Specific hand motion patterns correlate to miscommunications during dyadic conversations

Elif Ecem Özkan, Tom Gurion, Julian Hough, Patrick G.T. Healey, Lorenzo Jamone
School of Electronic Engineering and Computer Science
Queen Mary, University of London
London, UK
{e.ozkan, t.gurion, j.hough, p.healey, l.jamone}@qmul.ac.uk

2021 IEEE International Conference on Development and Learning (ICDL) | 978-1-7281-6242-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICDL49984.2021.9515613

Abstract—**Effective and natural communication is achieved by exchanging several multi-modal signals through highly coordinated communication mechanisms. These mechanisms are frequently subject to troubles of speaking in the form of disfluencies, typically followed by a self-repair from the speaker (i.e. to try to fix the misunderstanding): overall, these are signs of a possible miscommunication. Automatically detecting miscommunications is crucial to implement conversational agents, either digital or robotic, that could successfully interact with people. This can be done by searching for specific patterns across different communication channels, for example disfluencies in the speech signal or specific movements of the limbs. However, what are the motion patterns that correlate to miscommunications is still unclear. In this paper we report a human study in which we identify one of such patterns: in particular, we show that the hands of the speaker reliably move upwards during miscommunications. We performed a statistical analysis of synchronized speech and motion tracking data extracted from natural conversations of 15 dyads; our results show a statistically significant tendency of moving hands upwards during speech disfluencies, which are a clear sign of miscommunication.**

I. Introduction

The promise of artificial intelligence, the increasing demand for personalised embodied agents and enhancing health-care with technology motivate the development of new computational models in intelligent social interaction understanding. Topics investigating the interactive and dynamic nature of face-to-face human communication that have been explored through research in fields such as conversation analysis, computational linguistics and experimental psychology should inform human-agent/robot interaction research for more natural and interpersonal applications. In particular, the design of social robots in several applied situations, such as healthcare, education [1] and entertainment, require an interdisciplinary approach [2].

Here, we suggest to focus on a refined empirical representation that is useful in explaining how people incrementally build mutual understandings to inspire automatic detection and understanding of interactive human behaviour. Specifically, we aim to identify specific movement patterns that can be interpreted as signs of miscommunications during a natural face-to-face interaction between human dyads. If such patterns prove to exist, then automatic systems can be engineered to detect those motion patterns during conversations, for example using motion tracking based on computer vision. This can be used for example in social human-robot interaction, or even in human-robot collaboration in industry, to detect if the human has troubles understanding what the robot is trying to communicate. Also, similar systems could be used to monitor a conversation between a human and a digital avatar, or between two humans (such as patient doctor interactions), and to recommend specific interventions if miscommunications occur.

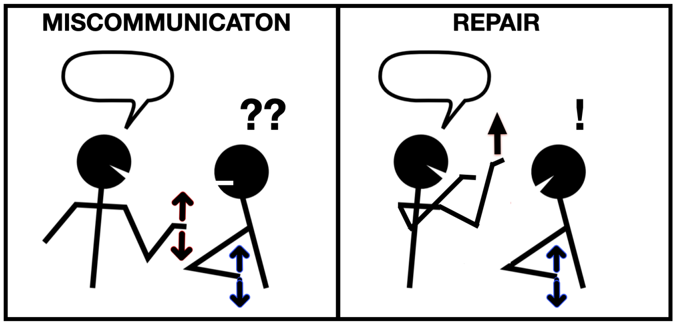


Fig. 1: By tracking hand positions during face-to-face dyadic dialogues, we discovered that people tend to move their hands upwards during miscommunications.

This paper aims to address the following research question:

1) Are there any specific generalisable non-verbal behaviour changes, which can be detected from the hand movement data, during miscommunication?
2) How could we compare these changes to other instances in interaction?

The rest of the paper is organised as follows. In the next section, II, we discuss the state-of-the-art from the perspectives of social and behavioural sciences, computational linguistics, social and cognitive robotics. In Sec. III we demonstrate our methodology, including the description of our dataset, automatically generated self-repair labels over speech data and a preliminary inspection over hand-height changes; in Sec. IV we examine the significance of the comparison between hand heights during instances containing self-repair versus

Work partially supported by the EPSRC UK through projects NCNR (EP/R02572X/1) and MAN^3 (EP/S00453X/1), and by the EPSRC and AHRC Centre for Doctoral Training in Media and Arts Technology (EP/L01632X/1).

instances that do not contain self-repair; in Sec. V we present our conclusions.

II. State-of-the-art

In the previous decades, the methods in designing mental processes in social robots for collaborative tasks were mostly based on human theory of mind, perspective taking, embodied cognition or highly integrated models (as can be seen in [3]). Furthermore, attempts at improving social intelligence in robots were dominated by human emotion recognition [4]. These approaches are also based on findings from the cognitive sciences but they do not account for the interpersonal and interactive aspect of social interaction. Lately, there is an understandable shift towards modelling human-robot interaction based on natural human-human communication research (as in the cases of [5], [6] [7], [8]) in correlation with the trends in cognitive sciences. More particularly, *the affective grounding* perspective aims to extend previous accounts in affective computing by defining interaction as a jointly coordinated interpersonal process, taking conversational analytic (CA) approaches into account [6]. Important concepts from CA such as back-channel responses and repair, which is a mechanism to fix misunderstandings, are implemented in regulating shared understanding and attention between human and robots.

The evidence from the fields of conversation analysis and linguistics [9], [10], experimental and social psychology [11] has shown shown that miscommunication has a significant effect on how shared meanings are created and maintained. The view of Miscommunication phenomena being a crucial mechanism of interaction demonstrates itself in the form of conversational partners or groups *repairing* misunderstandings on-the-go to update their mutual understanding. Although positive backchanneling (such as head-nods and utterences (“mmm”, “hmm”)) have been explored in many occasions as an important contribution to mutual understanding, there is still room for further exploration of the negative feedback in conversation that is misunderstanding. The variability of miscommunication types and the disparity of where it occurs in the dialog make it a difficult phenomena to disclose and model even though its systematical patterns are mostly discovered [12].

The mechanism that people use to deal with “troubles of speaking, hearing and understanding” is referred to as *Repair* in the field of Conversation Analysis [13]. It is very frequent [14] in everyday interaction, “the only type of turn with unrestricted privilege of occurrence” [15], and in some cases universal across languages [16] including sign language [17]. Inspired by these, the *Running Repairs Hypothesis* suggests that “the coordination of language use depends primarily on processes used to deal with misunderstanding on the fly and only secondarily on those associated with signaling understanding” [18]. The hypothesis assumes a dynamic approach to communication, negative interactional feedback being central in coordinating language over positive feedback [18] as the crucial points in interaction is about solving misunderstandings to achieve mutual understanding.

Repair mechanisms vary in possible different positions they occur in the dialog and who initiates them (speaker or recipients). The most common [12] and elementary form of repair is *self-repair*, when the speaker repairs their own speech in the process [19]. Self-repairs convey useful information for co-ordination in dialogue [12] through various means. As outlined in [20] self-repair incidents contribute to:

- compensating for misinformation and creating a warning in listener’s parsing process [21]
- synctactic analysis/re-analysis of language processing through grammaticality and ungrammaticality judgments [22]
- increase in the frequency of backchannel responses by the listener to express their attention and understanding towards the speaker [23]

The importance for automatic detection of self-repairs has been acknowledged and explored in the field of computational linguistics for robust natural language processing and understanding in interactive systems. Through advanced and state-of-the-art machine learning techniques low-latency [24] and even a real-time [20] detection of repairs (mostly in the case of disfluencies) through dialogue transcripts are effective. However, as reviewed in [20], these systems have their limitations and challenges. Moreover, humans in dialogue must decide how to respond to or whether to initiate repair as and when they encounter problems naturally. The mentioned algorithmic approaches in natural language processing cannot construe how humans actually identify these while speaking, listening and gesturing intuitively. In face-to-face human-human interaction, feedback in general, hence repair, concurs with multimodal non-verbal signals such as gaze [25], intonation [26], or gesture [27] [28] [23]. Evidence shows that, there are identifiable correlations between repair events in speech and non-verbal signals, such as increased self-repair associated with increased hand-gestures [27] [29]. Finding quantifiable differences in movement data can lead the way into creating robust applications that could automatically detect self-repair instances.

III. Methodology

In this study we used the dataset described in [30] (Fig. 2). Thirteen pairs of native English speakers were recorded in conversation by video cameras, radio microphones, and a motion capture system. They were asked to discuss the design of an apartment for them to share, for 15 minutes. This task is described in details in [31]. Their head position was tracked with HTC Vive trackers fitted on hats. HTC Vive handheld controllers held by the participants were used to track the position of their hands. The dataset is not annotated for miscommunication or repair, but it provides plentiful opportunities to observe miscommunication with its natural interaction task design.

Although the relevance and the extent of the movement readings are very assuring, we realise that the HTC Vive handheld controllers can be intrusive when it comes to capturing natural hand behaviour. Nonetheless, we still observe plentiful

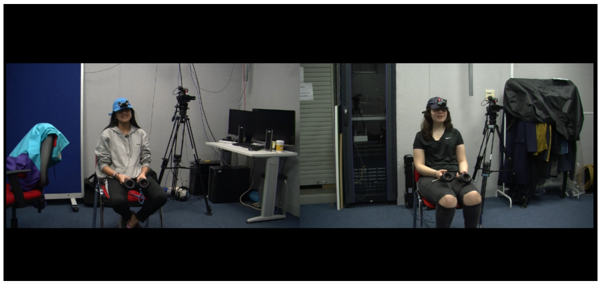


Fig. 2: Video snapshots of a conversation session from the dataset. Individuals holding handheld trackers and wearing hats with trackers sits across each-other during natural dialogue.

non-verbal communicative hand gesture use in all participants. Moreover, the effect from the controllers would be present in all participants for all conditions, hence does not prevent us to observe the difference in gesture during miscommunication instances.

A. Expected Outcomes

We aim to observe how the non-verbal behaviour in speakers and listeners change during miscommunication instances in order to find out if we can automatically detect miscommunication by looking at the movement data. Given that self-repair instances are crucial points in interaction and expected to coincide with observable changes in movement (such as hand gesture rate as explored in [27]), we predict to observe:

- quantifiable changes in the movement data within the temporal proximity of disfluency labels
- that such changes are generalisable over the whole dataset

B. Using Detected Disfluencies in the Speech to Label Miscommunication Events

Purver [20] suggests the importance of labeling miscommunication events with automatic tools, because their manual labeling is not only time-consuming, but also negatively affected but the high variability in the personal interpretation of what a miscommunication is. Although it is difficult to verify the validity of the performance of automatic detection over disparate data, it is worth inspecting the outputs first, instead of going through the prolonged manual annotation process. Hough [32] proposes a deep learning model that incrementally processes the output of an automatic speech recogniser to detect disfluencies (e.g. self-repairs). We have used this model, utilising IBM Watson [1] speech-to-text service for automatic speech recognition, to automatically label disfluencies (repair onset or related words) in the speech recordings of our dataset. Avoiding the infeasible annotation process, for our analysis, we have opted for using these labels for an estimation of where the miscommunication instances -in the case of self-repair- occur in the dialog. The disfluency detection tool outperforms previous state-of-the-art models when trained on the Switchboard dataset [33], F1 scores for *'e' (edit)* tags is above 0.9 and for *'rpS' (repair onset)* tags above 0.75 [32]. Following the assumption of *reparandum-interregnum-repair* structure (terms proposed by [34]) in speech repairs, the tool works on to incrementally detect these structures:

$$\texttt{John}\ \underbrace{\texttt{[likes}}_{\texttt{reparandum}} + \underbrace{\{\ \texttt{uh}\ \}}_{\texttt{interregnum}} \underbrace{\texttt{loves]}}_{\texttt{repair}}\ \texttt{Mary}$$

- *reparandum*: the word to be repaired
- *interregnum*: edit word *uh*, *I mean*, *you know*
- *repair*: repair onset word that initiates the repair.

In the case of constructing the labels in our dataset, the detection tool was used in the simple mode that labels utterances as *edit (e)*, *fluent (f)* and *repair onset (rpS)* terms. For example:

	| A	uh	flight	[to	Boston +	{ uh	I	mean }	to	Denver]	on	Friday |	Thank	you |
Disfluency (simple)	f	e	f	f	f	e	e	e	rpS	f	f	f	f	f

The utterances containing *edit (e)* or *repair onset (rpS)* tags were taken as disfluency labels with their corresponding time stamps. Over the whole dataset (6:37 hours), there are 3192 disfluency instances which gives us plenty samples to investigate the phenomenon. We have investigated the disfluent timestamps and the movement features during these instances and compared them with windows not containing any disfluency labels (hence fluent).

C. Automatic Detection of Speakers and Listeners

A simple floor control detection model [35] processes the audio from the participants' microphones to determine who is the speaker at any given moment. For the same timestamp, the other participant is considered a listener. This model, presented in Fig. 3, employs simple audio processing techniques, namely, low-pass filters and a thresholds, to operate. Audio from the participants' microphones is processed in buffers of 0.02 seconds. For each buffer the root mean square (RMS) value is calculated. These values are filtered by low pass filters with a cutoff frequency of $0.35Hz$. If the difference between the minimal and the maximal filtered RMS values is larger than 0.1 the participant with the maximal filtered RMS value is the speaker. Otherwise the previously reported speaker is the speaker again.

D. An Inspection of Hand Movement Patterns during Disfluent and Fluent Moments in Conversation

Automatic tagging resulted in 3192 disfluency tags over the whole dataset ($M = 122.769, sd = 40.078$, over 26 participants). The feature, hand heights, were reduced to one, selecting the highest one among right and left hand readings (Y-position) from the hand trackers. Maximum hand height feature windows of 1 seconds, 2 seconds and 4 seconds were created for disfluent and fluent cases. For disfluency windows, timestamps from the tags were taken as the centre ($time = 0$),

[1] https://www.ibm.com/watson/

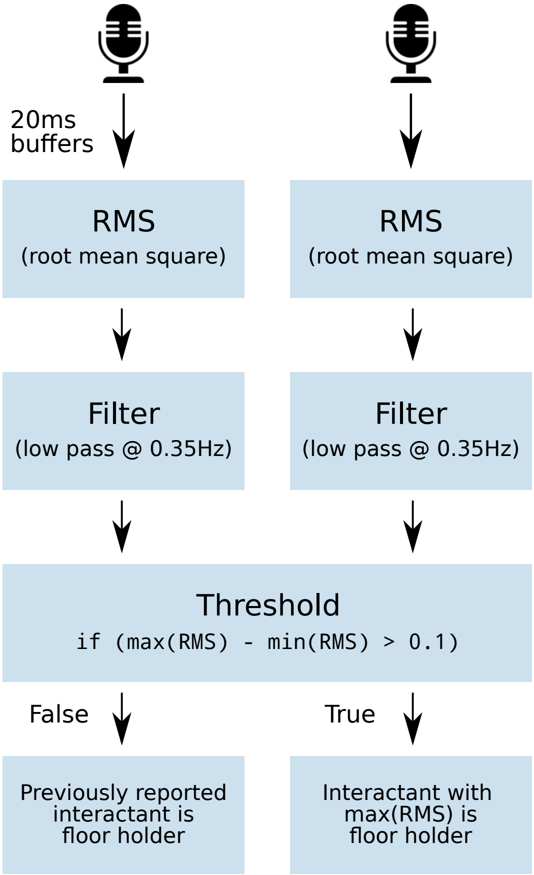


Fig. 3: Schematic diagram of the floor control detection model.

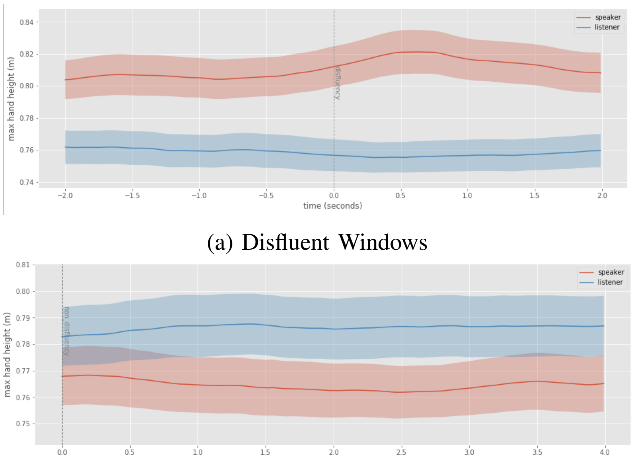
and equal number of maximum hand heights were taken before and after disfluency according to the window length (e.g. for 4 second windows, -2 seconds and $+2$ seconds).

For the inspection of movements accompanying fluent parts of the speech, we have extracted sections that are more than 6 seconds apart from the end time of a disfluency to the start time of next disfluency including a buffer of 1 second. This way, any movement that might have been affected from a previous disfluency would have been prevented to be included in the analysis of movements during fluent sections. 2 seconds and 4 seconds windows from fluent sections have resulted in 3557 fluent instances to be analysed.

For the construction of both windows', we have used a sampling period of 10 ms. This resulted in 400 corresponding hand-heights for each window in the case of 4 second windows. Lastly, each window is marked as a speaker or listener window based on the output of the floor control detection algorithm discussed in Sec. III-C at the middle timestamp of the window.

A preliminary inspection over maximum hand heights during disfluent and fluent windows showed differences in mean and variance over time (Fig. 4). For speakers, the mean of hand height starts increasing right before the disfluency and continues to get higher until 0.5 seconds after the disfluency. On the contrary, in fluent windows where speakers have shown no troubles of speaking, their hand heights look quite stationary. In both cases listeners show static hands heights.

For statistical analyses to determine the significance of these observations, the next step was to perform separate linear regressions and mixed linear regressions for speakers' hand height windows of 0.5 seconds right after disfluency ($disfluency = 1$) and fluent windows of 0.5 seconds ($disfluency = 0$), where we observed the substantial changes. The results of these analyses are presented in the next section. Additionally, for the statistical analysis, we have filtered the disfluency windows, by removing the windows that are less than 2 seconds apart, in order to prevent the overlaps between movement windows. This resulted in 2076 disfluency windows to be analysed.



(a) Disfluent Windows

(b) Fluent Windows

Fig. 4: Mean (line) and Variance (shades) of Hand Heights for Disfluency and Fluent Windows. In the case of speakers (in red), we observe an increase in hand height starting at the disfluency moment and continuing to increase approximately for 0.5 seconds. In comparison, during fluent instances the hand height seem to be stationary both in speakers and listeners.

The number of fluent windows were reduced to 2076 samples (same number as disfluency windows) by random selection. These equal sample sizes were used for separate linear regressions.

For the mixed model regression analysis, we have kept all the available windows (2076 disfluency, 3557 fluent).

IV. Results

A. Two Independent Linear Regressions Over Isolated Subgroups

Separate linear regressions for hand height based off of time offset were initially performed over two types of windows (both N = 2076). Time offset was constructed from 0 to 0.5 seconds, with a sampling period of 10 ms. In the case of fluent windows, time offset coefficient did not perform as a significant factor ($\beta_1 = 0.002658, p = 0.4$), meaning the change in hand-height as time progresses was not significant for fluent windows. The intercept which is the mean of the response, hand height, when the predictor, time offset equals to 0, significantly characterises the mean hand-height of the fluent windows at time 0 ($\beta_0 = 0.768965, p < 2e - 16 * **$). In the case of disfluent windows same regression analysis resulted in a significant and higher time coefficient ($\beta_1 = 0.020140, p = 9.5e - 10 * **$), and significantly higher mean initial hand-height ($\beta_0 = 0.800770, < 2e - 16 * **$). This suggests hands start at y-position of 0.8 meters and increases linearly with a slope of 0.02 meters per 10 ms.

B. Mixed Linear Regression Model

To explain these effects further, a mixed model regression analysis was used to model the hand height based as a function of two fixed factors, i.e. the presence or absence of a disfluency and time offset, where the participant number whose hand height was considered as a random factor. The results shown in Table I confirms that participants' hand heights were significantly higher during disfluencies (self-repairs) compared to maximum hand heights of fluent instances. This can be seen from the estimate values, intercept being the mean of hand height and others showing how it is affected in different conditions. The mean of hand height (0.7746), is 0.0139 higher in the case of disfluencies (disfluency present = Disfluency1). Furthermore, there is an increase in hand heights from the disfluency start at 0 seconds to 0.5 seconds. This is deducted from *Disfluency1:Time Offset* variable being 0.0208, meaning the slope of handheight/disfluency is 0.0208 higher when the disfluencies are present. In the fluent cases the hand heights are slightly decreasing but not significantly ($TimeOffset = -0.0010, p > 0.1$).

The following equations can be achieved to quantify the average initial hand height, increase and decrease amount over time in fluent (1) and disfluent (2) cases:

$$0.77 - 0.0010t \tag{1}$$

$$0.77 + 0.0139 + (0.0208 - 0.0010)t \tag{2}$$

The increase in hand height over time offset from the disfluency moment is significant ($p < 0.01$), suggesting that the speakers raise their hands right after disfluency. The decrease amount in hand heights over time in fluent windows are much lower and not significant, suggesting a more static hand height. These findings are consistent with our expectations regarding observing a difference in hand movements during repair instances compared to the parts of the interaction that does not contain troubles of speaking. Similarly, previous work had shown that hand gesture rate is positively correlated with repair rate [27]. Also, the statistical analyses are in line with our preliminary observations regarding the average hand heights during disfluencies versus fluent sections as presented in Sec. III-D Fig. 4.

It is important to note that, our fluent windows do not consider the presence of other types of repair (or repair initiation such as clarification questions), hence other possibilities of miscommunication instances. If we assume that some portion of our fluent windows in fact includes a small proportion of miscommunication (since the self-repair is the most frequent type of miscommunication compared to others), there is a likelihood of observing a higher discrepancy between miscommunication instances and the moments in interaction that do not contain any troubles of understanding between individuals.

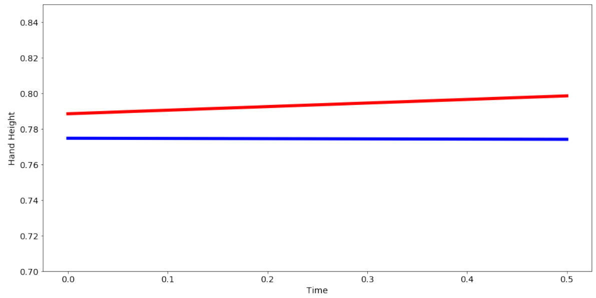


Fig. 5: Mixed Linear Regression Model, Hand Height (m) versus Time (sec). Regression lines, disfluency condition (red) and fluent condition (blue) over all samples, y-axis zoomed in within the vicinity of mean values for hand heights [0.7, 0.85].

V. Conclusion

Based on the conversation analytic approach that views miscommunication as a crucial part of interaction, we have investigated the signs for repair in non-verbal behaviour by comparing the speaker hand heights during disfluencies to other instances in face-to-face dialogue. Consistent with the finding that repair rate is positively correlated with hand gestures [27] and our expectations derived from the preliminary examinations of the dataset (III-D), our results show that the maximum hand heights for speakers are higher during disfluencies and increase after the disfluency instance compared to other moments in interaction. Finding an obvious cue in hand gestures contributes to the *Running Repairs Hypothesis* suggestions regarding repair being the crucial mechanism in coordinating language [18]. Moreover, a direct interpretation of speakers' hand heights can be useful from multiple perspectives when designing interactional systems. Our findings have a possible impact on:

- improving the performance of speech disfluency detection tools with the additional modality of hand heights;
- designing applications that can automatically detect miscommunication instances from the hand heights.

Conversational agents can communicate more effectively by integrating interactional strategies such as back-channelling more to signal being attentive when the speaker is in trouble. Automatic miscommunication detection systems could also recognise miscommunication between people, enabling applications in education and health-care to provide new measures of engagement and understanding.

TABLE I: Dependent Variable: Hand Height with Fixed Effects (Disfluency), (Time Offset), random effects (Participants)

Variable	***Estimate***	***SE***	***t***	***p***
(Intercept)	**0.7746**	0.0199	38.9106	$<$ **2e-16** ***
Disfluency1	**0.0139**	0.0009	15.7353	$<$ **2e-16** ***
Time Offset	-0.0010	0.0025	-0.3955	0.6946
Disfluency1:Time Offset	**0.0208**	0.0030	6.8605	**6.9e-12** ***

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$ Significance codes.

In future work, we will extend this study to include other movement features such as head positions, hand and head velocity to provide an inspection over multiple gestural cues. Especially, since this study suggests having a closer look at the repair and non-verbal behaviour, additional motion information should be investigated. We also plan to explore more sophisticated techniques such as classification, clustering and machine learning applications to automatically detect miscommunications from motion data.

References

[1] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka, "Social robots for education: A review," *Science Robotics*, vol. 3, no. 21, 2018.

[2] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and Autonomous Systems*, vol. 42, pp. 143–166, 03 2003.

[3] C. Breazeal, J. Gray, and M. Berlin, "An embodied cognition approach to mindreading skills for socially intelligent robots," *The International Journal of Robotics Research*, vol. 28, no. 5, pp. 656–680, 2009.

[4] C. Breazeal, "Emotion and sociable humanoid robots," *Int. J. Hum.-Comput. Stud.*, vol. 59, p. 119–155, July 2003.

[5] G. Mehlmann, M. Häring, K. Janowski, T. Baur, P. Gebhard, and E. Andre, "Exploring a model of gaze for grounding in multimodal hri," 11 2014.

[6] M. F. Jung, "Affective grounding in human-robot interaction," in *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, pp. 263–273, March 2017.

[7] H. W. Park, M. Gelsomini, J. J. Lee, T. Zhu, and C. Breazeal, "Backchannel opportunity prediction for social robot listeners," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, (Singapore, Singapore), pp. 2308–2314, IEEE, May 2017.

[8] H. W. Park, M. Gelsomini, J. J. Lee, and C. Breazeal, "Telling Stories to Robots: The Effect of Backchanneling on a Child's Storytelling," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17*, (Vienna, Austria), pp. 100–108, ACM Press, 2017.

[9] H. H. Clark and S. E. Brennan, "Grounding in communication," in *Perspectives on Socially Shared Cognition* (L. Resnick, L. B., M. John, S. Teasley, and D., eds.), pp. 13–1991, American Psychological Association, 1991.

[10] H. H. Clark, *Using Language*. 'Using' Linguistic Books, Cambridge University Press, 1996.

[11] F. C. Bartlett, "Remembering: A study in experimental and social psychology.," *Remembering: A study in experimental and social psychology.*, pp. xix, 317–xix, 317, 1932.

[12] M. Colman and P. Healey, "The distribution of repair in dialogue," *Cognitive Science*, vol. 33, 2011.

[13] E. Schegloff, *Recycled turn beginnings; A precise repair mechanism in conversation's turn-taking organization*. 01 1987.

[14] S. E. Brennan and M. F. Schober, "How listeners compensate for disfluencies in spontaneous speech.," *Journal of Memory and Language*, vol. 44, no. 2, pp. 274–296, 2001.

[15] E. A. Schegloff, "Reflections on quantification in the study of conversation," *Research on Language and Social Interaction*, vol. 26, no. 1, p. 99–128, 1993.

[16] M. Dingemanse, F. Torreira, and N. Enfield, "Is "huh?" a universal word? conversational infrastructure and the convergent evolution of linguistic items," *PloS one*, vol. 8, p. e78273, 11 2013.

[17] E. Manrique and N. Enfield, "Suspending the next turn as a form of repair initiation: evidence from argentine sign language," *Frontiers in Psychology*, vol. 6, p. 1326, 2015.

[18] P. G. T. Healey, G. Mills, A. Eshghi, and C. Howes, "Running repairs: Coordinating meaning in dialogue," *Topics in cognitive science*, vol. 10 2, pp. 367–388, 2018.

[19] E. Schegloff, G. Jefferson, and H. Sacks, "The preference for self-correction in the organization of repair in conversation," *Language*, vol. 53, pp. 361–382, 06 1977.

[20] M. Purver, J. Hough, and C. Howes, "Computational models of miscommunication phenomena," *Topics in Cognitive Science*, vol. 10, no. 2, pp. 425–451, 2018.

[21] S. Brennan and M. Schober, "How listeners compensate for disfluencies in spontaneous speech," *Journal of Memory and Language*, vol. 44, pp. 274–296, 02 2001.

[22] F. Ferreira, E. Lau, and K. Bailey, "Disfluencies, language comprehension, and tree adjoining grammars," *Cognitive Science*, vol. 28, pp. 721–749, 09 2004.

[23] P. Healey, M. Lavelle, C. Howes, S. Battersby, and R. McCabe, "How listeners respond to speaker's troubles," 07 2013.

[24] J. Hough and M. Purver, "Strongly incremental repair detection," *CoRR*, vol. abs/1408.6788, 2014.

[25] A. Hjalmarsson, C. Oertel, and K. Speech, "Gaze direction as a back-channel inviting cue in dialogue," 01 2012.

[26] A. Gravano and J. Hirschberg, "Backchannel-inviting cues in task-oriented dialogue," pp. 1019–1022, 01 2009.

[27] C. Howes, M. Lavelle, P. Healey, J. Hough, and R. McCabe, "Helping hands? gesture and self-repair in schizophrenia," 05 2016.

[28] M. Lavelle, C. Howes, P. Healey, and R. McCabe, "Speech and hand movement coordination in schizophrenia," 01 2013.

[29] P. Healey, N. Plant, C. Howes, and M. Lavelle, "When words fail: Collaborative gestures during clarification dialogues," 03 2015.

[30] T. Gurion, P. G. Healey, and J. Hough, "Comparing models of speakers' and listeners' head nods," in *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue*, (Whaltham, MA), SEMDIAL, jul 2020.

[31] J. Hough, Y. Tian, L. de Ruiter, S. Betz, S. Kousidis, D. Schlangen, and J. Ginzburg, "Duel: A multi-lingual multimodal dialogue corpus for disfluency, exclamations and laughter," in *10th edition of the Language Resources and Evaluation Conference*, 2016.

[32] J. Hough and D. Schlangen, "Joint, incremental disfluency detection and utterance segmentation from speech," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, (Valencia, Spain), pp. 326–336, Association for Computational Linguistics, Apr. 2017.

[33] M. Meteer and A. Taylor, *Dysfluency Annotation Stylebook for the Switchboard Corpus*. University of Pennsylvania, 1995.

[34] E. Schriberg, "Preliminaries to a theory of speech disfluencies," 1994.

[35] T. Gurion, J. Hough, and P. G. Healey, "A simple, reactive model of turn-taking in dialogue." Unpublished manuscript, 2021.