Speaker Motion Patterns during Self-repairs in Natural Dialogue

Elif Ecem Ozkan
Queen Mary University of London
London, UK
e.ozkan@qmul.ac.uk

Tom Gurion
t.gurion@qmul.ac.uk
Queen Mary University of London
London, UK

Julian Hough
j.hough@qmul.ac.uk
Queen Mary University of London
London, UK

Patrick G.T. Healey
p.healey@qmul.ac.uk
Queen Mary University of London
London, UK

Lorenzo Jamone
l.jamone@qmul.ac.uk
Queen Mary University of London
London, UK

ABSTRACT

An important milestone for any agent in interaction with humans on a regular basis is to achieve natural and efficient methods of communication. Such strategies should be derived on the hallmarks of human-human interaction. So far, the work in embodied conversational agents (ECAs) implementing such signals has been predominantly through imitating human-like positive back-channels, such as nodding, rather than active interaction. The field of Conversation Analysis (CA) focusing on natural human dialogue suggests that people continuously collaborate on achieving mutual understanding by frequently repairing misunderstandings as they happen. Detecting repairs from speech in real-time is challenging, even with state-of-the-art Natural Language Processing (NLP) models. We present specific human motion patterns during key moments of interaction, namely self initiated self-repairs, which would help agents to recognise and collaboratively solve speaker trouble. The features we present in this paper are the pairwise joint distances of head and hands which are more discriminative than the positions themselves.

CCS CONCEPTS

• **Human-centered computing** → *Collaborative interaction.*

KEYWORDS

multimodal interaction, non-verbal communication, human motion analysis

ACM Reference Format:
Elif Ecem Ozkan, Tom Gurion, Julian Hough, Patrick G.T. Healey, and Lorenzo Jamone. 2022. Speaker Motion Patterns during Self-repairs in Natural Dialogue. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22 Companion), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3536220.3563684

1 INTRODUCTION

Conversational agents that can utilise human-like communicative behaviours have been perceived as more successful in establishing relationships [32]. However, even though there are applications that employ social signal use by ECAs, such as mimicry [12] and nodding [24], the lack of satisfactory examples that can capture the efficiency and dynamic nature of human dialogue has been questioned [1, 21, 32]. Despite the major advancements in speech recognition, language processing and computer vision, systems have not been able to capture the effectiveness and flexibility of natural human dialogue. The proposed direction, therefore, is to put the hallmarks of human-human interaction (HHI) on the spotlight when designing HAI. It has been acknowledged by most recent reviews on the field of human-agent interaction (HAI) [1, 21] that an integrated approach in which multimodal and multifunctional feedback signals are mutually used and recognised is needed to advance HAI.

As human dialogue is rarely fluent and without errors [3], handling misunderstandings or *negative grounding* is a fundamental part of this requirement [2]. *Repair* in Conversation Analysis is the mechanism that people use to deal with "troubles of speaking, hearing and understanding" [28]. It is very frequent [4] in everyday interaction, "the only type of turn with unrestricted privilege of occurrence" [30], and universal across languages [8] including sign language [23]. Inspired by these, *Running Repairs Hypothesis* suggests that "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" [15], assuming an interactional approach to communication. Negative feedback is central in coordinating language [15] as the crucial points in interaction is about solving misunderstandings to achieve mutual understanding. Repairs in natural conversation can occur in multiple structures where they are systematically resolved over several conversational turns [29]. The most common type of repair, self-repairs (also referred to as *disfluencies*) are when a speaker modifies their utterances inside their turn [28, 31] by restarting, repeating, or changing their words. Self-repairs in particular have received significant attention in NLP and there are now systems that can recognise and parse them [16, 26, 27, 31], however they have not been used in live conditions.

Self-repairs often are accompanied by non-verbal signals in the form of filled pauses or gestures [6]. The human motion data that accompanies the repair have the potential of augmenting the engineering of interactive systems. In principle, these signals can easily

ICMI '22 Companion, November 7–11, 2022, Bengaluru, India
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9389-8/22/11…$15.00
https://doi.org/10.1145/3536220.3563684

be detected from a camera in real-time without relying on speech recognition and NLP. ECAs can exploit non-verbal signals in human motion data for a richer interaction.

Previous work has presented quantitative results proving that repairs co-occur along specific hand movement patterns [13, 14, 37] and head movements [13]. Linear regression and mixed regression analyses suggest that speakers' hand heights are significantly higher during self-repairs, and keep increasing for 0.5 seconds after a repair [37]. This paper further investigates the patterns in head and hand movement data (as pairwise distances) during self-initiated self-repairs. The results we report suggest that self-repairs manifest themselves with distinct motion characteristics of speakers: increase in distance between right hand and left hand, decrease in distance between head and right hand, and similarly, decrease in distance between head and left hand. Having such patterns in pairwise joint distances provide the promise of detecting these cues in a more discriminative way, and in different conditions such as height differences, sitting or in motion compared to the positions themselves.

2 BACKGROUND

A high number of approaches to embodied interaction in HAI have been criticised for lacking dynamic interactional capabilities as *surface-focused* approaches which employ limited methodologies for feedback generation in order to achieve desired listener behaviour in an attempt to show agents more engaged. Whereas the attempts at *grounding-focused* nuanced feedback that considers the cognitive states of the user have been encouraged (see [1] for an extensive comparison).

For a successful communication, participants need to coordinate in sharing some information or common ground, which is, "mutual knowledge, mutual beliefs and mutual assumptions" [5].The common ground needs to be kept track of and updated. This is called *grounding*. The mutual-understanding should either be confirmed by acknowledging the information with *back-channel responses* (such as nods or "mhm", "yeah") or be fixed in the case of negative evidence suggesting mishearing or misunderstanding (such as "huh", "what?", etc.). This is possible by exchanging a variety of verbal and non-verbal social signals such as gaze, gestures or facial expressions. Non-verbal signals can vary in meaning depending on context and interpretation, and it is difficult for artificial systems to capture this nuanced variation. Which behavioural channels or modalities (such as the facial and body gestures, prosody, etc.) are needed and how to handle context-sensitive information are among the main design problems for machine analysis of human behaviour [1, 25, 33].

Finding links between critical moments in interaction based on the hallmarks of HHI research and the embodied social signals that appear in these instances have the potential to significantly advance HAI. Identifying repair instances incrementally and in real time to collaborate on solving them is an important initial step in moving forward. Previous work has found that the speakers hand heights are significantly higher during disfluencies [37]. The deep disfluency detection tool used to label disfluencies in this study, or their state-of-the-art equivalents [26] can not achieve this task from raw audio data in real-time although there are promising developments in the field [27]. If an agent or system is designed to integrate the signs in real-time human movement as a feature for identifying self-initiated self-repairs, this would provide a multilingual and practical solution to the main issues presented.

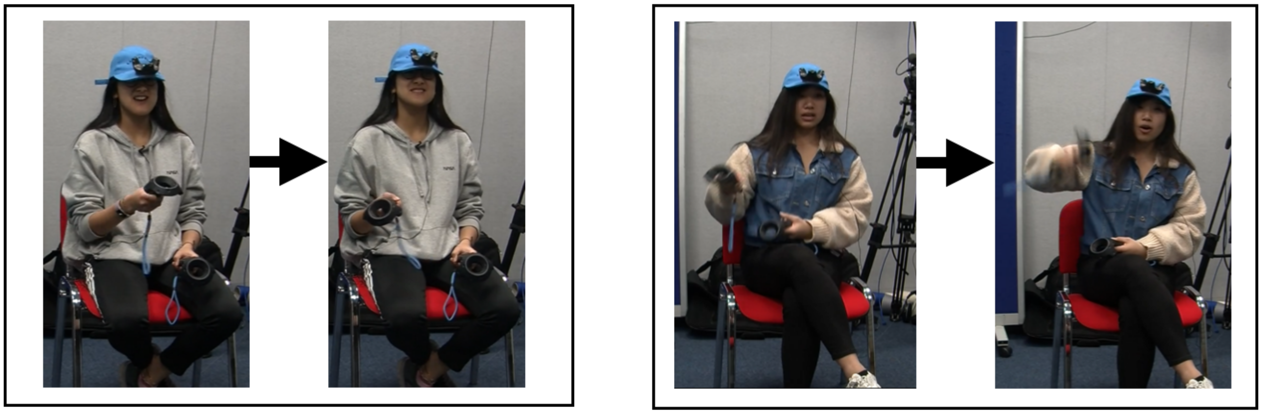
The open and challenging field of human daily activity recognition have explored the features of joint distances to employ human skeletal data for action recognition purposes. An important observation in terms of skeletal information for human motion analysis [36] has been that the pairwise relations of the joint positions are more discriminative than the 3 dimensional joint positions themselves [34]. The probabilistic approaches to achieve human daily activity recognition such as [9, 10] have employed various set of features based on Euclidean distances of the skeleton joints. The joint distances calculated in the same method were also employed for human social activity recognition [7] as an important feature. In a more recent work, the *distance descriptors* used in [35] that are used as a feature for activity recognition focuses on the distances between head and hands. Therefore, we have applied the analysis regarding hand heights in [37] to pairwise distances between 3D head and hand positions.

3 METHODOLOGY

Following the previous work which has shown speakers' hand heights are significantly higher during self-repair instances [37], we aimed for investigating motion patterns in miscommunication windows, using the dataset described in [11] (Fig. 1) that contains audio, video and 3D head and hand motion data of 13 dyads in natural face-to-face conversation. During their interaction the pairs were recorded with motion capture (head and hand 3D positions) and cameras while they discuss the design of an apartment for them to share, for 15 minutes. The details of this task are outlined in [18].

Here, we set out to analyse the relative distances between the 3D positions of head and hands for speakers and listener in the conditions of disfluent instances and other instances in interaction in which no disfluencies were detected. The dataset is not manually annotated for miscommunication due to the frequency and variety of repairs occuring in a natural dialogue for 15 minutes. Therefore, in order to identify the disfluency instances, we use the labels and the timestamps obtained from the automatic disfluency detection tool [17] as in previous work [37]. We construct windows starting from 2 seconds before to 2 seconds after the disfluency start timestamps. The mean and variance of distances between head - right hand, head - left hand and right hand - left hand in speakers and listeners (that are a dyad subjected to the detection of the floor control algorithm) are plotted within these windows for initial observations. The disfluency windows are also filtered 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.

The motion data that accompanies fluent (unrepaired) sections of speech is collected from the sections that are at least 6 seconds after the end time of a previous repair tag and before the next repair tag (including a buffer of 1 second) in order to exclude any movement that might have been related to a previous repair. The fluent sections that do not contain any repair tags were also split into 4-second windows, resulting in 3,557 instances to be analysed.



(a) Example of gesturing during disfluency

(b) Example 2 - different session and participant

Figure 1: Video snapshots of two different sessions from the dataset. Participants gesture after disfluencies to convey their message.

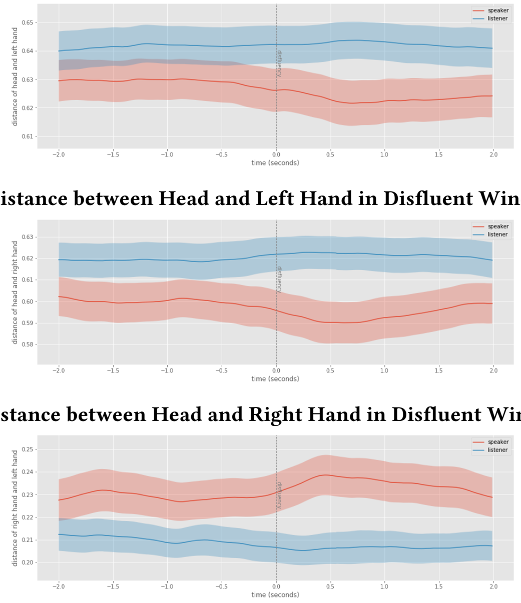
Both disfluency and fluent windows of 4 seconds have the sampling period of 10 ms., so 400 corresponding motion readings for each window. Each window is also labelled either as a speaker or a listener window determined on by the output of the floor control detection algorithm detailed in [37].

3.1 Preliminary Observations for Pairwise Distances between Head and Hand Movement Patterns during Fluent and Disfluent Moments in Conversation

The mean and variance of distances between head and hands over time in disfluency windows for speakers and listeners are presented in Fig. 2. For speakers, the mean for the distance of head and both right hand and left hand starts decreasing just before a repair and continues until 0.5 seconds after the repair. However, the distance between the hands for speakers increase within the same time reference from the disfluency moment. The same features for listeners appear to be stationary in these windows.

Fluent windows that are made of the instances that do not contain any disfluency labels (as per [17] as in the case of [37]) were also investigated for all the distance features. The mean and variance for head and hand distances in fluent windows are presented in Fig. 3 to be compared with fluent sections. It is important to note that in these instances there is no starting point since we do not have a labelled event. They are 4 second sections in the interaction that we do not observe a disfluency (as explained in Sec. 3). The comparison of the distance features during disfluent and fluent windows suggest a similar pattern to the one that was found in the study regarding hand heights [37].

To determine the significance of the observed discrepancies in the motion data in the disfluency condition, we performed mixed linear regressions for speakers' distance feature windows of 0.5 seconds right after disfluency ($disfluency = 1$) and fluent windows of 0.5 seconds ($disfluency = 0$). For this analysis, all windows have been kept (2076 disfluency, 3557 fluent).



(a) Distance between Head and Left Hand in Disfluent Windows

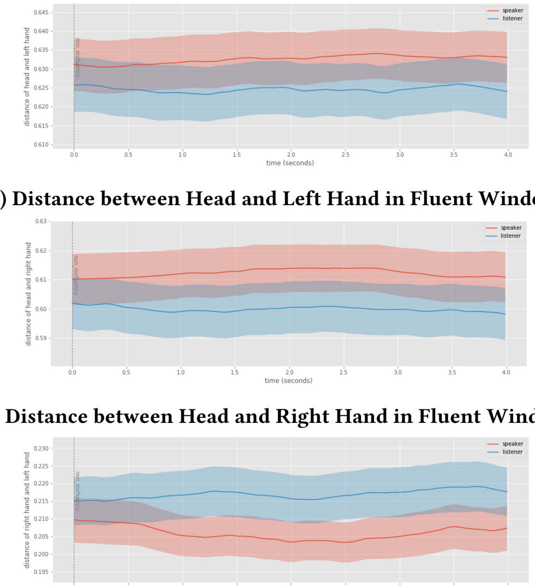
(b) Distance between Head and Right Hand in Disfluent Windows

(c) Distance between Right and Left Hand in Disfluent Windows

Figure 2: Mean (line) and Variance (shades) of Distance between Features for Disfluency Windows. The blue and red lines are for a dyad of speaker-listener (based on the floor-control detection algorithm at the middle of the window).

4 RESULTS

We have performed a mixed model regression analysis that models the distance of head and right hand based as a function of two fixed factors, i.e. the presence (*Disfluency1*) or absence of a disfluency and time offset, where the participant number of the motion readings was considered as a random factor. Table 1 shows the distances between head and left hand are significantly lower during disfluencies in comparison to the distances in fluent windows, as it can be seen



(a) Distance between Head and Left Hand in Fluent Windows

(b) Distance between Head and Right Hand in Fluent Windows

(c) Distance between Right and Left Hand in Fluent Windows

Figure 3: Mean (line) and Variance (shades) of Distance between Features for Fluent Windows.

from the estimate values. The overall mean of the distance between head and left hand (denoted by intercept) (0.6101), is 0.0112 lower in the case of disfluencies. The decrease in time in this distance feature between 0-0.5 seconds is proven by *Disfluency1:Time Offset* variable being -0.0113 meaning the slope of distance/timeoffset is lower when the disfluencies are present. The same analysis for the distance between head and left hand bears similar results; the mean of distance 0.6323 is 0.0055 lower during disfluencies, again suggesting that hands move closer to the head. The decrease in this distance between 0-0.5 seconds is 0.0052. For the distance between right hand and left hand, the increase suggesting the hands moving apart from each other is again significant. The mean distance o 0.2129 is 0.0090 higher in the case of disfluencies and the increase in this distance over time is 0.0216.

5 DISCUSSION

Video recordings of participants during disfluency instances revealed that they employ gestures when they can not find specific words for a certain type of furniture they desire in the apartment. In these cases they try to illustrate the object with their hands while describing. These instances mostly start with a disfluency when failing to find the word and continue with large gestures that correlate with the findings of larger distance between two hands and smaller distance between head and hands. For example, the participant in Fig. 1.a experiences difficulty to describe a sofa and after a disfluency draws an L-shape while verbally making the connection: "Ugh- can we get- uhh you know- those sofas that are L-shaped". The participant in Fig. 1.b similarly has difficulty to find a name for window seats: "It's one of those like window benches -uhh -the seat -the uhh". Again, large gestures with the hands to convey this message is observed. In such cases, ECAs could play an important role to cooperatively solve troubles of speaking either by displaying confusion until they are resolved or by asking questions such as "Do you mean a window seat?". This could be achieved by integrating simple computational models in state-of-the-art dialogue systems.

Even though the results are promising, the factor of constantly holding handheld controllers might have affected the gesture use of participants. We still observe plenty of gesture use, however, it is necessary to analyse the same features while participants are not holding anything that obstruct the hands. In future work, we will employ the same task without such devices and detect motion from video recordings using computer vision methods.

We underline the importance of utilising non-verbal signals of self-repair in ECAs, not only because that listening attentively or asking questions can correlate to perceived friendliness or intelligence of an agent [20] and improves rapport [12] but also, and more importantly, joint co-construction in which speaking and listening is at the core (over thinking and prediction) [21] results in richer and more robust interactions.

Table 1: Dependent Variable: Distance between Head and Right Hand with Fixed Effects (Disfluency), (Time Offset), random effects (Participants)

Variable	*Estimate*	*SE*	*t*	*p*
(Intercept)	**0.6101**	0.0152	40.0292	< **2e-16** ***
Disfluency1	**-0.0112**	0.0008	-13.3132	< **2e-16** ***
Time Offset	0.0010	0.0021	0.4894	0.6270
Disfluency1:Time Offset	**-0.0113**	0.0029	-3.9042	**9.46e-05** ***

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

Table 2: Dependent Variable: Distance between Head and Left Hand with Fixed Effects (Disfluency), (Time Offset), random effects (Participants)

Variable	*Estimate*	*SE*	*t*	*p*
(Intercept)	**0.6323**	0.0155	40.7758	< **2e-16** ***
Disfluency1	**-0.0055**	0.0007	-7.9054	**2.69e-15** ***
Time Offset	-0.0010	0.0020	-0.5069	0.6946
Disfluency1:Time Offset	**-0.0052**	0.0024	-2.1398	**0.0324** *

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

Table 3: Dependent Variable: Distance between Left and Right Hand with Fixed Effects (Disfluency), (Time Offset), random effects (Participants)

Variable	*Estimate*	*SE*	*t*	*p*
(Intercept)	**0.2129**	0.0105	20.1938	< **2e-16** ***
Disfluency1	**0.0090**	0.0009	10.3924	< **2e-16** ***
Time Offset	-0.0006	0.0025	-0.2599	0.7963
Disfluency1:Time Offset	**0.0216**	0.0030	7.2340	**4.72e-13** ***

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

6 CONCLUSION

In order to achieve intuitive and efficient HAI, the key findings and concepts from natural HHI research should inform social artificial agent applications [1, 21, 22]. The statement comes with the challenge of capturing the complex interplay of non-verbal multimodal signals humans display at the key moments of the interaction. A grounding-based system is considered to be feasible only if it can handle negative feedback [2]. Negative feedback is cued by non-verbal signals such as gestures [19, 37] and can be used to not only improve computational models for detecting self-repairs [26], but also to model the agent's contribution to the repair in a collaborative manner. Our results show that the pairwise distances in 3D positions of head and hands in speakers are significantly different during disfluencies, suggesting active gesturing and the use of larger gestures. The distances in head and hands could be used by ECAs as a detection method and improvement towards their contribution.

ACKNOWLEDGMENTS

Work supported by the EPSRC and AHRC Centre for Doctoral Training in Media and Arts Technology at Queen Mary University of London (EP/L01632X/1).

REFERENCES

[1] Agnes Axelsson, Hendrik Buschmeier, and Gabriel Skantze. 2022. Modeling Feedback in Interaction With Conversational Agents—A Review. *Frontiers in Computer Science* 4 (2022). https://doi.org/10.3389/fcomp.2022.744574

[2] Luciana Benotti and Patrick Rowan Blackburn. 2021. Grounding as a collaborative process. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, United States, 515–531.

[3] Susan Brennan. 2004. Conversation with and through computers. *User Modeling and User-Adapted Interaction* 1 (2004), 67–86.

[4] Susan E. Brennan and Michael F. Schober. 2001. How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language* 44, 2 (2001), 274–296. https://doi.org/10.1006/jmla.2000.2753

[5] Herbert H. Clark and Susan E. Brennan. 1991. Grounding in Communication. In *Perspectives on Socially Shared Cognition*, Lauren Resnick, Levine B., M. John, Stephanie Teasley, and D. (Eds.). American Psychological Association, 13–1991.

[6] Herbert H Clark and Jean E Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition* 84, 1 (2002), 73–111.

[7] Claudio Coppola, Diego R. Faria, Urbano Nunes, and Nicola Bellotto. 2016. Social activity recognition based on probabilistic merging of skeleton features with proximity priors from RGB-D data. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 5055–5061. https://doi.org/10.1109/IROS.2016.7759742

[8] Mark Dingemanse, Francisco Torreira, and N. Enfield. 2013. Is "Huh?" a Universal Word? Conversational Infrastructure and the Convergent Evolution of Linguistic Items. *PloS one* 8 (11 2013), e78273. https://doi.org/10.1371/journal.pone.0078273

[9] Diego R. Faria, Cristiano Premebida, and Urbano Nunes. 2014. A probabilistic approach for human everyday activities recognition using body motion from RGB-D images. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. 732–737. https://doi.org/10.1109/ROMAN.2014.6926340

[10] Diego R. Faria, Mario Vieira, Cristiano Premebida, and Urbano Nunes. 2015. Probabilistic human daily activity recognition towards robot-assisted living. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 582–587. https://doi.org/10.1109/ROMAN.2015.7333644

[11] Tom Gurion, Patrick G.T. Healey, and Julian Hough. 2020. Comparing models of speakers' and listeners' head nods. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue*. SEMDIAL, Whaltham, MA. http://semdial.org/anthology/Z20-Gurion_semdial_0013.pdf

[12] Joanna Hale and Antonia F. De C. Hamilton. 2016. Testing the relationship between mimicry, trust and rapport in virtual reality conversations. *Scientific Reports* 6, 1 (Oct. 2016), 35295. https://doi.org/10.1038/srep35295

[13] Patrick Healey, Mary Lavelle, Christine Howes, Stuart Battersby, and Rosemarie McCabe. 2013. How listeners respond to speaker's troubles.

[14] Patrick Healey, Nicola Plant, Christine Howes, and Mary Lavelle. 2015. When Words Fail: Collaborative Gestures During Clarification Dialogues.

[15] Patrick G. T. Healey, Gregory Mills, Arash Eshghi, and Christine Howes. 2018. Running Repairs: Coordinating Meaning in Dialogue. *Topics in cognitive science* 10 2 (2018), 367–388.

[16] Julian Hough and Matthew Purver. 2014. Strongly Incremental Repair Detection. *CoRR* abs/1408.6788 (2014). arXiv:1408.6788 http://arxiv.org/abs/1408.6788

[17] Julian Hough and David Schlangen. 2017. 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*. Association for Computational Linguistics, Valencia, Spain, 326–336. https://www.aclweb.org/anthology/E17-1031

[18] Julian Hough, Ye Tian, Laura de Ruiter, Simon Betz, Spyros Kousidis, David Schlangen, and Jonathan Ginzburg. 2016. Duel: A multi-lingual multimodal dialogue corpus for disfluency, exclamations and laughter. In *10th edition of the Language Resources and Evaluation Conference*.

[19] Christine Howes, Mary Lavelle, Patrick Healey, Julian Hough, and Rosemarie McCabe. 2016. Helping hands? Gesture and self-repair in schizophrenia.

[20] Maurits Kaptein, Panos Markopoulos, Boris Ruyter, and Emile Aarts. 2011. Two Acts of Social Intelligence: The Effects of Mimicry and Social Praise on the Evaluation of an Artificial Agent. *AI Soc.* 26, 3 (aug 2011), 261–273.

[21] Stefan Kopp and Nicole Krämer. 2021. Revisiting Human-Agent Communication: The Importance of Joint Co-construction and Understanding Mental States. *Frontiers in Psychology* 12 (2021). https://doi.org/10.3389/fpsyg.2021.580955

[22] Sebastian Loth and Jan P. De Ruiter. 2016. Editorial: Understanding Social Signals: How Do We Recognize the Intentions of Others? *Frontiers in Psychology* 7 (2016), 281. https://doi.org/10.3389/fpsyg.2016.00281

[23] Elizabeth Manrique and N. Enfield. 2015. Suspending the next turn as a form of repair initiation: evidence from Argentine Sign Language. *Frontiers in Psychology* 6 (2015), 1326. https://doi.org/10.3389/fpsyg.2015.01326

[24] Catharine Oertel, José Lopes, Yu Yu, Kenneth A. Funes Mora, Joakim Gustafson, Alan W. Black, and Jean-Marc Odobez. 2016. Towards Building an Attentive Artificial Listener: On the Perception of Attentiveness in Audio-Visual Feedback Tokens. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (Tokyo, Japan) *(ICMI '16)*. Association for Computing Machinery, New York, NY, USA, 21–28. https://doi.org/10.1145/2993148.2993188

[25] Maja Pantic, Anton Nijholt, Alex Pentland, and Thomas S. Huanag. 2008. Human-Centred Intelligent Human Computer Interaction (HCI2): How Far Are We from Attaining It? *Int. J. Auton. Adapt. Commun. Syst.* 1, 2 (Aug. 2008), 168–187. https://doi.org/10.1504/IJAACS.2008.019799

[26] Matthew Purver, Julian Hough, and Christine Howes. 2018. Computational Models of Miscommunication Phenomena. *Topics in Cognitive Science* 10, 2 (2018), 425–451. https://doi.org/10.1111/tops.12324 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/tops.12324

[27] Morteza Rohanian and Julian Hough. 2021. Best of Both Worlds: Making High Accuracy Non-incremental Transformer-based Disfluency Detection Incremental. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 3693–3703. https://doi.org/10.18653/v1/2021.acl-long.286

[28] Emanuel Schegloff. 1987. *Recycled turn beginnings; A precise repair mechanism in conversation's turn-taking organization.*

[29] Emanuel A Schegloff. 1992. Repair after next turn: The last structurally provided defense of intersubjectivity in conversation. *American journal of sociology* 97, 5 (1992), 1295–1345.

[30] Emanuel A. Schegloff. 1993. Reflections on Quantification in the Study of Conversation. *Research on Language and Social Interaction* 26, 1 (1993), 99–128. https://doi.org/10.1207/s15327973rlsi2601_5

[31] Elisabeth Schriberg. 1994. Preliminaries to a Theory of Speech Disfluencies.

[32] Michelle M.E. Van Pinxteren, Mark Pluymaekers, and Jos G.A.M. Lemmink. 2020. Human-like communication in conversational agents: a literature review and research agenda. *Journal of Service Management* 31, 2 (Jan. 2020), 203–225. https://doi.org/10.1108/JOSM-06-2019-0175 Publisher: Emerald Publishing Limited.

[33] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and Vision Computing* 27, 12 (Nov. 2009), 1743–1759. https://doi.org/10.1016/j.imavis.2008.11.007

[34] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. 2012. Mining actionlet ensemble for action recognition with depth cameras. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 1290–1297. https://doi.org/10.1109/CVPR.2012.6247813

[35] Dawid Warchoł and Tomasz Kapuściński. 2020. Human Action Recognition Using Bone Pair Descriptor and Distance Descriptor. *Symmetry* 12, 10 (2020). https://doi.org/10.3390/sym12101580

[36] Mao Ye, Qing Zhang, Liang Wang, Jiejie Zhu, Ruigang Yang, and Juergen Gall. 2013. *A Survey on Human Motion Analysis from Depth Data*. Springer Berlin Heidelberg, Berlin, Heidelberg, 149–187. https://doi.org/10.1007/978-3-642-44964-2_8

[37] Elif Ecem Özkan, Tom Gurion, Julian Hough, Patrick G.T. Healey, and Lorenzo Jamone. 2021. Specific hand motion patterns correlate to miscommunications during dyadic conversations. In *2021 IEEE International Conference on Development and Learning (ICDL)*. 1–6. https://doi.org/10.1109/ICDL49984.2021.9515613