Turn-taking cue delays in human-robot communication

R.H. Cuijpers and V.J.P. van den Goor

Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
r.h.cuijpers@tue.nl

Abstract. Fluent communication between a human and a robot relies on the use of effective turn-taking cues. In human speech staying silent after a sequence of utterances is usually accompanied by an explicit turn-yielding cue to signal the end of a turn. Here we study the effect of the timing of four turn-yielding cues relative to staying silent. In addition, we examine the relative strength of showing turn-yielding cues individually or in combination. We found that strong, meaningful cues perform considerably better but only if they precede the stay silent cue. The response times to synchronized cues match the stay silent control condition and delayed cues have longer response times. Cue combinations show a winner-take-all effect, as the strongest cue determines the performance of the set. In conclusion, effective human-robot conversation is obtained when the turn-taking cues are clear, meaningful and precede the stay silent cue.

1 Introduction

In order to engage in dialog with robots in social settings, it is vital that a robot can understand when we want its attention. Equally important for the robot to be an effective conversational partner is that we can understand when it requests our input. This is especially important in time-critical situations like emergencies. In human conversation, we use effective turn-taking cues to facilitate these dialogs [1]. Gaze and proximity are cues we commonly use for this purpose [2,3]. Previous research has shown that this approach can be used for human-robot communication as well. Examples of effective cues are gaze [4,5,6,7] and gestures [8] or a combination of both [9]. The salience of the information flow also affects the effectiveness of robot turn-taking [10], as unclear utterances result in long response times. Some promising experiments have been performed with multi-party dialogs in which an embodied conversational agent leads the conversation [11,12]. Other studies combine many cues into a multi-modal system [13] or use several perceptual cues to let multiple robots work as a team [14,15]. Van Schendel and Cuijpers [16] found that different isolated cues can influence response times in an alphabet citing task. However, most studies did not investigate the effect of timing in turn-taking cues or the relative strength. Moreover, it is unclear whether combining multiple cues can further improve performance.

1.1. Present study

The present study concentrated on finding effects of the timing of a turn-taking cue relative to the stay silent cue. We define the stay silent cue as the moment after speech has ended at which, without help of other cues, the mere duration of silence indicates a turn-yield.

In this research, the stay silent cue is fixed at a constant value by letting the robot wait for 0.6 seconds after each spoken utterance. This way, a rhythmic conversation emerged in which the human conversational partner expects each next letter to come at the same temporal offset. This is a vital part of the experiment, as it allows us to manipulate the timing of additional cues to occur in sync (delay=0.6s), before (0s) or after (1.2s) with the stay silent cue. We used the subject's response time as an objective measure of performance. We expect the zero delay condition to perform best for strong, meaningful cues because it provides salient turn-yield information prior to the stay silent cue. For weaker cues, we expect the stay silent cue to dominate our perception. We measured perceived strength and salience of the cues using a questionnaire. The long delay condition is expected to perform equal to the control condition (staying silent), as we expect the stay silent cue to dominate and provide sufficient information to perform the task. We were also interested in the effect of combined cues, by presenting two cues simultaneously. We expected that cue combinations perform (at least) similar to the strongest cue of the set, because cue integration has been shown to be statistically optimal in other perception domains (e.g. [17]).

The cues that are used in this study are the stay silent cue, the head turn cue, the stop arm movement cue, and the eye flash cue. When staying silent the robot performed no action and simply awaited human speech. In the head turn cue condition, the robot slowly turned away its head while it was citing letters, but, when yielding the turn, it quickly turned the head back to look at the participant. This is a natural cue that people use in everyday life, and therefore we expected it to perform well (i.e. reduce response times). In the stop arm movement cue, the robot moved its arms similar to the way humans do while walking. The arms stopped moving when the robot had finished its turn. This is usually not associated with turn-yielding, thus we expected it to perform worse than the head turn cue which is most natural. In the fourth eye flash cue condition, the robot flashed its eye Light Emitting Diodes (LEDs) with a bright white light to indicate its turn is over. This is an artificial cue that does not occur in nature. Additionally, two combination cues were used. Since Van Schendel and Cuijpers [16] found that the eye flash cue was most salient, it was decided to combine this cue with the head turn cue and the stop arm movement cue.

2 Methods

2.1 Participants and task

A total of 26 participants took part in the study of which 16 were female. The J.F. Schouten participant database was used to recruit 18 participants and the

others were recruited through word-of-mouth. Most of the participants (21) were between 19 and 27 years of age, the remaining 5 ranged from 45 to 78. Selection requirements were normal vision, normal hearing and knowledge of the alphabet in either the Dutch or English language. Each participant received monetary compensation for their time. The task of the participants was to cite the alphabet together with the robot. Both robot and participant took turns in citing several letters.

2.2 Experimental design

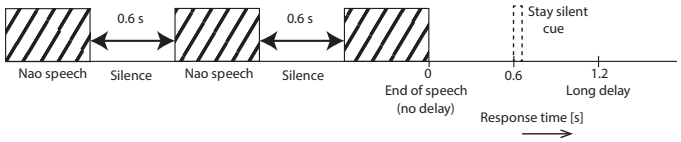

Fig. 1. Schematic overview of the time course of utterances defining the implicit "stay silent" cue.

The experiment had a 3 delays x 6 turn-yield cues within-subjects repeated measures design. The first independent variable was the time delay of the turn-taking cue relative to the stay silent cue. As shown in Fig. 1 the robot's utterances are interleaved with a silence duration of 0.6s. This value was found to result in a natural, pleasant speed of conversation. The first absence of an utterance demarcated the end of the robot's speech and, thus, a turn-yield event. The start time of the first absence of an utterance is used to define the 'stay silent' cue. The time delay of the other cues after the robot has finished the last utterance was varied between 0.0s (no delay, but 0.6s before the the stay silent cue); 0.6s (in sync with the stay silent cue) or 1.2s (0,6s delayed relative to the stay silent cue). The second independent variable was the turn-yielding cue used by the robot. This varied between stay silent (control condition), head turn, stop arm movement, eye flash, and the combinations stop arm movement + eye flash and head turn + eye flash. The dependent variable was the response time of the user, defined as the difference between the moment the Nao's speech had finished and the start of the first recognized utterance of the user.

2.3 Setup

The robot that was used for this research is the Nao robot (Softbank (Aldebaran) Robotics, Fr). To measure reaction times we used the 'SpeechDetected' event of the naoqi 1.14.5 Software Development Kit (Softbank Robotics, Fr) to detect when the participant started talking. The eye flash cue was implemented as a sudden change from low brightness (RGB color value [50,50,50]) to bright white (RGB color value [255,255,255]). The arm movements was a custom pre-recorded

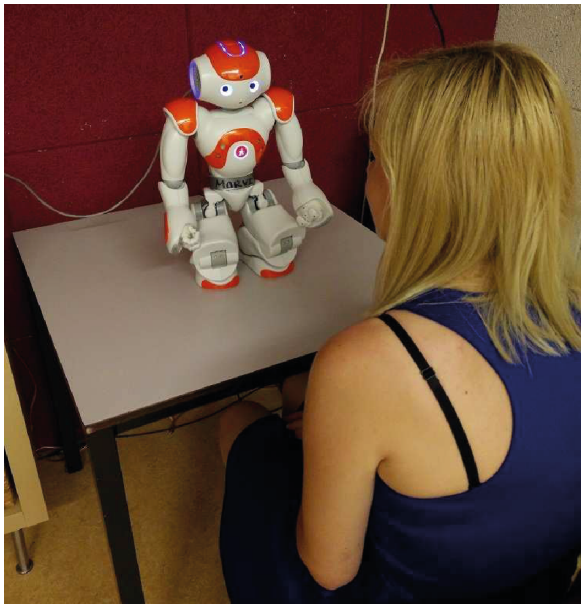

Fig. 2. Experimental setup

beat gesture with the right arm. During a turn-yield the ongoing movements were frozen instantly. The experiment was carried out at the GameXP lab at Eindhoven University of Technology, a laboratory that mimics a study room in a typical home. Participants took place in front of a table on which the Nao was positioned (see Fig. 2).

2.4 Procedure

After completing the informed consent form, participants were given a brief introduction about the experiment. At the start of the experiment, the Nao robot would speak about 5 letters of the alphabet after which it either initiated a turn-yielding cue or just stayed silent. The participant then continued citing the next letters of the alphabet. After a random number of letters (between 2 and 4), the robot would take over again and this process continued until the end of the experiment. Each time the end of the alphabet was reached, the participants/robot started citing from the beginning. In total 90 turn-yields were carried out by the robot. The experiment took about 15 minutes. Afterwards, the participants filled out a questionnaire indicating which turn yielding cues they had noticed and to which extent they believed it had influenced the conversation. The robot's speech recognition was not sophisticated enough to identify which letter was uttered. Therefore, the number of utterances were counted to help the robot keep track of the current position in the alphabet. In some cases, the robot 'missed' a letter. This happened when letters were spoken in such a rapid way

that the robot could not distinguish between the utterances. Participants were instructed to ignore that the robot could occasionally miss a letter. In order to avoid a recollection bias, each participant cited the alphabet in their primary language.

2.5 Data analysis

Before analyzing the data some outliers and data errors were removed (81 out of 2340; 2259 data points, or 3.5%). Since the reaction time data were right-tailed (skewness = 1.059 ± 0.052) and leptokurtic (kurtosis = 1.295 ± 0.103), a natural logarithm transformation was performed to reduce the deviation from normality. The resulting the data was much less right-tailed (skewness = 0.324 ± 0.052) and considerably less peaked (kurtosis 0.157 ± 0.103). All statistical analyses are done using the log-transformed data. All reported estimates and confidence intervals are transformed back to the time domain.

3 Results

3.1 Experiment results

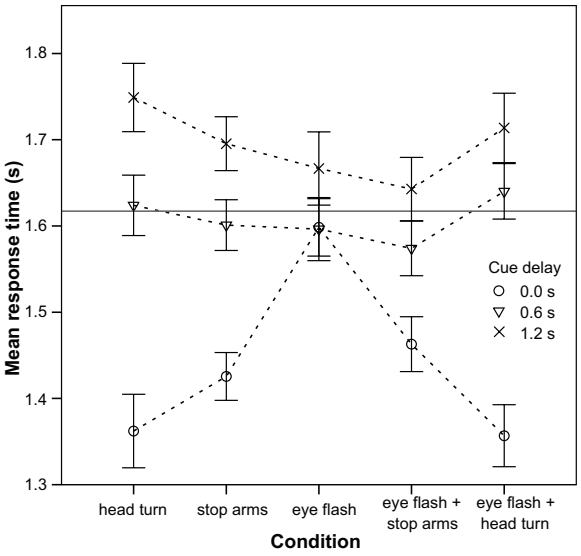

Fig. 3. Response times and error bars for each single (left) and combined (right) cue and delay condition. The horizontal line indicates the mean of the control condition (staying silent).

A univariate repeated measures ANOVA revealed significant main effects for both cue condition ($F(5, 138.782) = 3.281, p = 0.008$) and delay condition

($F(2, 53.184) = 83.554, p < 0.001$), as well as a significant interaction effect ($F(10, 307.959) = 6.690, p < 0.001$). Figure 3 shows the mean reaction times per cue and delay condition. It is clear that the 0.0s delay condition, where the turn-taking cue leads the stay silent cue by 0.6s, has a big impact on response time. The response time is much shorter when the head turn cue and stop arm movement cue are present than when the eye flash cue is present and in the stay silent condition. Surprisingly, performance is worse in the 1.2s delay condition compared to the 0.6s delay condition (simultaneous with the stay silent cue; $p = 0.001$). A One-way ANOVA was performed to verify that response times for the stay silent condition did not differ with changing delays ($F(2, 373) = 0.292, p = 0.747$). Consequently, we used its overall mean value (M=1.618 s $\pm$ 18.8 ms) as a baseline (dashed line in Fig. 3). Relative to baseline, the eye flash condition showed no significant difference ($F(2, 359) = 1.071, p = 0.344$). Apparently, this cue did not perform particularly well as the cue timing did not affect participant's performance. All of the other cue conditions differed significantly across delay conditions ($p < 0.001$).

For each of the three delay condition, an ANOVA was run to determine if there were differences between the cues. No differences were found for the 0.6s delay ($F(5, 736) = 0.564, p = 0.727$), nor for the 1.2s delay condition: ($F(5, 766) = 1.368, p = 0.234$). Indeed, post-hoc tests revealed no difference between any of the items in either delay condition. However, the 0.0s delay condition showed a considerable difference: $F(5, 739) = 14.298, p < 0.001$.

We were also interested in whether combining multiple cues provides additional benefits. More specifically, we tested whether performance in the cue combination conditions is equal to the combination's weak cue (first contrast) or to the strong cue (second contrast). In the first contrast performance of the eye flash cue is compared to the combined cue sets. Only for the 0.0s delay condition a significant effect is found ($p < 0.001$). This means that the performance for the combination cue conditions is better than just the eye flash cue, but only when the combined cues precede the stay silent cue. The second contrast compares the isolated head turn and stop arm movement cues with their combination counterparts (i.e. the corresponding cue including eye flash). None of the delay conditions show a significant difference, suggesting that the combinations indeed perform similarly to the best performing cue in the set.

3.2 Questionnaire results

The data gathered by the questionnaire was processed in SPSS (IBM corporation, USA). Some participants did not notice one or more of the cues. The data for the missed cues were excluded from analysis. Two pairs of polar opposite items were included in the questionnaire (i.e. improved the flow of the conversation and did not improve the conversation). These, along with other negatively framed questions, were reversed prior to analysis.

Data reduction was performed using principal component analysis. Since we do not require orthogonal factors, oblique rotation (direct-oblimin) was chosen. After reviewing the slopes in the scree plot, the number of factors was fixed

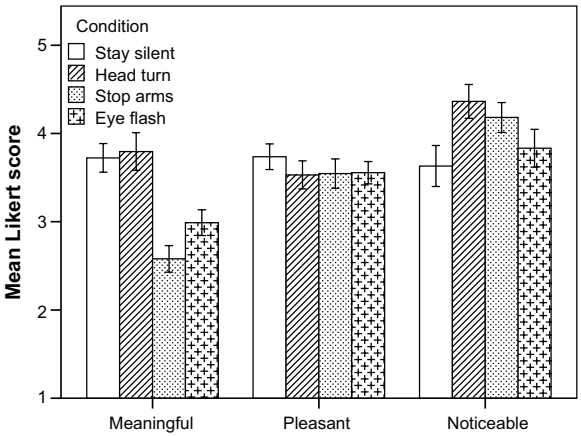

Fig. 4. Mean values for the extracted components in each cue condition. Error bars indicate 95% confidence interval.

Table 1. Rotated Component Matrix. Asterisk (*) indicates the item was reversed.

Questionnaire item	Comp. 1	Comp. 2	Comp. 3
Was friendly	-0.065	0.898	-0.105
Felt natural	0.252	0.702	0.143
Improved flow of the conversation	0.689	0.482	-0.014
Made it obvious it was my turn	0.851	0.013	-0.035
Was uncomfortable*	0.077	0.537	0.264
Had no clear meaning*	0.869	-0.07	0.043
Was hard to notice*	0.045	0.119	0.961
Did not improve the conversation*	0.753	0.348	0.258

Table 2. Overview of the components and the corresponding items. Asterisk (*) indicates the item was reversed.

Meaningful	Pleasant	Noticeable
Improved flow of the conversation	Was friendly	Was hard to notice*
Made it obvious it was my turn	Felt natural	
Had no clear meaning*	Was uncomfortable*	
Did not improve the conversation*		

at three (minimum initial eigenvalue of 0.974). The three components explain a total variance of 39.4%, 19.1% and 12.2%, respectively. Table 1 shows the resulting rotated component matrix. From the component matrix we determined the factors Meaningful, Pleasant and Noticeable (see Table 2). The factor scores for each cue condition are shown in Fig. 4. The stop arm movement and eye flash cues scored lowest on Meaningful, indicating they were not very informative and did not improve the conversation as much as the other cues. The three turn-taking cues scored equally on Pleasant, while the stay silent cue scored slightly higher. Noticeability was highest for the physical cues (stop arm movement and head turn), while eye flash and stay silent were less easily noticed.

4 Discussion and Conclusions

We measured the response time of subjects in a turn-based alphabet citing experiment. The turn-taking cue that the robot performed was manipulated in both type and delay.

4.1 Cue timing

We expected that the timing of turn-taking cues could improve performance (i.e. response times) when the cue was performed immediately after the robot had finished speaking. This effect was indeed quite apparent (Fig. 3), especially for the head turn and stop arm movement conditions, both separately and in combination with the eye flash cue. However, the isolated eye flash cue performed much worse than expected. In fact, the cue's performance did not differ from the 'stay silent' control condition. These observations suggest that not only timing, but also the type of cue has a strong influence on its effectiveness. Examining the questionnaire results, we found that both the eye flash and the stay silent cue scored lower on Noticeability than the other two cues. Eye flash also scored relatively low on Meaningful suggesting that it was unclear what was meant by the cue. Indeed, the cue is not one we use in natural human communication as it is physically impossible to turn on lights in our eyes. For robots, toys and other systems LEDs are frequently used for all sorts of signals (power on, stand-by etc.), but not usually for turn-taking. So the meaning of the eye flash cue is much more ambiguous than the other cues (even stop arms is akin to co-speech gestures). This was confirmed by subject's reports that the lighting of the eyes was confusing and it was not clear why the robot performed this action. This could also indicate that the timing is more critical for the eye flash cue, because of its ambiguous meaning. We expect that the same applies to other artificial cues. In contrast, the head turn cue – a common human way of indicating a turn-yield – scored highest on the Meaningful component and outperformed all other cues during the experiment. Indeed, participants indicated they recognized this behavior from their everyday human communication and found it clear it was their turn. Surprisingly, the analysis showed that although the head turn cue was superior in the zero delay condition, it was the worst for the long delay

condition. This is not what we expected, as the inherent presence of the stay silent cue should have resulted in a response time close to the control condition. Perhaps the clear, meaningful and obvious nature of the cue was the very reason the performance dropped, because the salient nature of the head turn cue dominated the stay silent cue even though it occurred later in time. Another possible explanation is that once the participant was ready to speak, they were disrupted by the late cue and waited for the head-turn to finish before continuing citing the alphabet. The resulting hesitation and/or confusion would presumably be more severe the more pronounced the cue is. This possibility is supported by the fact that the reduction in performance is least for the eye flash condition, which is the weakest turn-yield cue in our experiment.

4.2 Cue combinations

We also expected that cue combinations would perform as good as the best cue within the set when the cue was timed correctly (i.e. zero delay). We found this winner-take-all effect, as the head turn + eye flash combination performance was equal to the head turn only cue, but was significantly better than the eye flash only cue. The same was found for the stop arm movement + eye flash cue combination, where performance was equal to the stop arm movement only cue and was significantly better than the eye flash only cue. However, it is worth noting that this effect might be biased by the fact that the combinations in this research were made between one poor (eye flash) and one effective (head or arm) cue. Therefore, it is unclear whether the effect would also be observed when several effective cues were combined. Further research on combining different cue types and cue strengths could provide more insight in the exact workings of cue combinations in human-robot communication.

4.3 Limitations and future research

Overall the alphabet citing task worked very well in that it provided us with a huge number of turn-yield/turn-take events in a very short time. This resulted in plenty data for doing good statistical analyses. A limitation of this task is that participants sometimes (about 1 to 3 times out of 90 per participant) lost concentration and had trouble recollecting which letter was next. Apparently, the assumption that remembering the alphabet involved only highly automatized cognitive processes is not completely true. This might have had a small effect on the data in that the response times are slightly overestimated, although data points with very long delays were deleted as outliers. Follow-up studies may prevent this by choosing even easier speech elements as substitute for the alphabet. Also, subjects sometimes spoke too quickly, resulting in the Nao missing a letter. As the robot started speaking again, it would repeat one or more letters and the participant could notice this inconsistency. Although the results of the study were not directly influenced since only the first utterance was of importance, it might have cause annoyance or confusion. However, participants reported afterwards

that this was no problem. The inconsistency may be prevented by optimizing the speech recognition engine.

Future research can focus on combining more different groups of cues. This can include several weaker or stronger cues, or a combination of both. Comparing the performance of these sets can increase our understanding in how cues interact and improve the efficiency of communication between robots and humans. Also, adding more different delay settings can give us insight in how the influence of timing exactly influences people's performance. What is the shape of the performance curve? Can we find a sweet spot, perhaps somewhere in between the zero delay and stay silent cue-synced condition? An other interesting study would involve subjects from different cultural backgrounds. We know that there are cultural differences in gesture associations [18]. How does this influence the performance for the various turn-taking cues as used in this study?

The choice of cues not only depends on the human's perceptual abilities, but also on the choice of robot. Clearly, the Nao robot's humanoid shape admits natural non-verbal cues. Our results show that humanoid robots can exploit their humanoid form to improve communicative performance, although it still needs to be shown that our results also apply to other communicative tasks. For non-humanoid robots artificial cues like flashing eyes can still be very useful, because in noisy environments staying silent may not be an effective cue at all. Because of the inherent ambiguity of artificial cues the timing appears to be more critical. Based on our results it should deviate more than 0.6s from staying silent. Still future work should verify whether people interpret turn-taking cues as intended and whether they are timed correctly.

4.4 Conclusions

In summary, our results suggest that when the timing of a turn-yield cue precedes the stay silent cue, response performance improves proportional to the strength of the cue (i.e. meaningful and noticeable). If the cue occurs after this moment, the strength of the cue works counterproductive as it provides disruption and confusion. Combining multiple cues reveals a winner take all mechanism in favor of the most salient cue.

References

1. Sacks, H., Schegloff, E.A., Jefferson, G.: A simplest systematics for the organization of turn-taking for conversation. Language **50**(4) (1974) 696–735
2. Craig, H.K., Gallagher, T.M.: Gaze and proximity as turn regulators within three-party and two-party child conversations. Journal of Speech, Language, and Hearing Research **25**(1) (1982) 65–75
3. Novick, D.G., Hansen, B., Ward, K.: Coordinating turn-taking with gaze. In: Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on. Volume 3. (Oct 1996) 1888–1891 vol.3
4. Jokinen, K., Nishida, M., Yamamoto, S.: On eye-gaze and turn-taking. In: Proceedings of the 2010 workshop on Eye gaze in intelligent human machine interaction, ACM (2010) 118–123

5. Kose-Bagci, H., Dautenhahn, K., Nehaniv, C.L.: Emergent dynamics of turn-taking interaction in drumming games with a humanoid robot. In: Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on, IEEE (2008) 346–353
6. Mutlu, B., Shiwa, T., Kanda, T., Ishiguro, H., Hagita, N.: 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, ACM (2009) 61–68
7. Liu, C., Ishi, C.T., Ishiguro, H., Hagita, N.: Generation of nodding, head tilting and eye gazing for human-robot dialogue interaction. In: Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction. HRI '12, New York, NY, USA, ACM (2012) 285–292
8. Huang, C.M., Mutlu, B.: Modeling and evaluating narrative gestures for humanlike robots. In: Robotics: Science and Systems. (2013) 57–64
9. Ham, J., Cuijpers, R.H., Cabibihan, J.J.: Combining robotic persuasive strategies: the persuasive power of a storytelling robot that uses gazing and gestures. International Journal of Social Robotics **7**(4) (2015) 479–487
10. Thomaz, A.L., Chao, C.: Turn-taking based on information flow for fluent human-robot interaction. AI Magazine **32**(4) (2011) 53–63
11. Bohus, D., Horvitz, E.: Facilitating multiparty dialog with gaze, gesture, and speech. In: International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction. ICMI-MLMI '10, New York, NY, USA, ACM (2010) 5:1–5:8
12. Bohus, D., Horvitz, E.: Multiparty turn taking in situated dialog: Study, lessons, and directions. In: Proceedings of the SIGDIAL 2011 Conference. SIGDIAL '11, Stroudsburg, PA, USA, Association for Computational Linguistics (2011) 98–109
13. Matsusaka, Y., Tojo, T., Kubota, S., Furukawa, K., Tamiya, D., Hayata, K., Nakano, Y., Kobayashi, T.: Multi-person conversation via multi-modal interface-a robot who communicate with multi-user-. In: EUROSPEECH. Volume 99. (1999) 1723–1726
14. Mataric, M.J., Nilsson, M., Simsarin, K.T.: Cooperative multi-robot box-pushing. In: Intelligent Robots and Systems 95.'Human Robot Interaction and Cooperative Robots', Proceedings. 1995 IEEE/RSJ International Conference on. Volume 3., IEEE (1995) 556–561
15. Kube, C.R., Zhang, H.: The use of perceptual cues in multi-robot box-pushing. In: Robotics and Automation, 1996. Proceedings., 1996 IEEE International Conference on. Volume 3., IEEE (1996) 2085–2090
16. van Schendel, J.A., Cuijpers, R.H.: Turn-yielding cues in robot-human conversation. New Frontiers in Human-Robot Interaction (2015) 85
17. Hillis, J.M., Ernst, M.O., Banks, M.S., Landy, M.S.: Combining sensory information: mandatory fusion within, but not between, senses. Science (New York, N.Y.) **298** (November 2002) 1627–1630
18. Graham, J.A., Argyle, M.: A cross-cultural study of the communication of extra-verbal meaning by gestures. International Journal of Psychology **10**(1) (1975) 57–67