The Discontent with Intent Estimation In-the-Wild

The Case for Unrealized Intentions

Hung, Hayley; Li, Litian; Molhoek, Jord; Zhou, Jing

DOI
10.1145/3613905.3644055

Publication date
2024

Document Version
Final published version

Published in
CHI EA '24

The Discontent with Intent Estimation In-the-Wild: The Case for Unrealized Intentions

Hayley Hung
h.hung@tudelft.nl
Delft University of Technology
The Netherlands

Litian Li
wolilitian@gmail.com
Delft University of Technology
The Netherlands

Jord Molhoek
jordm@live.nl
Delft University of Technology
The Netherlands

Jing Zhou
jiczhou10@gmail.com
Delft University of Technology
The Netherlands

Figure 1: Illustration of the intention by outcome problem. (a): Common setups involve an individual grabbing objects with gaze and posture. (b): When 2 or more people operate with competing goals, realized as well as unrealized intentions can occur. (c): Example of a crowded in-the-wild setting where people operate with hidden individual goals; coordinating to speak, leave, and join conversations emerge simultaneously with (un)realized intentions. (d) Does E intend to speak to D?

ABSTRACT

The future of socially intelligent systems depends on developing abilities to anticipate and empathize with users. Whilst great strides have been made on developing systems for future behavior forecasting that sometimes also claim to do intention estimation, we argue that the predominant state-of-the-art treatment of these problems leads to a significant misunderstanding about this topic. This paper revisits intention estimation, describing the "intention by outcome" problem and how it severely limits a deeper understanding of the nature of the problem. We argue that without a deeper more nuanced understanding of how to develop intention estimation systems, we head into a severely biased world where intentions would only be considered valid by intelligent systems if they came true. Through a case study on estimating unrealized intentions to speak in-the-wild, we highlight open challenges of this largely unexplored topic.

KEYWORDS

intention estimation, in the wild, speaking

ACM Reference Format:
Hayley Hung, Litian Li, Jord Molhoek, and Jing Zhou. 2024. The Discontent with Intent Estimation In-the-Wild: The Case for Unrealized Intentions. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3613905.3644055

CHI EA '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0331-7/24/05
https://doi.org/10.1145/3613905.3644055

1 INTRODUCTION

This paper is about automated intention estimation. The goal of this paper is to highlight the challenges of developing intention estimation systems in Socially Intelligent Systems [2, 37, 54]. For some readers this may be considered already as a very well trodden research task…perhaps even a solved problem! There is a lot of work already published in this field. For example, intention estimation has been investigated frequently in human robot interaction (e.g. object picking tasks [23, 28, 35], service robots [1]), dialog systems where intent is typically one of the slots that need to be filled for understanding what action a user wants to do next, or what a user might be wanting to search for [27]. Those who have delved deeply into such a problem may also argue that the new name for intention estimation is behavior forecasting [41, 46]. After all, a holy grail of intelligent systems are their ability to anticipate and respond to our needs.

1.1 Defining Intent

Let's start with some formalities. Defining and understanding intention has been of interest in cognitive science and philosophy for a long time and remains a contemporary active field of study [3, 9, 31, 38]. In this paper, we take Bratman's [9] definition of intentions by considering them as part of future action planning coupled with a belief by the individual that they have the ability to carry out the action. Crucially for intelligent systems, these planning behaviours need to be perceivable externally; what Mele in his book on understanding intentional behavior describes as *overt intentional action* [31]. Crucially, Mele argued that there are two different types of intention; distal and proximal. The former describes a mental state related to the future which is not necessarily situated in the current situation. The latter refers to intentions related to immediate actions for which the closer we get to the action, the more likely it is that the behavior being observed is the action itself. One can easily get lost in the philosophical literature regarding intentions,

intentionality and action. Fortunately, the survey of Bellardinelli provides a pragmatic balance for the interested reader [8].

One final note on intention is its close relationship with action. As noted by Fuchs and Bellardinelli in the context of shared intention estimation in human-robot interaction [20], there is a blurred line between Where an intention stops and the intended action begins. However, what if the intended action never begins? Does that mean that the intention never existed? From the perspective of many existing research works on intention estimation, intentions are operationalized in such a way that if something didn't happen, it wasn't intended to begin with.

1.2 The Forgotten Intention and the Intention by Outcome Problem

We argue that this notion that something that doesn't happen was not intended in the first place is a major conceptual issue for Socially Intelligent Systems. As a side note, it when pitching this idea to Computer Scientists (CS) and Social Scientists (SS), there was a notable difference. The CS could not disentangle the difference between intentions and future behaviour whilst SS understood immediately the difference between intention, action, and outcome. CS folks appeared to be trapped in their own perspective bubble.

We can summarize the problem with the state of the art understanding of intention estimation by observing Figure 1(a). The classic trope for many intention estimation research works; a single hand reaches over the table, what will happen next? This is what the system typically sees at test time. However for the majority of research works on this topic, at data collection and training time, the development of ground truths for such settings is more complicated. Intentions can be enacted by asking subjects to perform a particular action which has been pre-determined and instructed to them [28], or the subject themselves decides on the outcome [23]. The researcher then observes the recorded data and determines post-hoc what the intentions were by effectively looking into the future [23, 35]. At test time however, the system does not have the possiblity to observe the future.

This paradigm of training pattern recognition and machine learning systems with more information at training time than is available at test time is a common technique which is used to ensure that good quality (potentially more objective) estimators are learned. However, this common practice shapes our perspective on how to design intention estimation systems. We call this *"the intention by outcome"* problem.

The perspective shaping of former works channels our research activities into trying to transform intention estimation into an objective task when in reality intentions are conceptually not the same as future outcomes and have subjective and dynamic qualities. These simplifying assumptions made by prior work become the basis of how we understand the context of intention estimation, which we argue leads to the formation of a perspective bubble. Let's try to break that now.

In each of these prior tasks, the user in those environments is the center of the world; systems are meant adapt to their needs. What if there are more users in these environments who are no longer interacting directly with an agent or interface but primarily with each other? A simple illustration of this point is seen in Figure 1(b) where we see an expanded version of Figure 1(a) to include a second hand on the right. Two hands reach over a table, what will happen next? We do not know definitively because we cannot predict the future. But we can make a good estimate. However, in this instantaneous moment, one would probably also not deny that both hands appear to indicate an intent to grab the apple.

1.3 From a Single User to a Constellation of Self-organizing Users

The example of Figure 1(b) gradually puts into focus how the context of use changes when building socially intelligent systems in more open-ended social situations like the social networking event show in Figure 1(c). In these settings, socially intelligent systems are observing and interpreting the behaviours of multiple people with hopefully the capacity to help them. How would such a system understand that you are not having a good time at a party because you have not yet managed to escape a conversation and talk to the very interesting person you see across the room? Or that you still haven't managed to get a word in edge-ways with a particularly interesting conversation despite having some very interesting stories to tell? The major difference in such settings is that each individual in these scenes is no longer the center of a world where everything in the environment exists purely to satiate whatever immediate need or want that they have. When multiple people are involved, their needs and wants may vary and compete against each other.

Once we move away from the rather controlled settings of object picking tasks and a single user interacting with an application or an artificial agent (e.g. robot), the notion of unrealized intentions becomes a lot more apparent. Multiple humans engage in conversation and self-organize into an orderly system of cooperation to exchange information. How people coordinate themselves into conversations, especially as the complexity grows from a single meeting to a cocktail party or professional networking event remains an open field of study in Social Science [16, 26].

Once we move out of the lab and into more in-the-wild settings, we come against other problems. How can we read the mind of a person while they are the middle of a conversation? Won't they get distracted if they are asked to report on their intentions? Couldn't this contaminate their spontaneous overt intentional behavior? More will be elaborated on this point later in Sec. 2.

1.4 The Risks of the Intention by Outcome Problem

While the earlier discussion highlights what may just be merely have been a slight conflation of conceptual terminology. We argue that scrutinizing a wider circle of potential applications of intelligent HCI systems, highlights significant challenges and also risks that must be overcome. This is because past work has only considered intention to be conceptualized as successful outcomes.

Whilst there has been work to address how one might mitigate when 'things don't work out,' e.g. frustration estimation in HRI settings [55], there are many settings where unrealized intentions may occur frequently where multiple users need to coordinate with each other and simply minimizing frustration due to unrealized intentions might be unrealistic or just not the right way to approach

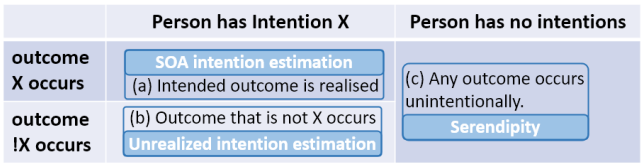

Figure 2: Taxonomy of intentions: (a): realized intentions; (b): unrealized intentions; (c): serendipity

the design of socially intelligent systems. After all, experiencing unrealized intentions may be an inevitable part of daily life.

Helping users to navigate the complexities associated with living in real life may be more effectively approached if intelligent systems could understand or at least have a hypothesis about someone's mental state such as their intention. Note that this goes beyond being able to estimate someone's mood or emotions to being ultimately able to provide explanations for states of mind. e.g. "You appear frustrated. Is it because you wanted to join in on the conversation and didn't manage to?"

In the worst case, ignoring unsuccessful intentions could pose a significant risk to the future development of intelligent systems and its potential risk of perpetuating the status quo rather than adapting to values of modern society such as efforts to reduce gender imbalance and reduce systematic racism. For example, it has been observed that female leaders emerge less frequently and appear to participate less in conversations [6, 21, 48]. Working with an "intent by outcome" model of the world, this would mean that females do not speak up because they never intended to. In cases of criminal arrests, any racial minorities would be labelled as loitering with intent because they get arrested irrespective of whether they had negative intent or not. If we do not revisit the intention estimation problem, we continue to develop systems that discriminate and exclude.

1.5 Relating Intentions and Outcomes: Predicting the Future is not the same as Intention Estimation

Perhaps the mere notion of intentions are still hard to grasp. Let us use an example based on social intentions in-the-wild as illustrated with Fig. 1(d). We observe a conversing group containing participants A – F at a professional networking event where F is speaking. If we observe the gaze of E, we see that they are the only person not looking at F but is instead gazing at D. An observer might perceive this gaze to indicate an intention to talk to D. While state-of-the-art approaches would conclude that a system is successful in predicting E's intention if D and E start having a conversation. The main point is that the perceived intention of E to speak to D is already observed and valid irrespective of whether D and E finally converse or not.

There needs to be more attention on investigating automated means to estimate and distinguish between *realized and unrealized intentions*. To do this more effectively it makes sense to think about the relationship between intentions and outcomes. Figure 2 illustrates the possible combinations; state-of-art approaches focus on (a); realized intentions. These are intentions that led to the desired outcome. Meanwhile unrealized intentions (intentions that did not lead to the desired outcome), as shown in (b) have received only cursory attention [43, 56]. For completeness we also include outcomes that can occur unintentionally, which can also be known as serendipity. For the simplicity of Figure 2, serendipity has been categorized to exist in settings where a person had no intention. A more nuanced definition of serendipity could define it actually as any unexpected outcome. In other words, an outcome could have occurred instead of an unrealized intention as well as an absence of intentions. When these potentially unrealized intentions are viewed from the lens of serendipity, it becomes again a hot topic in business [11, 30], organizations [10, 19], society [14], creativity and productivity [22, 32] to name but a few.

So why have unrealized intentions been of so little interest? Perhaps the focus of prior research work in intelligent systems were so focused on anticipating our intention in order to react to us, that it became synonymous with future forecasting. Or perhaps the problem is how do you identify and label something that was intended but did not happen?

2 THE CHALLENGES OF UNREALIZED INTENTION LABELLING IN-THE-WILD

We hope that by now we have convinced the reader that intention estimation is an important and under-explored topic. Particularly in the case of unrealized intentions. So let's start to label them! The discerning reader may already see some hurdles ahead of us once we start to explore intentions in the absence of outcomes. How do we read the mind of the user? And how do we do it at fine time scales? When does an intention start and end? How do we label unrealized intentions both in terms of the relevant cues and also an outcome that *did not occur?*.

The challenges of labelling intentions can be illustrated in Fig. 3. For many, (including initially the authors themselves) self-report seems the only reasonable method to obtain labels of intentions. Ideally, the person would report constantly (**self-reported in-situ continuous feedback**: Fig. 3(a)), leading to a very temporally precise measure of someone's intention. Some tools have been developed to enable in-situ and in-the-wild reporting of affect whilst watching mobile video that may be applicable in this case [58]. However, this could disturb the spontaneity of the social behavior. Another alternative is to have subjects watch an interaction they just had and rate it continuously [51]. However this is hard to scale. Another solution involves individuals reporting their intentions just after finishing an interaction (**self-reported in-situ posthoc feedback**: Fig. 3(b)). However reflective cognitive processes can occur during post-hoc reflection leading to a difference in reported intention compared to a spontaneous in-situ report (see [17] for a detailed discussion).

A final option uses external observers (**third-party continuous post-hoc feedback**: Fig. 3(c)). The temporal precision between the intention and sensor data is preserved, and the behaviour is not disturbed. However, whilst being much more scalable, it moves us further away from the truth of the individual being observed. There is perhaps still hope. While this may not be the same as the true intention of the person, learing plausible narratives of intention could still allow intelligent systems to reason about them, potentially updating and adjusting their understanding based on

feedback during user interactions. Meanwhile the individual's true intentions are kept private.

3 GUT REACTIONS AND COGNITIVE EMPATHY

It would appear that exploiting the perceptions of external observers would be a nice compromise for obtaining intentions at scale and with high temporal fidelity. But what makes this approach valid? By asking annotators to label plausible intentions, they are likely to simulate narratives plausible to their own life experiences [7, 15, 44]. On the other hand, given that the individual being observed is not the same as the observer, there is some distance that must be taken by an observer.

One can also consider such labelling tasks to be a form of Ethnomethodology, which is deliberately agnostic towards social theory in favour of a naïve eye, shaped towards the context of interest. This is an important notion as "..members of society must have some shared methods that they use to mutually construct the meaningful orderliness of social situations" [45]. Stated another way, the basic premise of Ethnomethodology assume that there is an emergent and meaningful orderliness to social situations that can be interpreted by external observers.

In the end it is likely that observers, and in particular lay observers if we take a crowd sourcing approach to the labelling process, will not necessarily be trained in the rigours of Ethnomethodological practice. And even if they were, disentangling the effect of subjective personal experience from a deliberate naïve eye approach may be impossible.

What is clear, however, is that these perceived intentions come with plausible narratives or explanations. It could well be that for a given constellation of behaviours, much like those seen in Figure 1(d), multiple plausible explanations for completely different perceived intentions for person E could exist. In the crowdsourcing literature, there has been a tendency to assume one single ground truth. Annotations are judged as being poor if they do not agree with other annotators and there are some people who are more expert than others [5].

Finally what happens in the case of unrealized intentions? We propose that gut instincts may be the way to identify them and in the following case study, we investigate this approach to label unrealized intentions to speak.

4 CASE STUDY: ESTIMATING UNREALIZED INTENTIONS TO SPEAK

As highlighted in Sec. 2, there are compelling benefits to using third party labels. To this end, we present a case study that investigates the estimation of proximal intentions. We focus specifically on two questions, focusing on the problem of speaking intention in-the-wild: (i) How feasible is it to label unrealized intentions using an external observer? Can we rely on Mele's idea of overt intentional actions [31]? (ii) How could we train a model to detect realized and unrealized intentions? (iii) Would there be a difference between the performance of realized and unrealized intentions?

4.1 Related Work on Intentions to Speak

To our knowledge, Wlodarczak and Heldner carried out the only research that has discussed unrealized intentions to speak [56]. They used sensors strapped to the chest and abdomen in seated discussions to measure breathing during conversations. The findings identified breath inhalations (that looked like preparations to speak) followed by long breath holds and a silent exhalation with no accompanying speech activity. They speculated that this behaviour could be indicative of unrealized intentions. Their findings are foundational but their sensing approach is hard to scale to dynamic in-the-wild ecologically valid settings. Moreover, their speculations were not validated with self-reports or third-party perceptions of the multi-modal data.

Other related work focuses mainly on behaviour forecasting, namely turn-change prediction and next-speaker prediction. Turn-change prediction estimates *when* a turn change is about to occur whilst next-speaker prediction estimates *who* will speak next. Petukhova and Bunt [36] found that gaze aversions, lip movements, and posture shifts are strongly correlated with receiving the next turn in conversations. According to Novick et al., 42% of the turn changes follow the pattern: the speaker first looks towards the listener as they complete the turn; then the speaker and the listener have a short moment of eye contact; lastly the listener looks away and begins speaking [34]. Automatic recognition of such patterns can be useful for next speaker prediction and possibly for prediction of intentions to speak. As noted by Schegloff, turn-taking is not always smooth, as an interruption could be interpreted as not letting the current speaker finish, which can be viewed conceptually as a form of unrealized intention [47].

Ishii et al. showed that gaze-transition patterns are useful for predicting the next speaker and when the next speaker starts to speak. They later found that a fusion model using both respiration and gaze behaviour performs better than using only one of them in isolation, and respiration was the more useful feature in their multimodal model [24]. A followup study also by Ishii et al. found that people tend to open their mouths slightly before they start speaking, which can be related to intentions to speak [25].

4.2 Our Approach

An overview of our experiments are shown in Figure 4. Developing systems that work outside of the lab in in-the-wild ecologically valid settings adds some additional constraints to our problem space. This includes exploiting a sensing setup that would maximise the detection of unrealized intentions and whilst being respectful of privacy concerns. Building on the promising results of estimating speaking status with accelerometers [39, 42] we leverage the ubiquity of the conference ID badge commonly hung around the neck. The badge form-factor can be exploited with sensors such as a tri-axial accelerometer which is most likely to capture the fine grained breathing activities observed by Wlodarczak and Heldner [56], as well as other intention related behavioural modalities (e.g. leaning, gaze via head pose). Given the loud background chatter associated with busy networking events, we decided against using the audio since detecting subtle noises could be more challenging. There are also privacy concerns when recording private conversations [13, 42]. Multi-modal analyses are left for future work.

	Spontaneous behaviour?	Association of feedback to intent
(a) Self-reported in-situ continuous feedback: e.g. "Do you want to speak?" ? ? ? ? t	Affected	Direct temporal association with spontaneous intent
(b) Self-reported in-situ posthoc feedback: e.g. "Did you want to speak?" episode ? t	Affected indirectly between interactions	Indirect time association, may not reflect spontaneous intent.
(c) Third-party continuous posthoc feedback: e.g. "Do they want speak?" ? ? ? ? t	Yes	Direct time association , may not reflect true intent.

Figure 3: Challenges to overcome for labelling realized and unrealized intentions. We present a study using (c) in Sec. 4.

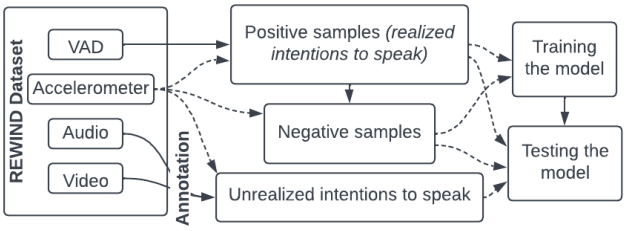

Figure 4: Overview of our approach. The dashed lines indicate the flow of only accelerometer data.

Our experimental approach is summarized in Fig. 4. A model is trained on the accelerometer data of automatically extracted *realized* intentions to speak (Sec. 4.5). These positive samples are automatically extracted from the voice activity detection (VAD) as explained in Sec. 4.5.1. The trained model is evaluated on the automatically generated realized and also manually annotated unrealized intentions to speak (Sec. 4.4). Results are presented in Sec. 4.6.

4.3 Experimental Data

For this case study, we used the REWIND dataset [53] which will soon be shared [39]. The REWIND dataset contains audio, video, and wearable accelerometer data of an indoor professional networking event with around 100 attendees who stood and were free to mingle as they pleased. Video data is recorded by elevated side-view cameras, as shown in Fig. 1(c). The event offered a mixed consent model where participants could choose whether to wear an accelerometer around their neck (like a conference ID badge), a wireless microphone attached to the face via specialised tape used for theatre productions, and appear under the cameras. A 10-minute segment (1:00:00 - 1:10:00) of the data is used for the exploratory study and annotations. In this interval, 13 participants had audio, video, and accelerometer data. Ethical approval from the university ethics board was granted.

4.4 Labelling Unrealized Intentions to Speak

To assess the quality of the model on unrealized intentions to speak, a sample of unrealized intentions to speak is needed. Before labelling the data, an initial exploratory study of unrealized intentions to speak was done. This involved critically examining our own behaviour in cases of intentions to speak and that of others. Finally, the data was observed to find cues that indicate intentions to speak.

The most important finding from this exploratory study was that mouth-opening patterns through lip or tongue smacks were audible in some cases. To the best of our knowledge, no literature associates these audible mouth-opening patterns with intentions to speak. We claim that we should! Another finding from this exploratory study was that due the loud background chatter, we found that people sometimes lean in or shift posture towards someone's ear when they want to say something. Furthermore, throat clearing seemed also to indicate an intention to speak. Lastly, we found an important conceptual distinction between intentions to start speaking and intentions to continue speaking; in the first case, the person is not yet speaking and attempts to take the turn, whereas in the latter, the person is already speaking and attempts to *keep* the turn.

After the exploratory study, during the 10-minute segment, start and end times of perceived unrealized intentions to speak were labelled manually. An important consideration is that these are annotations of *perceived* unrealized intentions to speak; when the annotator deemed it more likely than not the case that a person wanted to say something but did not (get the chance to), the segment is annotated. The end of an annotated unrealized intention is labelled based on when the person does not speak when the annotator would expect them to speak and also when they do not give any signals anymore that they have an intention. This highlights the inherent uncertainty in these annotations and the higher dependence on gut feeling of the annotator.

Annotations were performed by one of the co-authors using ELAN [33] using audio-visual data coming from the microphone and corresponding camera of a given subject. The annotator listened to one participant at a time while looking at them through the camera in which they are best visible. When the annotator perceived a (likely) unrealized intention to speak, this segment was annotated. This is primarily based on human intuition, whilst baring in mind the cues associated with (both realized and unrealized) intentions to speak that were identified in the literature survey and the exploratory study.

Following the findings of the exploratory study where we had identified two categories of unrealized intentions; starting and continuing to speak. We report from the chosen experimental data the number of observed unique individuals, samples, and mean, and standard deviation of the interval lengths respectively: intentions to take the turn (**UnrealStart**: 10, 22, $1.98 \pm 0.89s$) and intentions to

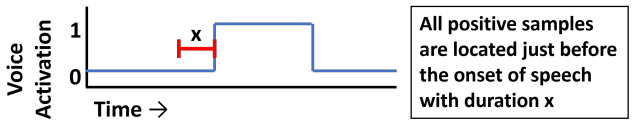

Figure 5: Positive sample generation; X is varied from 1-4s.

continue speaking (**UnrealCont**: 7, 17, 2.46 ± 0.97*s*). It is important to note that small sample size. This suggested that we would not be able to train a model from scratch with reasonable accuracy and that an alternative training approach was needed.

4.5 Training Intentions to Speak

Four models were trained with varying window lengths spanning 1-4s with 10 epochs under 3-fold cross-validation. To minimise variation in the model performance to discern differences between the different experiments, leave-person-out cross validation was not used. With so few test samples, we decided to train models for just *realized* intentions to speak and test on the samples identified in Sec. 4.4. These models were evaluated using the ROC AUC across 5 experiments using different types of positive test samples [1];**1. All**: realized and unrealized intentions, **2. Realized**, **3. Unrealized**, **4. UnrealStart**: unrealized intentions to start speaking, and **5. UnrealCont**: unrealized intentions to continue speaking.

Adapting the implementation of Vargas et al. [53], the structural architecture of our models comprises multiple residual neural networks embedded within convolutional neural networks. Non-overlapping training samples from individual accelerometer data were sampled from the mingling time outside of our chosen 10 minute segment and within the 10 minute segment for the test data. This procedure was repeated 100 times.

4.5.1 Voice Activity Pre-processing. For each speaker, we process the provided automatically detected voice activity generated by REWIND which computes voice activity by applying the following to each audio signal; loudness normalization, denoising, and Speaker diarization via NVIDIA NeMo [29]. This was then processed to remove extremely short turns (<1.5s) that are likely to be back-channels and short pauses (<1.5s) between speaking. The threshold of 1.5s was determined empirically.

4.5.2 Positive Sample Set Generation. Since there were so few samples of unrealised intentions identified, we needed to leverage the large number of realized samples that existed in the dataset. We hypothesized that the unrealized intentions might have similar characteristics to the realized intentions, though findings by Wlodarczak and Heldner suggest that humans do adapt their behaviour some time before the next speaker when they can see that they will not get to speak [56].

Positive realized intention samples were selected as the time windows prior to the activation of voice activity (interval X in Fig. 5). As the window length increases from 1 up to 4 seconds, the number of positive samples decreases due to a higher chance of overlapping with speaking segments. The length of the annotated positive unrealized intention samples also varies (as mentioned in Sec. 4.4). Since there were so few samples, we decided to use them all at test time irrespective of the window length. In all experiments, we aligned the end of the sample with the annotated end time of the positive unrealized intention sample and then computed back in time for the corresponding window length. The number of positive samples for training and test are shown in Tab. 1.

[1]code and data shared at https://github.com/TUDelft-SPC-Lab/UnrealisedIntentSpk

		Realized	Unrealized Start	Unrealized Continue
Train		2567 / 2053 / 1689 / 1487	unused	unused
Experiment	1	327 / 268 / 239 / 217		
	2	286 / 227 / 198 / 176	unused	unused
	3	unused	39/39/39/39	
	4	unused	22/22/22/22	unused
	5	unused	unused	17/17/17/17

Table 1: Numbers of positive samples by type: row 1. training; row 2-6: test samples generated at 1s/2s/3s/4s window lengths. Experiments with the different types of training data that was used; 1. All: realized and unrealized intentions, 2. Realized, and 3. Unrealized, 4. UnrealStart: unrealized intentions to start speaking, and 5. UnrealCont: unrealized intentions to continue speaking.

4.5.3 Negative Sample Set Generation. The negative samples for training were randomly sampled outside the period 1:00:00-1:10:00, where the length of the negative samples corresponds to the length of the positive samples (1-4s). Test samples were randomly sampled from the time interval of 1:00:00-1:10:00 taking care that they did not overlap with any of the positive samples (experiment 2). The negative test samples for experiments 3, 4, and 5 were sampled from this same segment and did not overlap with positive realized or unrealized intention samples. The negative samples generated for experiment 1 also included the negative samples of experiments 2 and 3. The ratio of positive to negative samples is always 1:20.

4.6 Results

A comparison of all the results is shown in Fig. 6 where the mean and standard error were computed over all folds over 100 runs. We observe that **Realized** had above average performance with better performance for longer windows and that **All** performs worse but still mostly above average. **Unrealized** shows performance comparable to **Realized** for 1s and 2s windows. However **Unrealized** has poor performance at windows of 3s or 4s, which could suggest that the unrealized intentional behavior is different further before the unrealized intention or that the labelling of the end of the unrealized intention was too far away from the start.

Finally, dividing up the data from **Unrealized**, we observe that **UnrealCont** has consistently worse performance while **UnrealStart** has surprisingly better performance than all other experiments for the 1s and 2s windows. This suggests that Unrealized intentions to start speaking exhibit extremely clear non-verbal behaviors that align well with the behavior of realized intentions. The performance drop as the window size increases could suggest that annotation of the end of the **UnrealStart** cases are shorter. These shorter window lengths also align with clear behaviours related to the intention to speak such as leaning.

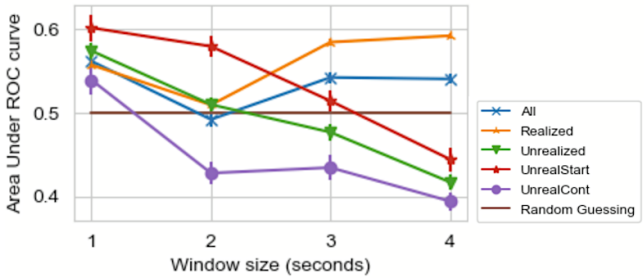

Figure 6: Mean and standard error for all experiments.

We speculate that unrealized intentions to continue speaking have more subtle behaviour since the subjects did not need to coordinate a turn change. The comparable performance of **UnrealCont** at 1s suggests that distinguishing behaviours are short such as for breathing. Finally, the general decreasing trend of experiments 3-5 as the window size increases may be due to samples being more and more contaminated by voice activity behavior.

5 SUMMARY AND OPEN CHALLENGES

We have proposed a reframing of intention estimation research to account for both realized and unrealized intentions to overcome the "intention by outcome" problem. We carried out a preliminary investigation showing that unrealized intentions can be labelled by an external observer. Our experiments using trained models demonstrate that realized and unrealized intentions are hard to disentangle which suggests potential opportunities to exploit transfer learning approaches. Beside this, many open questions remain:

Intention Detection vs. Segmentation. Our case study presented a simple first step where we used a window based approach to decide if it contains an intention to speak or not. This requires us to know the length of an overt intention behaviour beforehand. There is likely to be a preference for finer grained annotations where we can observe the precise time when an intentional behaviour starts and also ends. Performing the labelling for the segmentation task when there are realized intentions is fine; one chooses the moment within a window of time when the start of an overt intentional behaviour is observed. But how do we label the start *and end* of a intention that is never realized? Was the choice we made in the case study correct to rely on intuition to find the end of the intention rather than e.g. selecting the end of the first overt behaviour?

Associating Intentions with Outcomes. For the case of realized intentions, finding the associated moment where the overt intention starts to occur follows prior work. However, even if we can use our gut instincts to identify an unrealized intention, was it right that the unrealized intention needs to occur shortly after the intention? What if it happens much later on? Or for more complex intentions where the outcome might be far in the future, how do we associate intentions with outcomes if there are long temporal connections? How long should a system wait before declaring an intention to be (un)realized?

Intention by Outcome: From Problem to Paradox? Earlier, we described the intention by outcome problem in Sec. 1.2, where we critiqued the practice of looking into the future to train state-of-the-art intention estimators. Due to too little data on unrealized intentions to speak, our case study also relied on a training protocol that relied on training realized intentions to speak. Note that the negative training examples we selected could not necessarily guarantee not to capture unrealized intentions to speak. Our experimental results showed that there are similarities between the preparatory behaviours observed in unrealized as well as realized intentions. For an annotator to perceive realized or unrealized overt intentions, they may rely on many past observations of predominantly realized intentions to learn the relevant preparatory behaviors. It remains an open question whether unrealized intentions can be perceived without needing to have ever observed any realized examples, when e.g. more complex (non-overt) intentions are at play.

Subjectivity. Our case study used one annotator. Without an associated outcome, unrealized intention labelling requires an observer to fill the gap between the observed cues and the perceived intent using on their own life experiences. For example, for the example initially given in Fig. 1(d), if we were to ask multiple independent annotators, they may come up with multiple different intentions with all equally plausible explanations. Usually subjectivity is treated as noise but is there not a validity to any perception as long it is explained by observed cues? If so, how can we train models to account for this? What if all these equally valid perceptions do not agree despite each having very valid explanations? It is not clear how we should set up a learning problem based on multiple equally valid but potentially conflicting narratives.

Annotator Context. As argued by Dudzik et al., a lack of systematic treatment of annotator context could be limiting system performance for affect perceptions [18] and could apply to other labelling tasks. Thus far, the relevance of annotator context in annotation tasks is largely ignored. The Perspectivist Data Manifesto proposed by researchers in the Natural Language Processing Community could provide some useful insights [12]. If integrated into the labelling task, this could allow systems to be designed to take only certain types of biases or perspectives into model training based on a careful selection of the context of annotators. Being able to manipulate the type of perspective that an estimated intention should originate from may also provide opportunities to get closer to the intent of the actual subject in question. Could the perspective of annotators that are closer to that of the subject make them better estimators of the true intention of the subject?

Situation Context. The case study only addressed one type of intention. However, if we need to automatically infer multiple intentions, the context plays a significant role in how people perceive other people's overt intentions [4]. When multiple plausible narratives of intent can exist for the same stimulus (see earlier discussion on subjectivity), the way to differentiate them comes from the explanation that accompanies an intent label. The explanations are interpretations based on an understanding of the context. In the illustrated example, they relate to social aspects such as who is currently socially involved with whom and temporal aspects such as what is happening in the conversation(s) of interest. Detecting the social context is also a challenging problem as a model that

plausibly defines conversational interaction as a fine time scale remains and open and challenging problem [40, 49, 50, 52].

Multimodality. Labelling the unrealized intentions required multimodal stimuli, although the audio was the primary modality used to label the data. Further work is necessary to understand the interplay between different modalities and how they can be indicative of different types of intention. Moreover, one needs to consider both the digital modalities involved as well as different behavioural modalities that might be relevant. Finally, when the lexical aspects of spoken dialog are also taken into account, the complexity of intentions increases dramatically as intentions to steer conversations in different directions increase the possible space of intended behaviors even further.

The Dynamic Nature of Intentions. Our case study assumed that intentions are static. However, Wlodarczak and Heldner [57] observed circumstantial evidence that could indicate differences in breathing patterns that could be indicative of intentional planning behaviours to speak followed by a change of intention and aborting behaviours. The issue is that depending on the chosen sensor for a given application, the fidelity at which such subtleties can be detected remains an open question. If we assume that something of these changes in intention can be observed, then how should they be modelled? Conceptual questions also arise as to whether an aborted intention is already the onset of an unrealised intention or not. Considering also the hierarchical nature of intentions, someone might abandon an immediate intention if a serendipitous opportunity related to a longer term goal arises.

Accounting for Unobservable Intentions. In Sec. 1.1, we defined intention and motivated scoping the perceived intention estimation problem to one of detecting overt intentions. Could lead to unwanted biases.

Potential for Societal Benefit. Some examples of how we could use intention estimation to reduce unwanted bias in issues of gender bias in teams or racism in criminal cases was discussed in the introduction. However, simply making estimates is not enough. How should a system respond to and act upon automated perceptions of realized and unrealized intentions? The matter may need to be handled extremely carefully for applications that counter discrimination and exclusion, especially when making an incorrect inference could be extremely damaging. Finally, the potential benefits of using intention estimation systems to help highlight how certain perspectives might lead to particular biases in intention estimation needs further exploration.

ACKNOWLEDGMENTS

We would like to thank Tiffany Matej Hrkalovic, Stephanie Tan, and Catharine Oertel for comments and feedback on early versions of this draft. Hayley Hung was partially funded by ERC Consolidator Grant NEON:101089163.

REFERENCES

[1] Gabriele Abbate, Alessandro Giusti, Viktor Schmuck, Oya Celiktutan, and Antonio Paolillo. 2024. Self-supervised prediction of the intention to interact with a service robot. *Robotics and Autonomous Systems* 171 (2024), 104568. https://doi.org/10.1016/j.robot.2023.104568

[2] Zeynep Akata, Dan Balliet, Maarten De Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, et al. 2020. A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer* 53, 8 (2020), 18–28.

[3] Thomas Nadelhoffer Alfred Mele and Maria Khoudary. 2021. Folk psychology and proximal intentions. *Philosophical Psychology* 34, 6 (2021), 761–783. https://doi.org/10.1080/09515089.2021.1915471 arXiv:https://doi.org/10.1080/09515089.2021.1915471

[4] Ayeh Alhasan, Michael J Richardson, Nathan Caruana, and Emily S. Cross. 2023. Predicting intentions: How do we predict other's action intentions? Publication Date. *Proceedings of the Annual Meeting of the Cognitive Science Society* 45 (2023), 45. Issue 45.

[5] Lora Aroyo and Chris Welty. 2015. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Mag.* 36, 1 (mar 2015), 15–24. https://doi.org/10.1609/aimag.v36i1.2564

[6] Katie L Badura, Emily Grijalva, Daniel A Newman, Thomas Taiyi Yan, and Gahyun Jeon. 2018. Gender and leadership emergence: A meta-analysis and explanatory model. *Personnel Psychology* 71, 3 (2018), 335–367.

[7] L.F. Barrett. 2017. *How Emotions Are Made: The Secret Life of the Brain.* HarperCollins, New York, US. https://books.google.nl/books?id=hN8MBgAAQBAJ

[8] Anna Belardinelli. 2023. Gaze-based intention estimation: principles, methodologies, and applications in HRI. arXiv:2302.04530 [cs.RO]

[9] M. Bratman. 1987. *Intention, plans, and practical reason.* Harvard University Press, Cambridge, MA.

[10] Chloë Brown, Christos Efstratiou, Ilias Leontiadis, Daniele Quercia, and Cecilia Mascolo. 2014. Tracking Serendipitous Interactions: How Individual Cultures Shape the Office. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work* (Baltimore, Maryland, USA) *(CSCW '14)*. Association for Computing Machinery, New York, NY, USA, 1072–1081. https://doi.org/10.1145/2531602.2531641

[11] Christian Busch and Harry Barkema. 2020. Planned Luck: How Incubators Can Facilitate Serendipity for Nascent Entrepreneurs Through Fostering Network Embeddedness. *https://doi.org/10.1177/1042258720915798* 46, 4 (2020), 884–919. https://doi.org/10.1177/1042258720915798

[12] Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. *Proceedings of the AAAI Conference on Artificial Intelligence* 37 (6 2023), 6860–6868. Issue 6. https://doi.org/10.1609/AAAI.V37I6.25840

[13] Laura Cabrera-Quiros, Andrew Demetriou, Ekin Gedik, Leander van der Meij, and Hayley Hung. 2021. The MatchNMingle Dataset: A Novel Multi-Sensor Resource for the Analysis of Social Interactions and Group Dynamics In-the-Wild During Free-Standing Conversations and Speed Dates. *IEEE Transactions on Affective Computing* 12, 1 (2021), 113–130.

[14] Jeffrey K. H. Chan. 2019. *Serendipity.* Springer Singapore, Singapore, 99–126. https://doi.org/10.1007/978-981-13-0308-1_5

[15] Jean Decety and Philip L Jackson. 2004. The Functional Architecture of Human Empathy. *Behavioral and Cognitive Neuroscience Reviews* 3, 2 (jun 2004), 71–100. https://doi.org/10.1177/1534582304267187

[16] Bálint Diószegi, Anne Ter Wal, Valentina Tartari, and Daniella Laureiro Martinez. 2020. Deliberate vs. Spontaneous Networking Behaviors and Information Search: An Interactive Experiment. *Academy of Management Proceedings* 2020 (08 2020), 20055. https://doi.org/10.5465/AMBPP.2020.42

[17] Bernd Dudzik and Joost Broekens. 2023. A Valid Self-Report is Never Late, Nor is it Early: On Considering the "Right" Temporal Distance for Assessing Emotional Experience. arXiv:2302.02821 [cs.HC]

[18] Bernd Dudzik, Michel-Pierre Jansen, Franziska Burger, Frank Kaptein, Joost Broekens, Dirk KJ Heylen, Hayley Hung, Mark A Neerincx, and Khiet P Truong. 2019. Context in human emotion perception for automatic affect detection: A survey of audiovisual databases. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, Cambridge, UK, 206–212.

[19] Nathan Eagle. 2004. Can serendipity be planned? *MIT Sloan Management Review* 46, 1 (2004), 10.

[20] Stefan Fuchs and Anna Belardinelli. 2021. Gaze-Based Intention Estimation for Shared Autonomy in Pick-and-Place Tasks. *Frontiers in Neurorobotics* 15 (2021), 17 pages. https://doi.org/10.3389/fnbot.2021.647930

[21] Fabiola H. Gerpott, Nale Lehmann-Willenbrock, Jeroen D. Silvis, and Mark Van Vugt. 2018. In the eye of the beholder? An eye-tracking experiment on emergent leadership in team interactions. *The Leadership Quarterly* 29, 4 (2018), 523–532. https://doi.org/10.1016/j.leaqua.2017.11.003

[22] Lynda Gratton. 2020. *How to increase collaborative productivity in a pandemic.* MIT Sloan Management Review.

[23] Chien Ming Huang, Sean Andrist, Allison Sauppé, and Bilge Mutlu. 2015. Using gaze patterns to predict task intent in collaboration. *Frontiers in Psychology* 6 (7 2015), 1049. https://doi.org/10.3389/FPSYG.2015.01049/BIBTEX

[24] Ryo Ishii, Shiro Kumano, and Kazuhiro Otsuka. 2015. Multimodal Fusion Using Respiration and Gaze for Predicting Next Speaker in Multi-Party Meetings. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*

(Seattle, Washington, USA) *(ICMI '15)*. Association for Computing Machinery, New York, NY, USA, 99–106. https://doi.org/10.1145/2818346.2820755

[25] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, Ryuichiro Higashinaka, and Junji Tomita. 2019. Prediction of Who Will Be Next Speaker and When Using Mouth-Opening Pattern in Multi-Party Conversation. *Multimodal Technologies and Interaction* 3, 4 (2019), 24 pages. https://doi.org/10.3390/mti3040070

[26] Adam Kendon. 1990. *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Cambridge University Press, Cambridge, UK.

[27] Christoph Kofler, Martha Larson, and Alan Hanjalic. 2016. User Intent in Multimedia Search: A Survey of the State of the Art and Future Challenges. *ACM Comput. Surv.* 49, 2, Article 36 (aug 2016), 37 pages. https://doi.org/10.1145/2954930

[28] Julian F. Kooij, Fabian Flohr, Ewoud A. Pool, and Dariu M. Gavrila. 2019. Context-Based Path Prediction for Targets with Switching Dynamics. *Int. J. Comput. Vision* 127, 3 (mar 2019), 239–262. https://doi.org/10.1007/s11263-018-1104-4

[29] Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, Patrice Castonguay, Mariya Popova, Jocelyn Huang, and Jonathan M. Cohen. 2019. NeMo: a toolkit for building AI applications using Neural Modules. arXiv:1909.09577 [cs.LG]

[30] Jacqueline N. Lane, Ina Ganguli, Patrick Gaule, Eva Guinan, and Karim R. Lakhani. 2021. Engineering serendipity: When does knowledge sharing lead to knowledge production? *Strategic Management Journal* 42 (6 2021), 1215–1244. Issue 6. https://doi.org/10.1002/SMJ.3256

[31] Alfred R Mele. 1992. *Springs of action: Understanding intentional behavior*. Oxford University Press, New York, USA.

[32] John Meluso, Susan Johnson, and James Bagrow. 2020. *Making virtual teams work: Redesigning virtual collaboration for the future*. SocArXiv.

[33] Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive, 2022. *ELAN [Computer software]*. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive,. https://archive.mpi.nl/tla/elan Version 6.4.

[34] D.G. Novick, B. Hansen, and K. Ward. 1996. Coordinating turn-taking with gaze. In *Proceeding of Fourth International Conference on Spoken Language Processing*, Vol. 3. ICSLP '96, Philadelphia, PA, USA, 1888–1891 vol.3. https://doi.org/10.1109/ICSLP.1996.608001

[35] Tomislav Petković, Luka Petrović, Ivan Marković, and Ivan Petrović. 2022. Ensemble of lstms and feature selection for human action prediction. In *Intelligent Autonomous Systems 16: Proceedings of the 16th International Conference IAS-16*. Springer International Publishing, Cham, 429–441.

[36] Volha Petukhova and Harry Bunt. 2009. Who's next? Speaker-selection mechanisms in multiparty dialogue. In *Proceedings of the 13th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*. SEMDIAL, Stockholm, Sweden, 19–26. http://semdial.org/anthology/Z09-Petukhova_semdial_0009.pdf

[37] Rosalind W Picard. 2000. *Affective computing*. MIT press, Cambridge, MA, USA.

[38] Jason E Plaks and Jeffrey S Robinson. 2015. Construal level and free will beliefs shape perceptions of actors' proximal and distal intent. *Frontiers in psychology* 6 (2015), 777.

[39] Jose Vargas Quiros, Chirag Raman, Stephanie Tan, Ekin Gedik, Laura Cabrera-Quiros, and Hayley Hung. 2024. REWIND Dataset: Privacy-preserving Speaking Status Segmentation from Multimodal Body Movement Signals in the Wild. arXiv:2403.01229 [cs.CV]

[40] Chirag Raman and Hayley Hung. 2019. Towards automatic estimation of conversation floors within F-formations. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, Curran Associates, Inc., Cambridge, UK, 175–181.

[41] Chirag Raman, Hayley Hung, and Marco Loog. 2021. Social Processes: Self-Supervised Forecasting of Nonverbal Cues in Social Conversations. *CoRR* abs/2107.13576 (2021), 31 pages. arXiv:2107.13576 https://arxiv.org/abs/2107.13576

[42] Chirag Raman, Jose Vargas Quiros, Stephanie Tan, Ashraful Islam, Ekin Gedik, and Hayley Hung. 2022. ConfLab: A Data Collection Concept, Dataset, and Benchmark for Machine Analysis of Free-Standing Social Interactions in the Wild. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., New Orleans, Louisiana, USA, 23701–23715. https://proceedings.neurips.cc/paper_files/paper/2022/file/95f9ad2e251e9014697589037450f9bb-Paper-Datasets_and_Benchmarks.pdf

[43] Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John K. Tsotsos. 2019. PIE: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction. In *International Conference on Computer Vision (ICCV)*. IEEE, Seoul, South Korea, 6262–6271.

[44] K. Richard Ridderinkhof, Lukas Snoek, Geert Savelsbergh, Janna Cousijn, and A. Dilene van Campen. 2022. Action Intentions, Predictive Processing, and Mind Reading: Turning Goalkeepers Into Penalty Killers. *Frontiers in Human Neuroscience* 15 (1 2022), 789817. https://doi.org/10.3389/FNHUM.2021.789817

[45] George Ritzer. 2003. *The Blackwell Companion to Major Contemporary Social Theorists*. Blackwell Publishing Ltd, Malden, MA, USA, –384.

[46] Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M Kitani, Dariu M Gavrila, and Kai O Arras. 2020. Human motion trajectory prediction: a survey. *The International Journal of Robotics Research* 39, 8 (2020), 895–935. https://doi.org/10.1177/0278364920917446 arXiv:https://doi.org/10.1177/0278364920917446

[47] Emanuel A. Schegloff. 2001. Accounts of Conduct in Interaction: Interruption, Overlap, and Turn-Taking. In *Handbook of Sociological Theory*, Jonathan H. Turner (Ed.). Springer US, Boston, MA, 287–321. https://doi.org/10.1007/0-387-36274-6_15

[48] Sofia Schlamp, Fabiola H. Gerpott, and Sven C. Voelpel. 2021. Same talk, different reaction? Communication, emergent leadership and gender. *Journal of Managerial Psychology* 36, 1 (30 Jan. 2021), 51–74. https://doi.org/10.1108/JMP-01-2019-0062

[49] Viktor Schmuck and Oya Celiktutan. 2021. GROWL: Group Detection With Link Prediction. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE Press, Jodhpur, India, 1–8. https://doi.org/10.1109/FG52635.2021.9667061

[50] Stephanie Tan, David M.J. Tax, and Hayley Hung. 2022. Conversation Group Detection With Spatio-Temporal Context. In *Proceedings of the 2022 International Conference on Multimodal Interaction* (Bengaluru, India) *(ICMI '22)*. Association for Computing Machinery, New York, NY, USA, 170–180. https://doi.org/10.1145/3536221.3556611

[51] Emma M Templeton, Luke J Chang, Elizabeth A Reynolds, Marie D Cone LeBeaumont, and Thalia Wheatley. 2023. Long gaps between turns are awkward for strangers but not for friends. *Philosophical Transactions of the Royal Society B* 378, 1875 (2023), 20210471.

[52] Sydney Thompson, Abhijit Gupta, Anjali W. Gupta, Austin Chen, and Marynel Vázquez. 2021. Conversational Group Detection with Graph Neural Networks. In *Proceedings of the 2021 International Conference on Multimodal Interaction* (Montréal, QC, Canada) *(ICMI '21)*. Association for Computing Machinery, New York, NY, USA, 248–252. https://doi.org/10.1145/3462244.3479963

[53] Jose Vargas-Quiros, Laura Cabrera-Quiros, Catharine Oertel, and Hayley Hung. 2023. *Impact of Annotation Modality on Label Quality and Model Performance in the Automatic Assessment of Laughter In-the-Wild*. IEEE Transactions on Affective Computing (2023). https://doi.org/10.1109/TAFFC.2023.3269003

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

[55] Alexandra Weidemann and Nele Rußwinkel. 2021. The Role of Frustration in Human–Robot Interaction – What Is Needed for a Successful Collaboration? *Frontiers in Psychology* 12 (2021), 17 pages. https://doi.org/10.3389/fpsyg.2021.640186

[56] Marcin Wlodarczak and Mattias Heldner. 2020. Breathing in Conversation. *Frontiers in Psychology* 11 (2020), 17 pages. https://doi.org/10.3389/fpsyg.2020.575566

[57] Marcin Wlodarczak and Mattias Heldner. 2020. Breathing in Conversation. *Frontiers in Psychology* 11 (10 2020), 2574. https://doi.org/10.3389/FPSYG.2020.575566/BIBTEX

[58] Tianyi Zhang, Abdallah El Ali, Chen Wang, Alan Hanjalic, and Pablo Cesar. 2020. RCEA: Real-time, Continuous Emotion Annotation for Collecting Precise Mobile Video Ground Truth Labels. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (, Honolulu, HI, USA,) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376808