had no clear meaning*	
...felt natural	...did not improve the conversation*	
	...improved the flow of the conversation	

4.2 Artificial cue

While a decrease in response time can be a hint that the cue improves the conversation, this does not necessarily have to be the case. Results from the questionnaire, however, were in line with the results from the experiment when it came to the *flash eyes* cue. It was seen as an improvement to the conversation and to have a clearer meaning when compared to the *stop arms* and *stay silent* cues.

Some anecdotal evidence from the experiment pointed the same way. Several participants remarked that they appreciated the *flash eyes* cue, one of them explaining "It signals that he is done, and that he won't interrupt me". Multiple participants also described the cue as "natural", which is interesting for an artificial cue that human conversational partners are unable to perform.

Thus, one of the interesting things here is that the cue with the lowest response time was an artificial cue, as opposed to the *turn head* and *stop arms* cues, which were based on literature from human-human interaction. There appears to be a difference between a human being using such cues and the NAO doing the same. This could have several causes. One possible cause is that the NAO did not perform the cue correctly, and therefore its meaning was unclear to the participants. Results from the questionnaire are inconclusive on this point: these cues were not rated significantly lower on this point, and their means center around "Neither agree nor disagree". Another reason could be that the participants found the cues with movement to be unexpected and therefore hesitated in their responses.

4.3 Movement cues

The cues that were based on movement, namely *turn head* and *stop arms*, showed worse performance compared to *flash eyes*, which did not involve movement. The movements made by the robot could be a source of distraction or hesitation for the participants, which could explain the longer response times.

Some anecdotal evidence from the experiment pointed this way. Some of the participants talked about the *turn head* and *stop arms* cues, explaining that they found many of the robot's movements to be distracting, and were sometimes confused as to the meaning of these movements. The data from the questionnaire shows that the *stop arms* cue was rated as significantly higher on the *Noticeable* component. Could it have been too noticeable, thereby distracting the participant?

Additionally, during the experiment it often seemed that when the NAO started moving, the participant hesitated to continue, preferring to wait to see where the robot was going with this. One of them remarked that he did not recognize the movement of *turn head* as a cue to start speaking, so instead he "just waited until it was done".

The movements could have simply been unexpected. Linear regression showed that for at least the *stop arms* cue, the mean response time decreased after subsequent trials, suggesting the participants were faster to respond and perhaps got used to the cue. Perhaps after longer interaction with the robot, this cue could have lead to response times similar to *flash eyes*.

Whether these findings are specific to the NAO robot is unclear, but fact is that this particular robot makes distinct sounds during movements and that it remains completely static outside of the performed cues. This could make movement cues highly salient by default.

4.4 Improvements to the experiment

A critical component of the experiment was accurately measuring the response time. The external microphone made it possible to relatively accurately and precisely measure the points at which the participant started speaking. However the beginning of the measurement, defined as the point at which the NAO stopped speaking, was harder to measure accurately. In the experiment, the timer started running after the NAO signalled it was done. However further investigation revealed that there is in fact a pause between the actual end of the sound and this signal, of around 225 ms on average. Though this issue could unfortunately not be avoided during this experiment, it could have an impact on the results. In practice it means that the turn-yielding cue could be performed sooner after speaking, possibly leading to a larger decrease in response times and an even stronger effect. Indeed, if we subtract 225ms form the reaction times for all non-verbal cues except the *stay silent* cue in Figure 4, we obtain a graph where all non-verbal cues lead to a reaction time improvement

Table 5. Post-hoc test results for the *Improvement* component

(I) condition	(J) condition	Mean difference (I-J)	SE	Sig.
Flash eyes	Turn head	.449	.193	.096
	Stop arms	.921	.188	.000
	Stay silent	.724	.188	.001
Turn head	Flash eyes	-.449	.193	.096
	Stop arms	.472	.193	.072
	Stay silent	.275	.193	.488
Stop arms	Flash eyes	-.921	.188	.000
	Turn head	-.472	.193	.072
	Stay silent	-.197	.188	.720
Stay silent	Flash eyes	-.724	.188	.001
	Turn head	-.275	.193	.488
	Stop arms	.197	.188	.720

Table 6. Post-hoc test results for the *Noticeable* component

(I) condition	(J) condition	Mean difference (I-J)	SE	Sig.
Flash eyes	Turn head	-.393	.319	.608
	Stop arms	-.737	.310	.091
(p ¡ .001)	Stay silent	.105	.310	.986
Turn head	Flash eyes	.393	.319	.608
	Stop arms	-.344	.319	.704
	Stay silent	.498	.319	.406
Stop arms	Flash eyes	.737	.310	.091
	Turn head	.344	.319	.704
	Stay silent	.842	.310	.040
Stay silent	Flash eyes	-.105	.310	.986
	Turn head	-.498	.319	.406
	Stop arms	-.842	.310	.040

compared to the *stay silent* cue. However, the *flash eyes* cue would still be most salient and the relative effectiveness of these cues remains the same.

5 CONCLUSIONS

The present study explored the use of turn-yielding cues by a robot. We found that such turn-yielding cues can improve both performance and user experience during human-robot conversation. These results on turn-yielding are in line with earlier findings that show that non-verbal cues can influence turn taking in conversations [2, 18]. Our study adds to earlier research by specifically focusing on the relative effect of turn-yielding cues and it shows that the type of cue is of importance for both performance and user experience.

An important question is how these conclusions are to be used in the development of socially assistive robots. Should one, for instance, always make use of an eye-flashing cue? It is clear that turn-yielding cues have the potential to improve a conversation, but in our study at most one cue was presented at a time (in addition to the stay silent cue). While the eye-flashing cue showed the most promise during this experiment, its meaning is, in general, ambiguous. Flashing LEDs are used to signal all sorts of events. In that sense the *turn head* and *stop arms* cues are much better, because they not only inform the observer about the timing of an event but also that the event is a turn-yield. So we expect that these cues are more useful in complex interactions. Finally, it would be interesting to see how these cues interact. A head turn could disambiguate a LED flash, so that in combination the turn-yield cues are effective and robust.

REFERENCES

[1] Sean Andrist, Tomislav Pejsa, Bilge Mutlu, and Michael Gleicher, 'Designing effective gaze mechanisms for virtual agents', in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 705–714, (2012).

[2] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu, 'Conversational gaze aversion for humanlike robots', in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pp. 25–32, (2014).

[3] Maren Bennewitz, Felix Faber, Dominik Joho, Michael Schreiber, and Sven Behnke, 'Integrating vision and speech for conversations with multiple persons', in *International Conference on Intelligent Robots and Systems 2005*, pp. 2523–2528, (2005).

[4] Dan Bohus and Eric Horvitz, 'Multiparty turn taking in situated dialog: Study, lessons, and directions', in *Proceedings of the SIGDIAL 2011 Conference*, pp. 98–109, (2011).

[5] Joost Broekens, Marcel Heerink, and Henk Rosendal, 'Assistive social robots in elderly care: a review', *Gerontechnology*, **8**(2), 94–103, (2009).

[6] Sarah Chaytor and Uta Staiger, 'The future of healthcare in europe', *UCL European Institute*, (2011).

[7] Alex Colburn, Michael F. Cohen, and Steven Drucker, 'The role of eye

gaze in avatar mediated conversational interfaces', *Sketches and Applications, Siggraph'00*, (2000).
[8] Starkey Duncan, 'Some signals and rules for taking speaking turns in conversations.', *Journal of personality and social psychology*, **23**(2), 283, (1972).
[9] Maia Garau, Mel Slater, Simon Bee, and Martina Angela Sasse, 'The impact of eye gaze on communication using humanoid avatars', in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 309–316, (2001).
[10] Agustín Gravano and Julia Hirschberg, 'Turn-yielding cues in task-oriented dialogue', in *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 253–261, (2009).
[11] Jaap Ham, René Bokhorst, Raymond Cuijpers, David van der Pol, and John-John Cabibihan, 'Making robots persuasive: the influence of combining persuasive strategies (gazing and gestures) by a storytelling robot on its persuasive power', in *Social Robotics*, 71–83, Springer, (2011).
[12] D. K. J. Heylen, Ivo Van Es, Anton Nijholt, and E. M. A. G. van Dijk, 'Experimenting with the gaze of a conversational agent', (2002).
[13] Anna Hjalmarsson, 'The additive effect of turn-taking cues in human and synthetic voice', *Speech Communication*, **53**(1), 23–35, (2011).
[14] Chien-Ming Huang and Bilge Mutlu, 'Modeling and evaluating narrative gestures for humanlike robots.', in *Robotics: Science and Systems*, (2013).
[15] Martin Johansson, Gabriel Skantze, and Joakim Gustafson, 'Head pose patterns in multiparty human-robot team-building interactions', in *Social Robotics*, 351–360, Springer, (2013).
[16] Adam Kendon, 'Some functions of gaze-direction in social interaction', *Acta psychologica*, **26**, 22–63, (1967).
[17] Chaoran Liu, Carlos Toshinori Ishi, Hiroshi Ishiguro, and Norihiro Hagita, 'Generation of nodding, head tilting and eye gazing for human-robot dialogue interaction', in *Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on*, pp. 285–292, (2012).
[18] Bilge Mutlu, Toshiyuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita, 'Footing in human-robot conversations: how robots might shape participant roles using gaze cues', in *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pp. 61–68, (2009).
[19] David G. Novick, Brian Hansen, and Karen Ward, 'Coordinating turn-taking with gaze', in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pp. 1888–1891, (1996).
[20] T. Shiwa, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita, 'How quickly should communication robots respond?', in *Human-Robot Interaction (HRI), 2008 3rd ACM/IEEE International Conference on*, pp. 153–160, (March 2008).
[21] Candace L. Sidner, Cory D. Kidd, Christopher Lee, and Neal Lesh, 'Where to look: a study of human-robot engagement', in *Proceedings of the 9th international conference on Intelligent user interfaces*, pp. 78–84, (2004).
[22] Adriana Tapus, Mataric Maja, and Brian Scassellatti, 'The grand challenges in socially assistive robotics', *IEEE Robotics and Automation Magazine*, **14**(1), (2007).
[23] UN, 'World population prospects, the 2012 revision', *New York: Department for Economic and Social Affairs*, (2013).
[24] Elisabeth T. Van Dijk, Elena Torta, and Raymond H. Cuijpers, 'Effects of eye contact and iconic gestures on message retention in human-robot interaction', *International Journal of Social Robotics*, **5**(4), 491–501, (2013).
[25] Roel Vertegaal, Robert Slagter, Gerrit Van der Veer, and Anton Nijholt, 'Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes', in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 301–308, (2001).