Speakers Raise their Hands and Head during Self-Repairs in Dyadic Conversations

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

***Abstract*—People often encounter difficulties in building shared understanding during everyday conversation. The most common symptom of these difficulties are *self-repairs*, when a speaker restarts, edits or amends their utterances mid-turn. Previous work has focused on the verbal signals of self-repair, i.e. speech disfluences (filled pauses, truncated words and phrases, word substitutions or reformulations), and computational tools now exist that can automatically detect these verbal phenomena. However, face-to-face conversation also exploits rich non-verbal resources and previous research suggests that self-repairs are associated with distinct hand movement patterns. This paper extends those results by exploring head and hand movements of both speakers and listeners using two motion parameters: height (vertical position) and 3D velocity. The results show that speech sequences containing self-repairs are distinguishable from fluent ones: speakers raise their hands and head more (and move more rapidly) during self-repairs. We obtain these results by analysing data from a corpus of 13 unscripted dialogues, and we discuss how these findings could support the creation of improved cognitive artificial systems for natural human-machine and human-robot interaction.**

***Index Terms*—human behavior analysis, non-verbal communication, human motion analysis, human-robot interaction**

I. Introduction

THE growing interest in computational understanding of human social behaviour creates demand for new, efficient ways of detecting and quantifying different aspects of social interaction. One important area where current intelligent systems struggle is in detecting and recovering from misunderstandings [1], [2]. This is significant because misunderstandings are a ubiquitous, and arguably universal, feature of natural human interaction [3]–[6]. The basic interactive *repair* processes people use to deal with misunderstandings in conversation have been described in detail by conversation analysts [7]–[11] and there is experimental evidence that they underpin people's ability to adapt their communication to new situations and new tasks by revising and repairing the meaning of words and gestures on-the-fly [12], [13].

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).

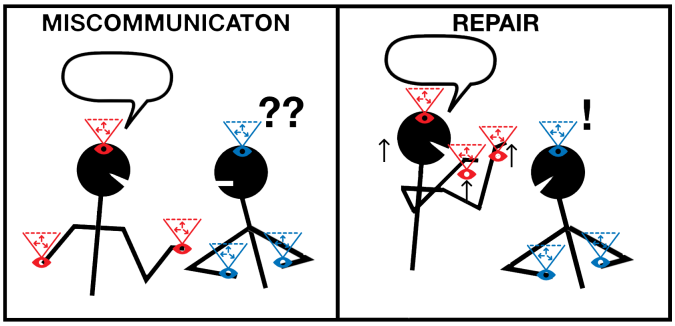

Fig. 1: Head and hands positions of speakers (red) and listeners (blue) during a dyadic conversation. We observed that the position of the head and hands of the speaker is generally higher during self-repair events.

Any intelligent system that aims to engage in natural interaction needs to be able to recognise and respond to repairs. The problem for engineering is that repairs require forms of real-time local adaptation that are difficult to achieve. For example, they require online, one-shot learning that is incompatible with large-scale offline learning from large numbers of examples of the kind used by contemporary machine learning approaches (e.g. large transformer models such as BERT [14]), current human-system and human-robot interaction is fragile especially when interaction follows an unpredicted trajectory and requires some form of collaborative recovery from a communication problem.

Repair processes in natural conversation have a systematic structure organised over sequences of multiple conversational turns [10]. They include complex abilities such as asking clarification questions and proposing paraphrases of another participant's turns. The most common forms of repair are self-repairs (sometimes referred to as *disfluencies* in the psycholinguistic literature) in which a speaker modifies what they are saying mid-turn [8], [15]. This can consist of restarts, repetitions, word substitutions, amendments and are often accompanied by filled pauses, such as "uhh" and "umm" that

provide a signal that a repair is in progress [16] (see below for more examples). Self-repairs in particular have received significant attention in natural language processing (NLP) and there are now systems that can recognise and parse them [15], [17], [18].

The NLP systems for capturing repairs focus on their verbal components. Recent work in Conversation Analysis has identified a number of non-verbal *embodied repair* processes. For example, the query or question face speakers produce to check ongoing understanding [19], the raised eyebrows a listener may produce to query something about what a speaker is currently saying [20], and various forms of distinctive hand movements such as 'cupping' the ear or a 'gesture freeze' as a form of clarification request [11], [21].

These signals are potentially attractive for engineering intelligent, interactive systems since, in principle, they can be detected from camera without requiring speech recognition. In addition embodied systems, such as robots and avatars, can potentially exploit non-verbal repair behaviours to make their interactions both richer and more robust, providing better routes to graceful failure. The recognition of specific facial expressions or gestures of embodied repair in the appropriate dialogue context is challenging.

Previous quantitative work has suggested that repairs are associated with distinctive patterns of hand movement [22]–[24] and, to a lesser extent, head movements [23]. More specifically, linear regression and mixed regression analyses suggest that speakers' hand heights are significantly higher during self-repairs, and height keeps increasing for 0.5 seconds after a repair [24]. This raises the question of whether simple motion parameters could be used to help identify the key moments in interaction when people are working to re-build shared understanding. This paper investigates the potential of measurements of changes in head and hand position and velocity as signals of self-repair. Three dimensional head and hand motion data obtained from a small corpus of task-oriented interactions is analysed to determine whether fluent speech can be distinguished from self-repaired speech on the basis of relative height and motion [24], [25]. The analysis addresses the following questions:

1) What are the specific non-verbal behaviour changes that are manifested in head and hand movement data during self-initiated self-repairs? Do they generalise to all such instances?
2) How do these changes compare to instances of fluent (un-repaired) interaction for a) speakers and b) listeners

The results we report suggest that self-repairs manifest themselves with distinct motion characteristics of speakers. Specifically, this work has three main contributions:

1) we extend our previous analysis of the speaker hand height [24] by analysing the 3D hand velocity as well (Sec. IV-A);
2) we perform a novel analysis of the head position and velocity of both speaker and listener (Sec. IV-B);
3) we discuss the observed hand and head motion patterns by proposing possible reasons for their emergence (Sec. IV-C).

The rest of the paper is organised as follows. In Sec. 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 describe the dataset and the techniques employed to analyse the data. Then, in Sec. IV we report the results; finally, in Sec. V we draw our conclusions and we outline possible cognitive artificial systems applications that could benefit from our results.

II. State-of-the-art in Social Sciences and Robotics

A. Multi-Modal Integration in Natural Conversation

Non-verbal signals are as connected to the meaning and message-to-be-conveyed as accompanying speech. They are an integral part of the *immediate communicative context* [26]–[28]. Together with words and prosody, embodied actions combine to form a *composite signal* [29], [30] or *integrated message* [28], [31]. The best known hand movements associated with communication are the content specific gestures, such as *iconic* and *metaphoric* gestures that help to convey the referential context of speech e.g. by manually drawing out a shape or depicting a spoken trajectory [32], [33]. These gestures are tightly connected with processes of speech production and are sometimes produced even when the recipient cannot see them e.g. when we are on the phone; even blind people addressing blind listeners sometimes gesture [33].

A different class of gestures, *interactive gestures*, play a key role not in articulating the content of speech but in managing the interaction [34]. These *conversational gestures* facilitate addressees' involvement by managing turn-taking, referring to previous contributions and, as introduced above, sometimes by using forms of embodied repair to signal problems or elicit help from interlocutors. The use of these conversational gestures is sensitive to whether people can see each other [34]; people used significantly more interactive gestures (at a higher rate) during dialogue than sequential monologue (in which they tell the story without help from each other).

For current purposes a useful feature of interactive gestures is they are much less dependent on the specific content of what is being communicated. For example, a gesture freeze or a turn-hand over 'you speak next' gesture can be identified without requiring an understanding of what the gesture freeze is querying or what the content of the next turn might be [21], [34]. This creates the possibility that useful information can be recovered from quite simple motion parameters. The simplest example of this is that speakers move their hands more, and more quickly, on average than listeners [22], [35]. This illustrates the potential for using simple non-verbal cues to determine important interaction states without requiring any speech processing.

Some motion paramaters have the potential to index quite complex features of interaction. Holler, Tutton and Wilkin [36] show how co-speech gestures are affected by the accumulation of common ground. Both use of words and gestures decrease overall as common ground – roughly, the level of mutual knowledge of common referents in a conversation– accumulates, however, the rate of gesturing increases (see also

[37]). More precisely, as the common ground accumulates the gesture/word rate increases, i.e. the use of gestures decrease less than words. This also underlines the important communicative functions of gesture use in adapting to the state of the conversation independently of content specific cognitive and lexical access theories of gesture production [36]. The findings regarding gestures in communication emphasise the significance of including the analyses of gestural information in the case of recognising, mimicking or computationally modelling natural communicative behaviours without forgetting to address language and interaction as composite, multimodal and social rather than a combination of distinct and static pieces of information. In a recent comprehensive overview, Holler [38] discusses the centrality of visual bodily signals in face-to-face human daily communication and how the vocal and visual modalities are tightly intertwined in language evolution and the coordination of minds.

B. Miscommunication Phenomena

Arguably, misunderstandings are one of the most critical moments in natural conversation [2]. Evidence from experimental and social psychology [13], [39], [40] highlights how patterns of misunderstanding and their resolution shape how shared meanings are created and maintained. People in conversation, in the form of conversational partners or groups *repair* misunderstandings on-the-go to update their mutual understanding, in other words common ground. Repair occurs in various forms and places in the dialog, which make it a difficult phenomena to investigate and model, but it has a systematic structure and distribution [4], [10].

An example of an incremental, collaborative, repair process involving non-verbal signals is provided by a simple experiment involving dyads in a collaborative task of building Lego blocks [41]. One participant is given the role of *Director* and provided with a paper-based drawing of a model whereas the other participant, the *Builder*, is in charge of constructing the model following the director's instructions and without seeing the drawing. The typical pattern of communication is: the *Director* gives an instruction; the *Builder* checks if they understand it by exhibiting and positioning the block before attaching it, possibly with a quizzical facial expression (clarification); if the position or the block selection is wrong, the *Director* discusses it with the *Builder* until the *Builder* finds the correct course of action. Notably, directors are able to change their sentences mid-way to correct builders actions, often in response to what the builder is doing. This concurrent, real-time monitoring of listener's actions and listener's non-verbal feedback enables them to adjust their instructions on-the-fly. In experiments such as these, people come up with creative solutions and short-cuts to the tasks they are given to collaborate efficiently by adapting their communication skills.

The procedures that people use to deal with "troubles of speaking, hearing and understanding" are referred to as *repairs* in Conversation Analysis [8]. We know that repairs are very frequent [42] in everyday interaction and they are "the only type of turn with unrestricted privilege of occurrence" [9]. They are also arguably universal across human languages (e.g. the utterence "huh" as a signal of misunderstanding [43]) and also occur in sign languages [44] and in graphical communication [12], [45]. The *Running Repairs Hypothesis* proposes 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" [13]. In effect, this treats negative interactional feedback as central to coordinating language use, and more important than positive feedback such as head-nods and utterences ("mmm", "hmm"), [13]. The idea is that the crucial points in natural interaction are about detecting and addressing misunderstandings in response to problems in the interaction.

C. Exploring Self-repair and its Detection as a Dynamic Strategy

The most frequent type of repair is when the speaker amends or modifies their own contribution while speaking, i.e. *self-repair* [4], [7]. Self-repairs provide useful information for co-ordination in dialogue [4], [40], and as outlined in [18] contribute to:

- compensating for misinformation and warning the listener of a change or amendment [40]
- assisting with syntactic analysis/re-analysis of language processing through grammaticality and ungrammaticality judgments [46]
- signal responses to the amount and type of listener's backchannel responses to the speaker [23], [27]

In the field of computational linguistics, self-repairs have been recognised as a significant component of spoken language, hence their automatic detection have been explored for robust natural language processing. State-of-the-art machine learning techniques have enabled [17] [18] effective detection of self-repairs (mostly in the case of disfluencies, i.e. self-initiated self-repairs) through dialogue transcripts with minimal error. However, it is important to acknowledge the limitations of these systems, as reviewed in [18]. The main challenge is to achieve the ability that enables humans to decide how to respond to or whether to initiate repair as and when they encounter problems naturally in dialogue. So far, the algorithmic approaches in natural language processing are not competent enough to decipher and/or replicate how humans recognise and intuitively act on these instances all the while speaking, listening and gesturing.

D. Investigating Non-verbal Behaviour during Repair

In face-to-face natural conversation, feedback –including repairs– can be displayed with multimodal non-verbal signals such as gaze [47], intonation [48], or gesture [49] [50] [23]. The relationship between embodied conduct (gaze and gestures) and the self-initiation of repair were examined in the case of word searches in Japanese conversation [51] [52]. Speech and visible acts by speakers during the course of word-finding difficulties demonstrate how the use of gestures can facilitate the recognition of missing words and promote relevant forms of co-participation. A more recent investigation of the combination of speech and iconic gestures during same-turn self-initiated repairs in Mandarin [53] discusses excerpts

from naturally occurring conversations in which repair operations consist of the integration of verbal and non-verbal components. "Gestural repairs" are defined as multi-stage self-initiated repair processes involving the use of iconic gestures. The investigated excerpts show that speakers involve hand gestures when they have trouble finding the right words in various ways. Iconic gestures show themselves as intentionally coincident repair solutions that speakers provide to their word-searching difficulties either within or after speech. A cross-linguistic study of Northern Italian, the Cha'palaa language of Ecuador, and Argentine Sign Language finds a generalised pattern of behaviours near repair instances: gesture holds (the hands, head, eyes, and upper body are held stationary for periods of time) occur until the onset of the repair solution; disengaging shortly after the solution [54].

Previous work has found strong correlation between repair and non-verbal behaviour. [49] revealed an association between increased self-repair and increased overall gesture in control groups containing healthy individuals. For the overall gesture calculations, they have worked with hand movement speeds. Encouraged by this finding, we previously studied and [24] have reported a significant increase in hand heights during self-initiated self-repairs.

Hand gestures are not the only indicator of repairs in natural conversation. A study investigating repair and non-verbal behaviour on a corpus of three person dialogues recorded with motion capture finds evidence of a correlation between head nods and self-repairs [23]. Cross-correlations of repair rate and rate of nodding shows speakers nod more than primary addressees or side participants in turns that include self-repairs. Both recipients also nod significantly more at the time the of the repairs, peaking at a 1-3 second offset. However, it is important to note that the detection and/or classification of specific movements such as head nods, particular hand gestures could prove to be problematic as it was pointed out by [25] that the methods for automatic detection of head nods are unreliable and the theories of head nods are generally underspecified.

Statistical evidence from discussed literature [49], [22], and [24] shows that, repair events in speech correspond to the utilisation of non-verbal signals in the same turn. Finding the quantifiable differences in movement data during repairs could provide advantages in automatically detecting self-repair instances as they happen and pave the way for creating more robust applications for social interaction.

E. State-of-the-art in Cognitive Robotics and HRI

The approaches in the previous decades for the design and implementations of mental processes in social collaborative robots were mostly rooted in concepts such as human theory of mind, perspective taking, embodied cognition or highly integrated models (as can be seen in [55]). Implementing an understanding of human attention and intention as the goal, human emotion recognition through facial expressions has gained significant interest. Facial expressions are an important component of non-verbal communication, and in fact multiple research efforts have been devoted to developing automatic recognition systems that detect facial expressions and associate them to specific emotions [56]. Recently, the focus on facial expressions in emotion recognition has extended to include full body movements, taking into account findings from cognitive sciences [57]. Along with feature selection for detection, systems should aid enhancing conversations with humans especially when creating a shared context between robots and humans, i.e. the common ground.

To further enable natural and effective interaction in a dialogue setting we also need to account for the interpersonal and interactive aspects of natural conversation as outlined in Sections II-A - II-B. Examples to such aspects are grounding [58], [59] and backchannelling [60], [61], in correlation with the findings from natural human-human interaction research. For instance, *the affective grounding* perspective for HRI, suggested by [59], aims to extend previous accounts in affective computing by recognising interaction as a jointly coordinated interpersonal process and taking conversational analytic (CA) notions into account by implementing back-channel responses and repair to regulate shared understanding and attention between human and robots. More recent work has tackled detecting grounding problems and miscommunication in a physically situated dialogue context that analyzes features from spoken language input for navigation [62].

Work in HRI has recognised the importance of findings from the literature in natural human dialogue and has begun to incorporate them in their applications. The technical possibilities afforded by substantially improved human motion tracking technologies using vision depth sensors in the past ten years will extend this further [63]. However, there has not yet been an attempt to use human motion data to detect instances of repair in natural dialogue through artificial (robotic or computer) systems.

If the artificial conversational partner needs to know what motion patterns to look for, it can generate specific interventions (e.g. recheck the content of a sentence, which might have been repaired on-the-go) or continue to update its processing; therefore, research that aims at identifying what motion patterns correlate to miscommunications is extremely important to advance the field. Moreover, obtaining and parsing motion data has advantages in situations where speech recognition fails (noisy environments) and easier for computational tools in real-time compared to advanced NLP modules.

III. METHODOLOGY

The finding that speakers' hand heights are significantly higher during self-repair instances [24] motivates the investigation of other motion patterns in miscommunication windows using the same dataset described in [25]. This consists of natural face-to-face conversation of 13 dyads (Fig. 2). 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. For details of this task, see: [64]. The profile of the participants is summarised in Table I. From the recorded sessions we have extracted *disfluency windows* and *fluent windows* for comparison (details of labelling, window extraction and filtering for

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

statistical analyses are explained in following sub-sections, previous work [24] and summarised in Fig. 3).

TABLE I: Dataset Details and Analysis Windows

Participant Details		
Gender	Female: 14	Male: 12
Age	Range: 18-26	Mean: 20.8, Std: 1.9
Number	Total: 26	Pairs: 13
Recorded Data		
Audio	44.1 kHz 24bit	2.5 GB (approx.)
Video	1280x720 at 25 fps	50 GB (approx.)
Motion capture	Avg. sample rate: 89 fps	
Length	Per Session: 15 min	No of sessions: 13
Analysis Windows		
	Disfluency	Fluent
Preliminary	6359	7117
Filtered	2076	3557

The previous study [24] is extended by analysing additional movement data: hand velocity, head height and head velocity. The methodology for selecting the windows of motion data to be analysed is identical to the previous method [24]. In summary, we use the disfluency labels obtained from the automatic disfluency detection tool [65] for the start of repair instances, as the dataset is not manually annotated for miscommunication. Taking the disfluency timestamps as the centre, we constuct windows starting from 2 seconds before to 2 seconds after the disfluency start timestamp. The movement features are then inspected by looking at the mean and variance of all these features in all speakers and listeners (that are a dyad subjected to the detection of the floor control algorithm) within these windows. The data analysis pipeline is further summarised in Fig. 3.

A. Expected Outcomes

In order to observe additional motion patterns during self-repairs in speakers and listeners, the same windows (i.e. disfluent and fluent) should be compared. The finding that there are changes in hand height [24] and hand gesture rate [49] during repairs suggests that other motion features (such as head motion data) could show quantifiable differences.

Low level features such as movement and velocity are less prone to problems of definition than semantic concepts such as head nods, shakes, and gestures. For example, up and down movement of the head, that is commonly associated with head nods, is often found in other behaviours such as laughter [66].

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

Manually annotating repairs is a difficult and time-consuming task, especially when we consider the high frequency and variability of repairs in natural conversation. Hence, labelling miscommunication events with automatic tools is a crucial step for more reliable and practical research [18]. Therefore, in the previous study [24] we have chosen to work with an automatic disfluency detection tool [65] to detect disfluencies (i.e. self-initiated self-repairs). This tool follows the assumption of *reparandum-interregnum-repair* structure [15] in speech repairs and incrementally detects these structures:

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

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

The disfluency detection tool which combines a deep learning system for sequence labelling and incremental decoding techniques outperforms previous state-of-the-art models when trained on the Switchboard dataset [67]: F1 scores for *'e' (edit)* tags is above 0.9 and for *'rpS' (repair onset)* tags above 0.75 [65]. We have used this model by inputting the IBM Watson [1] speech-to-text service outputs to automatically label disfluencies (repair onset or related words) in the speech texts. The timestamps of the disfluency labels were taken as the exact moment of disfluencies in the dialogue. Our previous method has provided statistical differences over speakers' hand heights during disfluency instances. Therefore, we are expanding our movement analysis of the disfluent instances constructed with the same timestamps of the automatic disfluency labels.

In order to construct the labels in our dataset, the detection tool was used in the simple mode and labelled utterances as *edit (e)*, *fluent (f)* and *repair onset (rpS)* term, detailed in [24]. *edit (e)* or *repair onset (rpS)* tags were taken as disfluency labels with their corresponding time stamps. There are 3192 disfluency instances ($M = 122.769, sd = 40.078$, over 26 participants) over the whole dataset which consists of 6:37 hours of natural conversation in total. As in the case of previous study [24], we have investigated the disfluent timestamps and the movement features during these instances and compared them with windows not containing any labels (hence fluent).

C. Automatic Detection of Speakers and Listeners

A simple floor control detection model [24] was used to process the audio captured from the participants' microphones.

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

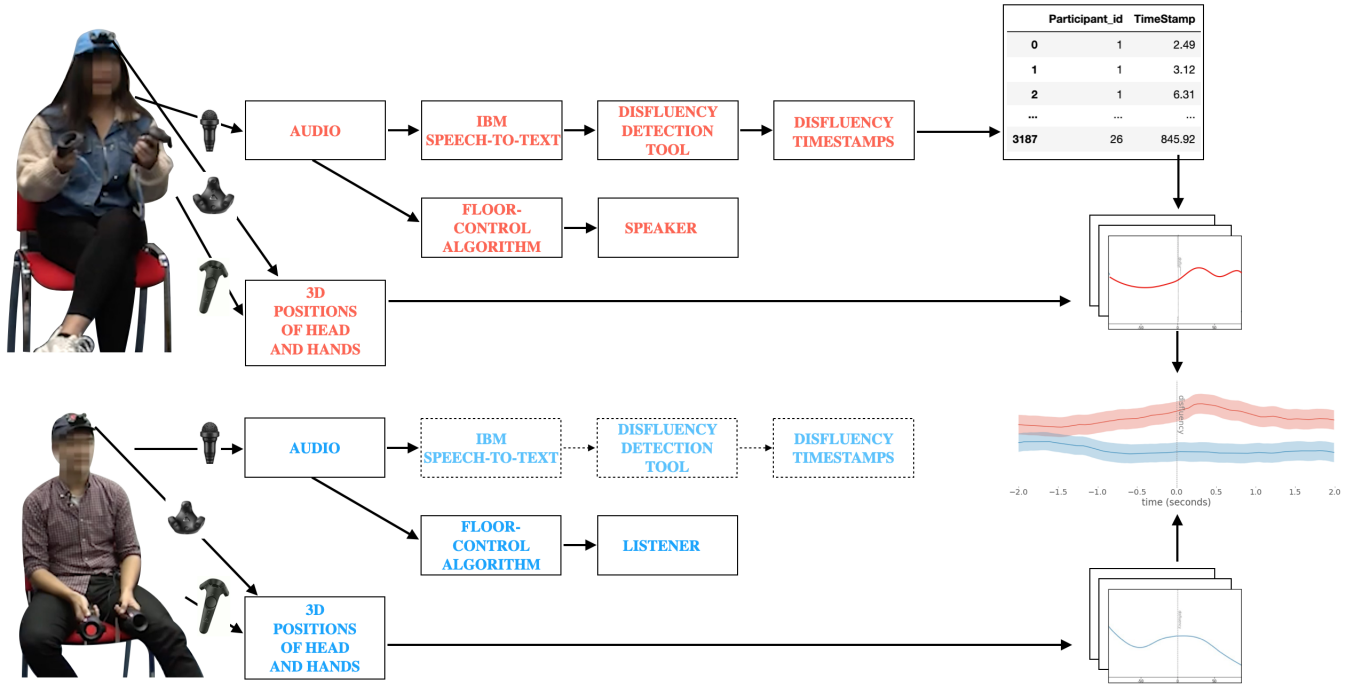

Fig. 3: Data analysis pipeline: The audio of the pairs is converted to text by IBM Speech-to-text and sent to Disfluency Detection Tool. The timestamps for the disfluency labels is used to analyse the motion data captured by HTC Vive trackers at those timestamps within windows of 4 seconds. The speaker and listeners are differentiated from the audio signals with floor control detection algorithm.

It determines who is the speaker at any given moment. For the same timestamp, the other participant that is the conversational partner (who is in the same session) is labelled to be the listener. This model, detailed in [24], employs simple audio processing techniques such as low-pass filters and a thresholds. All audio is processed in buffers of 0.02 seconds and for each buffer the root mean square (RMS) value is calculated. Then, these RMS values are filtered by low pass filters with a cutoff frequency of $0.35Hz$. When the difference between the minimal and the maximal filtered RMS values is larger than 0.1 the participant who has the maximal filtered RMS value is identified as the speaker. If this is not the case, the previously reported speaker is continued to be labelled as the speaker.

D. Hand and Head Movement Patterns during Fluent and Disfluent Moments in Conversation

In order to compare movement features in the presence and absence of self-initiated self-repairs (disfluencies), time-series windows of the hand height and velocity, head height and velocity features are generated for instances that contain self-repairs. Timestamps from the automatic tagging are taken as the centre of the repairs ($time = 0$), and an equal number of motion readings are taken either side depending on the window length (e.g. -2 seconds and $+2$ seconds for a four second window). In order to capture the motion data that accompanies fluent (unrepaired) sections of speech, sections that are at least 6 seconds after the end time of a previous repair tag and before the next are extracted (including a buffer of 1 second). Here, the aim was to exclude any movement that might have been related to a previous repair. The fluent sections were also split into 4-second windows, resulting in 3,557 instances to be analysed. Both repaired (disfluent) and fluent windows of 4 seconds have the sampling period of 10 ms., resulting in 400 corresponding motion readings for each window. Each window is also labelled either as a speaker or a listener window based on the output of the floor control detection algorithm discussed in Sec. III-C at the middle timestamp of the window.

In previous work [24], we explored the mean and variance of hand heights in repaired (disfluent) and fluent windows before proceeding to a statistical analysis. We observed an increase in mean hand heights within 0-0.5 seconds of a repair label and statistical analysis suggested this pattern was reliable. Following the same procedure, we report the mean and variance comparison figures for three additional movement features: hand velocity, head height and head velocity. The disfluency windows (the mean and variance of 6359 instances) and fluent windows (the mean and variance for 7117 instances) for all motion features are displayed in Fig. 4 for comparison.

The hand heights feature is calculated for the highest of the right and left hand readings (Y-position) from the hand trackers. The highest (also referred to as maximum) hand is chosen as the active hand for velocity calculation. Following [68], hand velocity is calculated as 3D position changes over a sample of each participant for each time t as:

$$V(t) = \sqrt{(\frac{x_{t+1} - x_{t-1}}{2S})^2 + (\frac{y_{t+1} - y_{t-1}}{2S})^2 + (\frac{z_{t+1} - z_{t-1}}{2S})^2} \quad (1)$$

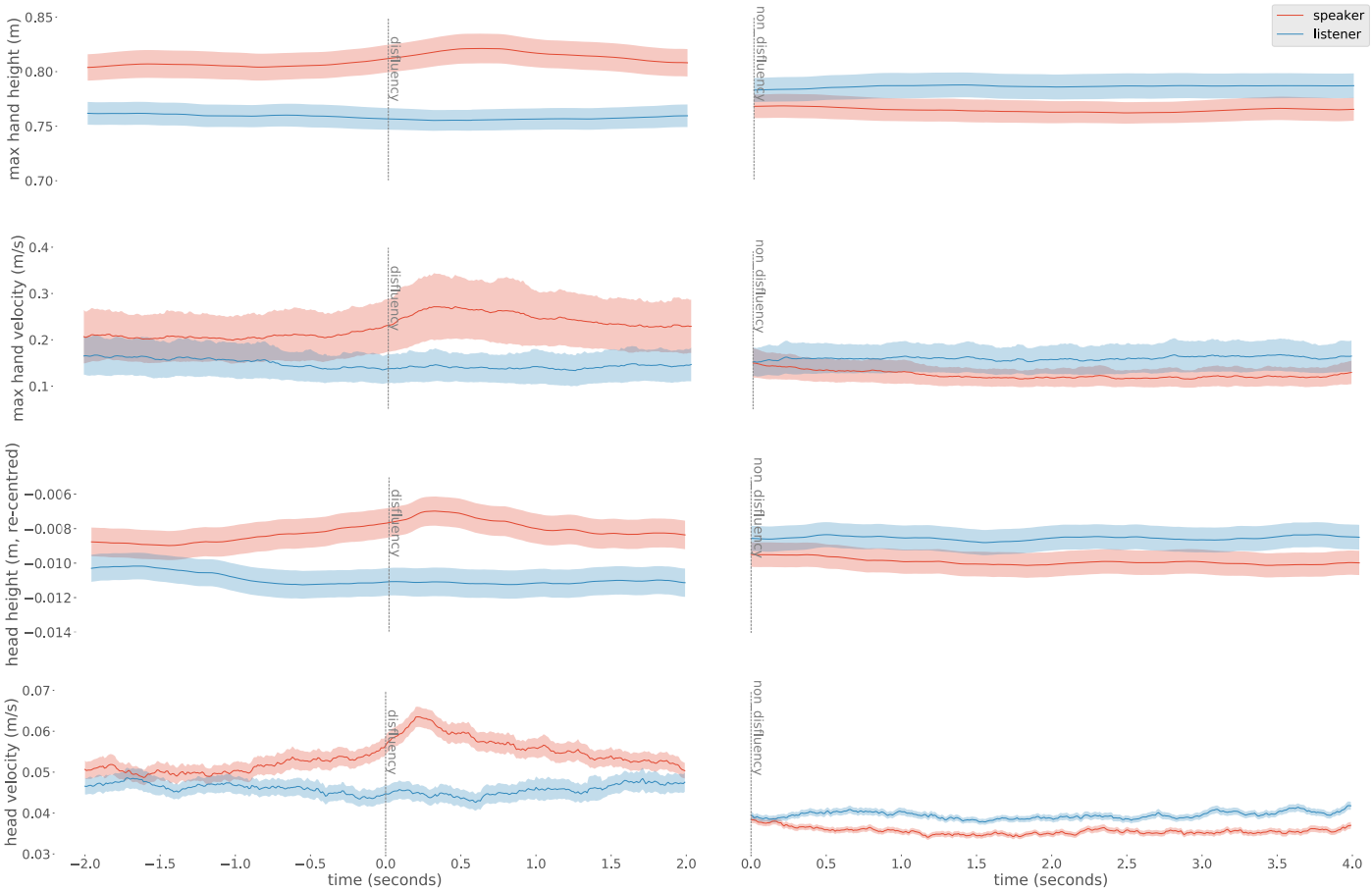

Fig. 4a: Disfluency Windows

Fig. 4b: Fluent Windows

Fig. 4: Mean (line) and Variance (shades) of Hand and Head Features for Disfluency and Fluent 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). In the case of speakers (in red), we observe an increase in hand features starting at the disfluency moment and continuing to increase approximately for 0.5 seconds. For head features, this increase is for approximately 0.25 seconds. In comparison, during fluent instances all hand and head features seem to be stationary both in speakers and listeners.

where x_{t-1}, y_{t-1}, and z_{t-1} are the positions along the three spatial axes at one sample prior to time t, and S is the interval of time between samples, $S = 1/100$ sec. The resulting velocity vectors are used as the motion feature.

Preliminary findings for maximum hand velocities during repaired and fluent windows show differences in mean and variance over time (Fig. 4) with a similar pattern to hand heights. For speakers, mean hand velocity starts increasing just before a repair and continues until 0.5 seconds after the repair. During fluent speech, speaker hand velocities are lower than listener hand velocities in both window types.

Head height and head velocity were also investigated within the fluent and disfluent instances. The windows of head motion data are generated identically to the hand motion. Head height measurements are normalised by subtracting each participant's initial height to prevent the differences in participant height intervening with the behavioural head height fluctuations we are looking for. Head velocity is calculated the same way as hand velocity, 3D position changes over a sample, as suggested [68] and presented in equation 1.

The comparison of head heights and head velocities during repaired and fluent windows shows a similar pattern but with more variation. Head heights in speakers in disfluency windows seem to start increasing from around 1.5 seconds before the disfluency label and continue increasing until around 0.25 seconds after the disfluency. Listeners' head heights decrease slightly between -2.0 and -1.0 seconds during disfluencies. Speaker head velocity shows a very similar pattern to head heights, congruent with the increase trend. Head velocities in listeners (both cases) and in speakers during fluent windows fluctuate throughout.

To determine the significance of these observations, we performed mixed linear regressions for speakers' head and hand feature windows of 0.5 seconds right after disfluency ($disfluency = 1$) and fluent windows of 0.5 seconds ($disfluency = 0$), where we have observed the substantial changes that are common in both heads and hands. 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. The number of fluent windows were reduced to

3557 samples by random selection. These samples were used for the mixed model regression analyses of all features (2076 disfluency, 3557 fluent).

IV. Results

The differences in head and hand movement features during disfluent and fluent instances were investigated by performing separate linear regressions on these subsets and mixed linear regression models. For all models, time offsets from 0 to 0.5 seconds were used, with a sampling period of 10 ms. In the case of disfluency windows, this correlates to the disfluency timestamp as the start. In the fluent cases, there is no particular starting point of interest so any window of 0.5 seconds length is sufficient for the comparison. For simplicity, we have selected the start of each fluent window. We present each movement feature along with their linear regression analyses.

A. Hand Movement Features

In our previous study, in order to analyse **hand height** of speakers in disfluent and fluent windows, we have performed two independent linear regressions over isolated sub-groups and a mixed linear regression over the whole data [24]. We have found that speakers' hand heights were significantly higher during disfluencies compared to maximum hand heights of fluent instances; this is confirmed by the results in Table II. The *Estimate* values, (intercept being the mean of hand height) show how these are 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; 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 ($TimeOffset = -0.0010, p > 0.1$).

In this study we add an analysis of **hand velocities**. To investigate the significance of observed patterns in hand velocities, we have performed a mixed model regression analysis that models the hand velocity 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 whose hand velocity was considered as a random factor. Table III shows hand velocities are significantly higher during disfluencies in comparison to hand velocities in fluent windows, as it can be seen from the estimate values. The overall mean of hand velocity (denoted by intercept) (0.1698), is 0.0377 higher in the case of disfluencies. The increase in hand velocity between 0-0.5 seconds is proven by *Disfluency1:Time Offset* variable being 0.1164, meaning the slope of handvelocity/disfluency is 0.1164 higher when the disfluencies are present.

The following equations quantify the average initial hand velocity, increase and decrease amount over time in fluent (2) and disfluent (3) cases:

$$0.1698 - 0.0181t \tag{2}$$

$$0.1698 + 0.0377 + (0.1164 - 0.0181)t \tag{3}$$

B. Head Movement Features

Similarly to the way in which we analyse hand features, we study head height and head velocity as the dependent variables for regression, and we report it below.

Mixed model regression analysis of the **head height** based as a function of two fixed factors, i.e. disfluency presence and time offset, where the participant number being a random factor shows similar characteristics of the results for hand height analysis of the same kind. Table IV substantiates that speakers' head heights were significantly higher during disfluencies. Looking at the estimate values, the mean (intercept) of head height (which is -0.0095) is higher when there is a disfluency (compared to fluent window mean) by (0.0019). Moreover, following the trend, there is an increase in head heights from the disfluency starting point as corroborated by *Disfluency1:Time Offset* variable, i.e. the slope of head-height/disfluency, is 0.0020 higher when the disfluencies are present. In the fluent cases the head heights are stationary but not significantly ($TimeOffset = 0, p > 0.1$).

Same analysis for **head velocity** has shown significantly higher head velocity means during disfluencies ($Disfluency1 = 0.0085, p < 0.01$) but no significant upward trend with the time offset that was present in the height feature. This might be due to our selection of time window. For the head movement features analyses, we have selected the same time window as the hands, which is 0-0.5 seconds after disfluency. However, in Sec. III-D we had observed the increasing trends in head features between (-1.5) and (0.25) time windows. There was a decrease in head features after 0.25. This results in not capturing the increasing trend in the regression analyses. We have updated our analyses by taking these sections for the head height feature, however only got significant effect for disfluencies ($Disfluency1 = 0.0017, p < 2e-16$) which is still an important finding.

C. Discussion

We discuss here the possible underlying reasons for the observed motion patterns, also in relationship to previous findings in the literature.

The increase in hand heights could indicate to speakers' use of gesture space in correlation with the visibility/attention field of the listener [32]. In other words, the speaker, aware of the current lack of understanding of the listener, instinctively moves the hands within the line of sight of the listener, to call for their attention, in an attempt to make the self-repair event more effective. Our analysis of the 3D hand velocity implies faster and more frequent hand movements during repair instances. This is consistent with general knowledge in the literature that hand movements are very important to support verbal communication, and in fact the number of hand gestures does not decrease as much as the number of spoken words as the common ground accumulates during conversations [36]; therefore it is to be expected that more hand movements appear when there are troubles in the communication, and this is consistent with previous findings showing the correlation between repair rates and hand gestures [49].

TABLE II: Dependent Variable: Hand Height with Fixed Effects (Disfluency), (Time Offset), random effects (Participants). The Estimates show an increase in the mean when Disfluency Variable is 1, they also increase further over time as shown by Disfluency1:TimeOffset variable.

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.

TABLE III: Dependent Variable: Hand Velocity with Fixed Effects (Disfluency), (Time Offset), random effects (Participants). The Estimates show an increase in the mean when Disfluency Variable is 1, they also increase further over time as shown by Disfluency1:TimeOffset variable.

Variable	***Estimate***	***SE***	***t***	***p***
(Intercept)	**0.1698**	0.0171	9.9571	**3.3e-10** ***
Disfluency1	**0.0377**	0.0022	16.7907	$<$ **2e-16** ***
Time Offset	-0.0181	0.0098	-1.8478	0.0746
Disfluency1:Time Offset	**0.1164**	0.0077	15.0669	$<$ **2e-16** ***

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

TABLE IV: Dependent Variable: Head Height with Fixed Effects (Disfluency), (Time Offset), random effects (Participants). The Estimates show an increase in the mean when Disfluency Variable is 1, they also increase further over time as shown by Disfluency1:TimeOffset variable.

Variable	***Estimate***	***SE***	***t***	***p***
(Intercept)	-0.0095	0.0069	-1.3765	0.1809
Disfluency1	**0.0019**	0.0001	12.6171	$<$ **2e-16** ***
Time Offset	0.0000	0.0003	0.0269	0.9786
Disfluency1:Time Offset	**0.0020**	0.0005	3.8317	**0.000128** ***

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

Our analysis of head movements shows an increase of head height (and higher movement speed) of the speaker during self-repairs. Note that participants are sitting while they engage in the conversation. Therefore, these results suggest a motion pattern in which the head is quickly lifted, probably as a consequence of the speaker straightening up the back or readjusting on the chair; an example of this is shown in Fig. 2, in which the speaker (images on the left) switches from the fluent (top image: relaxed on the chair, hands down) to the disfluent (bottom image: straight back, hands up) case. This behavior could be triggered by a desire to make themselves more visible to the listener (by straightening up the back) and to mentally reset (by readjusting on the chair) before re-starting one part of the conversation that needs to be explained again because of the lack of understanding by the listener. This would be a different motion pattern than e.g. head nods, that have been shown to be used by listeners as positive feedback to help speakers in trouble [25]. This interpretation of the head motion data is partially supported by another indication. The overall mean of the head height reported in Table IV is a negative value: (-0.0095). This was not unexpected given the data normalization procedure we execute with participants and their general behaviour during the procedure. In order to normalise the head height, as explained in Sec. III, we record the initial position of the head, when the participant is instructed to sit still for a few seconds after the recording is

TABLE V: Dependent Variable: Head Velocity with Fixed Effects (Disfluency), (Time Offset), random effects (Participants). The Estimates show an increase in the mean when Disfluency Variable is 1. The increase further over time as shown by Disfluency1:TimeOffset variable is not significant in the head velocity feature.

Variable	***Estimate***	***SE***	***t***	***p***
(Intercept)	**0.0485**	0.0042	11.6070	**9.84e-13** ***
Disfluency1	**0.0085**	0.0014	6.1591	**7.33e-10** ***
Time Offset	-0.0035	0.0023	-1.5342	0.133
Disfluency1:Time Offset	0.0028	0.0017	1.6227	0.105

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

started. However, participants sit in a more stiff manner, with their back straight, when instructed to "sit still", and they then relax on the chair after the conversation starts, with the head position being lower. This explains the negative overall mean of head heights. At the same time, it gives them room to lift the head higher if they want to, which they apparently do during self-repair events.

An important limitation of this study is that holding the large handheld controllers could have affected natural gesturing of the participants even though we observe plenty of gestures while holding the controllers. In future work, we plan to replicate the same task without the use of such large sensors. In the next data collection, the participants will be recorded using depth-cameras and the 3D positions of head and hands are to be extracted using computer vision methods.

V. Conclusions

This paper identifies some specific human non-verbal behaviors that appear to be correlated to self-repair events during natural dyadic conversations. In particular, we extend our previous study in which we showed how speakers raise their hands higher during self-repair events [24], and we report here a detailed analysis of the position and velocity of the hands and head of both speakers and listeners. Our results demonstrate that also the head position of the speaker raises during self-repair events, as well as the hands and head velocity. The results reinforce the idea, coming from previous findings in cognitive and behavioural sciences, that human conversation has to be treated as an interactive and multimodal process, and that artificial cognitive systems based on this approach could obtain a more effective human-machine and human-robot interaction in natural settings. This is facilitated by the fact that the automatic tracking (and even prediction) of human movements has improved substantially during the last two decades [69], and therefore it is reasonable to conceive artificial systems which integrate the detection of specific movements with the analysis of an audio voice signal. In particular, because self-repair events are a sign of some miscommunication happening (i.e. the speaker *repairs* their utterance because the listener has signalled a problem with understanding), artificial systems that could detect such events will be more aware of the context during a conversation, leading to better artificial conversational agents. Potential applications include social and service robots (in e.g. care facilities, education, entertainment) as well as computer programs that monitor conversations between a human and an automated agent, or between two humans (e.g. patient-doctor interactions), in order to recommend specific interventions or to generate reports. Furthermore, these findings could be specifically used to improve the performance of NLP tools detecting disfluencies by integration of head and hand positions modalities.

References

[1] P. Healey, "Human-like communication," in *Human-Like Machine Intelligence* (S. Muggleton and N. Chater, eds.), pp. 137–151, Oxford University Press, Oxford, England, 2021.

[2] P. G. T. Healey, J. P. de Ruiter, and G. J. Mills, "Editors' Introduction: Miscommunication," *Topics in Cognitive Science*, vol. 10, pp. 264–278, Apr. 2018.

[3] E. A. Schegloff, "Reflections on quantification in the study of conversation," *Research on language and social interaction*, vol. 26, no. 1, pp. 99–128, 1993.

[4] M. Colman and P. Healey, "The distribution of repair in dialogue,"

[5] M. Dingemanse, F. Torreira, and N. J. Enfield, "Is "huh?" a universal word? conversational infrastructure and the convergent evolution of linguistic items," *PloS one*, vol. 8, no. 11, p. e78273, 2013.

[6] M. Dingemanse, S. G. Roberts, J. Baranova, J. Blythe, P. Drew, S. Floyd, R. S. Gisladottir, K. H. Kendrick, S. C. Levinson, E. Manrique, *et al.*, "Universal principles in the repair of communication problems," *PloS one*, vol. 10, no. 9, p. e0136100, 2015.

[7] 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.

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

[9] 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.

[10] E. A. Schegloff, "Repair after next turn: The last structurally provided defense of intersubjectivity in conversation," *American journal of sociology*, vol. 97, no. 5, pp. 1295–1345, 1992.

[11] K. H. Kendrick, "Other-initiated repair in english," *Open Linguistics*, vol. 1, no. 1, 2015.

[12] P. Healey, "Interactive misalignment: The role of repair in the development of group sub-languages," *Language in Flux. College Publications*, vol. 212, 2008.

[13] 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.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018.

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

[16] H. H. Clark and J. E. F. Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, no. 1, pp. 73–111, 2002.

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

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

[19] S. McCullough and K. Emmorey, "Categorical perception of affective and linguistic facial expressions," *Cognition*, vol. 110, no. 2, pp. 208–221, 2009.

[20] M. L. Flecha-García, "Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in english," *Speech communication*, vol. 52, no. 6, pp. 542–554, 2010.

[21] F. Oloff, ""sorry?"/"como?"/"was?"–open class and embodied repair initiators in international workplace interactions," *Journal of Pragmatics*, vol. 126, pp. 29–51, 2018.

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

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

[24] E. E. Özkan, T. Gurion, J. Hough, P. G. Healey, and L. Jamone, "Specific hand motion patterns correlate to miscommunications during dyadic conversations," in *2021 IEEE International Conference on Development and Learning (ICDL)*, pp. 1–6, 2021.

[25] 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.

[26] J. B. Bavelas and N. Chovil, "Nonverbal and verbal communication: Hand gestures and facial displays as part of language use in face-to-face dialogue.," 2006.

[27] J. Bavelas, J. Gerwing, and S. Healing, *Hand and Facial Gestures in Conversational Interaction*, pp. 111–130. 09 2014.

[28] J. Bavelas and N. Chovil, "Visible acts of meaning: An integrated message model of language in face-to-face dialogue," *Journal of Language and Social Psychology - J LANG SOC PSYCHOL*, vol. 19, pp. 163–194, 06 2000.

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

[30] R. A. Engle, *Toward a theory of multimodal communication: Combining speech, gestures, diagrams, and demonstrations in instructional explanations*. Stanford University, 2000.

[31] L. Drijvers and A. Özyürek, "Visual context enhanced: The joint contribution of iconic gestures and visible speech to degraded speech comprehension," *Journal of Speech Language and Hearing Research*, vol. 60, pp. 1–11, 12 2016.

[32] D. Mcneill, "Hand and mind: What gestures reveal about thought," *Bibliovault OAI Repository, the University of Chicago Press*, vol. 27, 06 1994.

[33] D. Mcneill, *Gesture and Thought*. 01 2005.

[34] J. Bavelas, N. Chovil, L. Coates, and L. Roe, "Gestures specialized for dialogue," *Personality and Social Psychology Bulletin*, vol. 21, pp. 394–405, 04 1995.

[35] E. A. Schegloff, "On some gesture's relation to talk," *Structures of social action: Studies in conversation analysis*, pp. 266–296, 1984.

[36] J. Holler, M. Tutton, and K. Wilkin, "Co-speech gestures in the process of meaning coordination," 2011.

[37] J. Holler and K. Wilkin, "Communicating common ground: How mutually shared knowledge influences speech and gesture in a narrative task," *Language and Cognitive Processes*, vol. 24, no. 2, pp. 267–289, 2009.

[38] J. Holler, "Visual bodily signals as core devices for coordinating minds in interaction," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 377, 2022.

[39] 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.

[40] 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.

[41] H. H. Clark and M. A. Krych, "Speaking while monitoring addressees for understanding," *Journal of Memory and Language*, vol. 50, pp. 62–81, Jan. 2004.

[42] 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.

[43] 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.

[44] 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.

[45] P. G. Healey, N. Swoboda, I. Umata, and J. King, "Graphical language games: Interactional constraints on representational form," *Cognitive science*, vol. 31, no. 2, pp. 285–309, 2007.

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

[47] A. Hjalmarsson, C. Oertel, and K. Speech, "Gaze direction as a backchannel inviting cue in dialogue," 01 2012.

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

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

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

[51] M. Hayashi, "Language and the body as resources for collaborative action: A study of word searches in japanese conversation," *Research on Language and Social Interaction*, vol. 36, pp. 109–141, 04 2003.

[52] T. Ono and M. Hayashi, "Joint utterance construction in japanese conversation," *Japanese Language and Literature*, vol. 39, p. 419, 10 2005.

[53] R.-J. R. Wu, "Gestural repair in mandarin conversation," *Discourse Studies*, vol. 0, no. 0, p. 14614456211037451, 2021.

[54] S. Floyd, E. Manrique, G. Rossi, and F. Torreira, "Timing of visual bodily behavior in repair sequences: Evidence from three languages," *Discourse Processes*, vol. 53, no. 3, pp. 175–204, 2016.

[55] 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.

[56] N. Samadiani, G. Huang, B. Cai, W. Luo, C.-H. Chi, Y. Xiang, and J. He, "A review on automatic facial expression recognition systems assisted by multimodal sensor data," *Sensors (Basel, Switzerland)*, vol. 19, 2019.

[57] H. Gunes, C. Shan, S. Chen, and Y. Tian, *Bodily Expression for Automatic Affect Recognition*, ch. 14, pp. 343–377. John Wiley Sons, Ltd, 2015.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] M. Marge and A. I. Rudnicky, "Miscommunication detection and recovery in situated human–robot dialogue," *ACM Trans. Interact. Intell. Syst.*, vol. 9, feb 2019.

[63] L. Chen, H. Wei, and J. Ferryman, "A survey of human motion analysis using depth imagery," *Pattern Recognition Letters*, vol. 34, pp. 1995–2006, 11 2013.

[64] 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.

[65] 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.

[66] H. J. Griffin, M. S. Aung, B. Romera-Paredes, C. McLoughlin, G. McKeown, W. Curran, and N. Bianchi-Berthouze, "Laughter type recognition from whole body motion," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 349–355, IEEE, 2013.

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

[68] S. Boker, M. Xu, J. Rotondo, and K. King, "Windowed cross-correlation and peak picking for the analysis of variability in the association between

behavioral time series," *Psychological methods*, vol. 7, pp. 338–55, 10 2002.

[69] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras, "Human motion trajectory prediction: a survey," *The International Journal of Robotics Research*, vol. 39, no. 8, pp. 895–935, 2020.