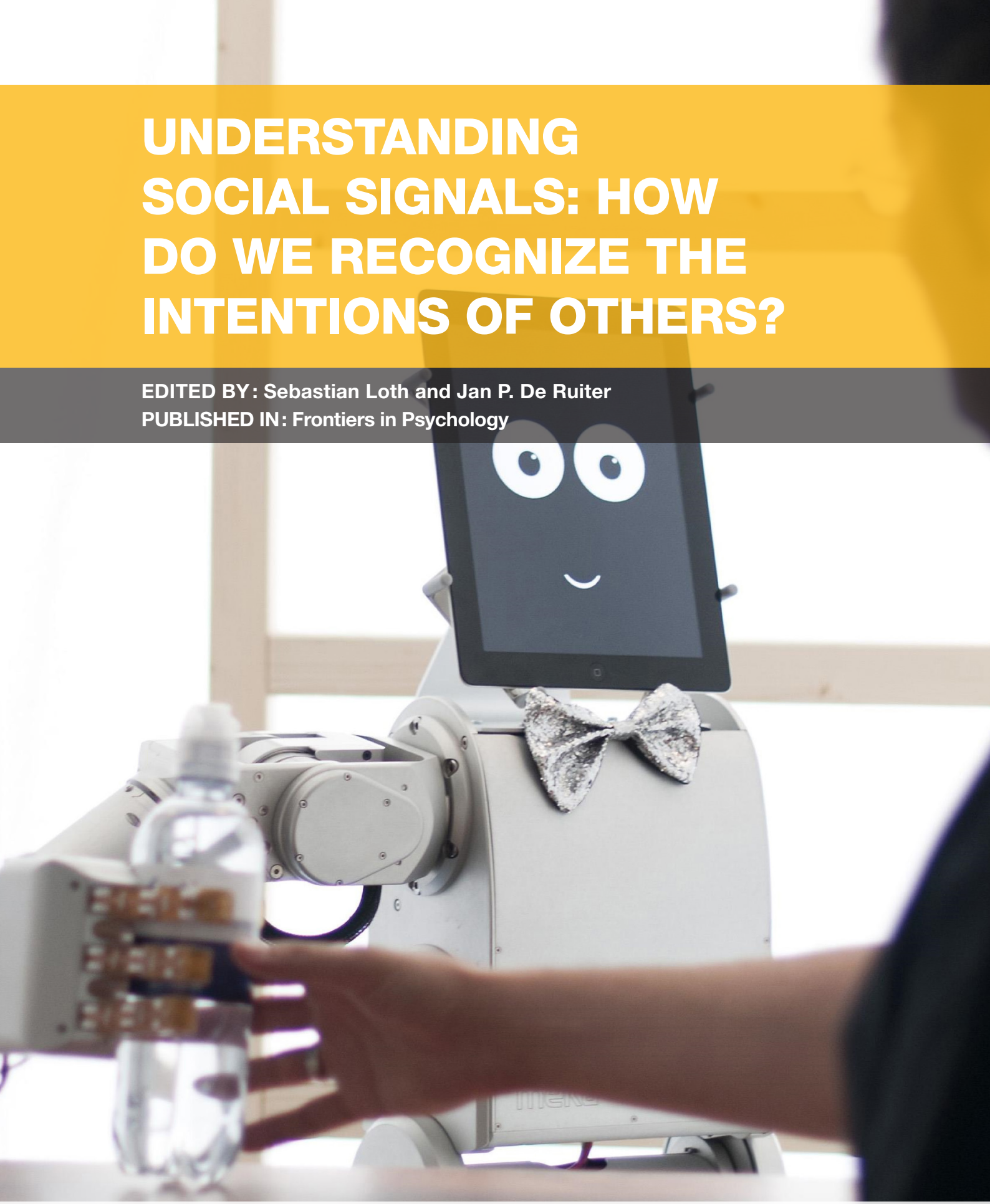
UNDERSTANDING SOCIAL SIGNALS: HOW DO WE RECOGNIZE THE INTENTIONS OF OTHERS?

EDITED BY: Sebastian Loth and Jan P. De Ruiter
PUBLISHED IN: Frontiers in Psychology



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88919-845-0
DOI 10.3389/978-2-88919-845-0

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: **researchtopics@frontiersin.org**

UNDERSTANDING SOCIAL SIGNALS: HOW DO WE RECOGNIZE THE INTENTIONS OF OTHERS?

Topic Editors:
Sebastian Loth, Bielefeld University, Germany
Jan P. De Ruiter, Bielefeld University, Germany

The bartending robot JAMES in a social interaction with its customer and serving a drink at fortiss in Munich, Germany. Copyright: fortiss www.fortiss.org

Powerful and economic sensors such as high definition cameras and corresponding recognition software have become readily available, e.g. for face and motion recognition. However, designing user interfaces for robots, phones and computers that facilitate a seamless, intuitive, and apparently effortless communication as between humans is still highly challenging. This has shifted the focus from developing ever faster and higher resolution sensors to interpreting available sensor data for understanding social signals and recognising users' intentions.

Psychologists, Ethnologists, Linguists and Sociologists have investigated social behaviour in human-human interaction. But their findings are rarely applied in the human-robot interaction domain. Instead, robot designers tend to rely on either proof-of-concept or machine learning based methods. In proving the concept, developers effectively demonstrate that users are able to adapt to robots deployed in the public space. Typically, an initial period of collecting human-robot interaction data is used for identifying frequently occurring problems. These are then addressed by adjusting the interaction policies on the basis of the collected data. However, the updated policies are strongly biased by the initial design of the robot and might not reflect natural, spontaneous user behaviour. In the machine learning approach, learning algorithms are used for finding a mapping between the sensor data space and a hypothesised or estimated set of intentions. However, this brute-force approach ignores the possibility that some signals or modalities are superfluous or even disruptive in intention recognition. Furthermore, this method is very sensitive to peculiarities of the training data. In sum, both methods cannot reliably support natural interaction as they

crucially depend on an accurate model of human intention recognition. Therefore, approaches to social robotics from engineers and computer scientists urgently have to be informed by studies of intention recognition in natural human-human communication.

Combining the investigation of natural human behaviour and the design of computer and robot interfaces can significantly improve the usability of modern technology. For example, robots will be easier to use by a broad public if they can interpret the social signals that users spontaneously produce for conveying their intentions anyway. By correctly identifying and even anticipating the user's intention, the user will perceive that the system truly understands her/his needs. Vice versa, if a robot produces socially appropriate signals, it will be easier for its users to understand the robot's intentions. Furthermore, studying natural behaviour as a basis for controlling robots and other devices results in greater robustness, responsiveness and approachability. Thus, we welcome submissions that (a) investigate how relevant social signals can be identified in human behaviour, (b) investigate the meaning of social signals in a specific context or task, (c) identify the minimal set of intentions for describing a context or task, (d) demonstrate how insights from the analysis of social behaviour can improve a robot's capabilities, or (e) demonstrate how a robot can make itself more understandable to the user by producing more human-like social signals.

Citation: Loth, S., De Ruiter, J. P., eds. (2016). Understanding Social Signals: How Do We Recognize the Intentions of Others? Lausanne: Frontiers Media. doi: 10.3389/978-2-88919-845-0

Table of Contents

frontiers in Psychology

EDITORIAL
published: 23 February 2016
doi: 10.3389/fpsyg.2016.00281

Editorial: Understanding Social Signals: How Do We Recognize the Intentions of Others?

Sebastian Loth and Jan P. De Ruiter*

Linguistics and Literary Studies, Psycholinguistics, Bielefeld University, Bielefeld, Germany

Keywords: social signals, social communication, intention recognition, human–robot interaction, human–human interaction, experimental methods, interaction design

The Editorial on the Research Topic

Understanding Social Signals: How Do We Recognize the Intentions of Others?

Humans interact with each other seamlessly, smoothly, and without obvious effort. Social signals are the basis of this highly effective communication. These signals are speech utterances, body movements such as gestures, manipulations of objects, and combinations thereof. For example, interlocutors typically position themselves in an F-formation (Goffman, 1963; Ciolek and Kendon, 1980; Kendon, 1990) and thereby signal to each other that they are part of that interaction. If another participant joins that interaction, the interlocutors integrate her in a new F-formation. The movements of each individual were comparably inconspicuous, but the intention for producing them was easily recognizable to the recipient. Humans use these signals intuitively and without conscious awareness. But in order to enable a robot to understand and respond appropriately to social signals, their form and function have to be made explicit. This research topic presents methods for identifying, understanding, and applying social signals in human–machine interaction.

Social signals are essentially multimodal but the analysis of human communication in human–machine interaction is often limited to the literal content of verbal utterances. For example, emotion has often been regarded as separate information that is specifically transferred through non-verbal signals, e.g., smiling. But Mehu argues that emotion is an inherent property of any social signal. The addressee would use the signal's emotional and literal content for determining how to respond to it. Identifying a signal's content requires combining and interpreting information from several modalities, taking into account the observer's prior experience. For example, Saegusa et al. show that a smiley next to a text message alters its perceived earnestness but its effect was more pronounced in hearing than non-hearing participants. Children also rely on multimodal signals for learning new words for objects. Hung et al. demonstrate that the children's strategic use of pointing gestures and spoken words depends on their linguistic experience, in particular if the gestures' reliability has been manipulated. Robotic recognizers have to combine data from sensors such as cameras and microphones for identifying objects and actions; a human is perceived as an entity with properties such as distance, body direction, and recent utterances. Similar to human observers, a robot requires detailed prior knowledge about social signals in order to interpret them. In a so-called "Ghost-in-the-Machine" study, Loth et al. show that human participants can identify social signals from the recognizer data of a robotic bartender. The study also shows that non-verbal signals were most important for initiating an interaction, whereas verbal signals were most important when placing an order. Multimodal signals unfold over time, and some features are available earlier than others. For example, the speaker's eye gaze reliably indicates the target of a selection task and preceded corresponding verbal utterances by almost 2 s in Huang et al.'s study of dyadic interactions. Thus, humans and robots can use this time for forming expectations about

OPEN ACCESS

Edited and reviewed by:
Eddy J. Davelaar,
Birkbeck, University of London, UK

****Correspondence:***
Sebastian Loth
Sebastian.Loth@uni-bielefeld.de

Specialty section:
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology

Received: *15 December 2015*
Accepted: *12 February 2016*
Published: *23 February 2016*

Citation:
Loth S and De Ruiter JP (2016) Editorial: Understanding Social Signals: How Do We Recognize the Intentions of Others?. Front. Psychol. 7:281. doi: 10.3389/fpsyg.2016.00281

the verbal utterance and planning ahead. Understanding social signals is important for fulfilling a task, e.g., serving a drink. In addition to task performance, Shalev and Oron-Gilead argue that social signals are also crucial in regulating the assertiveness of companion robots, i.e., should the robot take the initiative or wait for being prompted.

Industrial robots do not socially interact with humans but operate as a tool repeating precisely the same actions. Sciutti et al. suggest to use the robot's ability to exactly reproduce behavior in order to investigate social signals in controlled natural settings, e.g. what kind of object manipulation a participant expects from a hand position. Furthermore, this enables research in dynamic environments, allowing Katevas et al. to investigate social signals with a robot stand-up comedian. They found that the robot's gaze behavior was an important signal for eliciting laughter. In contrast to comedy, questionnaires of the US census have been standardized with the aim of eliciting accurate responses independently of the interviewer's performance. However, Conrad et al. show that the verbal skills but not the facial animations of a virtual interviewer influence the accuracy of the participants' responses.

Interactions can and often do go wrong. However, if problems are repaired swiftly, the interaction is still perceived as smooth. Schegloff (Schegloff et al., 1977; Schegloff, 1992) argued that the speaker can repair a problem in her own utterance immediately (first position repair) or the hearer would try to initiate the repair (second position repair). In a third position repair, the hearer's response revealed a problem to the speaker allowing her to repair this in her next turn. Importantly, repairs require that the problem has been detected in the first place. After analyzing video recordings of human–robot interactions, Giuliani et al. conclude that users initially stopped moving when they encountered a problem. This could be used as a signal for the robot to initiate a first position repair immediately. The user's second position repairs involved many head gestures and lots of smiling signaling the robot that there was a problem. Some of the speakers' behaviors typically synchronize during an interaction such that a de-synchronization can indicate a communication error. For example, Andrist et al. show that the speakers' eye gazes typically settle on particular objects in a selection task. A deviation from this pattern indicates that a problem in the communication had occurred which required an explicit repair later in the interaction. Thus, detecting this cue allows both humans and robots to initiate a first position repair and resolve the problem instantly. Similar to gaze behavior, body movements synchronize during an interaction. Avril et al. augment a play session of children and their care-givers with sensors typically used in human–machine interaction. They show that prolonged periods of avoidance behaviors and asynchrony of body movements could indicate severe conditions such as child neglect.

All studies in this research topic underscored the fact that human communication is based on the exchange of social signals that are essentially multimodal. If these signals deviated from expected patterns, this indicated problems in the communication. The pattern of deviation identified the type of problem suggesting how to repair it. Furthermore, the absence of social signals can indicate severe psychological conditions. Thus, social signals are highly diagnostic, both in "normal" and problematic communication. They provide intuitive means for controlling the robot's current task and its relation to its user. However, understanding and producing social signals depends on prior knowledge in both humans and robots. To summarize, this research topic combines the research of psychologists and robot designers to contribute to our understanding of social signals and applies these insights to human–machine interaction.

AUTHOR CONTRIBUTIONS

JDR and SL wrote, discussed, and revised several drafts before approving the final version.

FUNDING

This research/work was supported by the Cluster of Excellence Cognitive Interaction Technology "CITEC" (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).

ACKNOWLEDGMENTS

We thank all authors and reviewers for their time, effort, and particularly for contributing their research and thought to the success of this research topic.

REFERENCES

Ciolek, T. M., and Kendon, A. (1980). Environment and the spatial arrangement of conversational encounters. *Sociol. Inq.* 50, 237–271.

Goffman, E. (1963). *Behaviour in public places.* Galt, ON: Collier-Macmillan Canada Ltd. Available online at: http://solomon.soth.alexanderstreet.com/cgi-bin/asp/philo/soth/getdoc.pl?S10019969--D000001

Kendon, A. (ed.) (1990). "Spatial organization in social encounters: the F-formation system," in *Conducting Interaction: Patterns of Behavior in Focused Encounters* (Cambridge; New York, NY: Cambridge University Press), 209–238.

Schegloff, E. A. (1992). Repair after next turn: the last structurally provided defense of intersubjectivity in conversation. *Am. J. Sociol.* 97, 1295. doi: 10.1086/229903

Schegloff, E. A., Jefferson, G., and Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language* 53, 361–382.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

PERSPECTIVE
published: 07 July 2015
doi: 10.3389/fpsyg.2015.00961

The integration of emotional and symbolic components in multimodal communication

*Marc Mehu**

Department of Psychology, Webster Vienna Private University, Vienna, Austria

Human multimodal communication can be said to serve two main purposes: information transfer and social influence. In this paper, I argue that different components of multimodal signals play different roles in the processes of information transfer and social influence. Although the symbolic components of communication (e.g., verbal and denotative signals) are well suited to transfer conceptual information, emotional components (e.g., non-verbal signals that are difficult to manipulate voluntarily) likely take a function that is closer to social influence. I suggest that emotion should be considered a property of communicative signals, rather than an entity that is transferred as content by non-verbal signals. In this view, the effect of emotional processes on communication serve to change the quality of social signals to make them more efficient at producing responses in perceivers, whereas symbolic components increase the signals' efficiency at interacting with the cognitive processes dedicated to the assessment of relevance. The interaction between symbolic and emotional components will be discussed in relation to the need for perceivers to evaluate the reliability of multimodal signals.

Keywords: emotional communication, multimodal communication, social signals, ethology, non-verbal communication, pragmatics

OPEN ACCESS

Edited by:
Sebastian Loth, Universität Bielefeld, Germany

Reviewed by:
Han-Seok Seo, University of Arkansas, USA
Tom Ziemke, University of Skövde and Linköping University, Sweden
Shlomo Hareli, University of Haifa, Israel

***Correspondence:**
Marc Mehu, Department of Psychology, Webster Vienna Private University, 23 Praterstrasse, Vienna A-1020, Austria
marc.mehu@webster.ac.at

Specialty section:
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology

Received: *15 March 2015*
Accepted: *26 June 2015*
Published: *07 July 2015*

Citation:
Mehu M (2015) The integration of emotional and symbolic components in multimodal communication. Front. Psychol. 6:961. doi: 10.3389/fpsyg.2015.00961

Introduction

This article revolves around two ideas that have stayed, in my opinion, on the fringes of research in human communication. The first idea is that the primary function of social signals is to influence perceivers, i.e., to produce responses in others that are beneficial to signalers. This idea has been discussed extensively in the field of animal behavior (Owren et al., 2010; Stegmann, 2013), but less so in human communication (for an exception, see Owren and Bachorowski, 2003). The second idea defended in this paper is that multimodal communication has emerged as the result of an interaction, over human evolutionary history, between signaler and perceiver roles (for a similar argument in animal communication research, see Guilford and Dawkins, 1991; Rowe, 1999). In particular, I will argue that different elements of multimodal signals have evolved to address the selective pressures presented by two cognitive strategies perceivers use to process and respond to signals: an evaluation of relevance and an assessment of reliability. The goal of this article is to call for an integration of information transfer and social influence accounts of communication in a coherent framework informed by evolutionary theory.

Information Transfer and Social Influence in Human Communication

Social signals have been studied in many disciplines and the range of definitions for this concept is relatively broad (Mehu et al., 2012a). Borrowing the terminology developed by ethologists to define animal signals (see Maynard Smith and Harper, 2003), Mehu and Scherer (2012) have proposed a definition of human social signals that integrates the symbolic character of human communication with the more evolutionarily ancient properties of animal signals: "human social signals are acts or structures that influence the behavior or internal state of other individuals, that evolve because of that effect, and that are effective because the perceiver's response has also evolved; signals may or may not convey conceptual information or meaning" (p. 399). In this framework, human social signals are tangible units of communication that can be perceived as visual, auditory, olfactory, or tactile stimuli. As such, the physical properties of social signals constitute raw information on the basis of which perceivers take social decisions. These decisions can be adaptive for the perceiver if the signal's physical properties correlate with some psychobiological processes in which a perceiver has an interest (e.g., reproductive state, behavioral intentions, attitudes, cognitive evaluations, physiological reactions, subjective feelings, etc.). Although the material properties of social signals can be correlated with unobservable psychobiological processes, to conflate signals and their possible referents is inadequate as it does not help understand the complexity of communication. For example, it is hard to conceive that unobservable psychobiological processes are at the same time social signals and the referents of non-verbal communicative units. A more plausible assumption is that social signals are the means with which psychobiological processes like cognition, emotion, and attitudes are implemented in everyday social interactions.

Another aspect of social signals is their evolutionary function, i.e., how social signals increase survival and reproductive success of the individual who displays them. In the framework presented here, a signal's function is to produce a response in the perceiver that is adaptive to the signaler. I argue that a signal fulfills its function in a number of ways, and the diversity of these ways results from an evolutionary process whereby perceivers (the main targets of signals) have placed selective pressures on signalers in order to maximize adaptation to the social environment and to avoid social exploitation. The pairing between conceptual information and physical properties of non-verbal behavior is one way social signals achieve their function of social influence. Therefore, information transfer is a way social signals fulfill their function of social influence rather than a separate function in itself (see also Scarantino, 2013).

The mainstream view on non-verbal communication posits that signalers encode information relative to internal states (emotion, cognition, attitudes, social motives, and dispositions) in a signal, and that the signal is decoded by perceivers who then retrieve the information. Although the inference of a signaler's internal states is important for perceivers in most social situations, the faithful encoding of these states may not always be adaptive for signalers themselves because perceivers may act against a signalers' goals (Grammer et al., 1997). In my opinion, the fact that it is so important for perceivers to form a reliable representation of the social environment has inflated the importance, in the eyes of psychologists, of the disclosure of unobservable psychobiological processes by signalers. There are many situations in which signalers have an advantage either in concealing information that could be used by the perceiver to act at the expense of the signaler, or in using deception. As a general rule, when there is a conflict of interest between signalers and perceivers, it is expected that the signaler will (a) retain valuable information, (b) try to influence perceivers to its own advantage, or (c) use deceptive signals (Maynard Smith and Harper, 2003). On the other hand, when a given interactive outcome is advantageous for both signalers and perceivers it is expected that reliable transfer of information will take place because none of the parties involved would benefit from deceiving the other. Krebs and Dawkins (1984) have argued that the nature of signals should depend on these contextual aspects. For example, when signaler and perceiver have conflicting interests, signals will tend to be more intense in order to be more effective in producing a response in perceivers; while signals will be less conspicuous when signalers and perceivers both benefit from reliable disclosure of internal states and behavioral intentions. Therefore, contextual factors determine whether it is adaptive for a signaler to accurately convey internal states, or to make strategic efforts to either conceal their intentions or use manipulation tactics. This implies that inferences made by perceivers will not only be based on the signal itself but also on how the signal interacts with situational cues.

A model of communication that is purely based on information transfer is unlikely to help us understand the complexity of communication (Wilson and Sperber, 2006; Owren et al., 2010; Scott-Phillips and Kirby, 2013). More specifically, such a model would fail to recognize that the roles of signaler and perceiver, although complementary, have different functions. On the one hand, signalers produce signals that have a high impact on perceivers and, on the other hand, perceptual systems optimize the use of information that can be gleaned from the situation in which communication takes place. This idea is based on Owings and Morton's (1997) model of animal communication whereby communication is seen as a dynamic process that entails the management and assessment of the social environment by signalers and perceivers. In this model, information transfer is seen as secondary and is considered adaptive in only a fraction of the situations in which people communicate (Owren et al., 2010), namely when signalers and perceivers would both benefit from the reliable transfer of information. Therefore, depending on the situation, the communication process does not serve signalers and perceivers in the same way. Although in most cases the signaler should benefit from producing a desired response in the perceiver, the latter should mostly benefit from gaining adaptive social information. The signaler would only benefit from sending reliable signals when perceivers make their responses conditional on the acquisition of relevant and reliable information. It is the task of the researcher to determine whether the situation favors reliable transfer of information or social influence. This is likely to depend on the costs incurred by perceivers to respond to a signaler's displays. Consequently, the interaction between

information transfer and social influence will depend on the respective costs incurred by signalers and perceivers in a given context.

The Complexity of Human Communication is Reflected in Multimodal Signals

The concept of multimodal communication follows the observation that social signals are complex and cover several sensory modalities (Johnstone, 1996; Partan and Marler, 1999; Rowe, 1999). Signals can also have multiple components within a particular modality. For example, visual signals entail motor components (movements produced by muscular activity), morphological components (structure and shape of particular body parts), or color components (e.g., skin or hair coloration). Within the framework of information transfer, multimodal signals have been proposed to function in two different ways (Johnstone, 1996): By redundantly encoding the same information in several channels, and by varying the nature of the information conveyed in the different channels. The first solution (backup signals) ensures that the message is transmitted, even when environmental circumstances prevent one of the channels to operate (e.g., in poor light conditions, or in noisy environments). The second solution (multiple messages) increases the amount of information transferred by using different channels to convey additional information. From a social influence perspective, multimodal signals could be more efficient at influencing perceivers because their complex structure makes them better at interacting with perceivers' psychological mechanisms. Evidence from the field of animal communication suggests that multimodal signals are more easily detected, discriminated, and memorized (Rowe, 1999). In humans, the presentation of audio-visual signals appear to have a different impact on perceivers than the separate presentation of single modalities (Mehu and van der Maaten, 2014).

The present article defends the idea that the combination between different components or modalities of a signal has evolved to meet the requirement imposed by assessment systems. Perceivers' social decisions have relied increasingly on mental inferences involving the interaction between multiple indicators (mostly cues and signals emitted by signalers as well as situational features). Such inferences could function to resist social exploitation and manipulation by signalers and to optimize social decision making. Increasing cognitive complexity in primates (Dunbar and Shultz, 2007) placed a selective pressure on signals to become more efficient at interacting with perceiver's filtering mechanisms. On the other hand, evolutionary in-built robustness in primate signals could be a fertile bed for the evolution of more complex signals, which components could take on new functions in communication (Ay et al., 2007). I consider the transfer of abstract and conceptual information as one of these new functions, which is fulfilled by the symbolic components of multimodal signals. The symbolic components cut across the visual and auditory modalities (visual symbols and speech are two examples of these components in two different modalities) and their form is relatively arbitrary with regards to their function, which is to interact with representational structures of the mind. The evolution of the symbolic component of human communication has paralleled the development of voluntary motoric capabilities necessary for the production of communicative units at the acoustic (speech) but also the visual level (e.g., gestures). The increased voluntary control over this component facilitated signal production and the expression of intentions, but it also created a new opportunity for signalers to take advantage of perceivers' assessment systems and produce potentially deceptive signals.

The integration of symbolic components in communication enhances the efficiency of signals at producing desired responses in perceivers because such components allow signalers to provide information which perceivers have an interest in. By using symbolic components, signalers can also clarify the type of response sought in perceivers. By bringing elements that are absent in the current situation but nonetheless relevant to it, the symbolic component of a signal also broadens the communication context and the perceivers' opportunities to pose adaptive actions. Symbolic communication is therefore adaptive for signalers because it helps them influence perceivers more efficiently. Due to the increased potential it offers for assessment of the physical and social environments, this mode of communication is likely to have been selected by perceivers during evolutionary history. The increased voluntary control over signal production (in particular over the symbolic components) allows more flexibility in communication and a better relationship with cognitive executive functions such as memory and planning. It also gives more opportunities for signalers to deceive perceivers by sending false information. This created a selective pressure on perceivers to develop resistance mechanisms designed to evaluate the reliability of the source. It has been argued that, in addition to the evaluation of an utterance's relevance, humans have developed cognitive mechanisms to evaluate a signaler's reliability (Sperber et al., 2010). I argue that when evaluating the trustworthiness of a signal, perceivers use other cues or indices present in the signal that are difficult to manipulate or control voluntarily, for example emotional components.

Emotional expressions have been found to strongly influence person perception (Knutson, 1996; Hess et al., 2000). In line with the idea that emotions are essential to maintain commitment to social contracts (Hirshleifer, 1987; Frank, 1988) it was found that emotional expressions could function as reliable indicators of behavioral intentions and interpersonal dispositions such as prosociality (Brown et al., 2003; Mehu et al., 2007) or threat (Reed et al., 2014). The reason why perceivers would rely on emotional cues to make adaptive social decisions is that these cues reflect automatic psychobiological processes that are responsible for the production of adaptive behavior that may also have implications for the perceiver's adaptation. Therefore, the initial stages of emotion-based behavioral sequences are informative as they allow to predict future behavior and anticipate adjustment to a situation. It is therefore adaptive for perceivers to react emotionally to a range of observed emotional cues, and to consider these cues as important sources of social information (van Kleef, 2009). Inferences about the behavioral intentions of signalers is an important goal for perceivers and the accuracy of

these inferences likely determines whether reactions to emotional displays of others are adaptive in the long run. Facial signs of enjoyment displayed in parallel to verbally expressed intentions to cooperate can be predictive of cooperative moves (Reed et al., 2012), suggesting that emotional signals could be used to evaluate the reliability of verbal claims. Therefore, when evaluating multimodal signals that contain symbolic components, perceivers could give particular attention to the emotional components of the signals as the latter may ensure the reliability of the former (Mehu and Scherer, 2012).

Emotional signals lead to emotional reactions in perceivers (Forgas, 1998; Owren and Bachorowski, 2003; van Kleef, 2009), and these emotional reactions can modify the perceiver's thoughts and behavior to the advantage of the signaler (van Kleef et al., 2004). I argue that emotionality is a property of multimodal communication that makes it more efficient at producing responses in perceivers that are adaptive to signalers. In this view, emotion does not represent the content of a signal that is encoded by a signaler in order to be decoded by a perceiver, but one of a signal's properties, which is activated by a series of automatic cognitive and physiological processes that are difficult to control voluntarily. There are two ways emotional processes make a signal more efficient. The first is by modifying the physical properties of the signal and making it more intense, more salient, and more variable, hence more difficult for perceivers to resist to, to ignore, or to habituate to. Second, emotional processes make a signal more efficient by acting on the cognitive processes which function is to evaluate a signal's authenticity. In support to this view, perceived authenticity of an expression is related to the intensity of the facial cues that are more difficult to control voluntarily (Mehu et al., 2012b). In this context, emotional authenticity can be conceived as the likelihood that the signal is associated with the cognitive, physiological, and experiential processes involved in the coordination of adaptive responses (Scherer, 2005). In other words, an emotionally authentic signal is a good predictor of the signaler's tendency to react in a particular situation. In day-to-day communication, emotional signals act in parallel to symbolic signals to make the overall multimodal signal appear more salient in the eyes (or ears) of perceivers and to make the information content of the symbolic component more reliable. Rather than to transfer information about emotional states, the function of emotional signals is therefore to optimize the effect of multimodal signals on perceivers.

Conclusion

The question of what is transmitted in non-verbal communication has kept researchers busy for the last decades. Looking for information about emotion or its components (Ekman et al., 1980; Scherer and Grandjean, 2008), information about social motives (Fridlund, 1994; Parkinson, 2005), information about personality (Hall et al., 2005), or information about attitudes (Mehrabian, 1971), non-verbal communication research has been on an incessant quest for signal meaning. In my opinion, the strong focus on questions of meaning is based on excessive reliance on the view that communication mostly functions to transfer information. Models of information transfer are useful to understand certain aspects of symbolic communication, but they have to be complemented with models that emphasize social influence. Such integration implies that we recognize the different functions associated with the roles of signaler and perceiver in communication. Although these two roles interact to a great extent and have co-evolved during human evolutionary history, one cannot necessarily assume that signalers' goals are to serve perceivers' goals. With this in mind, research should pursue questions related to what is achieved by communicative signals and by perceivers' assessment mechanisms, along with a careful analysis of the contextual factors and interactive consequences of multimodal displays.

I propose that multimodal signals that include both symbolic and emotional components are advantageous for signalers in that they are more likely to produce the adequate response in perceivers because (a) they contain information necessary for perceivers to evaluate the signal in relation to context (they target perceiver's evaluations of relevance) and (b) they show appropriate correlation with social information adaptive to perceivers (they target perceivers' evaluation of the trustworthiness of the source). Future research needs to clarify the processes involved in the production of multimodal signals (for example the appraisal processes underlying emotional communication, Mortillaro et al., 2013) as well as the role of abstract, language-based, representational structures as possible meditators of perceivers' responses to signals. Finally, investigating the costs and benefits for signaler and perceiver that are inherent to the context in which communication takes place should also constitute an important element of future study designs in social signal processing research.

References

Ay, N., Flack, J. C., and Krakauer, D. C. (2007). Robustness and complexity co-constructed in multimodal signalling networks. *Philos. Trans. R. Soc. B Biol. Sci.* 362, 441–447. doi: 10.1098/rstb.2006.1971

Brown, W. M., Palameta, B., and Moore, C. (2003). Are there nonverbal cues to commitment? An exploratory study using the zero-acquaintance video presentation paradigm. *Evol. Psychol.* 1, 42–69.

Dunbar, R. I. M., and Shultz, S. (2007). Evolution in the social brain. *Science* 317, 1344–1347. doi: 10.1126/science.1145463

Ekman, P., Friesen, W. V., and Ancoli, S. (1980). Facial signs of emotional experience. *J. Pers. Soc. Psychol.* 39, 1125–1134. doi: 10.1037/h0077722

Forgas, J. P. (1998). On feeling good and getting your way: mood effects on negotiator cognition and bargaining strategies. *J. Pers. Soc. Psychol.* 74, 565–577. doi: 10.1037/0022-3514.74.3.565

Frank, R. H. (1988). *Passions Within Reason: The Strategic Role of the Emotions*. New York, NY: Norton.

Fridlund, A. J. (1994). *Human Facial Expression: An Evolutionary View*. San Diego, CA: Academic Press.

Grammer, K., Fivola, V., and Fieder, M. (1997). "The communication paradox and possible solutions: towards a radical empiricism," in *New Aspects of Human Ethology*, eds A. Schmitt, K. Atzwanger, and K. Schäfer (New York, NY: Plenum Press), 91–120.

Guilford, T., and Dawkins, M. S. (1991). Receiver psychology and the evolution of animal signals. *Anim. Behav.* 42, 1–14. doi: 10.1016/S0003-3472(05)80600-1

Hall, J. A., Coats, E. J., and Smith LeBeau, L. (2005). Nonverbal behavior and the vertical dimension of social relations: a meta-analysis. *Psychol. Bull.* 131, 898–924. doi: 10.1037/0033-2909.131.6.898

Hess, U., Blairy, S., and Kleck, R. E. (2000). The influence of facial emotion displays, gender, and ethnicity on judgments of dominance and affiliation. *J. Nonverbal Behav.* 24, 265–283. doi: 10.1023/A:1006623213355

Hirshleifer, J. (1987). "On the emotions as guarantors of threats and promises," in *On the Emotions as Guarantors of Threats and Promises*, ed. J. Dupré (Cambridge, MA: MIT Press), 307–326.

Johnstone, R. A. (1996). Multiple displays in animal communication: 'backup signals' and 'multiple messages'. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 351, 329–338. doi: 10.1098/rstb.1996.0026

Knutson, B. (1996). Facial expressions of emotion influence interpersonal trait inferences. *J. Nonverbal Behav.* 20, 165–182. doi: 10.1007/BF02281954

Krebs, J. R., and Dawkins, R. (1984). "Animal signals: mind-reading and manipulation," in *Behavioural Ecology: An Evolutionary Approach*, Vol. 2, ed. N. B. Davies (Oxford: Blackwell Scientific Publications), 380–402.

Maynard Smith, J., and Harper, D. G. C. (2003). *Animal Signals*. Oxford: Oxford University Press.

Mehrabian, A. (1971). *Silent Messages*. Belmont, CA: Wadsworth.

Mehu, M., D'Errico, F., and Heylen, D. (2012a). Conceptual analysis of social signals: the importance of clarifying terminology. *J. Multimodal User Interfaces* 6, 179–189. doi: 10.1007/s12193-012-0091-y

Mehu, M., Mortillaro, M., Bänziger, T., and Scherer, K. R. (2012b). Reliable Facial muscle activation enhances recognizability and credibility of emotional expression. *Emotion* 12, 701–715. doi: 10.1037/a0026717

Mehu, M., Grammer, K., and Dunbar, R. I. M. (2007). Smiles when sharing. *Evol. Hum. Behav.* 28, 415–422. doi: 10.1016/j.evolhumbehav.2007.05.010

Mehu, M., and Scherer, K. R. (2012). A psycho-ethological approach to social signal processing. *Cogn. Process.* 13(Suppl. 2), 397–414. doi: 10.1007/s10339-012-0435-2

Mehu, M., and van der Maaten, L. J. P. (2014). Multimodal integration of dynamic audio—visual cues in the communication of agreement and disagreement. *J. Nonverbal Behav.* 38, 569–597. doi: 10.1007/s10919-014-0192-2

Mortillaro, M., Mehu, M., and Scherer, K. R. (2013). "The evolutionary origin of multimodal synchronisation and emotional expression," in *Evolution of Emotional Communication: From Sounds in Nonhuman Mammals to Speech and Music in Man*, eds E. Altenmüller, S. Schmidt, and E. Zimmermann (Oxford: Oxford University Press), 3–25.

Owings, D. H., and Morton, E. S. (1997). "The role of information in communication: an assessment/management approach," in *Perspectives in Ethology: Communication*, Vol. 12, eds M. D. Beecher and N. S. Thompson (New York, NY: Plenum Press), 359–390.

Owren, M. J., and Bachorowski, J.-A. (2003). Reconsidering the evolution of nonlinguistic communication: the case of laughter. *J. Nonverbal Behav.* 27, 183–200. doi: 10.1023/A:1025394015198

Owren, M. J., Rendall, D., and Ryan, M. J. (2010). Redefining animal signaling: influence versus information in communication. *Biol. Philos.* 25, 755–780. doi: 10.1007/s10539-010-9224-4

Parkinson, B. (2005). Do facial movements express emotions or communicate motives? *Pers. Soc. Psychol. Rev.* 9, 278–311. doi: 10.1207/s15327957pspr0904_1

Partan, S. R., and Marler, P. (1999). Communication goes multimodal. *Science* 283, 1272–1273. doi: 10.1126/science.283.5406.1272

Reed, L. I., DeScioli, P., and Pinker, S. A. (2014). The commitment function of angry facial expressions. *Psychol. Sci.* 25, 1511–1517. doi: 10.1177/0956797614531027

Reed, L. I., Zeglen, K. N., and Schmidt, K. L. (2012). Facial expressions as honest signals of cooperative intent in a one-shot anonymous prisoner's dilemma game. *Evol. Hum. Behav.* 33, 200–209. doi: 10.1016/j.evolhumbehav.2011.09.003

Rowe, C. (1999). Receiver psychology and the evolution of multicomponent signals. *Anim. Behav.* 58, 921–931. doi: 10.1006/anbe.1999.1242

Scarantino, A. (2013). "Animal communication as information-mediated influence," in *Animal Communication Theory: Information and Influence*, ed. U. E. Stegmann (Cambridge: Cambridge University Press), 63–81.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Soc. Sci. Inf.* 44, 695–729. doi: 10.1177/0539018405058216

Scherer, K. R., and Grandjean, D. (2008). Facial expressions allow inference of both emotions and their components. *Cogn. Emot.* 22, 789–801. doi: 10.1080/02699930701516791

Scott-Phillips, T. C., and Kirby, S. (2013). "Information, influence and inference in language evolution," in *Animal Communication Theory: Information and Influence*, ed. U. E. Stegmann (Cambridge: Cambridge University Press), 421–442.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., et al. (2010). Epistemic vigilance. *Mind Lang.* 25, 359–393. doi: 10.1111/j.1468-0017.2010.01394.x

Stegmann, U. (2013). *Animal Communication Theory: Information and Influence*. Cambridge: Cambridge University Press.

van Kleef, G. A. (2009). How emotions regulate social life: the emotions as social information (EASI) model. *Curr. Dir. Psychol. Sci.* 18, 184–188. doi: 10.1111/j.1467-8721.2009.01633.x

van Kleef, G. A., De Dreu, C. K. W., and Manstead, A. S. R. (2004). The interpersonal effects of anger and happiness in negotiations. *J. Pers. Soc. Psychol.* 86, 57–76. doi: 10.1037/0022-3514.86.1.57

Wilson, D., and Sperber, D. (2006). "Relevance Theory," in *Handbook of Pragmatics*, eds L. R. Horn and G. Ward (Oxford: Blackwell Publishing), 607–632.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 02 April 2015
doi: 10.3389/fpsyg.2015.00383

Interpreting text messages with graphic facial expression by deaf and hearing people

Chihiro Saegusa[1]*, *Miki Namatame*[2] *and Katsumi Watanabe*[1]

[1] *Research Center for Advanced Science and Technology, The University of Tokyo, Tokyo, Japan,* [2] *Department of Synthetic Design, Tsukuba University of Technology, Tsukuba, Japan*

In interpreting verbal messages, humans use not only verbal information but also non-verbal signals such as facial expression. For example, when a person says "yes" with a troubled face, what he or she really means appears ambiguous. In the present study, we examined how deaf and hearing people differ in perceiving real meanings in texts accompanied by representations of facial expression. Deaf and hearing participants were asked to imagine that the face presented on the computer monitor was asked a question from another person (e.g., do you like her?). They observed either a realistic or a schematic face with a different magnitude of positive or negative expression on a computer monitor. A balloon that contained either a positive or negative text response to the question appeared at the same time as the face. Then, participants rated how much the individual on the monitor really meant it (i.e., perceived earnestness), using a 7-point scale. Results showed that the facial expression significantly modulated the perceived earnestness. The influence of positive expression on negative text responses was relatively weaker than that of negative expression on positive responses (i.e., "no" tended to mean "no" irrespective of facial expression) for both participant groups. However, this asymmetrical effect was stronger in the hearing group. These results suggest that the contribution of facial expression in perceiving real meanings from text messages is qualitatively similar but quantitatively different between deaf and hearing people.

Keywords: smileys, text interpretation, chat, social signals, earnestness, deaf, hearing

OPEN ACCESS

Edited by:
Sebastian Loth, Universität Bielefeld, Germany

Reviewed by:
Teresa Mitchell, University of Massachusetts Medical School, USA
Christine Howes, Queen Mary University of London, UK
Gregory J. Mills, University of Groningen, Netherlands

***Correspondence:**
Chihiro Saegusa, Research Center for Advanced Science and Technology, The University of Tokyo, 4-6-1, Komaba, Meguro-ku, Tokyo 153-8904, Japan
csaegusa@fennel.rcast.u-tokyo.ac.jp

Specialty section:
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology

Received: *11 January 2015*
Accepted: *18 March 2015*
Published: *02 April 2015*

Citation:
Saegusa C, Namatame M and Watanabe K (2015) Interpreting text messages with graphic facial expression by deaf and hearing people. Front. Psychol. 6:383. doi: 10.3389/fpsyg.2015.00383

Introduction

Interpreting verbal messages, perceiving others' real meaning, and responding to them appropriately are important in successful communication. In some cases, the meanings are communicated directly in a verbal form, but in most cases, we infer them by cues that are provided explicitly or implicitly (Duncan, 1969). Most of the cues that signal the real meanings might be in visual or auditory modalities. For instance, expressions of emotion in the face and through body movement would be cues in the visual modality, whereas prosody such as speed, intonation, and accent of the voice would be cues in the auditory modality (Scherer et al., 1991; Banse and Scherer, 1996).

The recent increase of human–computer interaction and human–human communication via computer requires a person to use similar yet slightly different communication styles than a face-to-face communication. The major difference is the amount of information and relative contribution

of different sensory modalities that are being accessed. For example, in a computer-mediated communication such as e-mails and online chats, we convey our thoughts mainly with text messages. Thus, there are fewer non-verbal cues for emotion that would otherwise play an important role in inferring real meaning in face-to-face communications. Emoticons and avatars are used to replace non-verbal cues in computer-mediated communications. It has been reported that emotional expression by such methods can modify inference of meanings from text messages (Derks et al., 2008).

Indeed, facial expression is a rich source of information on emotional states for the beholder and is considered the most important cue. It has been proposed that the perception of human facial expression is universal regardless of culture in most cases (Ekman et al., 1969; Ekman and Friesen, 1971), with some cognitive and behavioral differences in interpreting facial expressions, for example, with regard to perceiving the intensity of emotions (Ekman et al., 1987), integrating social context into emotion judgment (Masuda et al., 2008), mental representations (Jack et al., 2012), and fixation maps (Jack et al., 2009). Since there are many cultural differences in the cognitive process in addition to the difference in cognition of facial expression of emotion (for review, Nisbett and Masuda, 2003), the findings may reflect a general difference in cognitive process across cultures rather than differences specific to facial expression. Facial expression is useful not only for perceiving emotional states of the communicator but also in judging deception (Feldman et al., 1979). For instance, Ekman and Friesen (1974) demonstrated that people utilized both facial and body cues when detecting deception from videotaped interviews in which models acted out both honest and deceptive responses.

Similar to the cross-cultural commonalities and differences, deaf people and people with normal hearing share a common perception of expression of emotion, while using different eye movement paths in collecting information from the face (Watanabe et al., 2011). In addition, previous research has demonstrated differences between deaf people and hearing people in the perceptual and cognitive processing of faces when memorizing (Arnold and Murray, 1998) and discriminating faces, especially in discriminating the local features of faces (Bettger et al., 1997; McCullough and Emmorey, 1997). As McCullough et al. (2005) discussed, such differences might be due to deaf people's constant attention to componential facial features versus hearing people's constant attention to holistic facial information, and these differences might influence and/or be influenced by other cognitive processes, for example, how the facial expression information is integrated with information from other modalities.

Although facial expression is essential to understanding the emotional state of others, it is rarely used independently. Rather, it is integrated with other information. For example, in the perception of real intention based on verbal, vocal, and visual input, the perception of positivity in the affective message expressed in one modality is discounted when there is contradictory input from other modalities (Bugental et al., 1970; Friedman, 1979). However, the difference in the cognitive processing of facial expression between deaf and normal hearing people may result in a different usage of facial expression information when integrating it with information from other modalities to infer real meanings from text messages made by others.

Thus in the current study, we aimed to improve understanding of how the use of facial expression in perceiving real meanings from text messages differs between deaf and hearing adults, depending on combinations of verbal information presented as texts together with facial expressions of emotion to convey either consistent or contradictory contents.

We had two hypotheses for the current study. The first refers to the communication strategy in deaf people. In addition to the difference in gaze strategy during processing emotional expression of the face (Watanabe et al., 2011), there are a few reports suggesting a difference between deaf and hearing people in the usage of non-verbal cues when communicating with others (Barnett, 2002). For example, it was reported that differences in interpreting non-verbal gestures including body posture and facial expression may lead to misunderstandings between a deaf patient and his or her hearing physician (Barnett, 2002). However, to our knowledge, the exact contexts and situations for such misunderstandings remain unclear. In the current study, we investigated how facial emotion expression on a computer monitor would affect the inference of real meaning behind the explicitly presented text responses. Our prediction was that deaf people regard visual facial expressions as more useful sources for interpreting the text messages because they have less access to auditory cues (e.g., prosodic sounds). The second hypothesis refers to the politeness assumption (Brown and Levinson, 1987); that is, how a participant assumes the person/agent in the conversation as being polite may depend on the conversation context. The communication strategy might differ depending on the situation, especially when the response is a negative one. In order to examine this, we chose the following two questions: Asking someone for a favor and asking about liking another person. Asking someone for a favor occurs in a conversation between two persons. A negative response would not be desirable for the questioner. Such a situation requires the assumption that the answerer would avoid explicitly expressing a negative response but would employ an implicit way (e.g., negative facial expression). On the other hand, asking about liking another person who is not present in the conversation would threaten the relationship between the pair less although it might still not be socially desirable. Therefore, we expected that the influence of emotion expression as a non-verbal cue would be smaller. We further predicted that, if the response was positive, such a difference between emotion expressions would not be observed.

In our experiment, both realistic and schematic faces were investigated because we assumed that, irrespective of hearing ability and history, there might be a general difference in the amount of emotional signals that can be extracted from these types of faces (Wallraven et al., 2007) and a difference in strategy that observers take while seeing them. For example, it was reported that gaze behavior for recognizing schematically drawn faces and natural-looking faces is different, and that schematically drawn faces facilitate analytical processing (Schwarzer et al., 2005). Further evidence for the different strategies can be found in face recognition (Rosset et al., 2010) and in emotional processing of schematic faces in patients with autistic spectrum disorder (Rosset et al., 2008). In addition, understanding the possible differences

between face types would be informative when applying these findings to human–computer interaction because the agents on the computer are often abstract representations of a person.

Although a computer-generated (CG) face is not animated and thus may not have an intention as in the pragmatic and philosophical literature (Grice, 1969; Sperber and Wilson, 2002), humans tend to extract meaning from what is displayed on the face (Öhman, 2002). Thus in the current study, we investigated perception of the real meaning of what was conveyed via verbal message and emotion expression of the face with different levels of consistency. We were especially focused on whether participants' inferences of the meaning that is explicitly (e.g., verbally) explained would be affected by emotional valence that is displayed on the face.

Materials and Methods

Participants

The participants included 20 deaf Japanese people and 36 Japanese people with normal hearing function. All deaf participants were undergraduate students at the Tsukuba University of Technology, where hearing loss of 60 dB or more is one of the requirements for admission. Data from five hearing participants were excluded because the session for expression rating was not completed. The remaining data from 20 deaf participants (6 males and 14 females; mean age = 21.1 years old, SD = 1.0) and 31 participants with normal hearing function (20 males and 11 females; mean age = 21.2 years old, SD = 1.6) were used for the analyses.

Visual Stimuli

Schematic faces and CG faces with a stepwise emotional expression manipulation were used in the experiment (**Figure 1**). In the schematic faces, to express positive and negative emotions, the shapes and height of the eyebrows and mouth line were manipulated. For positive expressions, the middle points of the eyebrows were placed above the ends of the eyebrows, and the middle point of the mouth line was placed below the ends of the mouth. Conversely, for negative expressions, the middle points of the eyebrows were placed below the ends of the eyebrows, and that of the mouth line was placed above the ends. The heights of the middle points were systematically manipulated and connected with the end points (of eyebrows or mouth line) by using a spline curve. For CG faces, a face generated by the FaceGen Modeler 3.3 (Singular Inversions, Toronto, ON, Canada) with average race and average gender was used as default. Then, the face was morphed by changing to "SmileClosed" to generate positive expressions or changing to "Disgust" to generate negative expressions. "SmileClosed" and "Disgust" are parameters defined in the FaceGen Modeler.

To determine the optimal range of emotion expression to be used in the experiment, we conducted a preliminary experiment with faces with 11 levels of emotion expression. Thirteen hearing participants inferred the meaning behind the text messages displayed along with the face in an analogous way to the main experiment. For CG faces, negative emotion expressions were created with changing levels of "Disgust" in the FaceGen Modeler. Positive emotion expressions were created with changing levels of "SmileClosed." Thus, the set of CG faces consisted of 11 faces including 5 negative and 5 positive expressions and one neutral expression. The results of the preliminary experiment indicated that participants' evaluation drastically changed even with the mild expressions and that for the stronger expressions the evaluations tended to saturate. Thus, based on these results, we chose the range of emotion expressions being used in our main experiment. The range of expressions selected for CG faces consisted of emotion expression magnitudes of 0.13, 0.27, and 0.40 for both "Disgust" and "SmileClosed" within the settings of FaceGen Modeler in addition to the original neutral face. For the schematic faces, levels 5 and 8 used in the preliminary experiment (with 1 the most negative and 11 the most positive emotion expression of the faces that were used) were selected as the minimum and maximum expressions, respectively, and seven levels of emotional expressions were prepared to be distributed evenly within the range.

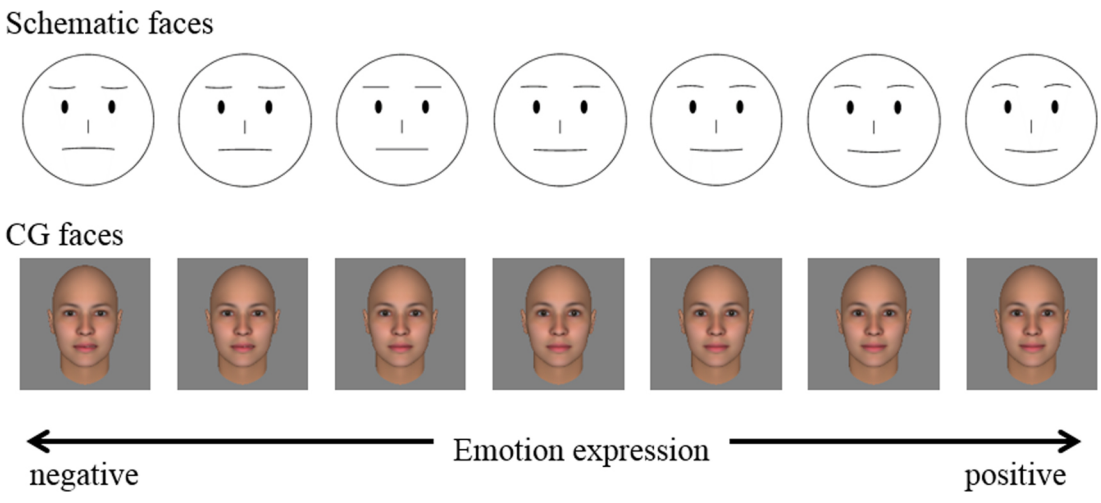


FIGURE 1 | Facial stimuli used in the experiment.

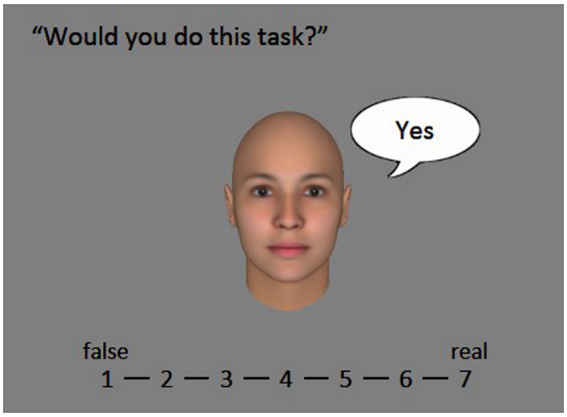


FIGURE 2 | Sample screen for evaluating earnestness. The texts were presented in Japanese in the experiment.

Procedure

The experiment consisted of two blocks. In the first experimental block, for each trial, a schematic or a CG face was presented at the center of the monitor, with a question directed to the face presented at the top and a response to the question in a cartoon balloon. Participants were asked to rate whether the response shown in a cartoon balloon represented the person's genuine feeling based on a 7-point Likert scale from 1 = false to 7 = real (perceived earnestness; see **Figure 2**). There were two sets of questions and responses (positive and negative) used in the experiment. All the questions and responses were presented in Japanese. In one set, the question was "Do you like her? (*kanojo no koto suki?* in Japanese)," and a positive response was "Yes (*suki*)" while a negative response was "No (*kirai*)" (negative). In the other set, the question was "Would you do this task? (*kono shigoto yatte kureru?*)," and a positive response was "Yes (*iiyo*)" while a negative response was "No (*iyada*)." The types of face stimuli (schematic or CG) and questions were fixed in sub-blocks in which seven levels of expression of emotion from negative to positive and two types of response (negative/positive) were presented in a randomized order. The order of sub-blocks was counterbalanced among participants. This was followed by the second experimental block, where the same faces as in the first block were presented one by one on the monitor, and participants were asked to judge how positive the emotion expressed on each face was, using a 7-point Likert scale ranging from 1 = negative to 7 = positive. This experimental block consisted of two sub-blocks, in which schematic and CG faces were presented separately. Each individual face was presented twice within a sub-block in randomized order. Experiments were written in Matlab using the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997; Kleiner et al., 2007). The instructions were given in written texts for both groups of participants. Although the specific situations of the contexts were not described in the instruction, most participants reported that they spontaneously took the situation as representing the messages and faces created by a third party in a face-to-face scenario. The procedure was approved by the internal review board of the University of Tokyo.

Results

Ratings of Emotion Expressed on the Faces

To check if the emotional expression manipulation was successful, effects of pre-assigned emotion expression level (1: the most negative to 7: the most positive), type of face (schematic/CG), and participants' hearing status (deaf/hearing) were examined by a three-way repeated measure analysis of variance (ANOVA), conducted on the ratings for positivity/negativity of emotions expressed in the faces.

Ratings of perceived positivity/negativity of emotion expressed in the schematic and CG faces increased as the pre-assigned level of expressed emotion increased (**Figure 3**). This indicated the manipulation of the expression of emotion was successful both in schematic faces and in CG faces. Results of the ANOVA demonstrated that the main effect of pre-assigned level of expressed emotion on the ratings was significant [$F(3.31, 162.2) = 607.6$, $p < 0.001$, $\eta_p^2 = 0.93$, using Greenhouse–Geisser corrected degrees of freedom], such that faces manipulated to appear more positive were perceived as more positive. A significant interaction effect was found between stimulus type and emotional expression level [$F(4.69, 229.9) = 4.08$, $p < 0.01$, $\eta_p^2 = 0.078$] while the main effect of stimulus type was not significant [$F(1, 49) = 0.63$, $p = 0.43$, $\eta_p^2 = 0.013$]. This indicates that the perceived positivity of the expressed emotions might differ between CG and schematic faces. Neither main effect nor interactions associated with participants' hearing status were significant: $F(1, 49) = 1.46$, $p = 0.23$, $\eta_p^2 = 0.029$ for the main effect of hearing status, $F(1, 49) = 1.78$, $p = 0.19$, $\eta_p^2 = 0.035$ for the interaction between stimulus type and hearing status, $F(3.31, 162.2) = 1.41$, $p = 0.24$, $\eta_p^2 = 0.028$ for the interaction between pre-assigned emotion expression level and hearing status, and $F(4.69, 229.9) = 1.03$, $p = 0.40$, $\eta_p^2 = 0.021$ for the three-way interaction, using Greenhouse–Geisser corrected degrees of freedom in calculating the latter two F-values.

To further examine if there were differences in emotion expression recognition between deaf and hearing participants, we performed a regression analyses within each participant on the ratings of perceived emotion with the pre-assigned emotion expression level a descriptive factor, separately for schematic and CG faces. Then, the coefficients of the pre-assigned level were compared across stimulus type and participant groups with a repeated-measure Bayesian ANOVA with using JASP 0.5 (Love et al., 2014). The results indicated neither significant effects of stimulus type ($BF_{10} = 0.25$; substantial evidence for H_0), participant group ($BF_{10} = 0.12$; substantial evidence for H_0), nor a significant interaction between these two factors ($BF_{10} = 0.11$; substantial evidence for H_0). The results supported that deaf and hearing participants did not differ in interpreting facial emotional expression of the faces used in the experiment.

Inferring Real Meaning from Text Messages Accompanied with Facial Expression

For perceived earnestness, the ratings for the trials where the facial character responded negatively to the questions (i.e., response was "No") were inverted before being used in the analyses. Thus, in the ratings after this manipulation, one indicates that participants estimated the response's real meaning as negative, while seven

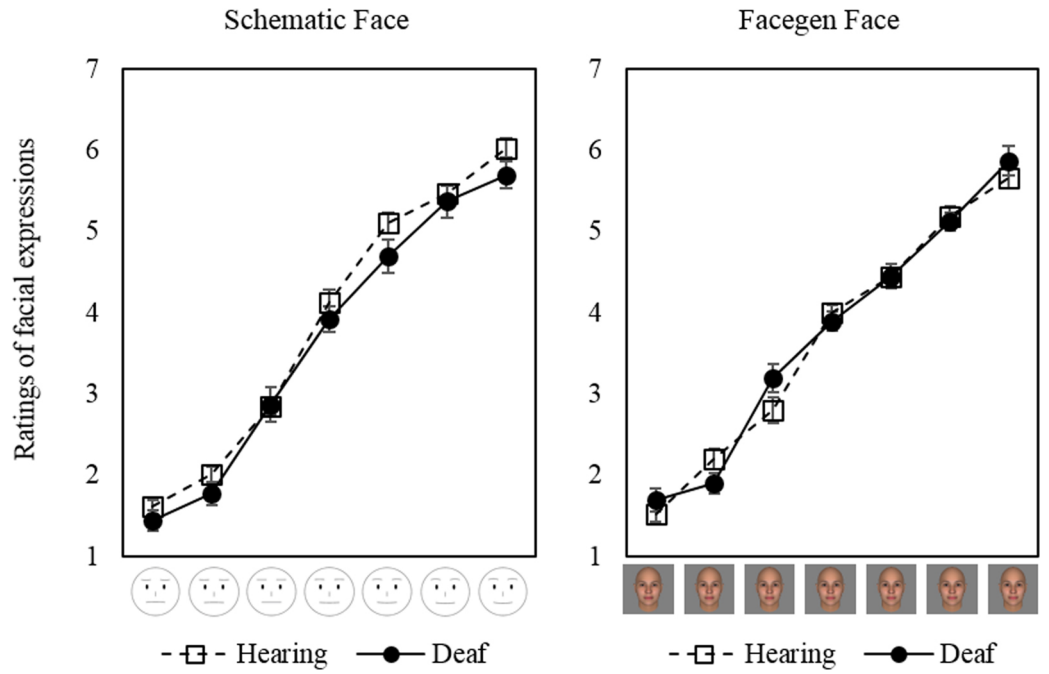


FIGURE 3 | Ratings of facial expressions. Error bars represent standard errors of means.

indicates that participants estimated the response's real meaning as positive, regardless of congruency between the response shown in the balloon and the estimated real meaning. Then, the influence of the facial expression on the participants' interpretation of the text response shown in the balloon (positive or negative) and the influence of the participants' hearing status were examined by a mixed-design ANOVA with stimulus type (schematic or CG), question asked, text response (positive or negative), and pre-assigned level of expressed emotion as within-participant factors and participants' group (deaf or hearing) as between-participants factor.

Generally, as shown in **Figure 4**, the texts with positive facial expression were interpreted as having more positive real meaning, regardless of stimulus type, question, or response presented in the balloon. The ANOVA results that demonstrated the significant main effect of the pre-assigned level of emotion expressed on the face [$F(2.12, 103.8) = 189.3$, $p < 0.001$, $\eta_p^2 = 0.79$, using Greenhouse–Geisser corrected degrees of freedom] supported this finding. Regarding the effect of response type, the main effect of response type and the interaction between response type and expressed emotion level were both significant [$F(1, 49) = 5.10$, $p < 0.05$, $\eta_p^2 = 0.094$, for the main effect; $F(2.30, 112.7) = 18.4$, $p < 0.001$, $\eta_p^2 = 0.273$ for the interaction, using Greenhouse–Geisser corrected degrees of freedom], suggesting that the effect of facial expression differed depending on whether the response was positive or negative. When the response was negative (dashed lines in **Figure 4**), the ratings tended to be low. This indicates that if the response was negative, the real meaning was judged as negative irrespective of the facial expression. Further, this interaction significantly differed between the participant groups [$F(2.30, 112.7) = 6.15$, $p < 0.005$, $\eta_p^2 = 0.112$, using Greenhouse–Geisser corrected degrees of freedom]. Thus, this indicates a difference between deaf people and hearing people in how facial expression was integrated into the evaluation of perceived earnestness (and negativity of the text messages). The ANOVA results also demonstrated a significant interaction between face type and expressed emotion level [$F(6, 294) = 7.55$, $p < 0.001$, $\eta_p^2 = 0.13$], but this might be an artifact from the different interpretation of emotion depending on face type found in the positivity/negativity ratings of the expressed emotion. The interaction between face type, context, and participants was also significant [$F(1, 49) = 5.06$, $p < 0.05$, $\eta_p^2 = 0.094$]. All remaining main effects and interactions, including the main effect of participants' hearing status [$F(1, 49) = 0.067$, $p = 0.80$, $\eta_p^2 = 0.001$], were not significant or only marginally significant.

We also conducted separate ANOVAs for each group to interpret the significant interactions. The main effect of facial expression was significant both in hearing [$F(1.97, 59.1) = 94.2$, $p < 0.001$, $\eta_p^2 = 0.76$] and deaf participants [$F(1.97, 37.5) = 102.2$, $p < 0.001$, $\eta_p^2 = 0.84$; both using Greenhouse–Geisser corrected degrees of freedom]. Thus, it was confirmed that the emotion expression had a significant influence on how the text response was interpreted.

The main effect of the response type was significant only in hearing participants [$F(1, 30) = 6.61$, $p < 0.05$, $\eta_p^2 = 0.18$ for hearing; $F(1, 19) = 0.63$, $p = 0.44$, $\eta_p^2 = 0.032$ for deaf]. This might reflect that a significant interaction between response type and expression level was found in hearing participants [$F(2.07, 62.0) = 23.3$, $p < 0.001$, $\eta_p^2 = 0.44$], while the interaction was only marginally significant in deaf participants [$F(2.67, 50.8) = 2.41$, $p = 0.084$, $\eta_p^2 = 0.11$]. These results may indicate that the rating was differently influenced by emotional expressions depending on the content of verbal response in hearing participants and resulted in the significant main effect of the response type.

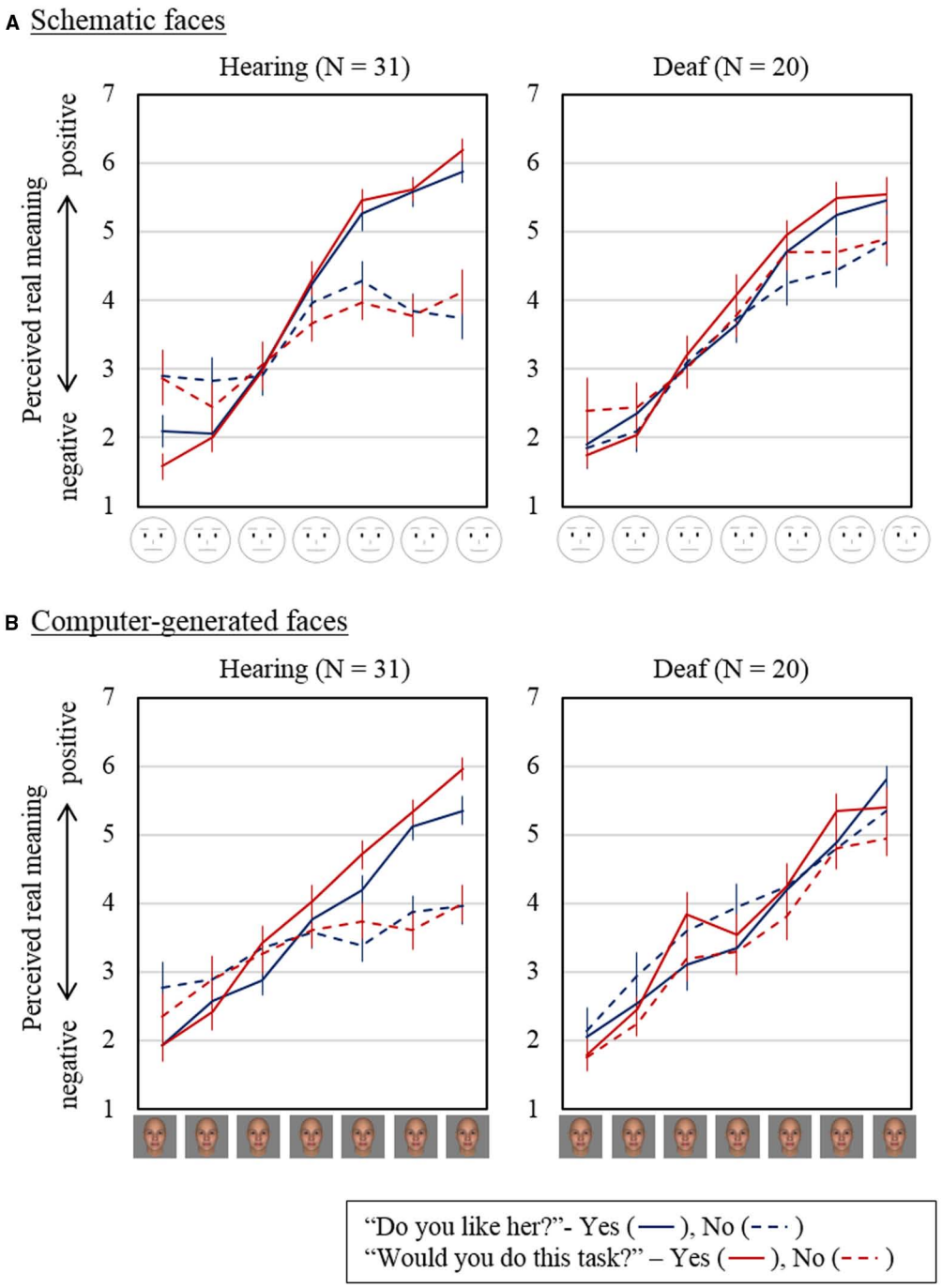


FIGURE 4 | Perceived real meaning inferred from the combination of text and facial expression for schematic faces (A) and for computer-generated faces (B). Blue lines are for the question "Do you like her?" with a positive response "Yes" (blue solid lines) and a negative response "No" (blue dashed lines). Red lines are for the question "Would you do this task?" with a positive response "Yes" (red solid lines) and a negative response "No" (red dashed lines). Error bars represent standard errors of the means.

Significant interactions between face type and expression level were found in both participant groups [$F(6, 180) = 5.19, p < 0.001$, $\eta_p^2 = 0.15$ for hearing; $F(6, 114) = 3.26, p < 0.01$, $\eta_p^2 = 0.15$ for deaf]. As already discussed, the influence of pre-assigned expression level on perceived positivity/negativity differed between face types. Thus, the interactions between face type and expression level in the rating might reflect the significant interaction in evaluation of facial expression itself rather than the difference in the process of integrating the facial emotion expression to interpret the real meaning.

Only in deaf participants, significant interactions between face type, context, and response type [$F(1, 19) = 5.03$, $p < 0.05$, $\eta_p^2 = 0.21$] and between context, response type, and emotion expression of the face [$F(6, 114) = 2.44$, $p < 0.05$, $\eta_p^2 = 0.11$] were found. *Post hoc* comparisons of these interactions indicated that the ratings of schematic faces in the situation where a positive response was given to the question "Do you like her?" were significantly higher than those of CG faces [difference of mean = (−0.39, 95% CI (−0.70, (−0.075), $p < 0.05$], while the ratings for different face types were not significantly different for the question "Would you do this task?" [difference of mean = 0.27, 95% CI (−0.20, 0.73), $p = 0.231$, both with Bonferroni correction]. The ratings with the two most positive emotion expressions were significantly higher when interpreting positive verbal responses than when interpreting negative responses but only in the trials with CG faces [difference of mean = 0.68, 95% CI (0.14, 1.21), $p < 0.05$ for the second-most positive expression; difference of mean = 0.55, 95% CI (0.019, 1.08), $p < 0.05$ for the most positive expression]. These differences were not found when interpreting the verbal responses presented with schematic faces.

An Ordinal Logistic Regression Model for Predicting Perceived Real Meaning of the Verbal Responses

To investigate possible factors that affected the ratings of positivity of the real meaning, an ordinal logistic regression analysis was performed with all the possible factors (participant group, context, type of face, and type of text response), covariate (emotion expression level of face), as well as all possible interactions between them. Then, we restructured the model by using the factors that had significant impacts on our first model. Extracted factors were participant group (i.e., hearing ability), emotion expression level, interaction between participant group and emotion expression level, interaction between participant group and type of text response (i.e., "yes" or "no"), interaction between hearing ability, type of text response, and emotion expression level. The results confirmed what we found in the ANOVAs.

Overall, the ratings were more positive in hearing participants than in deaf participants [odds ratio = 2.54, 95% CI (1.66, 3.89), Wald $\chi^2(1) = 18.3$, $p < 0.001$]. The higher ratings of positivity were associated with more positive emotion expression with an odds ratio of 1.94 [95% CI (1.80, 2.10), Wald $\chi^2(1) = 280.9$, $p < 0.001$]. The effect was smaller in hearing participants than in deaf participants with an odds ratio of 0.73 [95% CI (0.67, 0.81)], Wald $\chi^2(1) = 40.4$, $p < 0.001$.

Hearing participants perceived the positive text response (i.e., "yes") as more positive than they perceived negative (i.e., "no") as negative with an odds ratio of 0.21 [95% CI (0.14, 0.31), Wald $\chi^2(1) = 63.1$, $p < 0.001$]. In contrast, deaf participants did not show such an asymmetry [odds ratio = 0.75, 95% CI (0.47, 1.20), Wald $\chi^2(1) = 1.45$, $p = 0.23$].

Furthermore, significant interactions between response type and emotion expression level were found both in hearing and deaf participants. In both groups, the increase of ratings with increasing emotion expression level was steeper for positive than for negative text response [odds ratio = 1.75, 95% CI (1.60, 1.90), Wald $\chi^2(1) = 158.1$, $p < 0.001$ for hearing participants; odds ratio = 1.15, 95% CI (1.03, 1.27), Wald $\chi^2(1) = 6.41$, $p < 0.05$ for deaf participants].

Discussion

The present results showed that there was no significant effect related to participant hearing status in the judgment of facial expression, suggesting that the way hearing and deaf participants interpreted expression of emotion on faces did not differ. Past research has also suggested no difference between deaf and hearing participants in interpreting the emotional valence of facial expression using human facial pictures depicting various emotions (Watanabe et al., 2011). Our findings are consistent with these results and extended the understanding to non-realistic human faces (i.e., schematic faces and CG faces).

The findings from the present study also indicate that in terms of inferring real meaning of the verbal response, the emotion expressed on the face might qualify the meaning of what is explicitly stated as a verbal response to the question. For example, when the verbal response was "yes," the real meaning was rated at approximately two and thus interpreted as "no" for the faces expressing the highest levels of negative emotion. Facial expressions serve as strong non-verbal cues in recognizing others' intention (Ekman and Friesen, 1974; Friedman, 1979). The significant interactions between response type and facial emotion expression in the ANOVA and the ordinal regression model indicate an asymmetry in the contribution of facial expression depending on the response. In other words, we rely on facial expression in interpreting the text messages more when interpreting a positive than a negative response. This may indicate that we spontaneously assume that others may hide their real feeling in order to behave kindly or politely to us (politeness assumption), and thus in such a situation, we may tend to integrate non-verbal cues other than their direct response presented verbally. When others respond negatively, we tend to interpret their responses as their real meaning and thus make less use of non-verbal cues such as facial expression, as there is no reason for others to pretend to be unsociable.

As for the commonality and difference between deaf and hearing participants, the current results showed that (1) there was no difference in interpreting emotional valence from faces, (2) both groups were influenced by the facial expressions to infer the real meaning behind the text response, (3) the influence of facial expression was smaller when interpreting the text response that was expressing negative contents to the questioner in hearing participants, and (4) there was no such difference in deaf participants. For the influence of face type and conversational context, the ordinal logistic regression analysis showed that (5) no influence of facial type or conversational context was found in both participant groups, while (6) the interactions between facial type, context, and response or between context, response, and expression level were suggested for deaf participants only. *Post hoc* analyses following ANOVAs suggested that the influence of response type was observed only in CG faces in deaf participants.

In our results, the most pronounced difference in communication style between deaf and hearing people was the effect of positive emotion expression on interpreting the negative responses.

Hearing participants tended to interpret negative response as having negative meaning, irrespective of the positivity of emotion expression (i.e., "no means no"). However, deaf participants tended to be influenced more by positive facial expression when interpreting the negative responses.

One possible reason could be that deaf people consider nonverbal visual cues (including facial expression) as more useful sources for interpreting verbal messages because they usually have less access to auditory cues. Hearing people integrate face and voice information in understanding others in everyday situations (Campanella and Belin, 2007), while the degree of cross-modal influence between facial expression and voice depends on culture (Tanaka et al., 2010). In the current study, our experimental condition provided verbal information as written texts presented on the monitor and thus did not provide prosodic sound information that could be used to infer emotion. However, this did not prevent participants from imagining the prosody of each verbal stimulus. Hearing people may weigh visual information differently than deaf people because they usually have access to auditory cues (e.g., prosodic sounds). More specifically, visual facial expressions might be more useful sources for deaf people for understanding emotions. This in turn might explain the smaller asymmetry (i.e., the relatively larger effect of positive facial expression on the negative messages). However, there are other possibilities that might explain the current findings, such as difference in conceptualization of politeness and exploratory strategy (e.g., eye-movement). Further research will be required to clarify this issue.

Our results suggested that there was no significant difference between face types. This implies that even simple schematic faces can be as strong non-verbal cues for modifying interpretations of text messages as realistic CG faces, which is consistent with research on emoticons and avatars (Walther and D'Addario, 2001; Derks et al., 2008). However, our results also showed that the influence of facial expression on interpretation of the text message differed depending on hearing experience or ability of participants and that this difference was found in particular when the text response was expressing negative content to the questioner. These findings indicate the inhomogeneous effect of facial emotion information on text messages and its interaction with the communication strategy of the receiver. Therefore, caution should be exercised when emoticons or expressive avatars are used as nonverbal cues in human–computer interaction and human–human interaction via information systems. Although, in the current study, we focused on the difference between hearing and deaf people, our findings that the integration of emotion expression might rely on the presumption of politeness might be extended to possible differences between cultures. Perception or expectation of politeness and how it is conceptualized in the conversation might differ between cultures (Matsumoto, 1988; Haugh, 2004). In particular, as Matsumoto (1988) reported, the concept of "face" (in pragmatics) in Japanese culture may differ from that of other cultures, and this might represent a consideration for the present findings.

In conclusion, facial expressions influenced the interpretation of the response that was verbally presented as text. The influence of positive facial expressions on the perception of negative verbal response was smaller compared to that of negative facial expressions on the perception of positive verbal response. Although the perception of facial expression did not differ depending on hearing status, the influence of positive/negative emotion expressions on the perception of negative/positive verbal response was less asymmetrical in deaf participants compared to that in participants with normal hearing. This difference might be due to the difference in availability and usage of prosodic sound and facial expression (i.e., feature/holistic processing of faces in deaf/hearing participants) in inferring the real meanings from verbal messages. Although we focused on the effect of facial expression on interpretation of text messages in the current study, our results could also be interpreted in other ways, that is, text messages may affect the interpretation of the facial emotion expressions. These possibilities require further investigations.

Acknowledgments

This work was partly supported by JSPS KAKENHI 24300279, Cosmetology Research Foundation, Japan, and CREST, Japan Science and Technology Agency.

References

Arnold, P., and Murray, C. (1998). Memory for faces and objects by deaf and hearing signers and hearing nonsigners. *J. Psycholinguist. Res.* 27, 481–497. doi: 10.1023/A:1023277220438

Banse, R., and Scherer, K. (1996). Acoustic profiles in vocal emotion expression. *J. Pers. Soc. Psychol.* 70, 614–636. doi: 10.1037/0022-3514.70.3.614

Barnett, S. (2002). Communication with deaf and hard-of-hearing people: a guide for medical education. *Acad. Med.* 77, 694–700. doi: 10.1097/00001888-200207000-00009

Bettger, J., Emmorey, K., McCullough, S. H., and Bellugi, U. (1997). Enhanced facial discrimination: effects of experience with American Sign Language. *J. Deaf Stud. Deaf Educ.* 2, 223–233. doi: 10.1093/oxfordjournals.deafed.a014328

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision* 10, 433–436. doi: 10.1163/156856897X00357

Brown, P., and Levinson, S. C. (1987). *Politeness: Some Universals in Language Usage*. Cambridge: Cambridge University Press.

Bugental, D. E., Kaswan, J. W., and Love, L. R. (1970). Perception of contradictory meanings conveyed by verbal and nonverbal channels. *J. Pers. Soc. Psychol.* 16, 647–655. doi: 10.1037/h0030254

Campanella, S., and Belin, P. (2007). Integrating face and voice in person perception. *Trends Cogn. Sci.* 11, 535–543. doi: 10.1016/j.tics.2007.10.001

Derks, D., Fischer, A. H., and Bos, A. E. R. (2008). The role of emotion in computer-mediated communication: a review. *Comput. Hum. Behav.* 24, 766–785. doi: 10.1016/j.chb.2007.04.004

Duncan, S. J. (1969). Nonverbal communication. *Psychol. Bull.* 72, 118–137. doi: 10.1037/h0027795

Ekman, P., and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *J. Pers. Soc. Psychol.* 17, 124–129. doi: 10.1037/h0030377

Ekman, P., and Friesen, W. (1974). Detecting deception from the body or face. *J. Pers. Soc. Psychol.* 29, 288–298. doi: 10.1037/h0036006

Ekman, P., Friesen, W., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., et al. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *J. Pers. Soc. Psychol.* 53, 712–717. doi: 10.1037/0022-3514.53.4.712

Ekman, P., Sorenson, E., and Friesen, W. (1969). Pan-cultural elements in facial displays of emotion. *Science* 164, 86–88. doi: 10.1126/science.164.3875.86

Feldman, R., Jenkins, L., and Popoola, O. (1979). Detection of deception in adults and children via facial expressions. *Child Dev.* 50, 350–355. doi: 10.2307/1129409

Friedman, H. (1979). The interactive effects of facial expressions of emotion and verbal messages on perceptions of affective meaning. *J. Exp. Soc. Psychol.* 15, 453–469. doi: 10.1016/0022-1031(79)90008-8

Grice, H. (1969). Utterer's meaning and intention. *Philos. Rev.* 78, 147–177. doi: 10.2307/2184179

Haugh, M. (2004). Revisiting the conceptualisation of politeness in English and Japanese. *Multilingua* 23, 85–109. doi: 10.1515/mult.2004.009

Jack, R. E., Blais, C., Scheepers, C., Schyns, P. G., and Caldara, R. (2009). Cultural confusions show that facial expressions are not universal. *Curr. Biol.* 19, 1543–1548. doi: 10.1016/j.cub.2009.07.051

Jack, R. E., Garrod, O. G. B., Yu, H., Caldara, R., and Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proc. Natl. Acad. Sci. U.S.A.* 109, 7241–7244. doi: 10.1073/pnas.1200155109

Kleiner, M., Brainard, D., and Pelli, D. (2007). What's new in psychtoolbox-3? *Perception* 36, ECVP Abstract Supplement. doi: 10.1068/v070821

Love, J., Selker, R., Verhagen, J., Smira, M., Wild, A., Marsman, M., et al. (2014). JASP (Version 0.5) [Computer software].

Masuda, T., Ellsworth, P., Mesquita, B., Leu, J., Tanida, S., and Van de Veerdonk, E. (2008). Placing the face in context: cultural differences in the perception of facial emotion. *J. Pers. Soc. Psychol.* 94, 365–381. doi: 10.1037/0022-3514.94.3.365

Matsumoto, Y. (1988). Reexamination of the universality of face: politeness phenomena in Japanese. *J. Pragmatics* 12, 403–426. doi: 10.1016/0378-2166(88)90003-3

McCullough, S., and Emmorey, K. (1997). Face processing by deaf ASL signers: evidence for expertise in distinguishing local features. *J. Deaf Stud. Deaf Educ.* 2, 212–222. doi: 10.1093/oxfordjournals.deafed.a014327

McCullough, S., Emmorey, K., and Sereno, M. (2005). Neural organization for recognition of grammatical and emotional facial expressions in deaf ASL signers and hearing nonsigners. *Cogn. Brain Res.* 22, 193–203. doi: 10.1016/j.cogbrainres.2004.08.012

Nisbett, R. E., and Masuda, T. (2003). Culture and point of view. *Proc. Natl. Acad. Sci. U.S.A.* 100, 11163–11170. doi: 10.1073/pnas.1934527100

Öhman, A. (2002). Automaticity and the amygdala: nonconscious responses to emotional faces. *Curr. Dir. Psychol. Sci.* 11, 62–66. doi: 10.1111/1467-8721.00169

Pelli, D. G. (1997). The videotoolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision* 10, 437–442. doi: 10.1163/156856897X00366

Rosset, D. B., Rondan, C., Da Fonseca, D., Santos, A., Assouline, B., and Deruelle, C. (2008). Typical emotion processing for cartoon but not for real faces in children with autistic spectrum disorders. *J. Autism Dev. Disord.* 38, 919–925. doi: 10.1007/s10803-007-0465-2

Rosset, D. B., Santos, A., Da Fonseca, D., Poinso, F., O'Connor, K., and Deruelle, C. (2010). Do children perceive features of real and cartoon faces in the same way? Evidence from typical development and autism. *J. Clin. Exp. Neuropsyc.* 32, 212–218. doi: 10.1080/13803390902971123

Scherer, K. R., Banse, R., Wallbott, H. G., and Goldbeck, T. (1991). Vocal cues in emotion encoding and decoding. *Motiv. Emot.* 15, 123–148. doi: 10.1007/BF00995674

Schwarzer, G., Huber, S., and Dümmler, T. (2005). Gaze behavior in analytical and holistic face processing. *Mem. Cogn.* 33, 344–354. doi: 10.3758/BF03195322

Sperber, D., and Wilson, D. (2002). Pragmatics, modularity and mind-reading. *Mind Lang.* 17, 3–23. doi: 10.1111/1468-0017.00186

Tanaka, A., Koizumi, A., Imai, H., Hiramatsu, S., Hiramoto, E., and de Gelder, B. (2010). I feel your voice. Cultural differences in the multisensory perception of emotion. *Psychol. Sci.* 21, 1259–1262. doi: 10.1177/0956797610380698

Wallraven, C., Bülthoff, H. H., Cunningham, D. W., Fischer, J., and Bartz, D. (2007). Evaluation of real-world and computer-generated stylized facial expressions. *ACM Trans. Appl. Percept.* 4. doi: 10.1145/1278387.1278390

Walther, J., and D'Addario, K. (2001). The impacts of emoticons on message interpretation in computer-mediated communication. *Soc. Sci. Comput. Rev.* 19, 324–347. doi: 10.1177/089443930101900307

Watanabe, K., Matsuda, T., Nishioka, T., and Namatame, M. (2011). Eye gaze during observation of static faces in deaf people. *PLoS ONE* 6:e16919. doi: 10.1371/journal.pone.0016919

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 10 June 2015
doi: 10.3389/fpsyg.2015.00796

Bilingual children weigh speaker's referential cues and word-learning heuristics differently in different language contexts when interpreting a speaker's intent

*Wan-Yu Hung, Ferninda Patrycia and W. Q. Yow**

Humanities, Arts and Social Sciences, Singapore University of Technology and Design, Singapore, Singapore

OPEN ACCESS

Edited by:
Sebastian Loth,
Universität Bielefeld, Germany

Reviewed by:
David Sobel,
Brown University, USA
Diane Poulin-Dubois,
Concordia University, Canada

****Correspondence:***
W. Q. Yow,
Humanities, Arts and Social Sciences,
Singapore University of Technology and Design, 8 Somapah Road,
Singapore 487372, Singapore
quin@sutd.edu.sg

Specialty section:
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology

Received: *18 February 2015*
Accepted: *27 May 2015*
Published: *10 June 2015*

Citation:
Hung W-Y, Patrycia F and Yow WQ (2015) Bilingual children weigh speaker's referential cues and word-learning heuristics differently in different language contexts when interpreting a speaker's intent. Front. Psychol. 6:796. doi: 10.3389/fpsyg.2015.00796

Past research has investigated how children use different sources of information such as social cues and word-learning heuristics to infer referential intents. The present research explored how children weigh and use some of these cues to make referential inferences. Specifically, we examined how switching between languages known (familiar) or unknown (unfamiliar) to a child would influence his or her choice of cue to interpret a novel label in a challenging disambiguation task, where a pointing cue was pitted against the mutual exclusivity (ME) principle. Forty-eight 3-and 4-years-old English–Mandarin bilingual children listened to a story told either in English only (No-Switch), English and Mandarin (Familiar-Switch), English and Japanese (Unfamiliar-Switch), or English and English-sounding nonsense sentences (Nonsense-Switch). They were then asked to select an object (from a pair of familiar and novel objects) after hearing a novel label paired with the speaker's point at the familiar object, e.g., "Can you give me the *blicket*?" Results showed that children in the Familiar-Switch condition were more willing to relax ME to follow the speaker's point to pick the familiar object than those in the Unfamiliar-Switch condition, who were more likely to pick the novel object. No significant differences were found between the other conditions. Further analyses revealed that children in the Unfamiliar-Switch condition looked at the speaker longer than children in the other conditions when the switch happened. Our findings suggest that children weigh speakers' referential cues and word-learning heuristics differently in different language contexts while taking into account their communicative history with the speaker. There are important implications for general education and other learning efforts, such as designing learning games so that the history of credibility with the user is maintained and how learning may be best scaffolded in a helpful and trusting environment.

Keywords: mutual exclusivity, bilingualism, code-switching, word-learning, communicative signals, pragmatic cues

Introduction

Existing research in developmental psychology has shown that children use various strategies such as word-learning heuristics to help narrow down potential objects when trying to identify referents (Clark, 1988; Markman and Wachtel, 1988; Markman, 1994; Landau et al., 1998). An example of a word-learning heuristic is the *mutual exclusivity (ME) principle* that assumes a one-to-one correspondence between a label and an object, such that a novel label refers to a novel object rather than a familiar object (ME hereafter; Markman and Wachtel, 1988). Children by the age of two are also able to use a speaker's cues, such as point and gaze, as clues to understanding the speaker's referential intents (e.g., Lempers, 1979; Leung and Rheingold, 1981; Baron-Cohen, 1989; Franco and Butterworth, 1996; Povinelli et al., 1997). More recently, research has shown that a change in language contexts, such as code-switching (the alternate use of two languages in a single discourse), can heighten bilingual children's use of a speaker's point and gaze when determining a target referent object (Yow and Hung, 2013).

Research suggests that language environment such as growing up bilingually may affect how children weigh the importance of ME to understand referential intents, and that ME is typically more relaxed among bilingual children than monolingual children (Davidson et al., 1997; Byers-Heinlein and Werker, 2009; Houston-Price et al., 2010). By comparing infants from different language backgrounds, Byers-Heinlein and Werker (2009) found that infants from a monolingual background showed the strongest use of ME (e.g., looking significantly longer at an unknown object than at a familiar object when a novel label was used), followed by infants from a bilingual background and, finally, infants from a trilingual background, who showed the weakest use of ME (see Davidson et al., 1997, for a similar pattern in bilingual and monolingual preschool children). Other studies also found that when ME was pitted against other referential cues, such as pointing, bilingual children were more likely to violate ME in favor of pointing, as opposed to monolingual children who showed a more robust use of ME (Jaswal and Hansen, 2006; Yow and Markman, 2007; Healey and Skarabela, 2008; but cf. Grassmann and Tomasello, 2010).

There have been theoretical attempts to explain why bilingual children are more willing to suspend ME in the presence of other referential cues. One account suggests that living in a bilingual environment involves frequent encounter of situations where the same object can be named differently in different languages (e.g., *house* in English vs. *casa* in Spanish). It is believed that such experiences violate the assumption of the one-to-one word-referent mapping in ME and hence bilingual children are likely to perceive ME as a less helpful word-learning heuristic compared with referential cues such as pointing (Au and Glusman, 1990; Davidson et al., 1997; Healey and Skarabela, 2008). Another account suggests that bilingual children are flexible in perspective-taking and are adept at taking another person's referential cues to learn about a novel situation or determine a target referent when there is a conflict with their own assumptions such as ME (Genesee et al., 1975; Rosenblum and Pinker, 1983; Healey and Skarabela, 2008). Thus, this bilingual advantage of perspective-taking is largely due to the demands of living in a bilingual environment, which requires assimilation and accommodation of different linguistic perspectives that are unique to individual languages (e.g., every French noun has a grammatical gender but English nouns have no grammatical gender associated with them). For instance, Bassetti (2007) found that while Italian monolingual children tended to attribute female voices for objects that are feminine in Italian, Italian–German bilingual children tended to hold more balanced views toward object gender, especially for objects with conflicting grammatical gender in different languages (e.g., *clock* is masculine in Italian but feminine in German).

We propose that bilingual children's inclination to suspend ME could also be related to their frequent encounters with complex conversations in either language and code-switching (the alternate use of two or more languages in the context of a single conversation). Bilingual children have to often figure out what language a speaker is using and how to interact appropriately to avoid a potential communication breakdown. They may pay greater attention to a speaker's referential cues (e.g., point and gaze) to determine the speaker's communicative intent (Yow and Markman, 2011; Brojde et al., 2012). Bilingual children may thus rely more on a speaker's referential cues than general word-learning heuristics (e.g., ME) to determine the speaker's referential intent in a challenging communicative context. Yow and Hung (2013) found that bilingual preschoolers who heard a speaker code-switched in a mixture of known and unknown languages were better able to utilize the speaker's point and gaze than those who did not, possibly to accommodate the extra communicative demands. Our study seeks to examine how an exposure to a code-switching scenario would influence bilingual children's use of referential cues and ME in a challenging context where these two cues are pitted against each other.

Research has shown that exposure to code-switching may influence word learning and language processing in a number of aspects, such as receptive vocabulary (e.g., Byers-Heinlein, 2013), speed of lexical access (e.g., Macnamara, 1967; Grosjean, 1988), reading comprehension (e.g., Beauvillain and Grainger, 1987; Thomas and Allport, 2000), and naming and reading aloud (e.g., Kolers, 1966; Meuller and Allport, 1999). Nonetheless, the existing studies on this topic are predominantly about code-switching in languages that are spoken in one's family (i.e., *familiar* code-switching). Little is known about how such language processes may be influenced by exposure to languages that are unknown to the listener (i.e., *unfamiliar* code-switching). This distinction between *familiar* and *unfamiliar* code-switching from a listener's perspective could be important because there may be significant differences in the efforts required for comprehending these two types of code-switching. In the present study, we distinguished code-switching as either familiar or unfamiliar from the listener's point of view (i.e., switch between languages known to the listener, or switch between a language known to the listener and another language unknown to the

listener, respectively). As unfamiliar code-switching involves a language unknown to the listener, it is likely to incur some form of communication breakdown. The extent of efforts required for comprehending unfamiliar code-switching may be much greater than comprehending familiar code-switching. Therefore, unfamiliar code-switching may trigger children to pay attention to other referential cues (e.g., point) than solely depend on language-related heuristic, such as ME, in a conflicting situation.

To date, it remains relatively unknown whether these two types of code-switching would influence children's word-learning in a context where both referential cues and ME are available. This study attempts to address this question by using a word disambiguation task to examine bilingual children's choice of cue (ME or point) under different code-switching conditions. English–Mandarin bilingual children first heard a storytelling episode either in English only (No-Switch), English and Mandarin (Familiar-Switch), or English and Japanese (Unfamiliar-Switch), followed by a disambiguation task, where pairs of familiar and novel objects were presented to them and an experimenter requested for an object using a novel label while pointing to the familiar object. Since communication breakdown may be incurred in the unfamiliar code-switching condition but not the other two conditions, children may weigh and use the ME and point differently when trying to figure out the target referent. We predicted that children who heard unfamiliar code-switching would make use of the speaker's point more than ME when interpreting the novel label compared to children who heard familiar or no code-switching.

Study 1

Participants

Thirty-six 3- and 4-years-old English–Mandarin bilingual children from three different childcare centers in Singapore participated in this study (17 females, 19 males; M_{age} = 3;11, range 3;0–4;10). Prior to the experiment, parents filled a language background questionnaire that asked about their general demographic information and their children's language use at home (see **Table 1**). Children were randomly assigned to one of three code-switching groups: No-Switch, Familiar-Switch, or Unfamiliar-Switch (see section on Storytelling), with the constraint that in the Unfamiliar-Switch group, we only included children who did not have any exposure to Japanese language to ensure that the children were indeed unfamiliar with the language.

TABLE 1 | Demographic information and language use: means (SD).

	No-Switch	Familiar-Switch	Unfamiliar-Switch
Age	3;11 (0;7)	4;1 (0;8)	3;10 (0;8)
SES (Father)[a]	3.67 (0.99)	3.50 (0.67)	3.42 (0.67)
SES (Mother)[a]	3.50 (1.17)	3.50 (0.67)	3.42 (0.67)
Exposure to English (%)[b]	58.67 (20.64)	65.00 (12.25)	62.92 (16.58)
Exposure to Mandarin (%)[b]	34.17 (20.76)	32.08 (11.57)	29.42 (11.62)
Parental code-switching[c]	2.71 (0.40)	2.34 (0.54)	2.43 (0.38)
Working memory[d]	4.75 (1.36)	5.83 (1.34)	5.08 (1.44)
Inhibitory control[e]	13.08 (3.73)	13.58 (2.84)	12.75 (2.53)

Total N = 36. Seven children had exposure to a third language for at least 10% of their time, which includes Tamil (n = 1), Hindi (n = 1), Malay (n = 1), Thai (n = 1), Japanese (n = 1), Cantonese (n = 1), and Hokkien (n = 1). These children were distributed across the three conditions.

[a]*SES = Socioeconomic status measured by education level, in a range from 0 (none or no formal education) to 5 (postgraduate degree).*

[b]*Average amount of exposure in a typical week in percentage.*

[c]*Two parents did not complete this questionnaire (one from the No-Switch condition and one from the Unfamiliar-Switch condition).*

[d]*Working memory was measured by the forward digit-span test.*

[e]*Inhibitory control was measured by the day–night Stroop task.*

Materials

Parents' Code-Switching Questionnaire

This questionnaire was used to obtain information about parents' code-switching behavior during their daily communication with the child. It contained eight items and was constructed based on the *Bilingual Switching Questionnaire* (Rodriguez-Fornells et al., 2012). The items asked parents how frequently they code-switch both in general and within a sentence, how frequently they code-switch for certain topics or issues, and how frequently they think they unintentionally code-switch during their conversation with their child. For each item, parents were asked to rate on a 5-point frequency scale (1 = *never* to 5 = *always*). A mean score of these items was calculated for each child.

Picture Books

We created two A5 size wordless color picture books (pictures were modified from de Bezenac, 2010a,b; http://www.freekidsbooks.org). Each picture book consisted of five pictures printed on five separate pages. The two picture books were matched on contents and the number of characters involved.

Objects and Labels

Six pairs of objects and six novel labels were used in the disambiguation task (see Appendix A). Each pair consisted of a familiar object and a novel object of similar size and of comparable visual attractiveness.

Forward Digit Span Task

This task was adapted from the Wechsler Intelligence Scale for Children-Revised (Wechsler, 1974) and was used to ensure that children in the three conditions were comparable in their working memory capacity. In this task, an experimenter read out a string of digits one at a time, and the child was asked to repeat them in the same order as the experimenter had recited them. The length of the digit strings started from two and increased by one digit after every two trials. The trials continued until two consecutive errors were made in trials of the same digit length. A list of 16 strings of digits was used and the longest string consists of eight digits. The total score reflected the number of strings the child repeated correctly, and ranged from 0 to 16.

Day–Night Stroop Task

We used the day–night Stroop task (adapted from Gerstadt et al., 1994) to ensure that children from the three conditions did not differ in inhibitory control capacity. We presented each child with a series of cards in a pre-determined random order (Siegal et al., 2009), each with either a picture of a moon or a sun on it. The child was instructed to say "day" on seeing a moon card and "night" for a sun card. There were two practice trials, followed by 14 test trials. The experimenter explained the rule again and restarted with the first two trials if the child failed either of the first two trials. Once the child successfully answered both practice trials, the experimenter continued to administer the remaining 14 trials. The total score ranged from 2 to 16.

Procedure

This study was approved by Institutional Review Board (IRB) of the Singapore University of Technology and Design (SUTD). Children whose parents had given informed consent for their participation were tested individually in a quiet room at their childcare center. Each of them received a session of storytelling, a disambiguation task, a forward digit span task, and a day–night Stroop task, in this order.

Storytelling

An experimenter first introduced the child to one of the two wordless picture books by saying, "Look at this picture book! I am going to tell you a story." She then proceeded with a story that consisted of five sentences, which corresponded to each of the five pictures in the picture book. For the No-Switch group, the experimenter told the story completely in English. For the Familiar-Switch group, the experimenter alternated the descriptions of the pictures in English and Mandarin (i.e., in this sequence: English–Mandarin–English–Mandarin–English). For the Unfamiliar-Switch group, the experimenter alternated the descriptions in English and Japanese in the same sequence as in the Familiar-Switch group. All the sentences were of comparable length (see Appendix B). The experimenter presented the picture book in front of the child on the table they shared, and helped turn the pages without pointing to any part of the pictures so as not to prime the child to attend to the experimenter's point in the subsequent disambiguation task. The experimenter did not provide any feedback to the child throughout the storytelling episode. The child was given sufficient time to glance through the picture on each page before the experimenter continued to the next page.

Disambiguation Task

After the storytelling session, the same experimenter conducted six trials of the disambiguation task adapted from Jaswal and Hansen's (2006) procedures. For each trial, the experimenter first presented the child with a pair of one familiar object and one novel object, and directed the child's attention to both objects equally without labeling them (e.g., "Look at these!"). The experimenter then placed the two objects on the table half way between herself and the child, slightly more than shoulder length apart, and asked the child to give her one of the objects by using a novel label, "Can you give me the *blicket*?" The task was made challenging by the experimenter pointing subtly but unambiguously to the familiar object while providing the novel label (see **Figure 1**). To draw the child's attention, the experimenter made a gentle tap on the table twice every time before making the request. The experimenter kept her gaze direction neutral by looking straight at the child until a response was made. We counter-balanced the pairings of novel labels and object pairs, and the presentation order of the novel labels. For half of the children, the task started with the familiar object on the left. For each child, the familiar objects appeared on the child's left side half of the times.

FIGURE 1 | An experimenter pointing to the familiar object of a pair of familiar and novel objects (familiar object: clock; novel object: mosquito coil).

Results

One-way between-subjects Kruskal–Wallis tests confirmed that children of the three code-switching groups were matched on age, amount of exposure to English and Mandarin, parental education level, reported amount of parental code-switching with child, working memory, and inhibitory control, all $ps > 0.10$. Non-parametric tests were used because the scores of the control variables were not normally distributed.

We hypothesized those bilingual children who heard unfamiliar code-switching would likely use the speaker's point more often than children who heard familiar or no code-switching to interpret the novel label. Thus, responses of the disambiguation task were coded as "1" if the child chose the familiar object according to the experimenter's point, or "0" if the child used ME to choose the novel object instead. The total score across the six trials ranged from 0 to 6.

A one-way between-subjects ANOVA showed that the three groups of children performed differently in the disambiguation task [$F(2,33) = 4.33$, $p = 0.021$, Cohen's $d = 1.73$]. Bonferroni *post hoc* comparisons revealed that there was a significant difference between the Familiar-Switch and Unfamiliar-Switch groups, but not between the No-Switch and Familiar-Switch groups, or between the No-Switch and Unfamiliar-Switch groups (see **Table 2**). This suggests that bilingual children's choice of cue (ME or point) was not influenced by the presence of code-switching *per se*, but rather by the *type* of code-switching used

TABLE 2 | Average frequency of accepting the pointing cues to pick familiar objects (out of six trials).

Condition	Mean	SD
No-Switch	2.92	1.73
Familiar-Switch	4.08	2.07
Unfamiliar-Switch	2.00	1.35

Total N = 36.

to communicate with them. Contrary to our hypothesis, the Unfamiliar-Switch group was less likely than the Familiar-Switch group to use the experimenter's point to interpret the novel label. Two-tailed one-sample *t*-tests showed that while the No-Switch group performed at chance level [$t(11) = -0.17$, $p = 0.87$, Cohen's $d = -0.048$], the Familiar-Switch group tended to use the experimenter's point over ME to disambiguate the novel label, [$t(11) = 1.82$, $p = 0.097$, Cohen's $d = 0.53$], and the Unfamiliar-Switch group significantly chose ME rather than the experimenter's point above chance when disambiguating the novel label, [$t(11) = -2.57$, $p = 0.026$, Cohen's $d = -0.74$]. A closer look at the distribution of children's responses revealed that across the six trials, 33.3% of the No-Switch group used ME and point equally (three trials each), 41.7% used mostly ME (in four or more trials), and 25% used mostly point (in four or more trials). On the other hand, 8.3% of the Familiar-Switch group used ME and point equally, 25% used mostly ME and 66.7% used mostly point. In contrast, 41.7% of the Unfamiliar-Switch group used ME and point equally, 50% used mostly ME, and only 8.3% used mostly point.

An additional analysis was conducted on the children's looking time toward the experimenter during the storytelling session. If extra efforts were required to comprehend the foreign sentences, we would expect the Unfamiliar-Switch group to look at the experimenter for interpretation more often than the other groups when the code-switched sentences were uttered. We calculated how long a child spent on looking at the experimenter when she code-switched. Two independent coders coded offline the proportion of time a child looked at the experimenter when the second and fourth sentences were uttered, i.e., where instances of code-switching took place for the Familiar-Switch and Unfamiliar-Switch groups (inter-rater reliability $r = 0.99$, $p < 0.001$). Looking time of one participant from the Unfamiliar-Switch group and one from the No-Switch group could not be coded due to technical problems during recording. A one-way between-subjects ANOVA showed that the three groups were significantly different in their looking time [$F(2,31) = 7.59$, $p = 0.002$, Cohen's $d = 1.65$], with the Unfamiliar-Switch group showing the longest look ($M = 43.12\%$, SD = 25.95%), followed by the Familiar-Switch group ($M = 19.14\%$, SD = 16.00%), and the No-Switch group ($M = 12.01\%$, SD = 15.64%). Bonferroni *post hoc* comparisons revealed that the looking time difference was significant between the Unfamiliar-Switch and Familiar-Switch groups and between the Unfamiliar-Switch and No-Switch groups, but not between the Familiar-Switch and No-Switch groups. This finding reveals that the Unfamiliar-Switch group indeed paid more attention to (i.e., looked longer at) the experimenter when they heard unfamiliar code-switching compared to the Familiar-Switch and No-Switch groups. This supports our speculation that children in the Unfamiliar-Switch group are looking for some clarification or assistance from the experimenter when they do not understand the foreign utterances. It is to be noted that the experimenter in our study remained focused on telling the story based on her scripts and did not respond to the children at all. If children had expected the experimenter to clarify or provide clues to her foreign utterances but were "ignored" (i.e., experimenter did not respond), then it is possible that the Unfamiliar-Switch group subsequently chose to use ME to determine the referent in the disambiguation task as they believed that the experimenter's point would not be helpful anyway.

In summary, this study showed that the type of code-switching differentially influenced children's choice of cue (ME or point) in a disambiguation task. Unexpectedly, bilingual children in the Unfamiliar-Switch condition showed a significant tendency to use ME over the point compared to bilingual children in the Familiar-Switch condition. Children's increased looking time to the experimenter during the unfamiliar code-switched sentences implied that they might have expected the experimenter to clarify her utterances when they did not understand her. Hence, when the experimenter failed to repair the breakdown in the communication during the storytelling session, children in the Unfamiliar-Switch condition might have subsequently chosen to rely on other strategies (i.e., ME) instead of her point to interpret the novel label in the disambiguation task. Nonetheless, it is also possible that this preference could be due to an abrupt phonological change involved in unfamiliar code-switching. The sudden change in phonological makeup of the utterances may have prompted children to default to word-learning heuristics to select a referent. To tease apart these two possibilities, Study 2 used a nonsense English storytelling condition to induce comparable semantic barriers as unfamiliar code-switching. If communication barriers dictated children's performance, we predicted that children who heard nonsense English would similarly choose to rely on ME over the speaker's point to interpret the novel label as those in the Unfamiliar-Switch group in Study 1. Alternatively, if the type of code-switching provides a unique communicative signal over and beyond semantic familiarity and comprehension, the two groups would differ in their choice of cues in the disambiguation task.

Study 2

Participants

Twelve other 3- and 4-years-old English–Mandarin bilingual children from the same childcare centers as Study 1 participated in this study (six females, six males; $M_{age} = 4;0$, range = 3;6–4;10; see **Table 3**).

Materials

The children were presented with the same materials as in Study 1, except that a similar but different picture book was used (pictures were modified from de Bezenac, 2010c; http://www.freekidsbooks.org). This picture book and the picture books

TABLE 3 | Demographic information and language use: means (SD).

	Nonsense-Switch (Study 2)	Unfamiliar-Switch (Study 1)
Age	4;0 (0;6)	3;10 (0;8)
SES (Father)	4.00 (0.43)	3.42 (0.67)
SES (Mother)	3.92 (0.67)	3.42 (0.67)
Exposure to English (%)	65.00 (14.62)	62.92 (16.58)
Exposure to Mandarin (%)	32.67(13.93)	29.42 (11.62)
Parental code-switching	2.31 (0.56)	2.43 (0.38)
Working memory	5.92 (0.79)	5.08 (1.44)
Inhibitory control	11.00 (4.71)	12.75 (2.53)

Total N = 24. Five children had exposure to a third language for at least 10% of their time, which includes Tamil (n = 1), Malay (n = 1), Teochew (n = 1), Hokkien (n = 1), and other Chinese dialect (n = 1). These children were distributed across the two conditions.

used in Study 1 were matched on contents and the number of characters involved.

Procedure

This study was approved by IRB of the SUTD. Study 2 followed the same procedure as Study 1 except that the story was told in alternate English and English-sounding nonsense sentences (Nonsense-Switch). The sentences were comparable to those sentences used in Study 1 in length (see Appendix B). The nonsense words were chosen from two nonsense poems, Jabberwocky (Carroll, 1872), and The Faulty Bagnose (Lennon, 1965).

Results

Mann–Whitney *U* tests using Bonferroni adjusted alpha levels of 0.01 per test (0.05/8) showed that children in the Nonsense-Switch group in Study 2 did not differ significantly from the Unfamiliar-Switch group in Study 1 on all the control variables. Non-parametric tests were used because the scores of the control variables were not normally distributed.

An independent-samples *t*-test between the Nonsense-Switch group in Study 2 and the Unfamiliar-Switch group in Study 1 confirmed that there was no significant difference in performance between the two groups [$t(22) = -1.22$, $p = 0.24$, Cohen's $d = -0.50$; see **Table 4**]. Two-tailed one-sample *t*-tests also found that the Nonsense-Switch group performed at chance level [$t(11) = -0.30$, $p = 0.77$, Cohen's $d = -0.086$]. Recall that in Study 1, the Unfamiliar-Switch group significantly used ME more than the experimenter's point [$t(11) = -2.57$, $p = 0.026$, Cohen's $d = -0.74$]. This suggests that although overall, the Nonsense-Switch group did not differ significantly from the Unfamiliar-Switch group, they, in fact, used ME and the experimenter's point equally often to interpret the novel label, compared to the Unfamiliar-Switch group who used ME over the experimenter's point. A more detailed examination of the children's responses revealed that 41.7% of the Nonsense-Switch group used ME and point equally (three trials each), 33.3% used mostly ME (in four or more trials out of six), and 25% used mostly point (in four or more trials of six). While 41.7% of the Unfamiliar-Switch group in Study 1 also used ME and point equally, 50% of them used mostly ME, and only 8.3% used mostly point.

TABLE 4 | Average frequency of accepting the pointing cues to pick familiar objects (out of six trials).

Condition	Mean	SD
Nonsense-Switch (Study 2)	2.83	1.95
Unfamiliar-Switch (Study 1)	2.00	1.35

We also coded the proportion of time each child looked at the experimenter at both instances of code-switching (during the second and fourth sentence of the story). An independent-samples *t*-test revealed that the difference between the two groups were marginally significant [$t(21) = 1.75$, $p = 0.094$, Cohen's $d = -0.73$; $M_{Unfamiliar\text{-}Switch} = 43.12\%$, $SD_{Unfamiliar\text{-}Switch} = 25.95\%$, $M_{Nonsense\text{-}Switch} = 24.81\%$, $SD_{Nonsense\text{-}Switch} = 24.19\%$). The Nonsense-Switch group tended to look less at the experimenter when they heard the nonsense sentences compared to the Unfamiliar-Switch group.

In summary, while children in the Nonsense-Switch group seemed to perform similarly as those in the Unfamiliar-Switch group in their choice of cue in a disambiguation task, their behavior was less consistent than the Unfamiliar-Switch group in relying on ME over the experimenter's point. They also looked less at the experimenter when hearing the nonsense sentences compared to the Unfamiliar-Switch group when hearing the unfamiliar sentences. This result suggests that the unfamiliar code-switching effect found in Study 1 cannot be attributed to semantic barriers *per se*, and there is something unique about the communicative intent of a speaker when switching between familiar and foreign utterances.

General Discussion

This research sought to answer whether exposure to a language switch, in particular, the specific *types* of switch, would influence bilingual children's choice of cue (ME or point) in understanding referential intents. Our study showed that, indeed, the *type* of code-switching influenced the children's choice of cue. The No-Switch and Nonsense-Switch groups were equally likely to use the experimenter's point and ME to interpret a novel label. This finding of the No-Switch group was consistent with Yow and Markman (2007) where they found a proportionate use of the speaker's point and ME among bilingual children in an analogous disambiguation task without prior episodes of code-switching. While the Familiar-Switch group showed a tendency to use the speaker's point instead of ME, the Unfamiliar-Switch group significantly used ME instead of the speaker's point. Although this seems to contradict our prediction that children who heard unfamiliar code-switching would pay more attention to a speaker's referential cues to overcome communicative challenges and thus rely on the speaker's point over ME to interpret a novel label, our analysis of the children's looking time revealed otherwise. The Unfamiliar-Switch group did look at the experimenter significantly longer when they heard the unfamiliar code-switched sentences than those who heard only English sentences, familiar English–Mandarin sentences, or English and nonsense English sentences. This suggests that unfamiliar code-switching provides a distinctive signal in the

communication process, possibly above and beyond the semantic difficulties in comprehension experienced in other types of language use (such as nonsense English words).

We reasoned that the bilingual children looked at the experimenter longer when hearing the unfamiliar code-switching because they were expecting the experimenter to provide some clarification to help them understand her utterances, or at least some cues as to what these unfamiliar utterances were about. This is because code-switching usually serves as a way to contextualize daily conversation, for example, in quoting someone (Gardner-Chloros et al., 2000), to acquire the conversational turn in overlap multiparty play episodes (Cromdal, 2001), or to mark topic changes and text-to-text connection during book-reading activities (Kabuto, 2010). It is likely that children in the Unfamiliar-Switch group were expecting the experimenter to contextualize the unfamiliar language switch during the storytelling episode.

Yet, the experimenter gave no feedback to the child and provided no explanation to the code-switched sentences throughout the storytelling episode. The Unfamiliar-Switch group might have perceived that the experimenter was unhelpful and unreliable because the communication breakdown was left unresolved. The Unfamiliar-Switch group might then assume that the experimenter's point during the disambiguation task would not be helpful or reliable in interpreting the novel label after all. Studies have shown that children tended to judge a person as unhelpful or tended to avoid choosing the person as a source of help if the person had previously provided insufficient or incomplete information to them (Gweon et al., 2011; Gillis and Nilsen, 2013). Consistent with our results, Krogh-Jesperson and Echols (2012) also found that children's willingness to accept second labels depended on the perceived credibility of the speakers. This could explain why the Unfamiliar-Switch group chose to use their own ME assumptions over the experimenter's point to interpret the novel label instead, even though they have paid more attention to her earlier.

One possible interpretation of our results is that because the Unfamiliar-Switch group assumed the experimenter's point would not be helpful, they chose to *avoid* following the referential cue rather than chose to use the ME principle. We argue that the Unfamiliar-Switch group was more likely to use the ME principle rather than choose to avoid using the cue because word-learning heuristics are robust assumptions that children use to help narrow down potential objects (e.g., Jaswal and Hansen, 2006). That said, further studies could tease these two possible interpretations apart. For example, a three-object-choice paradigm could be used with the disambiguation task instead of a two-object-choice, that is, children are asked to choose between two familiar and one novel objects. If children were using ME rather than avoiding the experimenter's cue, then they would choose the novel object significantly more often than the other familiar object not pointed at. If children were avoiding following the experimenter's cue rather than using ME, then they would be equally likely to choose the novel object or the other familiar object not pointed at.

Nevertheless, our studies showed that there are nuances in the use of different cues when trying to understand a speaker's referential intent. Bilingual children are generally willing to relax ME and use the speaker's point to label a familiar object with a novel name. But this strategy may change, depending on the social communication process bounded by the context of a language switch. Bilingual children may perceive the social cues of the speaker as unhelpful or unreliable if the speaker did not behave according to the social rules surrounding language use. In this case, children may default to using word-learning heuristics to select a referent instead. Earlier unresolved social communication challenges may impact on how the social cues given by the same person will be interpreted and used later.

Our study demonstrated how the same information (e.g., gesture) might be utilized differently based on the experiences people previously had (e.g., violation of social expectation). Children tend to return to their default learning strategy as compared to possibly more effective methods provided by the speaker if they perceive the speaker as not helpful. This provides important implications for other domains that involve interactions between people and even those that involve learning applications. For example, the initial trust between learners and learning software may be undermined with a few instances of violation of expectation. The entire learning process may then lose its projected effectiveness as the learner starts to perceive the software as not helpful or unreliable. Thus, learning software and learning games may have to be designed in such a way that their credibility with the user is not lost as learning strategy changes.

In summary, we found that bilingual children were selective in their choice of cue to interpret a novel label depending on the surrounding language context (e.g., familiar or unfamiliar code-switching). We argued that bilingual children pay increased attention to the speaker when hearing unfamiliar code-switching partly for the purpose of overcoming communication challenges. Despite this, we found that bilingual children did not necessarily use the speaker's point to interpret a novel label. They would weigh the various sources of information available to them and rely more on their own ME assumptions if they regarded the speaker as unhelpful according to their past interaction with the speaker. Future studies could examine whether bilingual children would regard a speaker who code-switches in an unfamiliar language as an unhelpful informant, and how this perception of unhelpfulness might influence their willingness to accept the speaker's communicative cues to interpret a novel label. Further studies could also examine how children's perceived helpfulness of the speaker would generalize to other learning contexts, such as from adults vs. from educational software, pointing cues vs. paralinguistic cues, etc. This may have important implications on general education and how learning can be best scaffolded in a helpful and trusting environment.

Acknowledgments

This research was supported by the SUTD SRG grant (SRG HASS 2011 011) and the SUTD-MIT IDC grant (IDG31100106 and IDD41100104) awarded to the last author. We thank all parents and children who participated in this study. We are also grateful

for Ms. Beng Luan Tan (Principal of Creative O), Ms. Rachel Ding (Principal of Red SchoolHouse), Ms. Emily Ho (Principal of Columbia Academy) and the teachers and administration staff at the above three preschools for their support throughout our data collection.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2015.00796/abstract

References

Au, T., and Glusman, M. (1990). The principle of mutual exclusivity in word learning: to honor or not to honor? *Child Dev.* 61, 1474–1490. doi: 10.2307/1130757

Baron-Cohen, S. (1989). Perceptual role taking and protodeclarative pointing in autism. *Br. J. Dev. Psychol.* 7, 113–127. doi: 10.1111/j.2044-835X.1989.tb00793.x

Bassetti, B. (2007). Grammatical gender and concepts of objects in Italian-German bilingual children. *Int. J. Biling.* 11, 251–273. doi: 10.1177/13670069070110030101

Beauvillain, C., and Grainger, J. (1987). Accessing interlexical homographs: some limitations of a language-selective access. *J. Mem. Lang.* 26, 658–672. doi: 10.1016/0749-596X(87)90108-2

Brojde, C. L., Ahmed, S., and Colunga, E. (2012). Bilingual and monolingual children attend to different cues when learning new words. *Front. Psychol.* 3:155. doi: 10.3389/fpsyg.2012.00155

Byers-Heinlein, K. (2013). Parental language mixing: its measurement and the relation of mixed input to young bilingual children's vocabulary size. *Bilingualism Lang. Cogn.* 16, 32–48. doi: 10.1017/S1366728912000120

Byers-Heinlein, K., and Werker, J. F. (2009). Monolingual, bilingual, trilingual: infants' language experience influences the development of a word learning heuristic. *Dev. Sci.* 12, 815–823. doi: 10.1111/j.1467-7687.2009.00902.x

Carroll, L. (1872). *Through the Looking-Glass and WAlice Found There*. London: Macmillan & Co.

Clark, E. V. (1988). On the logic of contrast. *J. Child Lang.* 15, 317–335. doi: 10.1017/S0305000900012393

Cromdal, J. (2001). Overlap in bilingual play: some implications of code-switching for overlap resolution. *Res. Lang. Soc. Int.* 34, 421–451. doi: 10.1207/S15327973RLSI3404_02

Davidson, D., Jergovic, D., Imami, Z., and Theodos, V. (1997). Monolingual and bilingual children's use of the mutual exclusivity constraint. *J. Child Lang.* 24, 3–24. doi: 10.1017/S0305000996002917

de Bezenac, A. (2010a). *Cookie Rookie*. Available at: http://www.freekidsbooks.org

de Bezenac, A. (2010b). *That Worked!*. Available at: http://www.freekidsbooks.org

de Bezenac, A. (2010c). *Turtle Trouble*. Available at: http://www.freekidsbooks.org

Franco, F., and Butterworth, G. (1996). Pointing and social awareness: declaring and requesting in the second year. *J. Child Lang.* 23, 307–336. doi: 10.1017/S0305000900008813

Gardner-Chloros, P., Charles, R., and Cheshire, J. (2000). Parallel patterns? A comparison of monolingual speech and bilingual codeswitching discourse. *J. Pragmat.* 32, 1305–1341. doi: 10.1016/S0378-2166(99)00120-4

Genesee, F., Tucker, G. R., and Lambert, W. E. (1975). Communication skills of bilingual children. *Child Dev.* 46, 1010–1014. doi: 10.2307/1128415

Gerstadt, C. L., Hong, Y. J., and Diamond, A. (1994). The relationship between cognition and action: performance of children 3 ½–7 years old on a Stroop-like day-night test. *Cognition* 53, 129–153. doi: 10.1016/0010-0277(94)90068-X

Gillis, R. L., and Nilsen, E. S. (2013). Children's use of information quality to establish speaker preferences. *Dev. Psychol.* 49, 480–490. doi: 10.1037/a0029479

Grassmann, S., and Tomasello, M. (2010). Young children follow pointing over words in interpreting acts of reference. *Dev. Sci.* 13, 252–263. doi: 10.1111/j.1467-7687.2009.00871.x

Grosjean, F. (1988). Exploring the recognition of guest words in bilingual speech. *Lang. Cogn. Process.* 3, 233–274. doi: 10.1080/01690968808402089

Gweon, H., Pelton, H., and Schulz, L. E. (2011). "Adults and school-aged children accurately evaluate sins of omission in pedagogical context," in *Proceedings of the 33rd Annual Conference of Cognitive Science Society*, Boston, MA, 1342–1347.

Healey, E., and Skarabela, B. (2008). "Are children willing to accept two labels for a single object? A comparative study of mutual exclusivity in monolingual and bilingual children," in *Proceedings of the 2007 Child Language Seminar: 30 Anniversary*, eds T. Marinis, A. Papangeli, and V. Stojanovik (Reading: University of Reading), 49–58.

Houston-Price, C., Caloghiris, Z., and Raviglione, E. (2010). Language experience shapes the development of the mutual exclusivity bias. *Infancy* 15, 125–150. doi: 10.1111/j.1532-7078.2009.00009.x

Jaswal, V. K., and Hansen, M. B. (2006). Learning words: children disregard some pragmatic information that conflicts with mutual exclusivity. *Dev. Sci.* 9, 158–165. doi: 10.1111/j.1467-7687.2006.00475.x

Kabuto, B. (2010). Code-switching during parent-child interactions: taking multiple theoretical perspectives. *J. Early Child. Lit.* 10, 131–157. doi: 10.1177/1468798409345109

Kolers, P. (1966). Reading and talking bilingually. *Am. J. Psychol.* 79, 357–376. doi: 10.2307/1420877

Krogh-Jesperson, S., and Echols, C.-H. (2012). The influence of speaker reliability on first versus second label learning. *Child Dev.* 83, 581–590. doi: 10.1111/j.1467-8624.2011.01713.x

Landau, B., Smith, L. B., and Jones, S. S. (1998). Object shape, object function and object name. *J. Mem. Lang.* 38, 1–27. doi: 10.1006/jmla.1997.2533

Lempers, J. (1979). Young children's production and comprehension of nonverbal deitic behaviors. *J. Genet. Psychol.* 135, 93–102. doi: 10.1080/00221325.1979.10533420

Lennon, J. (1965). *A Spaniard in the Works*. New York, NY: Simon & Schuster.

Leung, E., and Rheingold, H. (1981). Development of pointing as a social gesture. *Dev. Psychol.* 17, 215–220. doi: 10.1037/0012-1649.17.2.215

Macnamara, J. (1967). The linguistic independence of bilinguals. *J. Verbal Learning Verbal Behav.* 6, 729–736. doi: 10.1016/S0022-5371(67)80078-1

Markman, E. M. (1994). Constraints on word meaning in early language learning. *Lingua* 92, 199–227. doi: 10.1016/0024-3841(94)90342-5

Markman, E. M., and Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cogn. Psychol.* 20, 121–157. doi: 10.1016/0010-0285(88)90017-5

Meuller, R. F. I., and Allport, A. (1999). Bilingual language switching in naming: asymmetrical costs of language selection. *J. Mem. Lang.* 40, 25–40. doi: 10.1006/jmla.1998.2602

Povinelli, D. J., Reaux, J. E., Bierschwale, D. T., Allain, A. D., and Simon, B. B. (1997). Exploitation of pointing as a referential gesture in young children, but not adolescent chimpanzees. *Cogn. Dev.* 12, 423–461. doi: 10.1016/S0885-2014(97)90017-4

Rodriguez-Fornells, A., Kramer, U. M., Lorenzo-Seva, U., Festman, J., and Munte, T. F. (2012). Self-assessment of individual differences in language switching. *Front. Psychol.* 2:388. doi: 10.3389/fpsyg.2011.00388

Rosenblum, T., and Pinker, S. A. (1983). Word magic revisited: monolingual and bilingual children's understanding of the word-object relationship. *Child Dev.* 54, 773–780. doi: 10.2307/1130064

Siegal, M., Iozzi, L., and Surian, L. (2009). Bilingualism and conversational understanding in young children. *Cognition* 110, 115–122. doi: 10.1016/j.cognition.2008.11.002

Thomas, M. S. C., and Allport, D. A. (2000). Switching costs in bilingual visual word recognition. *J. Mem. Lang.* 43, 44–66. doi: 10.1006/jmla.1999.2700

Wechsler, D. (1974). *Manual for the Wechsler Intelligence Scale for Children—Revised*. New York, NY: Psychological Corporation.

Yow, W. Q., and Hung, W. Y. (2013). Impact of bilingual (code-switching) experience on preschoolers? Sensitivity to pragmatic cues. *Paper Presented at the Society of Research in Child Development*, Seattle.

Yow, W. Q., and Markman, E. M. (2007). "Monolingual and bilingual children's use of the mutual exclusivity assumption and pragmatic cues in word learning," in *Poster Presented at 2007 SRCD Biennial Meeting*, Boston.

Yow, W. Q., and Markman, E. M. (2011). Young bilingual children's heightened sensitivity to referential cues. *J. Cogn. Dev.* 12, 12–31. doi: 10.1080/15248372.2011.539524

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 04 November 2015
doi: 10.3389/fpsyg.2015.01641

Ghost-in-the-Machine reveals human social signals for human–robot interaction

Sebastian Loth[1], Katharina Jettka[1], Manuel Giuliani[2] and Jan P. de Ruiter[1]*

[1] Psycholinguistics, Faculty of Linguistics and Literary Studies, Bielefeld University, Bielefeld, Germany, [2] Center for Human-Computer Interaction, Department of Computer Sciences, University of Salzburg, Salzburg, Austria

We used a new method called "Ghost-in-the-Machine" (GiM) to investigate social interactions with a robotic bartender taking orders for drinks and serving them. Using the GiM paradigm allowed us to identify how human participants recognize the intentions of customers on the basis of the output of the robotic recognizers. Specifically, we measured which recognizer modalities (e.g., speech, the distance to the bar) were relevant at different stages of the interaction. This provided insights into human social behavior necessary for the development of socially competent robots. When initiating the drink-order interaction, the most important recognizers were those based on computer vision. When drink orders were being placed, however, the most important information source was the speech recognition. Interestingly, the participants used only a subset of the available information, focussing only on a few relevant recognizers while ignoring others. This reduced the risk of acting on erroneous sensor data and enabled them to complete service interactions more swiftly than a robot using all available sensor data. We also investigated socially appropriate response strategies. In their responses, the participants preferred to use the same modality as the customer's requests, e.g., they tended to respond verbally to verbal requests. Also, they added redundancy to their responses, for instance by using echo questions. We argue that incorporating the social strategies discovered with the GiM paradigm in multimodal grammars of human–robot interactions improves the robustness and the ease-of-use of these interactions, and therefore provides a smoother user experience.

Keywords: human–robot interaction, social behavior, eye tracking, interaction strategies, social signals, intention recognition

OPEN ACCESS

Edited by:
Claire Marie Fletcher-Flinn, University of Auckland, New Zealand

Reviewed by:
Yuichi Yamashita, National Center of Neurology and Psychiatry, Japan
Paul Dickerson, University of Roehampton, UK

****Correspondence:***
Sebastian Loth
sebastian.loth@uni-bielefeld.de

Specialty section:
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology

Received: *12 March 2015*
Accepted: *12 October 2015*
Published: *04 November 2015*

Citation:
Loth S, Jettka K, Giuliani M and de Ruiter JP (2015) Ghost-in-the-Machine reveals human social signals for human–robot interaction. Front. Psychol. 6:1641. doi: 10.3389/fpsyg.2015.01641

INTRODUCTION

Robotic agents are increasingly used for interacting with humans in public spaces, e.g., for providing information as a museum guide (Yousuf et al., 2012) or serving snacks (Lee et al., 2009). We used the bar scenario as challenging example for a social environment. The robot acts as bartender that accepts drink orders from human customers and serves drinks (see **Figure 1**). Thus, the robot has to complete the task (i.e., serving the correct drink) and, importantly, it has to understand and produce socially acceptable behavior. The bartending robot is located at a fixed position behind the bar. Typically multiple customers are in close proximity in front of the bar. First, the robot has to identify the customers who would like to initiate an interaction

FIGURE 1 | Robotic bartender JAMES serving drinks to a customer.

(Loth et al., 2013). Once the interaction has been established, the robot has to sense the customer's dialog moves, reason about them and produce an appropriate response (Petrick and Foster, 2012). That means that the robot has to have an understanding of the user's engagement behavior (Sidner and Lee, 2003; Sidner et al., 2005), recognize the user's intentions (Gray et al., 2005), and produce socially appropriate responses (Petrick and Foster, 2012; Breazeal et al., 2013). Thus, reliable, robust, and social interaction policies are crucial for enabling users to interact intuitively with a robot (Goodrich and Schultz, 2007). Additionally, users enjoy interacting with a social robot to a greater extent than with a purely task-oriented system (Foster et al., 2012; Giuliani et al., 2013). In order to develop empirically driven and socially appropriate interaction policies for the robotic bartender, we tested (a) whether the recognizer data are sufficient for entertaining a socially credible interaction, (b) which recognizer modality was the most informative at each stage of the interaction, and (c) what kind of repair strategies humans employ in a social interaction. We used the Ghost-in-the-Machine paradigm (GiM; Loth et al., 2014) because the results can be transferred directly into robot policies as the human participants are presented the same recognizer data as the robotic planner.

RELATED WORK

Human–human interaction is highly fluent and can be regarded as the gold standard for human–robot interaction. Thus, we briefly review the mechanisms involved in human–human interaction and how they can be modeled in a robotic agent. Typically empirical studies were designed for investigating particular aspects of robotic interaction policies. We review previous studies with respect to how transferable their results are. In particular, we focus on whether the data that the human participants observed in the study were comparable to the kind of data that the robotic planner has access to. We highlight potential problems in these studies before describing our GiM study in more detail.

Social Signals

Interacting with other humans is perceived as most natural and intuitive compared to robotic or virtual agents. Thus, in order to improve the interaction with the robotic bartender, we have to understand how humans communicate their intentions in a social environment. Levinson (1995) argued that humans recognize the intentions of others from communicative actions. These are composed of one or more observable, basic actions

in several modalities (e.g., Levinson, 1995; Vinciarelli et al., 2012). We refer to these observable actions as *social signals*. These basic actions are the starting point for human and robotic recognition. Humans identify basic actions such as walking by matching the percept against a representation in memory (Jeannerod, 2006). But it is not clear how humans understand the intention of somebody who is walking (Levinson, 1995). In robots, sensor data are typically categorized by trained classifiers into one type of action. For example, the computer vision recognizes dynamic actions such as waving, walking, and running (for review, see Poppe, 2010). Also, the user's pose (Shotton et al., 2013), hands and faces can be tracked (Baltzakis et al., 2012; Gaschler et al., 2012) for identifying deictic gestures in (close to) real-time (Pateraki et al., 2014). The automatic speech recognition (ASR) aims to recognize the user's utterance by matching it against a dictionary (or a grammar and a dictionary). In general, recognizers transform a constant stream of data from the sensors (e.g., microphone, camera) into distinct events such as an instance of waving or a specific speech utterance. However, robotic recognizers generally require a substantial amount of computation. Additionally, a dimly lit and noisy bar location challenges them such that their results tend to be more error-prone. Human bartenders face a similar problem as they cannot constantly monitor each potential customer in a busy place given that their cognitive resources are limited (e.g., Broadbent, 1969). This holds especially for monitoring within a single sensory modality[1] (Allport et al., 1972; Mcleod, 1977). Thus, the human bartenders have to employ heuristics, for instance by focussing on distinctive aspects of the scene (e.g., the distance of customers to the bar).

Humans select the relevant aspects by relying on prior shared knowledge about the expected behavior and signals of both partners in the interaction (Levinson, 1995). These expectations also determine the attentional focus of the partners. Once the signal is identified, humans evaluate plausible intentions, i.e., the human recipient tries to attribute a plausible social intention to the signal (Grice, 1957; Levinson, 1995). This is essential as it makes an action meaningful. But correctly identifying social signals and understanding other's intentions is logically intractable and thus, prior shared knowledge and heuristics are required (Levinson, 1995). For a robotic agent, this knowledge has to be explicated and formalized, e.g., in scripts that capture the conventionalized sequence of events (Schank and Abelson, 1977; Abelson, 1981) or the computational AIRBUS model that combines prior expectations, knowledge about conventions, and recognized signals during interactions (De Ruiter and Cummins, 2012). By explicating this implicit social knowledge, we can improve the robustness and the perceived quality of human–robot interaction. At the same time, the computational efforts can be limited to extracting only the necessary information by identifying the relevant recognizer modalities. For example, in the bar scenario customers signal to a member of staff that they would like to place an order by positioning themselves very close to the bar and turning toward the counter or a member of staff (Loth et al., 2013). Thus, only these two modalities have to be attended in order to identify new customers reliably. Furthermore, the participants only attended the body posture of potential customers if they were close to the bar whereas the body posture was irrelevant for customers who were further away. This reduces the cognitive load of understanding the scene for a human observer even further. Using this hierarchical rule-set is also advantageous to the robot. By analysing the body posture of customers who are close to the bar only, the line of sight is less likely to be obstructed by objects and other customers and thus, the recognizer works more reliably with less computational efforts. Additionally, the robotic recognizers are subject to noise. By reducing the number of noisy data sources, the amount of potentially misleading recognizer data is also reduced. Thus, our central aim is to provide an empirical method for reliably identifying social signals and the relevant recognizer modalities they are signaled in.

Our review of human social cognition suggests that using prior knowledge and focussing on particular aspects of the scene (recognizer modalities) can reduce errors and computational efforts. However, this is achieved by ignoring substantial amounts of data which may sound counter-intuitive. But this is a general finding in human cognition. Humans focus on task-relevant aspects of the scene and ignore other events in the visual (inattentional blindness; Mack and Rock, 1998) and auditory domain (inattentional deafness; Dalton and Fraenkel, 2012). For example, Simons and Chabris (1999) asked their participants to count the number of passes played by a basketball team and argued that the frequent failures to notice a man in a gorilla costume who walked through the scene were due his irrelevance to the task. Thus, by selecting the aspects of the scene (recognizer modalities) appropriately, the robot's performance becomes more human-like and more predictable to its human users. In turn, we aimed to identify which aspect of the scene is relevant before and during an interaction at the bar.

In a social interaction, producing socially acceptable behavior is equally important as understanding it. For example, in a task requiring users to sort blocks that were handed to them by a robot, they sorted the blocks on their own strategy, e.g., by color. Only if a short delay was included between stretching the robot's arm and releasing the block, the users attended the robot's gaze and used it as a sorting instruction (Admoni et al., 2014). Thus, the delay formed a social signal to attend the robot's gaze direction. Also, users smile more often if the robot smiles at them (Krämer et al., 2013). In general, interacting with a robot that acts socially appropriately is perceived as more pleasing than with a purely task-oriented robot (Giuliani et al., 2013). Thus, we aim to identify social signals to be displayed by robotic bartenders that can be reliably interpreted by its customers.

Methods of Deriving Interaction Models

Interaction models can be hand-crafted but are often partly based on empirical data. For example, hand-crafted models are typically

[1] In Psychology, the term modality refers to sensory modalities such as vision. In robotics, a modality tends to refer to a particular variable of the recognizer output. For example, the customer's distance to the bar and her/his body orientation form two robotic modalities even though both are derived from the human visual modality. In order to avoid confusion, we distinguish sensory modality (e.g., vision) and recognizer modality (e.g., distance to the bar).

adapted after an initial testing period in the wild such that the first model serves as test and data collection device. Other methods of gathering empirical data are computer games and the Wizard-of-Oz paradigm (WOz). In this review, we focus on how the relevant recognizer modalities were identified.

For detecting whether visitors intended to interact with a robotic receptionist, Michalowski et al. (2006) based their interaction model on proxemics (Hall, 1969). This hand-crafted model triggered a greeting as soon as a potential user was sufficiently close to the robot. But passers-by who accidentally came close to the robot felt disturbed when the robot greeted them out of the blue (Michalowski et al., 2006, p. 766). Thus, Rich et al. (2010) and Holroyd et al. (2011) used several multimodal cues that were partly inspired by research on human–human interaction (Schegloff and Sacks, 1973), e.g., the point of gaze. This is a highly informative cue but it can be difficult to measure in the wild. Importantly, it might not be accessible to humans in a busy environment and thus, not be part of the conventionalized social signals that we aim to identify. For example, in the bar setting less fine grained aspects of the scene such as the distance to the bar and the body or head orientation were most relevant (Loth et al., 2013). An initially hand-crafted interaction model can also be adapted to the user behavior during a test period of real-world interactions. For example, Bohus and Horvitz (2009a,b,c,d, 2010, 2011) implemented a number of sensors and recognizers in their static receptionist and trivia quiz platform, and more recently in a direction-giving robot (Bohus et al., 2014). They refined their engagement models constantly but they could not accommodate all user behavior (Bohus and Horvitz, 2009a). In particular, multiple users formed a challenge for these accounts (Michalowski et al., 2006; Bohus and Horvitz, 2009c) whereas our bar scenario typically involves multiple customers. Goodrich and Schultz (2007) classified these accounts as *proof-of-concept* because the users interacted with a given system and adapted their behavior. This was illustrated by the graphically simple WAITER game (Xu et al., 2010). Even though the manager participant had only indirect evidence, this participant adapted quickly to the abilities of the waiter participant that were manipulated by the game engine. This suggests that proof-of-concept approaches do not investigate what is intuitive to the users but how well they adapt to a given system. However, identifying the underlying psychological principles of natural behavior and designing the robot's policies around them is more useful (Goodrich and Schultz, 2007).

Games with a purpose (GWAP; von Ahn and Dabbish, 2008) and in particular online games allow acquiring large data sets, e.g., as training data for machine learning accounts. In The-Restaurant-Game, users could engage online as waitress or customer (Orkin and Roy, 2007, 2009). Orkin and Roy (2009) derived a sequential graph of actions that was argued to reflect collective intelligence. After training the virtual agents on these data, they worked reasonably but also produced some errors, e.g., asking for selecting a starter after starters had just been served. Even though the players had an intentional structure in mind, this method did not capture this structure from the surface behavior (Orkin and Roy, 2009, p. 392).

The WOz paradigm is typically used for investigating the user behavior while s/he believes to interact with a real robot. But in fact, an informed assistant or another participant acts as a 'wizard' that controls the robot (Kelley, 1984; Fraser and Gilbert, 1991; Dahlbäck et al., 1993). For maintaining the illusion of a real robot and providing swift responses, the workload of controlling the robot sometimes has to be divided between several wizards which may cause inconsistencies in the robot's behavior (Green et al., 2004; Rieser and Lemon, 2009). Several WOz studies also investigated the behavior of a single wizard. For example, for investigating when wizards asked for clarifications (Rieser and Lemon, 2009) and which mode of information presentation they selected (Rieser et al., 2011). In these studies, the distortion of an ASR was simulated by a typist translating the user's speech into text and deleting or replacing words. However, in more than 80% of WOz studies, the wizards had access to immediate, unfiltered video and audio data of their users (Riek, 2012). In contrast, the robotic planner has to rely on the robot's recognizers introducing delays, losses, and misinterpretations of data. This difference can impair the transferability of the findings into robotic decision policies. For example, Lee et al. (2009) collected WOz data and designed a script for their Snackbot. But the real-life evaluation showed that half of their script phrases were unsuitable (Lee et al., 2009, p. 11). Thus, it is important to ensure that the wizards and the robotic planner operate on the same type of data. For example, semantically analyzed data of the ASR component was presented to the wizards of a restaurant information system (Liu et al., 2009). This method is similar to our GiM approach (Loth et al., 2014) and we expand on this in our study.

Lichtenthäler et al. (2013) introduced the *Inversed Oz of Wizard* for investigating how a wizard would avoid a collision between a confederate and the robot under her/his control. In this setting, the wizard observed the confederate and the robot from the same room. Thus, the human observer could have subconsciously interpreted subtle cues in the motion patterns of the confederate that the robotic recognizers are not able to interpret reliably, e.g., by observing the motion preceding an attack in volleyball (Schorer et al., 2013) or a penalty kick in football (Noël et al., 2014), athletes can anticipate the actions of their opponents (also see Abernethy et al., 2007). This is more pronounced in everyday behaviors of groups as they tend to synchronize by subtly communicating their next movements to each other (Néda et al., 2000; Richardson et al., 2007; Lakens and Stel, 2011). Thus, especially in settings with multiple users such as the bar scenario, the robotic planner would not have access to the information that was essential to the human performance. In order to avoid this missing link, we carefully designed the GiM interface such that the human participant has access to the same information as the robotic planner.

MATERIALS AND METHODS

We aimed to (a) identify the social signals and relevant recognizer modalities in the bar scenario, (b) learn how the robotic bartender should respond to its customers in a socially appropriate way,

and (c) combine these insights for developing strategies for recovering from false or inconclusive recognizer data that are socially acceptable and, specifically, less annoying to the customer than repeatedly asking for clarifications. Thus, we used the GiM paradigm (Loth et al., 2014). In this paradigm, the main participant (*ghost*) observes the scene through the eyes and ears of the robot, i.e., the ghost has access to the recognizer data but no direct video or audio link to the customers. Hence, the ghost and the robotic planner use the same data. In order to interact with the customers, the ghost has to select actions from the robot's repertoire. In contrast to the typical WOz studies that focus on the user's behavior, we are primarily interested in the behavior of the ghosts. For assessing the reliability of this paradigm, we compared our findings to earlier empirical studies that relied on real world observations (Brouwer et al., 1979) and experiments using natural stimuli (Loth et al., 2013, 2015).

In order to avoid confusion, we refer to the main participants as *ghosts* and to the participants who ordered drinks as *customers*.

Participants

Thirty-one participants were recruited as main participants from the departmental participant pool (formed of linguistics and other students as well as university staff) in Bielefeld, Germany. They received €5 and a chocolate bar in exchange for their time and effort. The eye tracker could not be calibrated with two participants and their data were not included in the results.

The experiment and all procedures were approved by University Bielefeld's Ethics Committee (EUB) under approval No4807. An informed written consent was collected prior to the experiment.

Apparatus

The participants were seated in front of a typical office computer screen (50 cm × 32 cm, 1920 × 1200 pixel) with a viewing distance of approximately 70 cm. Their eye gaze was recorded using a head-free faceLAB Eye Tracker (2009) positioned below the center of the screen. A dedicated JAVA application (Java Runtime Environment, 2012) presented the recognizer data and recorded the ghosts' responses entered through a standard keyboard and mouse. We positioned the control screen for the eye tracker such that the participants could not see the display in order to avoid distraction. An experimenter checked whether the eye tracker worked as intended but was seated such that it was obvious to the participant that s/he was not observed. The setup is shown in **Figure 2**.

Ghost-in-the-Machine Design

The ghosts were presented the output of the robotic recognizers by visualizing the variables using arrows and traffic lights. However, we were careful not to add any information that the planner cannot access. The ghosts responded to their customers by selecting actions from the robot's repertoire that met their own expectations of appropriate behavior. In the following, we describe the control and information panels, their content and how this relates to a typical robotic architecture in more detail.

The user interface for the ghosts consisted of three frames on a computer screen. On the left and right hand side of the screen, an information panel for each of the two customers was presented. At the bottom center of the screen, the control panel showed the robot's repertoire (see **Figure 3**).

In the architecture of the robot, the sensors (e.g., camera or microphone) transmit their data to recognizers. These software

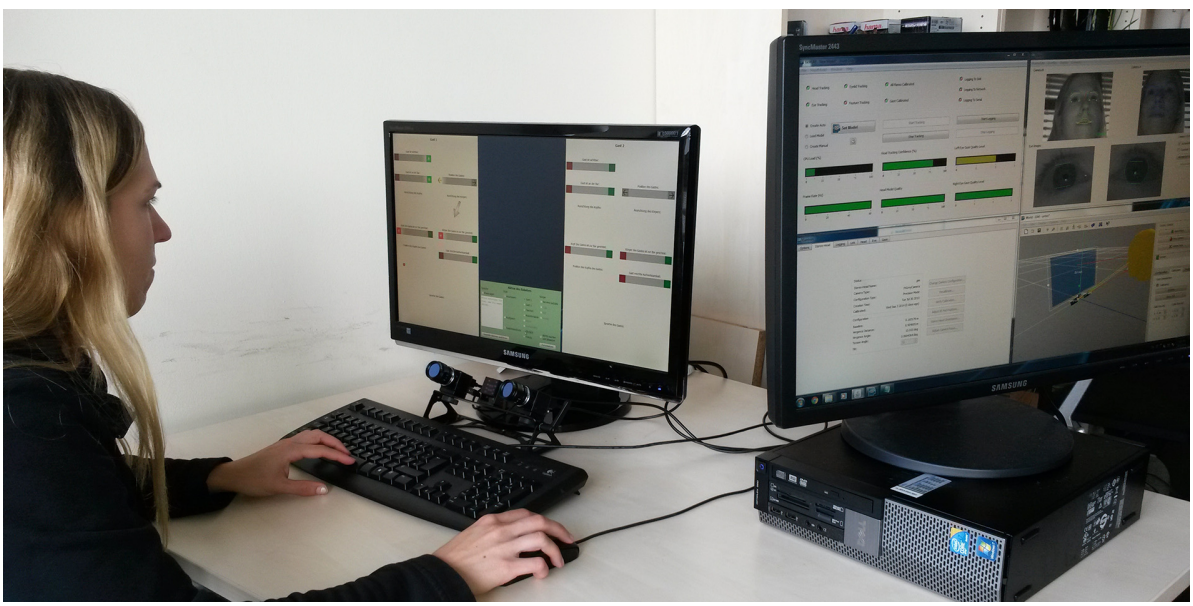


FIGURE 2 | Setup of the Ghost-in-the-Machine (GiM) study with the ghost participant, eye tracker and GiM user interface on the left hand side, and the eye tracking control screen on the right hand side.

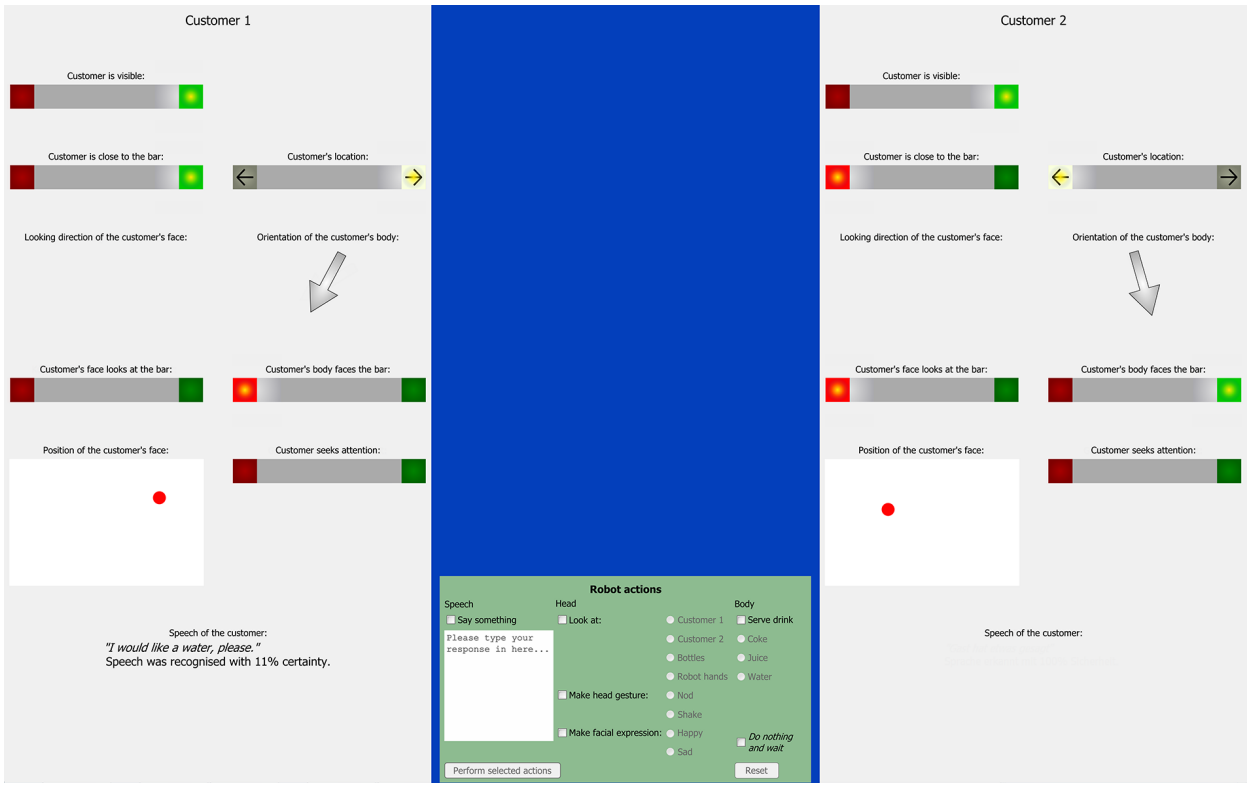


FIGURE 3 | Full screen shot of the GiM interface showing information panels for up to two customers and the control panel for selecting actions.

programs analyze the raw data and extract information, e.g., the presence of a face or the words spoken. A component called the social state estimator collects these data and produces an updated state representation to the robotic planner if a major change was detected (Foster et al., 2013; Petrick and Foster, 2013). The updates slice the continuous data from the sensors and recognizers into distinct, temporally ordered updates of the scene. Each update formed a *turn* in this GiM study. A turn comprised of (a) an update of the information panels, and (b) a response by the ghosts. The next update was presented after the ghosts confirmed their selected actions (or explicitly selected no action) without time limit. Thus, the time span between recorded updates and presented updates could differ but their temporal order was unchanged. This turn-by-turn cycle continued until the trial was terminated. Since we used pre-recorded customer data in this study, the ghosts' actions were recorded but never enacted by a robot and we did not try to convey otherwise. Thus, there was a potential discrepancy between the customer's and ghost's actions. This was addressed according to the experimental condition (see Materials and Conditions for details).

The user interface aimed at presenting the abstract recognizer data intuitively to the ghosts. The recognizer updates for each customer comprised of six binary variables (*is visible*, *is close to bar*, *location to left/right*, *face looks at bar*, *body faces bar*, *seeks attention*), three continuous numeric variables (*body orientation* and *face orientation* in degrees of angle, and the coordinates of the customer's *face position*) and one variable dedicated to the customer's speech. The values of the binary variables are computed by the social state estimator. For this, the social state estimator has built-in knowledge about the geometry of the robot's bar and an interpretation mechanism that computes whether a customer is seeking the robot's attention based on the his/her body posture and location at the bar (Foster et al., 2013).

The binary variables were presented in the style of a traffic light indicating that these variables could be true (green) or false (red). For the *customer's location*, the same design was used with a left and right pointing arrow. If data for one indicator were not available, both lights were switched off. For example, if only one customer was visible in the scene, the other information panel was 'switched off'. The angles of the body and face orientation (if available) were presented as arrows such that pointing downward represented a face/body looking straight at the counter. The position of the customer's face was represented as a dot in a rectangle representing the space above the bar counter. Finally, the output of the ASR was presented at the bottom of each information panel. If speech was detected, this component showed the final speech hypothesis and its confidence level.

The control panel listed the robot's repertoire in several groups with radio button selectors. The ghosts could use a free text field for speaking to the customers. The ghosts could make the robot look at one of the customers, the bottles, or the robot's

hands. They could select to nod or to shake the robot's head and select a happy or a sad face. The panel also offered to serve one out of the three available drinks. Finally, the panel offered an option to do nothing and wait for the next update. This was a check box that had to be explicitly selected in order to hinder the ghosts from just clicking through the trials. In order to proceed to the next turn, the ghosts had to select at least one action or tick the *Do nothing and wait*-check box and confirm their selection. The interface hindered the ghosts from using impossible combinations (e.g., making a happy and a sad face at the same time). The selected action could be as complex as desired, e.g., looking at a customer, saying 'Here is your drink,' smiling and serving the drink.

Materials and Conditions

The recognizer data were pre-recorded during an evaluation of the real robot in Munich (Foster et al., 2012). The evaluation trials included up to two customers in several configurations: both customers order drinks, one of the customers orders both drinks, and only one customer orders a drink with a bystander. For the evaluation, naïve participants were recruited and instructed to order a drink from the robotic bartender in English. The menu consisting of water, coke, and juice was introduced to them but no further instructions were given, i.e., there was no directive with regards on how to approach, speak to or take the drinks from the robot. After the participants placed an order, they evaluated the robot in a questionnaire (for further details, see Foster et al., 2012). Examples of the recognizer recordings are presented in the Supplementary Material.

This GiM study included an intention and a speech recognition condition. The intention recognition trials focussed on how interactions between customers and the robotic agent were initiated. We assessed the validity of the GiM paradigm by directly comparing its results to findings from an experiment with natural data (Loth et al., 2013). The speech recognition trials investigated how the ghosts identified which drink to serve. We were especially interested in socially appropriate repair strategies if the ASR hindered the robot from identifying an order or caused long delays.

We selected two practice and six experimental trials per condition. Based on the video recordings of the evaluation in Munich, the practice and two experimental trials per condition were selected to be relatively easy. That means that the recognizers provided clear data and in turn, the robot performed well without producing long delays or repeatedly asking the customers for their orders. The remaining four trials represented difficult cases that aimed at eliciting repair strategies. They included long delays in the interaction, failures to gather correct sensor data and/or failures of the robot's decision policies. Alternating easy and difficult trials aimed at hindering the ghosts from treating less accurate data (e.g., very low confidence levels of the ASR) as if they were normal rather than thinking about strategies. All data presented to the ghosts were real recognizer data and thus, subject to noise, inaccuracies and sensor failures. This was explained to ghosts in the instructions in order to make clear that the displayed information was not ground truth but that data can be false or conflicting.

The intention and speech recognition trials differed in how the trials were organized. Since the main focus of the intention recognition trials was at the very beginning of the interaction, the respective recognizer data were presented starting from the first update of the recorded customer–robot interaction. Our aim was to establish how the ghosts identified whether a customer intended to place an order and how the existing computational account should be adapted. Since the indicator *Seeking attention* reflected the existing computation and could have biased the ghosts, it was switched off. The trials were terminated as soon as the ghost selected an action other than *Do nothing and wait*. Thus, we tracked when and how the ghosts first acknowledged a new customer. Since we used pre-recorded data, the ghosts may have selected an action that differed or occurred earlier than the robot's actions during the evaluation. In turn, the customer's response would not match the ghost's actions. We minimized this risk by terminating the trials quickly. In contrast, the speech recognition trials aimed at a later stage in the interaction. Thus, we had to ensure that the ghost did not undertake any actions that mismatched with their customer until the order was placed. At the same time, the ghost had to be informed about what has happened until then. Thus, we presented the recognizer updates from the beginning but altered the control panel such that only a *Continue* button was available. Clicking *Continue* updated the indicators to the next update. As a result, the ghosts observed what has happened during the trial but were unable to deviate from the pre-recorded actions. As soon as one of the customers made a speech utterance, the control panel was rolled back and allowed the ghosts to select any of the actions. The trials terminated as soon as the ghost served a drink or there were no more updates to display. This possibly long interaction increased the risk of a discrepancy between the ghost and its customers. However, we suspected that the ghosts would aim to understand the drink order and we knew from the recordings that the customers repeated their drink order in various ways. We report on this in the results and discussion sections. In all trials, the ghost was informed about the end of a trial by a pop-up message and removing all panels from screen. Once the message was confirmed, the panels appeared on screen and the next trial started.

RESULTS

We report the results of the intention and speech recognition trials separately. In each section, we summarize (a) the recognizer data displayed in the information panels, (b) the turn duration and summed dwell time on screen, (c) the dwell time on each indicator, and (d) the selected response. The analyses of the recognizer and eye tracking data refer to the addressed customer. For example, if the ghost selected to look at *Customer 1*, the recognizer data of *Customer 1* were analyzed.

In general, the ghosts experienced the experiment as very immersive. This was the case even though the customer data were pre-recorded, there was no actual customer feedback and the interface was comparably simple. For example, some ghosts apologized for having made jokes to their customers after the

experiment or complained about the unresponsiveness of their customers if the recognizers did not show any new information. However, some ghost participants trialed how well our design would respond to unexpected behaviors and tried to take advantage of the pre-recorded nature of the data. These trials are listed in detail with regards to each condition.

Intention Recognition Trials

The experiment comprised a total of 174 intention recognition trials from 29 participants. Three ghosts repeatedly selected *Do nothing and wait* until the ASR unequivocally identified an order. Their data and two additional trials showing a similar pattern were excluded (in total 20 trials, 11.5%). One additional trial was excluded because the ghost ignored the customers and did not respond at all. Thus, the following analyses cover the 153 remaining trials. Each trial was a customer–staff interaction of one or more turns. The customers' actions were presented through an update of the indicators and the ghosts responded by selecting a robot action which completed the turn. The intention recognition trials continued with another turn if the ghosts selected *Wait and do nothing* (*No response*) and were terminated with the ghosts' first selected action (*Response*). Thus, each of the 153 trials comprised one *Response* turn. The total of 117 *No response*-turns distributed over 69 trials.

Recognizer Data

The data in **Table 1** summarize the recognizer data by listing the state of the traffic light and arrow indicators as well as the presence of detected speech in the information panel of the addressed customer (see **Figure 3**). Please note that the indicator *Seeks attention* was switched off in all intention recognition trials (see Materials and Conditions). The arrow shaped indicator showing the *Face orientation* was never active due to a technical problem during the data acquisition. For the same reason, the binary indicator *Face to bar* either showed no value or *false*, i.e., this indicator never showed *true*. Thus, the information from both indicators was potentially misleading and we return this when discussing the results.

The continuous indicator *Body orientation* was recoded as a binomial variable such that we distinguished whether the arrow was displayed (*known*) or not (*unknown*). We opted for this simplification because the recognizer was only able to compute the angles from the camera image if the customers faced the camera to some degree. If the customer turned away especially when turning outward, the recognizer could not determine the angle. The recorded angles ranged between 76° and –36°, i.e., the indicator never showed that a customer was turned away from the bar. Thus, by recoding the variable into *known* and *unknown*, we created a very lenient version of the *Body to bar* indicator. Entering both variables in the analysis allowed us to distinguish whether a stricter metric as applied by the social state estimator for the *Body to bar* variable or a more lenient coding as in the recoded *Body orientation* variable was more appropriate. Similarly, we recoded the *Face position* indicator's values into *known* and *unknown*. This indicator was only active if the customer's face was directed toward the bar and if it was within the observable area in front of the camera.

By grouping the state of the indicators into *No response* and *Response* updates, the data in **Table 1** shows a summarized history of the trials. The ghosts acknowledged customers in the *Response* turns whom they have not acknowledged in the preceding *No Response* turns. Thus by identifying how these groups differ, we can understand which indicator changes were crucial to the ghosts to initiate a customer–staff interaction. Most of the indicators were highly interdependent, e.g., the body orientation could only be measured if the customer was visible to the system. Thus, we designed a multinomial regression model using the *nnet* package (Ripley and Venables, 2014) of R development core team (2007). The binary variable distinguishing between *Response* and *No response* was used as the dependent variable and the variables coding the state of the indicators (see **Table 1**) served as predictors (independent variables). Thus, the regression used the state of all indicators to distinguish whether an update was part of the history (*No response*) or whether it triggered an acknowledgment (*Response*). The predictor variables were excluded from the regression model if the more parsimonious model did not differ statistically significantly from the full model. Thus, only the set of predictors that could distinguish most effectively between a *No response* and a *Response* turn would remain in the model, i.e., the indicators that had the greatest influence in the ghosts' decision.

TABLE 1 | State of the indicators of the addressed customer as a function of whether the ghosts acknowledged the new customer (Response) or not (No response).

Indicator	State	No response		Response	
		Number	Percent	Number	Percent
Visible	Unknown	48	41%	4	3%
	False	2	2%	2	2%
	True	67	57%	147	95%
Close to bar	Unknown	48	41%	4	3%
	False	42	36%	52	34%
	True	27	23%	97	63%
Location	Unknown	48	41%	4	3%
	Known	69	59%	149	97%
Body orientation	Unknown	58	50%	19	12%
	Known	59	50%	134	88%
Face orientation	Unknown	117	100%	153	100%
	Known	0	0%	0	0%
Body to bar	Unknown	48	41%	4	3%
	False	56	48%	57	37%
	True	13	11%	92	60%
Face to bar	Unknown	48	41%	4	3%
	False	69	59%	149	97%
	True	0	0%	0	0%
Seeks attention	Unknown	117	100%	153	100%
	Known	0	0%	0	0%
Face position	Unknown	51	44%	11	7%
	Known	66	56%	142	93%
Speech	Said nothing	117	100%	153	100%
	Said something	0	0%	0	0%

The data in **Table 1** show that the customers never said anything, i.e., the ghosts always acknowledged the customer before s/he said something. Thus, the speech was excluded as triggering the ghosts' response and did not enter the regression model. After excluding all predictors but the *Close to bar*, *Body to bar*, and the *Face position* indicators, the multinomial model had a *Cox and Snell* $R^2 = 0.334$ compared to $R^2 = 0.335$ of the full model. Excluding the *Face position* resulted in a statistically significant difference. But the model based on the *Close to bar* and the *Body to bar* indicators explained almost as much of the variance $R^2 = 0.321$ as the model including these three variables[2]. We concluded that the *Close to bar* and the *Body to bar* indicators had the greatest impact on the ghosts' decisions in the intention recognition trials.

Turn Duration and Dwell Time

The user interface measured the time span between an update and the corresponding response of the ghost, i.e., the time required to complete a turn (see **Table 2**). The data reflect a comparison of 153 acknowledgments (*Response*) and 117 intermediate updates (*No response*). If a trial included several intermediate turns, their duration and dwell times were averaged before entering the analysis. Thus, 69 intermediate updates contributed to the turn duration. Three trials (one *No response* and three *Response* turns) were excluded from the analysis of the dwell times because the eye tracker was unable to record any data. The dwell times were determined by mapping the point of gaze and duration provided by the faceLAB software to the components of the display. The dwell time on the control panel is possibly underestimated due to its position at the bottom center of the screen. First, the noise of the eye tracker could have resulted in falsely detecting gazes at lower parts of the panel as outside the screen. Secondly, glasses are more likely to reflect the IR illuminator such that the eyes are covered by the reflection if the participant looks straight toward the illuminator below the center bottom region of the screen. However, this design allowed us to position the information panels that we analyzed in more detail with a maximum distance to this area.

The turn duration and dwell times were analyzed with JASP (Love et al., 2014). We report the *BayesFactors* from a Bayesian *t*-test (Rouder et al., 2009; Morey et al., 2014) alongside the respective standard *t*-test statistics. A Cauchy distribution with scale parameter $\frac{1}{\sqrt{2}}$ served as the prior for the effect size (see Rouder et al., 2009). The advantage of using Bayesian *t*-tests is that they also allow researchers to evaluate the amount of evidence for the null hypothesis, which is not possible with standard, frequentist statistical tests. The effect sizes of the standard *t*-tests were computed using G*Power (Faul et al., 2007). The independent samples comparison of the turn durations showed that if the ghosts acknowledged a customer they took statistically significantly longer compared to selecting to wait for the next update [$t(220) = 4.054$, $p < 0.001$, $BF_{10} = 246.6$, $d = 0.57$]. Also, the ghosts dwelled longer on the information panel of the customer whom they addressed [$t(216) = 3.983$, $p < 0.001$, $BF_{10} = 190.2$, $d = 0.56$] and the control panel [$t(216) = 5.033$, $p < 0.001$, $BF_{10} = 12765$, $d = 0.70$] if they acknowledged the new customer. In contrast, the ghosts attended the information of the other customer less if they made an acknowledgment [$t(216) = 7.158$, $p < 0.001$, $BF_{10} = 6.03{*}10^8$, $d = 0.94$].

Dwell Time on Indicators

The data in **Table 3** summarize the ghosts' dwell times on each indicator of the information panel corresponding to the addressed customer. For accommodating the absolute differences in turn duration, we computed the relative dwell time on each indicator by normalizing with the summed dwell time of the respective information panel (see **Table 2**).

We analyzed which indicators received more or less of the ghosts' attention in the *Response*-turns, i.e., the relative dwell times during their decision to acknowledge a new customer. If the ghosts looked randomly at the information panel, we would expect an even distribution of the relative dwell time of 0.1 across the ten indicators. Thus, one-sample *t*-tests and corresponding Bayesian tests were performed against an expected mean of 0.1. There was a statistically significant difference for the *Body to bar*-indicator [$t(149) = 7.061$, $p < 0.001$, $BF_{10} = 1.23{*}10^8$, $d = 0.58$] indicating that the ghosts attended this indicator longer than expected. The *Face orientation* [$t(149) = 22.466$, $p < 0.001$, $BF_{10} = 2.20{*}10^{46}$, $d = 1.81$] and the *Speech* indicators [$t(149) = 17.076$, $p < 0.001$, $BF_{10} = 4.58{*}10^{33}$, $d = 1.39$] were avoided compared to a random gaze pattern. There was no statistical difference for all other indicators [all $t(149) < 2.0$, all $p > 0.05$] and the *BayesFactor* indicated their relative dwell times were equal to a random distribution (all $BF_{10} < 0.3$).

Responses

The ghosts acknowledged their customers by selecting a response from the control panel (see **Figure 3**). The options that the ghosts selected in 153 trials are summarized in **Table 4**.

In the vast majority of cases, the ghosts selected to look at their customer and in one third of the cases made a happy face. Only one quarter of the responses included a verbal utterance. This was either a greeting (e.g., "Hello?"), a prompt to place an order (e.g., "What would you like?"), or both.

[2]The other combinations showed the following results: *Face position* and *Body to bar* $R^2 = 0.303$ with no statistically significant difference to the single predictor *Body to bar* $R^2 = 0.296$, *Face position* and *Close to bar* $R^2 = 0.243$ with no statistically significant difference to the single predictor *Close to bar* $R^2 = 0.244$.

TABLE 2 | Average turn duration, dwell time on the information and control panels as a function of whether the ghost acknowledged a new customer (Response) or not (No response).

Time	No response		Response	
	Time in ms	*SD in ms*	*Time in ms*	*SD in ms*
Turn duration	12728	6540	18105	10095
Dwell time addressed customer	2179	1988	3774	3017
Dwell time other customer	2304	1870	828	1145
Dwell time on control panel	2089	1741	4609	3956

TABLE 3 | Mean dwell times for each indicator of the addressed customer as a function of whether the ghosts acknowledged a new customer (Response) or not (No response).

Indicator	No response				Response			
	Dwell time in ms	*SD in ms*	*Relative dwell time*	*SD in pp*	*Dwell in ms*	*SD in ms*	*Relative dwell time*	*SD in pp*
Visible	328	513	15.5%	17.9	399	678	11.2%	17.3
Close to bar	367	545	16.4%	16.9	485	775	11.2%	12.7
Location	181	268	7.3%	9.2	329	543	10.8%	15.0
Body orientation	160	246	6.1%	8.3	290	407	9.8%	14.4
Face orientation	68	90	4.0%	5.8	100	174	2.4%	4.2
Body to bar	401	491	18.2%	14.5	816	983	20.7%	18.5
Face to bar	205	301	9.8%	10.2	395	609	9.1%	12.1
Seeks attention	191	384	9.2%	12.1	386	560	10.7%	12.9
Face position	239	343	12.6%	15.2	520	887	12.3%	17.2
Speech	16	45	0.9%	2.1	56	109	2.0%	5.8

TABLE 4 | Number and percent of the selected actions for acknowledging a new customer.

Action	Number	Percent
Say something	40	26%
Greeting	25	63%
Prompt to order	19	48%
Looking at something	142	93%
At customer	136	96%
At bottles	4	3%
At hands	2	1%
Make head gesture	4	3%
Nodding	4	100%
Shaking	0	0%
Make facial expression	59	39%
Happy	58	98%
Sad	1	2%
Serve a drink	0	0%

Speech Recognition Trials

In total 174 speech recognition trials were presented to 29 participants. In two trials, the ghost did not serve a drink and the trial was terminated after the pre-recorded customer data ran out. These trials were excluded from all further analyses. In sum 172 drinks were served (one per valid trial) and 553 intermediate updates and their corresponding *No serving*-responses (turns) were recorded. They were distributed unevenly across the trials: $M = 3.2$, $SD = 5.9$, $Mdn = 1.0$, $Max = 38$, $Min = 0$. In 12 trials the ghosts served a drink in their first response. Thus, *No serving*-responses occurred in 160 trials. In 98 trials one *No serving*-response occurred and another 37 trials included three *No serving*-responses.

Recognizer Data

The recognizer data of the addressed customer in the speech recognition trials are summarized in **Table 5**. These recognizer updates were either followed by the ghost serving a drink (*Serving*) or the ghost decided to continue the interaction without a serving (*No serving*), e.g., by asking a question. Please note that the *Face orientation* and *Face to bar indicators* did not work as a result of a failure to record the data during the evaluation. The variable *Body orientation* was recoded into *known* and *unknown* as in the intention recognition trials. The range of the recorded angles was between 22° and −59° and was smaller compared to the intention recognition trials.

The data in **Table 5** compare the state of all indicators when the ghosts served a drink to an average of earlier updates during the course of their interaction. This comparison can identify which change in the available information made the ghosts serve a drink. The data show that the customers were almost always detected as seeking attention, their face position was known and a speech utterance was recognized when the ghosts served a drink. The majority of customers was close to the bar. But the data also suggest that customers were less likely to be served if they were visible. In order to determine which of the indicators influenced the ghosts' decision to serve a drink (*No serving* vs. *Serving*), we designed a multinomial regression model using the state of all indicators as predictors and eliminated them if the more parsimonious model did not differ significantly from the full model. This regression model aimed at identifying the indicators that can distinguish most effectively between an update that occurred at some point in the interaction and the update that triggered the ghosts to serve a drink. After removing all predictors but the *Body orientation* and the *Speech* indicators, the multinomial model had a *Cox and Snell* $R^2 = 0.266$ compared to $R^2 = 0.269$ of the full model. Removing the *Body orientation* indicator resulted in a statistically significant difference, but the loss of explained variance was about one percent $R^2 = 0.256$. We concluded that the customer's speech had the greatest impact on the ghost's decision to serve a drink.

The customer's speech was presented together with the confidence level of the ASR. We compared the confidence levels of the customers' orders ($N_{total} = 232$, $M_{total} = 49.43$, $SD_{total} = 30.03$, $Mdn_{total} = 42.00$) when the ghosts served a drink ($N_{serving} = 154$, $M_{serving} = 59.73$, $SD_{serving} = 28.54$, $Mdn_{serving} = 73.00$) and when they did not ($N_{noserving} = 78$,

TABLE 5 | State of the indicators of the addressed customer as a function of whether the ghosts served a drink.

Indicator	State	No serving		Serving	
		Number	Percent	Number	Percent
Visible	Unknown	1	0%	1	1%
	False	95	17%	54	31%
	True	457	83%	117	68%
Close to bar	Unknown	1	0%	1	1%
	False	80	14%	32	19%
	True	472	85%	139	81%
Location	Unknown	1	0%	1	1%
	Known	552	100%	171	99%
Body orientation	Unknown	8	1%	5	3%
	Known	545	99%	167	97%
Face orientation	Unknown	553	100%	172	100%
	Known	0	0%	0	0%
Body to bar	Unknown	1	0%	1	1%
	False	250	45%	85	49%
	True	302	55%	86	50%
Face to bar	Unknown	1	0%	1	1%
	False	552	100%	171	99%
	True	0	0%	0	0%
Seeks attention	Unknown	1	0%	1	1%
	False	14	3%	5	3%
	True	538	97%	166	97%
Face position	Unknown	1	0%	1	1%
	Known	552	100%	171	99%
Speech	Said nothing	375	68%	18	10%
	Greeting	100	18%	0	0%
	Order	78	14%	154	90%

$M_{noserving} = 29.09$, $SD_{noserving} = 21.35$, $Mdn_{noserving} = 24.00$). This reflects a comparison of the 78 orders without a serving and the 154 servings in the bottom row of **Table 5**. The independent samples test revealed a statistically significant difference [$t(230) = 8.366$, $p < 0.001$, $BF_{10} = 1.01^{*}10^{12}$, $d = 1.02$] indicating that the confidence level was higher when the ghosts served a drink compared to when they did not.

Turn Duration and Dwell Time

The turn durations (time between update presented on screen and response) are presented in **Table 6**. The data reflect a comparison of 172 servings (*Serving*) and 553 intermediate updates (*No Serving*). If a trial included several intermediate turns, the duration and dwell times were averaged before entering the analysis such that 160 data points contributed to *No serving*-data.

TABLE 6 | Average turn duration, dwell time on the information and control panels as a function of whether the ghost served a drink.

Time	No serving		Serving	
	Time in ms	*SD in ms*	*Time in ms*	*SD in ms*
Turn duration	25374	15632	25250	16809
Dwell time addressed customer	3083	1561	2623	2664
Dwell time other customer	1540	2066	791	1094
Dwell time on control panel	7081	6149	8754	7496

The turn duration and dwell times were analyzed as above. The independent samples comparison of the turn durations showed that there was no statistically significant difference between servings and intermediate updates [$t(330) = 0.069$, $p = 0.945$, $BF_{10} = 0.087$]. Also, there was no such difference in the dwell time on the information panel of the addressed customer [$t(330) = 1.599$, $p = 0.111$, $BF_{10} = 0.302$]. However, the ghosts dwelled statistically significantly less on the information panel of the other customer if they served a drink [$t(330) = 4.167$, $p < 0.001$, $BF_{10} = 350.9$, $d = 0.45$]. There was a tendency indicating that the ghosts attended the control panel longer if they served a drink [$t(330) = 2.214$, $p = 0.028$, $BF_{10} = 0.941$, $d = 0.24$]. The *t*-test indicated a statistically significant difference. But the *BayesFactor* did not and the effect size was comparably small. Thus, we do not consider this difference as significant.

Dwell Time on Indicators

The eye tracking data were analyzed as in the intention recognition trials. The data in **Table 7** reflect the information panel of the addressed customer. The data and analyses below refer to the average of 172 servings and intermediate updates in 160 trials.

The relative dwell times of the *Serving*-turns were analyzed as above using a one-sample *t*-test against a mean value of 0.1 across the ten indicators. The ghosts attended the indicators *Visibility* [$t(171) = 18.791$, $p < 0.001$, $BF_{10} = 1.36^{*}10^{40}$, $d = 1.43$], *Close to bar* [$t(171) = 5.642$, $p < 0.001$, $BF_{10} = 1.22^{*}10^{5}$, $d = 0.43$], *Body orientation* [$t(171) = 3.939$, $p < 0.001$, $BF_{10} = 97.379$, $d = 0.30$], *Face orientation* [$t(171) = 11.832$, $p < 0.001$, $BF_{10} = 7.34^{*}10^{20}$, $d = 0.90$], and *Face to bar* [$t(171) = 9.081$, $p < 0.001$, $BF_{10} = 2.06^{*}10^{13}$, $d = 0.68$] statistically significantly less than expected by a random distribution. The relative dwelling times on the indicators for *Location* [$t(171) = 1.646$, $p = 0.102$, $BF_{10} = 0.230$], *Body to bar* [$t(171) = 0.683$, $p = 0.495$, $BF_{10} = 0.076$] and *Seeking attention* [$t(171) = 1.374$, $p = 0.171$, $BF_{10} = 0.154$] did not differ from a random distribution. In contrast, the indicators for the *Face position* [$t(171) = 4.982$, $p < 0.001$, $BF_{10} = 6113.811$, $d = 0.38$] and the *Speech* [$t(171) = 7.497$, $p < 0.001$, $BF_{10} = 1.88^{*}10^{9}$, $d = 0.57$] received more attendance than at random. It should be noted that the face coordinates of *Customer 2* were closely located to the control panel. Their distance was the shortest on the entire screen. Thus, if the ghosts dwelled on the serving options of the control panel a misattribution of the point of gaze could occur between the panel and the *Face position* of *Customer 2* but not of *Customer 1*. The difference in the relative dwell times of this indicator of *Customer 1* [$M = 16.7\%$, $SD = 26.8$pp] and *Customer 2* [$M = 36.9\%$, $SD = 34.2$pp] during the *Serving*-turn supported this assumption. Thus, we repeated the one-sample analysis in Servings to *Customer 1* only [$t(131) = 2.844$, $p = 0.005$, $BF_{10} = 3.392$, $d = 0.25$]. After excluding the potentially misattributed points of gaze, the effect size was smaller but the

TABLE 7 | Mean dwell times for each indicator of the addressed customer as a function of whether the ghost served a drink.

Indicator	No serving				Serving			
	Dwell time in ms	*SD in ms*	*Relative dwell time*	*SD in pp*	*Dwell in ms*	*SD in ms*	*Relative dwell time*	*SD in pp*
Visible	161	354	4.5%	7.7	131	344	2.7%	5.1
Close to bar	161	375	4.4%	6.7	198	465	5.3%	11.0
Location	189	268	9.2%	15.5	166	315	7.9%	16.7
Body orientation	167	254	5.2%	6.1	111	233	5.4%	15.3
Face orientation	57	120	2.1%	5.4	53	157	2.1%	8.8
Body to bar	481	591	16.7%	17.5	331	581	10.8%	15.0
Face to bar	130	215	4.2%	8.4	116	299	3.5%	9.5
Seeks attention	428	517	16.1%	17.2	316	316	11.7%	16.0
Face position	608	974	17.3%	20.5	560	980	21.3%	29.9
Speech	700	1093	20.3%	26.6	641	807	29.4%	33.9

result was still compatible with our initial analysis indicating that the ghosts dwelled longer on the *Face position* indicator than expected with a random distribution.

The greater number of intermediate turns in the speech recognition trials allowed us to address whether the ghosts' attention changed in terms of relative dwell times during the course of the trials. We compared the relative dwell times of the information panel of the addressed customer in the *No serving* and *Serving*-turns. An independent samples test showed that the relative dwell time on the *Speech* was larger in the *Serving*-turns [$t(330) = 2.714$, $p = 0.007$, $BF_{10} = 3.082$, $d = 0.29$]. The relative dwell times reduced for the *Body to bar* indicator [$t(330) = 3.329$, $p < 0.001$, $BF_{10} = 18.173$, $d = 0.36$]. This tendency was also found in the *Visibility* [$t(330) = 2.482$, $p = 0.014$, $BF_{10} = 1.726$, $d = 0.26$] and *Seeks attention* indicators [$t(330) = 2.426$, $p = 0.016$, $BF_{10} = 1.512$, $d = 0.28$]. In these cases, the *t*-test showed a statistically significant difference but the *BayesFactor* was not conclusive. There was no statistical difference for all other indicators [all $t(330) < 2.0$, all $p > 0.05$] and the *BayesFactor* provided evidence in favor of the relative dwell times being equal in *Serving* and *No serving*-turns (all $BF_{10} < 0.3$).

Responses

The ghosts responded to their customers by selecting a response from the control panel (see **Figure 3**). The data in **Table 8** summarize the responses to 553 intermediate updates (*No serving*) and 172 servings.

The ghosts made the robot look at the customer in the majority of their responses. In particular, when serving a drink almost all ghosts selected that the robot should look at the customer. About half the responses were accompanied by speech during the interaction. These utterances were mainly prompting the customers either to place an order (e.g., "What would you like?") or asking the customers to repeat their order using one out of two strategies. First, the ghosts just asked their customer to repeat their utterance (e.g., "Could you say this again?"). Secondly, they repeated the name of the drink that the ASR presented as the most likely guess (e.g., "A water for you?"). Both strategies were used in about half of the cases (see **Table 9**). The ghosts used similar utterances when serving a drink. Either they said something friendly to confirm that the order is about to be served (e.g., "Here you are.") or they included the name of the drink in their utterance (e.g., "Here is your water."). Both options were used in about half the cases (see **Table 9**). The servings were also

TABLE 8 | Number and percent of the selected actions during the interaction and accompanying the serving of a drink.

Response	No serving		Serving	
	Number	*Percent*	*Number*	*Percent*
Say something	290	52%	121	70%
Greeting	48	17%	0	0%
Prompt to order	118	41%	0	0%
Prompt to repeat	123	42%	0	0%
Confirming serving	17	6%	116	96%
Looking at something	395	71%	161	94%
At customer	390	99%	156	97%
At bottles	4	1%	4	2%
At hands	1	0%	1	1%
Make head gesture	59	11%	77	45%
Nodding	49	83%	77	100%
Shaking	10	17%	0	0%
Make facial expression	209	38%	121	70%
Happy	199	95%	119	98%
Sad	10	5%	2	2%
Total	553		172	

TABLE 9 | Number and ratio of echo questions and statements in prompts to repeat the order and confirmations to serve the drink as a function of whether the ghosts served a drink.

Response	No serving		Serving	
	Number	*Percent*	*Number*	*Percent*
Prompt to repeat	123		0	
Echo question	*56*	*46%*	*0*	*0%*
Confirming serving	17		116	
Echo statement	*2*	*12%*	*52*	*45%*

accompanied by more expressive face and head movements. Two thirds of the servings included a happy face compared to only one third of the intermediate responses. Also, almost half the servings included a nodding but only 10% of the intermediate responses was accompanied by a head gesture at all.

The data in **Table 8** suggest that in 17 trials the serving of a drink was verbally confirmed but not actually served. Thus, we inspected these cases more closely. In eight cases, the drink was served in the next turn, i.e., not immediately after confirming the order but at the next opportunity. The other nine trials involved an utterance that is ambiguous if used without punctuation marks ("Bitteschön"). In **Table 8**, this was categorized as a polite German confirmation ("Bitteschön." [Here you are.]). This would imply that the ghosts have forgotten to serve the drink. However, in the seven cases that did not repeat the name of the drink, it could also invite to place an order ("Bitteschön?" [What can I do for you?]). This implies that the ghosts ignored the customer's utterance and used an expression for inviting to place an order out of the blue. We cannot decide whether one of these interpretations was intended by the ghosts. But the closer analysis showed that the ghosts rarely used a verbal confirmation to serve a drink without actually serving it. In all cases, the trial continued until the ghost served a drink. It should be noted that the customers had to repeat their orders several times with the real robot. Thus, if the ghosts did not use the next turn for serving a drink, they served it with another drink order later in the trial.

DISCUSSION

Most of the ghosts reported that they experienced GiM as very immersive and experienced a turn-by-turn role-game. They invested great efforts into establishing a social interaction with their customers despite the fact that their behavior was pre-recorded and displayed in a number of indicators. First, the number of trials and the time spent illustrates the ghosts' efforts. The majority of trials involved three or less turns but the ghosts used up to 38 turns if necessary. The variance in the number of turns illustrates that they adapted to each customer in order to entertain a socially credible interaction. Secondly, the human ghosts were more efficient on the same data than the actual robot. Since each trial of the pre-recorded data had a maximum number of turns defined by the original robot–customer interaction, the ghosts would have been unable to complete the trial if they required more data than the robot. This occurred only three times compared to the 325 trials that entered our analyses. During the evaluation, the robot had a real-time interaction with its customers such that it could ask questions and elicit a direct response. In contrast, the ghosts communicated with pre-recorded customers whom could not respond to, e.g., a clarification question. Thus, the ghosts used their social knowledge to outperform the robotic bartender, e.g., the ghosts' responses indicated that they interpreted their customers' responses in the context of their own questions and utterances which were not present at the time of recording. Thirdly, the results of the intention recognition trials are compatible with findings from observations in the real world and experiments using natural stimuli. Thus, we conclude that the ghosts made credible efforts and that the results reflect human social behavior that can reveal strategies for improving human–robot interactions. We discuss the results in more detail starting with the intention recognition trials and secondly, the speech recognition trials.

Intention Recognition Trials

The results of the intention recognition trials showed that ghosts relied on the *Close to bar* and *Body to bar* indicators for identifying new customers. This finding replicates the results of an experiment using natural videos and snapshots from real bars where the participants detected that customers bid for the attention of bar staff if they were close to the bar and their body and/or head was directed to the bar (Loth et al., 2013). This behavior was also observed at ticket counters in Amsterdam Centraal station. Similar to the bartending robot, a member of staff sits at a fixed position behind the counter and waits for customers. The interactions were initiated if a customer approached the counter and looked at the assistant (Brouwer et al., 1979; Clark, 2012). As in our results, the distance to the counter and head/body direction were essential in this setting. That means that the interactions were initiated by the placement (Clark, 2003) of the customer's body. More specifically, this was described as asking a wordless question (Clark, 2012). Furthermore, implementing this strategy for detecting customers with the intention to place an order produced more reliable and more stable results than other classifiers (Foster, 2014). Thus, the social signal for initiating an interaction is formed by these two components. The results of this GiM study supported that finding and demonstrated that we can obtain reliable and valid results with this paradigm (also see Loth et al., 2014).

The ghosts' detection strategy relied on only two recognizer modalities (distance to bar and body orientation) whereas other modalities were not relevant including the customer's speech. However, this finding could be attributed to the customer's speech being (a) irrelevant, or (b) relevant, but there was no speech detected during the data recording and speaking coincided with other cues in the natural data experiments. The design of this GiM study enabled us to distinguish between these possibilities. First, in the natural data experiment the participants had to judge whether a particular snapshot showed a customer bidding for attention. In contrast, the control panel of the GiM interface offered the ghosts to wait for another update that may include additional cues such as a speech utterance. Thus, the ghosts decided when they responded to a new customer. However, the ghosts never waited for a speech utterance. Secondly, the eye tracking data allowed us to identify which recognizer modality was attended by the ghosts. They dwelled on the *Speech* indicator less than expected with a random gaze pattern. Rather, they focussed on the information about the customer's pose and position, especially the binary indicator *Body to bar*. It could be argued that the ghosts did not gaze at the *Speech* indicator because speech was not displayed and thus, they looked at something else. But this was not the case. The *Seeks attention* indicator would have provided a

straight forward hint for the ghosts but it was disabled. Hence, the *Seeks attention* and *Speech* indicators equally showed no information. However, the ghosts dwelled on the potentially relevant *Seeks attention* indicator about 10% of their relative dwell time but only 2% on the *Speech* indicator (see **Table 3**). Thus, the ghosts deliberately ignored the *Speech* indicator whereas there was no clear pattern of ignoring or focussing on the *Seeks attention* indicator. Together, this provided converging evidence that modalities other than the distance to the bar and the head/body orientation were not relevant for detecting the intention to place an order, and generally to initiate an interaction.

Using this strategy indicates that the ghosts subconsciously accessed their knowledge of initiating an interaction and specifically scanned the panels for the expected social signal. In turn, they could ignore most of the recognizer data without risking to ignore a customer. However, it appears counter-intuitive to ignore information since there was no time pressure that could have hindered the ghosts from scanning the entire display. This could be attributed to the fact that the human cognitive resources are limited in general (Broadbent, 1969) and in particular within one sensory modality such as vision (Allport et al., 1972; Mcleod, 1977). Thus, the ghosts used their social knowledge for limiting the information that they attended to a few relevant indicators. For example, the GiM interface included two indicators for the customer's body orientation. The arrow shaped indicator *Body orientation* provided an analog display of recognizer data and was larger than the *Body to bar* indicator which depicted a binary value computed by the social state estimator. Despite the fact that the binary indicator was smaller on the display, the ghosts attended and relied on this to a greater extent compared to the analog version. First, one of the indicators was sufficient and thus, the ghosts limited their attention to one of them. Secondly, the ghosts consistently selected the binary indicator. One of the differences between the two indicators is the required effort for using the information. For interpreting the arrow indicator the ghosts would have to evaluate the angle of the customer's body orientation themselves whereas the binary indicator was simpler and provided this interpretation.

The ghosts not only ignored redundant information and relied on the most convenient display, they also ignored irrelevant data. The results showed that they almost exclusively focussed on the customer's distance to the bar and their body orientation. This pattern was not an artifact of our GiM design. For example, the participants in the natural data experiment only analyzed the body posture of customer who were close to the bar but not of other customers (Loth et al., 2013). A similar focus on task-relevant aspects was observed in intentional blindness in the visual (Simons and Chabris, 1999) and auditory domain (Dalton and Fraenkel, 2012). Thus, focussing on those aspects that are relevant for detecting an expected social signal reflects general cognitive processes in social interactions. Identifying these strategies is crucial for human-robot interaction as it allows to discard possibly misleading data, e.g., a speech utterance from another customer. Using these social strategies saves computational effort, improves the robot's reliability and makes its performance more predictable by being more human-like.

The GiM paradigm also allows the manipulation of very specific pieces of information, e.g., for investigating the relevance of a particular modality and for eliciting recovery strategies in sensor failures. The customer's face data were not recorded during the robot evaluation resulting in an apparent sensor failure. Thus, the indicator *Face orientation* never worked and the binary *Face to bar* indicator either indicated that the face was not detected or that it did not look toward the bar. Thus, attending and using this information could have misled the ghosts. They could have assumed that the customer looked away from the bar and has not intended to interact with them. However, the ghosts did acknowledge their new customers. Thus, we concluded that the ghosts recognized that the face related information was unreliable, discarded this information and recovered from that sensor failure by relying on data about the customer's body instead, specifically the *Body to bar* indicator. These results do not allow us to decide whether the head or body orientation took priority if both sensors operated as desired. However, a deliberate manipulation can reveal repair strategies if sensors fail and thereby, provide insight into the structure and redundancies in human social signals. In this experiment, the available information was sufficient to the ghosts to identify and serve customers. Thus, a robot could rely on the body orientation only and would not require a high resolution camera and face tracking. For example, a mobile robot could save on energy by using cameras and trackers only when needed.

In addition to understanding the user's behavior, the GiM paradigm allows us to determine which actions constitute a socially appropriate response. In the intention recognition trials, the ghosts had to communicate that they have noticed the customers and are ready to take their drink orders. Almost all ghosts decided to look at their customers, i.e., they visibly shifted the robot's attention to the customer. This reflects the first part of a visual handshake. The customer can accept this invitation and complete the visual handshake by looking at the (robotic) bartender. The first part of offering a visual handshake and the second part of accepting it form an adjacency pair (Schegloff, 1968; Schegloff and Sacks, 1973) in a non-verbal modality. If completed, the handshake ensures that both sides are ready to begin a verbal communication. Argyle and Dean (1965) argued that mutual eye contact signals to both sides that the channel of verbal communication is open. Furthermore, establishing eye contact puts some pressure on the assistant to respond to the customer who has caught their eye (Goffman, 1963, p. 94). Vice-versa, avoiding eye contact is an effective method of avoiding a conversation in the first place (Goffman, 1963). However, looking at the customers could also be attributed to a visual inspection of the scene. But if the ghosts decided to look at something it was coherently the customer (96% of cases, see **Table 4**). Additionally, the dwell times provided evidence in favor of an intended action. First, the time spent on the control panel doubled if the ghosts acknowledged a customer. Secondly, the dwell times doubled on the addressed customer and reduced to one third for the other customer just before the ghosts initiated the handshake.

Thirdly, 40% of the responses included a happy face that was directed toward the customer. This indicates that the ghosts invested additional efforts in a meaningful action rather than a casual visual inspection. Finally, the ghosts rarely selected actions other than a visual handshake. Only 19 times (12% of cases) a customer was prompted to place an order and only four times (3% of cases) a nodding head gesture was selected. In sum, a socially appropriate response to a new customer is to smilingly offer a visual handshake. The customer is then free to accept it by looking at the (robotic) bartender or to ignore it. This is very effective and at the same time less annoying than (repeatedly) inviting customers to place an order. Furthermore, this finding resembles observations in natural scenes and strengthens our conclusion that the GiM paradigm provides reliable insights. Thus, a robotic agent should employ this simple, effective but not annoying socially appropriate signal.

Speech Recognition Trials

The speech recognition trials posed a greater challenge to the ghosts than intention recognition as evidenced by more and longer turns as well as longer dwell times on the panels (see **Tables 2** and **6**). We attributed this to the difficulty of interacting with pre-recorded customers and eliciting their orders. The pre-recorded nature of the customers also included the risk that the customers appear ignorant to the ghosts' actions, specifically if they asked questions. However, the ghosts were as efficient as or better than the real robot and managed to serve a drink in 172 out of 174 interactions. This shows that (a) the ghosts performed well under challenging conditions, and (b) their responses can reveal useful strategies that improve interactions with service robots.

The analysis of the recognizer updates and the eye tracking data showed that once the interaction was initiated, the attention focus shifted from physical properties to the customer's *Speech* (see **Figure 4**). For example, *Body to bar* was the most attended indicator in the intention recognition trials. In the intermediate speech recognition turns, its relative dwell time was reduced and reduced further during the *Serving*-turns such that it was not different from a random gaze pattern. At the same time, the dwell times on the *Speech* indicator increased. Thus, the closer the ghosts were to serving a drink, the more they shifted their attention away from physical properties in the visual sensory modality toward the customer's speech in the verbal modality.

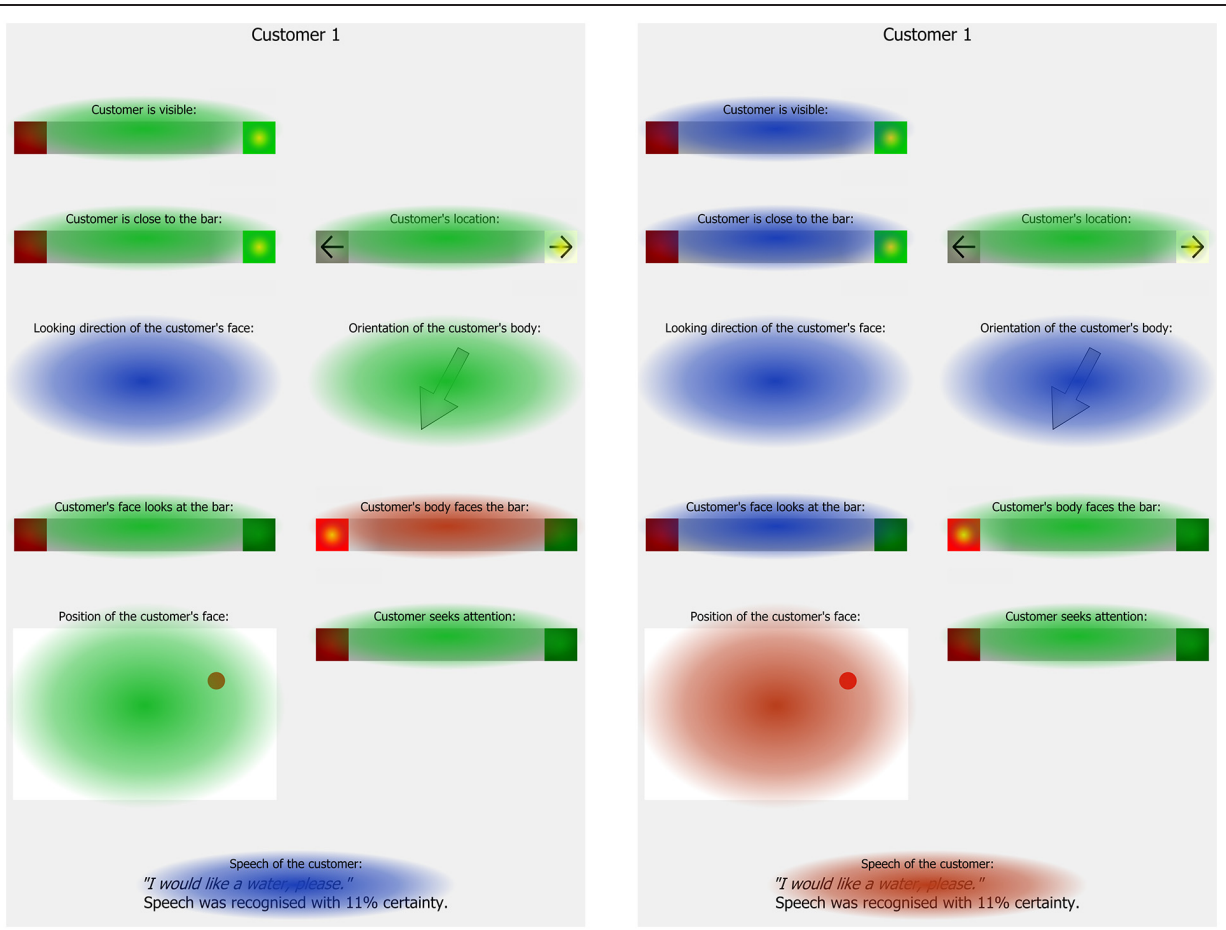


FIGURE 4 | Comparison of relative dwell times on each indicator in the Response-turns in the intention recognition trials (left hand side) and the Serving-turns in the speech recognition trials (right hand side). The color denotes whether the indicator attended less than (blue), equal to (green), or more than (red) expected by a random gaze distribution.

As a result the customer's speech was the single most attended indicator (see **Table 7**). As in the intention recognition trials, the ghosts subconsciously identified the most relevant modality from their social knowledge. In case of the orders, the social signal is essentially verbal and thus, the ghosts reduced their cognitive load by focussing on the *Speech*. The ghosts further reduced their load by focussing on the customer whom they would serve and spending significantly less time on the other customer especially when serving a drink (see **Table 6**). This adds converging evidence to our conclusion that the ghosts specifically scan for the expected social signals and thereby reduce their cognitive load.

Focussing on the socially relevant modality not only reduces the workload, it also prevents mistakes, e.g., abrupt terminations of the interaction. In one fifth of the servings the recognizers suggested that the customer was not close to the bar, in one third of the servings s/he was not visible, and her/his body was not oriented to the bar in half of the servings. In these cases, the ghosts would not have acknowledged a new customer but yet they served them a drink. In contrast, the robotic bartender at the evaluation assumed that customers must be visible and did not serve a drink. Instead, the robot terminated the interaction, waited until the customer was visible again and treated her/him as a new customer. Thus, the ghosts achieved a greater efficiency than the robot by continuing their interaction and serving the drink. We attribute this to the fact that the ghosts expected some closing to their conversation (Schegloff and Sacks, 1973), e.g., saying "Thank you", rather than a sudden disappearance of the customer. Thus, the ghosts accepted the order even if the misleading data suggested the customer was (temporarily) not visible to the recognizers. In conclusion, a robot cannot expect to detect the customers as bidding for attention throughout the interaction. For example, if the customer moves or leans onto the bar, the recognizers can fail temporarily. However, the robot can expect some closing to the interaction and should not abruptly terminate the interaction as in the evaluation (Foster et al., 2012) and in a direction giving robot (Bohus et al., 2014). The robot still maintains its ability to detect whether the customer has actually left, e.g., if there is no speech input and the recognizers cannot detect the customer. Although it may be counter-intuitive to discard data, a smart weighting and ignoring some data can improve the robustness of a robotic agent and prevent abrupt terminations. In addition to improving the robot, focussing on the socially relevant modalities reflects cognitive principles in social interactions.

The ghosts strongly focussed on the *Speech* indicator. But another comparably large share of the relative dwell time was spent on the *Face position* indicator (see **Table 7**). Our analysis showed that this was partly due to the fact that the selector for serving a drink was spatially very close to the *Face position* of *Customer 2*. After accommodating for this confound, the ghosts attended the *Face position* reliably more than expected if their gaze randomly distributed across the information panel. It could be argued that the ghosts tried to establish eye contact to the customer by looking at a dot that depicted the customer's face. This could be attributed to the fact that maintaining some level eye contact is important in a conversation because markedly looking away could signal that one is not an interested recipient (Schegloff and Sacks, 1973; Goodwin, 2000). However, maintaining eye contact would have been reasonable throughout the interaction and, importantly, whether or not the ghosts observed the dot was not visible to the customers. Thus, we cannot identify how the ghosts have particularly benefitted from the *Face position* indicator immediately before serving a drink (*Serving*-turns).

In the speech recognition trials, the ghosts predominantly tried to elicit which drink the customers ordered using verbal utterances in particular if the customer's verbal utterance was unclear or recognized with a low confidence level. That means that the ghosts responded verbally to a verbal customer request. In contrast, the ghosts acknowledged a new customer by changing physical properties of the robot in the intention recognition trials, e.g., they manipulated the robot's looking direction, but they did not speak. Thus, the ghosts preferred to respond in the same modality that was used by the customer. That means that the ghosts responded non-verbally to non-verbal actions and verbally to verbal actions. There was only one exception from this rule. If the ghosts served a drink, they responded with a physical action to a verbal request. However, this action was often accompanied by a verbal utterance (70% of the cases) and the customer specifically asked the bartender to serve a drink. In sum, the ghosts showed a strong preference to respond to a request within the same modality. Thus, a robotic agent should copy this human preference unless the user asked for a specific action.

The analysis of the ghosts' responses after the customer placed an order revealed two strategies that contributed to their greater efficiency compared to the robotic bartender. As the robotic bartender, the ghosts decided in accordance with the confidence level of the ASR whether to serve the drink or to ask for a clarification. But their threshold for servings ($M_{serving} = 59\%$, $M_{clarification} = 29\%$) was lower than the 80% of the robotic bartender (Foster et al., 2012). Thus, firstly this threshold should be lowered to about 50% in order to serve the drink quicker. Secondly, the ghosts used echo questions in about half of their 123 clarification questions, i.e., they repeated the most likely guess of the ASR as a question (e.g., "A coke for you?"). A typical response would be a short confirmation (e.g., "Yes, please.") or a correction (e.g., "No, I have ordered a juice."). This strategy is particularly useful if the ASR has low confidence levels because the next challenge is to correctly identify the customer's reply. Corrections tend to be delayed, prefaced, qualified and/or mitigated by an apology or an indirect form (Schegloff et al., 1977; Heritage, 1984). Thus, detecting whether the customer responded affirmative or with a correction could be achieved by simply analysing the length of the customer's response. In this study, we used pre-recorded customer data. Thus, the customers could not respond to an echo question. But the data included the responses to the robot's repeated prompts for an order (i.e., "What would you like?"). In turn, the ghosts perceived that their customers responded by repeating or slightly reformulating the original order with repeatedly low ASR confidence levels. Since the next turn after a question is typically perceived as a response (Schegloff, 1972; Sacks et al., 1974), a repetition such

as "A coke, please. Confidence level 15%." was perceived as more meaningful in the context of an echo question. Thus, the ghosts accepted the repetitions as positive answer and served the drink. Effectively, the ghosts retrospectively loaded their customer's answer with an additional social meaning (Clark, 2012). But this strategy increased the redundancy in the interaction by repeating what has been said. However, this did not delay the serving but speeded the interaction and offered the customer to detect and correct communication errors. With a similar effect, the ghosts repeated the name of the drink in about half the servings (e.g., "Here is your coke.") This allows the customer to silently accept this or to correct the robot in the last minute while the actions are already in preparation. In sum, the analysis of the confidence levels of ASR showed that about 50% is sufficient as a threshold. Furthermore, clarification requests and utterances accompanying (robot) actions such as servings should be formulated as echo questions or statements. Introducing redundancy by echoing essential information is socially appropriate and helps in achieving a smooth interaction, especially if the ASR confidence levels are low or in a noisy environment.

In addition to verbal utterances, the ghosts selected to look at the customer in most of their responses. This was the case if they asked for clarification, served a drink and even if they selected no other action. They maintained visual contact despite the fact that the control panel was reset after each update and required the ghosts to explicitly select this option in each response. Thus, we concluded that this option was important. Since the ghosts initiated the interaction by establishing visual contact to their customers, removing it could be interpreted as ending the interaction (Schegloff and Sacks, 1973; Goodwin, 2000). The ghosts almost constantly selected to look at their customers, but other options such as smiling and nodding were used more restrictedly. In particular, the ghosts nodded and smiled when confirming an order. Thus, they used these actions as an additional signal to confirm that the order was understood. As a result, confirming a drink order was a highly multimodal signal comprising of facial expressions, head gestures, verbal utterances and the serving itself. In this rich signal, the head gestures and facial expressions are redundant from a task-oriented perspective. However, they served the social purpose of clearly marking the serving to their customers. In sum, the ghosts maintained visual contact to their customers throughout the interaction. In contrast, nodding and smiling were used more restrictedly to confirm that the order was understood.

CONCLUSION

The GiM paradigm is a reliable method for understanding the social behavior of users and the responses that they expect in a human–robot interaction setting. We demonstrated that results obtained with the GiM paradigm replicate findings that were obtained in the analyses of natural scenes, video recordings of natural scenes and in experiments using natural stimuli (Goffman, 1963; Argyle and Dean, 1965; Brouwer et al., 1979; Goodwin, 2000; Clark, 2012; Loth et al., 2013, 2015). In addition to experimenting with natural stimuli, the GiM paradigm allowed us to separately identify each single aspect of the scene (represented by a recognizer modality) that the ghost participants dedicated their overt attention to and its impact on their actions.

Our results showed that our ghost participants focussed on a small number of socially relevant modalities and ignored other, potentially misleading data. We argued that this is due to the ghosts scanning for particular social signals for recognizing the user's intention and a general limitation of their cognitive resources (Broadbent, 1969; Allport et al., 1972; Mcleod, 1977). We also found that ignoring other data is advantageous as it hinders being distracted by misleading information that can lead to e.g., abrupt terminations of the interaction (Bohus et al., 2014). Thus, we demonstrated how fundamental principles of human cognition operate in social settings and also showed how a robotic agent can be improved by incorporating these principles.

Our study investigated two aspects of the bar setting: initiating the interaction, and ordering and serving the drink. The relevance of the modalities shifted as ghosts expected different social signals at each stage from their prior knowledge. When the customer tried to get the attention of the robotic bartender, her/his position and pose were most important. In contrast, the verbal modality was the most important for orders and servings. Thus, we have to identify which social signals are expected at each stage of an interaction and adapt the robotic policies to attend the relevant modalities. Furthermore, our findings showed that the ghosts preferred to respond in the same modality that the customer has used, i.e., changing the robot's pose if the user signaled to them through their pose and position, and speaking if the user spoke to the robot. Thus, a multimodal grammar has to incorporate: (a) a method for focussing on the expected social signals and the relevant modalities, (b) keeping track of changes in expected signals and modalities, and (c) a preference to respond in the same modality as the user's signal.

This GiM study revealed communication strategies that are simple, effective and socially appropriate. In acknowledgments, we showed that the robot should offer a visual handshake to the customers by looking at them and inviting them to join the interaction by looking at the robot, rather than annoy them by repeatedly inviting them to place an order. During the interaction, our ghost participants created redundancy by echoing salient parts of the customer's utterance such as the drink order. Even though redundancy implies longer and more turns, this socially appropriate strategy required fewer turns and fewer clarification questions (cf. Giuliani et al., 2013) especially in cases involving inconclusive recognizer data. In sum, we found simple strategies for a smoother human–robot interaction that can enhance the robot's multimodal grammar.

The ghost participants enjoyed the game-like interface of our GiM software and invested efforts and time into building a socially appropriate interaction with their customers. Thus, we concluded that our results reflect reliable, replicable insights in human social behavior and cognitive principles. Our initial study delivered substantial improvements for human–robot interaction policies by making the robot's performance more

robust, human-like and in turn, more predictable and enjoyable to its users. In our study, we used pre-recorded user data. But the GiM paradigm can be advanced into a real-time research tool in order to investigate the entire interaction. Furthermore, specific pieces of information or modalities can be manipulated in order to elicit repair and compensation strategies. The GiM interface can be adapted to various settings and its game-like experience makes it an ideal research tool for deriving multimodal grammars including strategies for recovering from inconclusive sensor data. Thus, the GiM paradigm is an effective, simple, and highly versatile method for understanding human social behavior that has the potential to revolutionize the field of social robotics.

FUNDING

This research was funded by the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 270435.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2015.01641

REFERENCES

Abelson, R. P. (1981). Psychological status of the script concept. *Am. Psychol.* 36, 715–729.

Abernethy, B., Maxwell, J. P., Jackson, R. C., and Masters, R. S. W. (2007). "Skill in sport," in *Handbook of Applied Cognition*, eds F. T. Durso, R. S. Nickerson, S. T. Dumais, S. Lewandowsky, and T. J. Perfect (Chichester: John Wiley & Sons Ltd.), 333–359.

Admoni, H., Dragan, A., Srinivasa, S. S., and Scassellati, B. (2014). "Deliberate delays during robot-to-human handovers improve compliance with gaze communication," in *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction* (Bielefeld: ACM Press), 49–56. doi: 10.1145/2559636.2559682

Allport, D. A., Antonis, B., and Reynolds, P. (1972). On the division of attention: adisproof of the single channel hypothesis. *Q. J. Exp. Psychol.* 24, 225–235. doi: 10.1080/00335557243000102

Argyle, M., and Dean, J. (1965). Eye-Contact, distance and affiliation. *Sociometry* 28, 289–304. doi: 10.2307/2786027

Baltzakis, H., Pateraki, M., and Trahanias, P. (2012). Visual tracking of hands, faces and facial features of multiple persons. *Mach. Vis. Appl.* 23, 1141–1157. doi: 10.1007/s00138-012-0409-5

Bohus, D., and Horvitz, E. (2009a). "Dialog in the open world: platform and applications," in *Proceedings of the 2009 International Conference on Multimodal Interfaces (ICMI-MLMI)* (Cambridge, MA: ACM Press), 31. doi: 10.1145/1647314.1647323

Bohus, D., and Horvitz, E. (2009b). "Learning to predict engagement with a spoken dialog system in open-world settings," in *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (London: Association for Computational Linguistics), 244–252.

Bohus, D., and Horvitz, E. (2009c). "Models for multiparty engagement in open-world dialog," in *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (London: Association for Computational Linguistics), 225–234.

Bohus, D., and Horvitz, E. (2009d). "Open-world dialog: challenges, directions, and prototype," in *Proceedings of the IJCAI'2009 Workshop on Knowledge and Reasoning in Practical Dialogue Systems,* Pasadena, CA.

Bohus, D., and Horvitz, E. (2010). "On the challenges and opportunities of physically situated dialog," in *Proceedings of the AAAI Fall Symposium Series*, Arlington, VA.

Bohus, D., and Horvitz, E. (2011). "Multiparty turn taking in situated dialog: study, lessons, and directions," in *Proceedings of the SIGDIAL 2011 Conference* (Portland, OR: Association for Computational Linguistics), 98–109.

Bohus, D., Saw, C. W., and Horvitz, E. (2014). "Directions robot: in-the-wild experiences and lessons learned," in *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems* (Paris: International Foundation for Autonomous Agent and Multiagent Systems).

Breazeal, C., DePalma, N., Orkin, J., Chernova, S., and Jung, M. (2013). Crowdsourcing human-robot interaction: new methods and system evaluation in a public environment. *J. Human-Robot Interact.* 2, 82–111. doi: 10.5898/JHRI.2.1.Breazeal

Broadbent, D. E. (1969). *Perception and Communication*. Oxford: Pergamon Press.

Brouwer, D., Gerritsen, M., and De Haan, D. (1979). Speech differences between women and men on the wrong track? *Lang. Soc.* 8, 33–50. doi: 10.1017/S0047404500005935

Clark, H. H. (2003). "Pointing and placing," in *Pointing?: Where Language, Culture, and Cognition Meet*, eds S. Kita (Mahwah, NJ: L. Erlbaum Associates), 243–268.

Clark, H. H. (2012). "Wordless questions, wordless answers," in *Questions: Formal, Functional and Interactional Perspectives*, ed. J. P. De Ruiter (Cambridge: Cambridge University Press), 81–102.

Dahlbäck, N., Jönsson, A., and Ahrenberg, L. (1993). "Wizard of Oz studies," in *IUI '93 Proceedings of the 1st international conference on Intelligent User Interfaces* (New York, NY: ACM Press), 193–200. doi: 10.1145/169891.169968

Dalton, P., and Fraenkel, N. (2012). Gorillas we have missed: sustained inattentional deafness for dynamic events. *Cognition* 124, 367–372. doi: 10.1016/j.cognition.2012.05.012

De Ruiter, J. P., and Cummins, C. (2012). "A model of intentional communication: AIRBUS (Asymmetric Intention Recognition with Bayesian Updating of Signals)," *Presented at the Semantics and Pragmatics of Dialogue (SemDial)*, Paris.

faceLAB Eye Tracker (2009). *(Version 5). Tucson*. Arizona: Seeing Machines Inc.

Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 175–191. doi: 10.3758/BF03193146

Foster, M. E. (2014). "Validating attention classifiers for multi-party human-robot interaction," in *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction: Workshop on Attention Models in Robotics* (Bielefeld: ACM Press).

Foster, M. E., Gaschler, A., and Giuliani, M. (2013). "How can I help you? Comparing engagement classification strategies for a robot bartender," in *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI 2013)* (Sydney: ACM Press), 255–262. doi: 10.1145/2522848.2522879

Foster, M. E., Gaschler, A., Giuliani, M., Isard, A., Pateraki, M., and Petrick, R. P. A. (2012). "Two people walk into a bar: dynamic multi-party social interaction with a robot agent," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction (ICMI 2012)* (Santa Monica, CA: ACM Press). doi: 10.1145/2388676.2388680

Fraser, N. M., and Gilbert, G. N. (1991). Simulating speech systems. *Comput. Speech Lang.* 5, 81–99. doi: 10.1016/0885-2308(91)90019-M

Gaschler, A., Huth, K., Giuliani, M., Kessler, I., De Ruiter, J. P., and Knoll, A. (2012). "Modelling state of interaction from head poses for social human-robot interaction," in *Proceedings of the Gaze in Human- Robot Interaction Workshop held at the 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2012)*, Boston.

Giuliani, M., Petrick, R. P. A., Foster, M. E., Gaschler, A., Isard, A., Pateraki, M., et al. (2013). "Comparing task-based and socially intelligent behaviour in a robot bartender," in *Proceedings of the 15th ACM on International Conference on Multimodal Interaction* (Sydney: ACM Press), 263–270. doi: 10.1145/2522848.2522869

Goffman, E. (1963). *Behaviour in Public Places*. Galt, ON: Collier-Macmillan

Goodrich, M. A., and Schultz, A. C. (2007). Human-robot interaction: a survey. *Found. Trends Hum. Comput. Inter.* 1, 203–275. doi: 10.1561/1100000005

Goodwin, C. (2000). Action and embodiment within situated human interaction. *J. Pragmat.* 32, 1489–1522. doi: 10.1016/S0378-2166(99)00096-X

Gray, J., Breazeal, C., Berlin, M., Brooks, A., and Lieberman, J. (2005). *Action Parsing and Goal Inference Using Self as Simulator*. Cambridge, MA: IEEE, 202–209. doi: 10.1109/ROMAN.2005.1513780

Green, A., Hüttenrauch, H., and Eklundh, K. S. (2004). "Applying the wizard-of-Oz framework to cooperative service discovery and configuration," in *RO-MAN 2004: 13th IEEE International Workshop on Robot and Human Interactive Communication: Proceedings: September 20-22, 2004* (Kurashiki: IEEE).

Grice, H. P. (1957). Meaning. *Philos. Rev.* 66:377. doi: 10.2307/2182440

Hall, E. T. (1969). *The Hidden Dimension: An Anthropologist Examines Humans' Use of Space in Public and Private*. New York, NY: Anchor Books, Doubleday & Company Inc.

Heritage, J. (1984). *Garfinkel and Ethnomethodology*. New York, N.Y: Polity Press.

Holroyd, A., Rich, C., Sidner, C. L., and Ponsler, B. (2011). "Generating connection events for human-robot collaboration," in *Proceedings of the 20th IEEE International Symposium on Robot and Human Interactive Communication* (Atlanta, GA: IEEE), 241–246. doi: 10.1109/ROMAN.2011.6005245

Java Runtime Environment (2012). *(Version 7). 500 Oracle Parkway*. Redwood Shores, CA: Oracle Corporation. Available at: http://www.java.com

Jeannerod, M. (2006). "Representations for actions," in *Motor Cognition: What Actions Tell the Self,* eds M. D'Esposito, J. Driver, T. Robbins, D. Schacter, A. Treismann, and L. Weiskranz (Oxford, New York: Oxford University Press), 1–21.

Kelley, J. F. (1984). An iterative design methodology for user-friendly natural language office information applications. *ACM Trans. Inf. Syst.* 2, 26–41. doi: 10.1145/357417.357420

Krämer, N., Kopp, S., Becker-Asano, C., and Sommer, N. (2013). Smile and the world will smile with you—The effects of a virtual agent"s smile on users" evaluation and behavior. *Int. J. Hum. Comput. Stud.* 71, 335–349. doi: 10.1016/j.ijhcs.2012.09.006

Lakens, D., and Stel, M. (2011). If they move in sync, they must feel in sync: movement synchrony leads to attributions of rapport and entitativity. *Soc. Cogn.* 29, 1–14. doi: 10.1521/soco.2011.29.1.1

Lee, M. K., Forlizzi, J., Rybski, P. E., Crabbe, F., Chung, W., Finkle, J., et al. (2009). "The snackbot: documenting the design of a robot for long-term human-robot interaction," in *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction* (La Jolla, CA: ACM Press), 7–14.

Levinson, S. C. (1995). "Interactional biases in human thinking," in *Social Intelligence and Interaction: Expressions and Implications of the Social Bias in Human Intelligence*, ed. E. N. Goody (New York, NY: Cambridge University Press), 221–260.

Lichtenthäler, C., Peters, A., Griffiths, S., and Kirsch, A. (2013). "Social navigation - identifying robot navigation patterns in a path crossing scenario," in *Social Robotics*, Vol. 8239, eds G. Herrmann, M. J. Pearson, A. Lenz, P. Bremner, A. Spiers, and U. Leonards (Cham: Springer International Publishing), 84–93.

Liu, X., Rieser, V., and Lemon, O. (2009). "A wizard-of-oz interface to study information presentation strategies for spoken dialogue systems," in *Proceedings of the First International Workshop on Spoken Dialogue Systems (IWSDS)*, Irsee.

Loth, S., Giuliani, M., and De Ruiter, J. P. (2014). "Ghost-in-the-machine: initial results," in *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, Bielefeld, 234–235. doi: 10.1145/2559636.2563696

Loth, S., Huth, K., and De Ruiter, J. P. (2013). Automatic detection of service initiation signals used in bars. *Front. Psychol.* 4:557. doi: 10.3389/fpsyg.2013.00557

Loth, S., Huth, K., and De Ruiter, J. P. (2015). "Seeking attention: testing a model of initiating service Interactions," in *A Multidisciplinary Approach to Service Encounters*, Vol. 14, eds M. de la O. Hernández-López and L. Fernández Amaya (Amsterdam: Brill), 229–247.

Love, J., Selker, R., Verhagen, J., Smira, M., Wild, A., Marsman, M., et al. (2014). *JASP (Version 0.5)*. Amsterdam: JASP. Available at: https://jasp-stats.org/

Mack, A., and Rock, I. (1998). *Inattentional Blindness*. Cambridge, MA: MIT Press.

Mcleod, P. (1977). A dual task response modality effect: support for multiprocessor models of attention. *Q. J. Exp. Psychol.* 29, 651–667. doi: 10.1080/14640747708400639

Michalowski, M. P., Sabanovic, S., and Simmons, R. (2006). "A spatial model of engagement for a social robot," in *Proceedings of the 9th IEEE International Workshop on Advanced Motion Control* (Istanbul: IEEE), 762–767. doi: 10.1109/AMC.2006.1631755

Morey, R. D., Rouder, J. N., and Jamil, T. (2014). *Package "BayesFactor" (Version 0.9.9) [R]*. Groningen, NL: Rijksuniversiteit Groningen. Available at: http://bayesfactorpcl.r-forge.r-project.org/

Néda, Z., Ravasz, E., Brechet, Y., Vicsek, T., and Barabási, A.-L. (2000). Self-organising processes: the sound of many hands clapping. *Nature* 403, 849–850. doi: 10.1038/35002660

Noël, B., Furley, P., van der Kamp, J., Dicks, M., and Memmert, D. (2014). The development of a method for identifying penalty kick strategies in association football. *J. Sports Sci.* 33, 1–10. doi: 10.1080/02640414.2014.926383

Orkin, J., and Roy, D. (2007). The restaurant game: learning social behavior and language from thousands of players online. *J. Game Dev.* 3, 39–60.

Orkin, J., and Roy, D. (2009). "Automatic learning and generation of social behaviour from collective human gameplay," in *Proceedings of the 8th International Conference on Autonomous Agents and Multimagent Systems?: May 10-15, 2009* (Budapest: International Foundation for Autonomous Agent and Multiagent Systems).

Pateraki, M., Baltzakis, H., and Trahanias, P. (2014). Visual estimation of pointed targets for robot guidance via fusion of face pose and hand orientation. *Comput. Vis. Image Understand.* 120, 1–13. doi: 10.1016/j.cviu.2013.12.006

Petrick, R. P. A., and Foster, M. E. (2012). "What would you like to drink? Recognising and planning with social states in a robot bartender domain," in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence* (Toronto: AAAI Press), 69–76.

Petrick, R. P. A., and Foster, M. E. (2013). "Plan-based social interaction with a robot bartender," in *Proceedings of the ICAPS 2013 Application Showcase*, eds N. Policella and N. Onder (Rome: ICAPS), 10–13.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image Vis. Comput.* 28, 976–990. doi: 10.1016/j.imavis.2009.11.014

R development core team (2007). *R: A Language And Environment For Statistical Computing (Version 2.12.0)*. Wien: R Foundation for Statistical Computing.

Rich, C., Ponsler, B., Holroyd, A., and Sidner, C. L. (2010). "Recognizing engagement in human-robot interaction," in *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction* (Osaka: ACM Press), 375–382. doi: 10.1145/1734454.1734580

Richardson, M. J., Marsh, K. L., Isenhower, R. W., Goodman, J. R. L., and Schmidt, R. C. (2007). Rocking together: dynamics of intentional and unintentional interpersonal coordination. *Hum. Move. Sci.* 26, 867–891. doi: 10.1016/j.humov.2007.07.002

Riek, L. (2012). Wizard of Oz studies in hri: a systematic review and new reporting guidelines. *J. Hum. Robot Inter.* 1, 119–136. doi: 10.5898/JHRI.1.1.Riek

Rieser, V., Keizer, S., Liu, X., and Lemon, O. (2011). "Adaptive information presentation for spoken dialogue systems: evaluation with human subjects," in *Proceedings of the 13th European Workshop on Natural Language Generation* (Nancy: Association for Computational Linguistics), 102–109.

Rieser, V., and Lemon, O. (2009). Learning human multimodal dialogue strategies. *Nat. Lang. Eng.* 16, 3–23. doi: 10.1017/S1351324909005099

Ripley, B., and Venables, W. (2014). *Package "nnet" (Version 7.3-8) [R]*. Oxford, Oxforshire: University of Oxford. Avilable at: http://www.stats.ox.ac.uk/pub/MASS4/

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.* 16, 225–237. doi: 10.3758/PBR.16.2.225

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language* 50, 696–735. doi: 10.2307/412243

Schank, R. C., and Abelson, R. P. (1977). *Scripts, Plans, Goals and Understanding: An Inquiry Into Human Knowledge Structures*. Hillsdale, NJ: L. Erlbaum.

Schegloff, E. A. (1968). Sequencing in conversational openings. *Am. Anthropol.* 70, 1075–1095. doi: 10.1525/aa.1968.70.6.02a00030

Schegloff, E. A. (1972). "Notes on a conversational practice: formulating place," in *Studies in Social Interaction*, ed. D. Sudnow (New York, NY: The Free Press and Collier-Macmillan Limited), 75–119.

Schegloff, E. A., Jefferson, G., and Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language* 53, 361–382. doi: 10.2307/413107

Schegloff, E. A., and Sacks, H. (1973). Opening up closings. *Semiotica* 8, 289–327. doi: 10.1515/semi.1973.8.4.289

Schorer, J., Rienhoff, R., Fischer, L., and Baker, J. (2013). Foveal and peripheral fields of vision influences perceptual skill in anticipating opponents' attacking position in volleyball. *Appl. Psychophysiol. Biofeedback* 38, 185–192. doi: 10.1007/s10484-013-9224-7

Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., et al. (2013). Real-time human pose recognition in parts from single depth images. *Commun. ACM* 56, 116–124. doi: 10.1145/2398356.2398381

Sidner, C. L., and Lee, C. (2003). "Engagement rules for human-robot collaborative interactions," in *IEEE International Conference on Systems, Man and Cybernetics*, Vol. 4 (Washington, DC: IEEE), 3957–3962. doi: 10.1109/ICSMC.2003.1244506

Sidner, C. L., Lee, C., Kidd, C. D., Lesh, N., and Rich, C. (2005). Explorations in engagement for humans and robots. *Artif. Intell.* 166, 140–164. doi: 10.1016/j.artint.2005.03.005

Simons, D. J., and Chabris, C. F. (1999). Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception* 28, 1059–1074. doi: 10.1068/p2952

Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., et al. (2012). Bridging the gap between social animal and unsocial machine: a survey of social signal processing. *IEEE Trans. Affect. Comput.* 3, 69–87. doi: 10.1109/T-AFFC.2011.27

von Ahn, L., and Dabbish, L. (2008). Designing games with a purpose. *Commun. ACM* 51, 58–67. doi: 10.1145/1378704.1378719

Xu, Y., Ohmoto, Y., Ueda, K., Komatsu, T., Okadome, T., Kamei, K., et al. (2010). Active adaptation in human-agent collaborative interaction. *J. Intell. Inf. Syst.* 37, 23–38. doi: 10.1007/s10844-010-0135-2

Yousuf, A. M., Kobayashi, Y., Yamazaki, A., and Yamazaki, K. (2012). "Development of a mobile museum guide robot that can configure spatial formation with visitors," in *Intelligent Computing Technology Berlin, Heidelberg*, Vol. 7389, eds D.-S. Huang, C. Jiang, V. Bevilacqua, and J. C. Figueroa (Berlin: Springer), 432–432.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 24 July 2015
doi: 10.3389/fpsyg.2015.01049

Using gaze patterns to predict task intent in collaboration

*Chien-Ming Huang *, Sean Andrist, Allison Sauppé and Bilge Mutlu*

Department of Computer Sciences, University of Wisconsin–Madison, Madison, WI, USA

In everyday interactions, humans naturally exhibit behavioral cues, such as gaze and head movements, that signal their intentions while interpreting the behavioral cues of others to predict their intentions. Such intention prediction enables each partner to adapt their behaviors to the intent of others, serving a critical role in joint action where parties work together to achieve a common goal. Among behavioral cues, eye gaze is particularly important in understanding a person's attention and intention. In this work, we seek to quantify how gaze patterns may indicate a person's intention. Our investigation was contextualized in a dyadic sandwich-making scenario in which a "worker" prepared a sandwich by adding ingredients requested by a "customer." In this context, we investigated the extent to which the customers' gaze cues serve as predictors of which ingredients they intend to request. Predictive features were derived to represent characteristics of the customers' gaze patterns. We developed a support vector machine-based (SVM-based) model that achieved 76% accuracy in predicting the customers' intended requests based solely on gaze features. Moreover, the predictor made correct predictions approximately 1.8 s before the spoken request from the customer. We further analyzed several episodes of interactions from our data to develop a deeper understanding of the scenarios where our predictor succeeded and failed in making correct predictions. These analyses revealed additional gaze patterns that may be leveraged to improve intention prediction. This work highlights gaze cues as a significant resource for understanding human intentions and informs the design of real-time recognizers of user intention for intelligent systems, such as assistive robots and ubiquitous devices, that may enable more complex capabilities and improved user experience.

Keywords: intention, eye gaze, support vector machine, gaze patterns, intention prediction

OPEN ACCESS

Edited by:
Sebastian Loth,
Universität Bielefeld, Germany

Reviewed by:
Dimitri Ognibene,
King's College London, UK
Oskar Palinko,
Istituto Italiano di Tecnologia, Italy

****Correspondence:***
Chien-Ming Huang,
Department of Computer Sciences,
University of Wisconsin–Madison,
1210 West Dayton Street, Madison,
WI 53706, USA
cmhuang@cs.wisc.edu

Specialty section:
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology

Received: *23 March 2015*
Accepted: *09 July 2015*
Published: *24 July 2015*

Citation:
Huang C-M, Andrist S, Sauppé A and Mutlu B (2015) Using gaze patterns to predict task intent in collaboration. Front. Psychol. 6:1049. doi: 10.3389/fpsyg.2015.01049

1. Introduction

In daily interactions, humans frequently engage in *joint action*—a collaborative process that involves parties working together to coordinate attention, communication, and actions to achieve a common goal (Clark, 1996; Sebanz et al., 2006). For example, movers carrying a large piece of furniture, an instructor training students in a chemistry lab, or a server taking customer orders at a deli counter must coordinate their behaviors with one another. To achieve successful joint action, people monitor each others' actions and task progress, predict each others' intentions, and adjust their own actions accordingly (Sebanz and Knoblich, 2009). Such action monitoring and intention prediction are integral to the establishment of common ground between parties engaged in joint action. As a result, parties consciously and subconsciously exhibit

behavioral cues, such as eye gaze and gestures, to manifest intentions for others to read while interpreting others' behavioral cues to understand their intention, thereby facilitating joint action. These behavioral cues are a gateway to understanding a person's mental states, including attention, intentions, and goals. Moreover, increasing evidence from neuroscience and developmental psychology has shown that action monitoring allows people to use their behavior repertoire and motor system to predict and understand others' actions and intentions (Blakemore and Decety, 2001; Buccino et al., 2001; Rizzolatti and Craighero, 2004).

Among other behaviors, gaze cues are particularly informative in the manifestation of mental states. Deictic gaze toward an object, for instance, may signal the person's interest in the object and has been found to be temporally coupled with the corresponding speech reference to the object (Meyer et al., 1998; Griffin, 2001). Moreover, people use gaze cues to draw others' attention toward an intended object in the environment in order to establish perceptual common ground (Sebanz et al., 2006). The ability to understand and follow such cues is critical for sharing mental states in an interaction (Butterworth, 1991). Gaze cues may also signal planned actions; empirical evidence has shown that gaze cues indicate action intent and lead motor actions that follow (Land et al., 1999; Johansson et al., 2001).

While prior research has highlighted the link between gaze cues and intention, the current work aims to develop a model quantifying how patterns of gaze cues may characterize and even predict intentions. To this end, we collected data of dyadic interactions in which a "customer" and a "worker" engaged in a sandwich-making task and analyzed how the customers' gaze patterns indicated their intentions, which we characterized as the ingredients they chose. Conceptually, this interaction can be characterized as involving three processes: (1) the customer looks at possible ingredients to make a decision about which ingredient to request (Hayhoe and Ballard, 2014); (2) the customer signals their decision via behavioral cues (Pezzulo et al., 2013); and (3) the worker observes the customer's gaze behaviors to predict their intentions (Doshi and Trivedi, 2009; Ognibene and Demiris, 2013; Ognibene et al., 2013). Our goal is to quantify how much information the customer's gaze provides about their intentions in the first two processes. We built and tested a machine learning model that predicted customer intentions from tracked eye gaze data. Specifically, we developed a support vector machine-based approach that predicted the customers' intention—choice of ingredients—based on their exhibited gaze patterns. The effectiveness of the predictor was evaluated using the collected gaze data. Our model and findings contribute to our understanding of the relationship between gaze cues and intent and to design guidelines for emerging technologies, such as assistive robots and ubiquitous devices, that utilize real-time intention prediction to provide their users with effective and anticipatory assistance.

This paper is organized as follows. Section 2 reviews behavioral signals of human intentions and action monitoring for intention understanding. We present a computational model that quantifies the relationship between gaze cues and human intentions and an evaluation of the effectiveness of the model in Section 3. We discuss our results, potential applications, and limitations of this work in Section 4.

2. Background

In everyday interactions, from carrying furniture to successfully navigating in a crowded space, people engage in an implicit form of coordination (Sebanz et al., 2006). This coordination relies on the successful communication and recognition of intent by the parties involved in the interaction and enables each person to adapt their behavior to accommodate their partner's intentions. While communicating intent can be achieved through a number of behavioral channels (Morris and Desebrock, 1977; White, 1989; Clark and Brennan, 1991; Shibata et al., 1995; Bangerter, 2004), gaze has been identified as crucial in understanding the intentions of others, as the direction of gaze indicates where a person is directing their attention and the actions that they may subsequently perform (Baron-Cohen et al., 2001; Meltzoff and Brooks, 2001). Below, we review research into how humans develop an understanding of intent in themselves and others and utilize gaze cues to communicate intent.

2.1. Human Intent

The concept of intentionality is defined as the commitment of a person to executing a particular action (Malle and Knobe, 1997). The formulation of an intent is often driven by the individual's desire to achieve a particular goal (Astington, 1993). This formulation requires a variety of other skills, including forethought and planning, to appropriately fulfill an intention (Bratman, 1987). What differentiates an intent from a desire is this level of planning in preparation to turn the intention into an achievable reality (d'Andrade, 1987).

From an early age, children begin to attribute intent to the actions of others. For example, children at 15 months of age are capable of understanding the intentions of others in physical tasks, even when the goal is not achieved (Meltzoff, 1995). Later, children learn how behaviors are driven by intent (Feinfield et al., 1999), contributing to the development of an ethical system where intentionality is used as a factor to establish the culpability of an individual.

Prior work suggests that, after developing a capacity for understanding intent, humans also develop *Theory of Mind* (ToM)—the ability to attribute mental states to others (Leslie, 1987). The development of ToM enables people to understand that other humans they interact with may have intents that can differ from their own (Leslie, 1987; Blakemore and Decety, 2001). ToM then shapes the way people interact with one another in a way that is most easily observable in physical tasks, such as moving a table together or navigating through a crowd. In these scenarios, humans rely on ToM abilities to attribute intent to other participants and to adapt their own behaviors to accommodate the intent of others, resulting in seamless interactions.

2.2. Communicating Intent via Gaze

While the ability to attribute intent to others is important in joint action, discerning what the intentions of other participants are

with a high degree of reliability can be difficult without some amount of evidence. One approach people subconsciously use to infer the intent of others is by observing their behavioral cues (Blakemore and Decety, 2001). Humans employ a number of behavioral cues, such as gaze and gestures, when working with others on a task (Morris and Desebrock, 1977; White, 1989; Clark and Brennan, 1991; Shibata et al., 1995; Baron-Cohen et al., 2001; Meltzoff and Brooks, 2001; Bangerter, 2004). These cues aid in their partner's understanding of and fluency in the task, enabling their partner to adjust their behavior accordingly to accommodate intended actions (Blakemore and Decety, 2001). While a number of behavioral channels can be used to understand intent, gaze is considered preeminent among them due to the clarity with which it can indicate attention; for instance, partners would assume that an area being gazed toward will be the next space to be acted upon (Baron-Cohen et al., 2001; Meltzoff and Brooks, 2001).

Gaze behavior is crucial to human communication of intent throughout the development of social behavior. During infancy, children can follow the gaze cues of adults, which serve as the basis of joint attention (Butler et al., 2000), and use their own gaze to communicate an object of interest (Morales et al., 1998). Older preverbal children can employ gaze in conjunction with gestures to communicate more concretely (Masur, 1983). The use and understanding of gaze becomes more complex and nuanced with age, allowing humans to better identify targets of joint attention (Heal, 2005). This development of gaze understanding mirrors the development of understanding of intent and ToM discussed above, allowing humans to gradually develop a more complex intuition of others and their intentions.

During an interaction, gaze behavior can indicate one's intent in a variety of ways, such as communicating a future action or an emotional state. During a joint task, awareness of a partner's gaze behavior helps enable effective task coordination between participants (Tomasello, 1995). Prior work by Brennan et al. (2008) used head-mounted eye trackers to examine gaze patterns during a joint search task. Awareness of a partner's gaze behavior was not only sufficient for completing the task, but it also resulted in significantly faster search times than verbal coordination did. Additionally, participants who were aware of their partner's gaze behavior offered more precise help during the task when it was necessary. Adams and Kleck (2005) conducted a controlled laboratory study where participants were presented with photographs of people who were either gazing toward or away from the participant. Results showed that participants' perceptions of the photographed person's emotional state were affected by the person's gaze direction.

Gaze behavior can be used in conjunction with other attributes or behavioral cues to more accurately predict intent. Ordering of gaze fixations has been used to infer the type of visual task a person is performing, such as memorizing a picture vs. counting the number of people photographed in a picture (Haji-Abolhassani and Clark, 2014). Prior work used eye gaze and its associated head movements as input for a sparse Bayesian learning model (McCall et al., 2007) to predict a driver's future actions when operating a motor vehicle (Doshi and Trivedi, 2009). Additionally, work by Yi and Ballard (2009) built a dynamic Bayesian network from a user's gaze and hand movements to predict their task state in real time during a sandwich-building task.

While prior work has examined the connection between gaze and intent in a variety of situations, the current work aims to provide an empirical approach to modeling gaze behavior to predict task intent during collaboration. Specifically, it extends prior work in two ways. First, the current work investigates the relationship between gaze cues and task intent in a collaborative context, whereas prior work employed tasks that involved only one person completing them, e.g., making a sandwich (Yi and Ballard, 2009) or driving a car (Doshi and Trivedi, 2009). Second, the prior predictive models utilized multiple sources of information, while this present work focuses on using gaze cues only. A related problem to the focus of the present work is how to use the predicted intention of others to direct one's own focus (e.g., gaze fixation). For example, Ognibene and Demiris (2013) and Ognibene et al. (2013) utilized people's motions to predict their intentions and used these predictions to control the attention of a robotic observer.

3. Prediction of Human Intentions

In this section, we describe our process for understanding and quantifying the relationship between gaze cues and human intentions. This process includes collecting human interaction data, modeling the characteristics of gaze patterns from our data, and evaluating the effectiveness of the computational model. In addition to the quantitative evaluation, we provide qualitative analyses of the circumstances under which our model succeeds and fails in predicting user intentions.

3.1. Data Collection and Annotation

Our data collection involved pairs of human participants engaged in a collaborative task. We used this study both to collect data for our model as well as to build an intuition as to how joint attention is coordinated through both verbal and non-verbal cues in day-to-day human interactions. During the data collection study, participants performed a sandwich-making task in which they sat across from each other at a table that contained 23 possible sandwich ingredients and two slices of bread. The initial layout of the ingredients was the same for each pair of participants (**Figure 1**). One participant was assigned the role of "customer," and the other was assigned the role of "worker." The customer used verbal instructions to communicate to the worker what ingredients he/she wanted on the sandwich. Upon hearing the request from the customer, the worker immediately picked up that ingredient and placed it on top of the bread.

We recruited 13 dyads of participants for the data collection study. All dyads were recruited from the University of Wisconsin–Madison campus and were previously unacquainted. The protocol for the data collection study was reviewed and approved by the University of Wisconsin–Madison's Education and Social/Behavioral Science Institutional Review Board (IRB). Prior to the experiment, participants completed a written consent of participation. Each dyad carried out the sandwich-making task twice so that each participant acted as both customer and

FIGURE 1 | Data collection of dyadic interactions in a sandwich-making task. Left: Two participants, wearing gaze trackers, working together to make a sandwich. **Middle**: The participant's view of the task space from the gaze tracker. The orange circle indicates their current gaze target. **Right**: The layout of ingredients on the table. The ingredients, from top to bottom, left to right, are *lettuce1, pickle1, tomato2, turkey, roast beef, bacon2, mustard, cheddar cheese, onions, pickle2, ham, mayo, egg, salami, swiss cheese, bologna, bacon1, peanut butter, lettuce2, pickle3, tomato1, ketchup, jelly*.

worker. The customer was instructed to request 15 ingredients for their sandwich. Participants kept their own count of the number of ingredients ordered, stopping when they had reached 15. The customer was further instructed to only request a single ingredient at a time and to refrain from directly pointing to or touching the ingredients. Upon completing the first sandwich, an experimenter entered the study room and reset the ingredients back to their original locations on the table, and the participants switched roles for the second sandwich.

Throughout the data collection study, both participants wore mobile eye-tracking glasses developed by SMI[1]. These eye-trackers perform binocular dark-pupil tracking with a sampling rate of 30 Hz and gaze position accuracy of 0.5°. Each set of glasses contains a forward-facing high-definition (HD) camera that was used to record both audio and video at 24 fps. The gaze trackers were time-synchronized with each other so that the gaze data from both participants could be correlated.

Following data collection, the proprietary BeGaze software created by SMI was used to automatically segment the gaze data into fixations—periods of time when the eyes were at rest on a single target—and saccades—periods of time when the eyes were engaged in rapid movement. Fixations were labeled with the name of the target fixated upon. Possible targets included the sandwich ingredients (**Figure 1**), the slices of bread, the conversational partner, and elsewhere in space. Speech was also transcribed for each participant. Customer requests for specific objects were tagged with the ID of the referenced object.

3.2. Intention Modeling

In this work, we considered the customers' intentions to be their chosen ingredients. Informed by the literature, we hypothesized that the customers' gaze patterns would signify their intent of which ingredients they wanted on their sandwich and aimed to develop a model to accurately predict intentions based on their gaze patterns. Our data collection resulted in a total of 334 episodes of ingredient requests. We excluded episodes where more than 40% of the gaze data was missing before verbal requests, yielding 276 episodes for data analysis and modeling.

[1]http://www.smivision.com/en/gaze-and-eye-tracking-systems/home.html

A naive, but plausible, strategy to predict a person's intent is solely based on his or her current gaze, which may indicate the person's current attention and interest (Frischen et al., 2007). To evaluate the efficacy of this strategy, we built an *attention-based* intention predictor that performed predictions according to which ingredient the customer most recently fixated on. An evaluation of the 276 episodes showed that the attention-based predictor achieved 65.22% accuracy in predicting the customers' choice of ingredient. This strategy outperformed random guesses of the ingredient, which were between 4.35 (i.e., 1/23) and 11.11% (i.e., 1/9), depending on how many potential ingredients were still available at that point in the interaction.

While the attention-based method was reasonably effective in predicting the intended ingredients, it only relied on the most recently glanced-at ingredient and omitted any prior gaze cues. However, the history of gaze cues may provide richer information for understanding and anticipating intent. In particular, we made two observations from the 276 episode analysis. First, participants seemed to glance at the intended ingredient longer than other ingredients. Second, participants glanced multiple times toward the intended ingredient before making the corresponding verbal request. These observations, along with significance of attention, informed our selection of characteristic features, as listed below, to represent patterns of participant's gaze cues. Each of the four features was computed for all potential ingredients in every episode of an ingredient request.

Feature 1: Number of glances toward the ingredient before the verbal request (Integer)

Feature 2: Duration (in milliseconds) of the first glance toward the ingredient before the verbal request (Real value)

Feature 3: Total duration (in milliseconds) of all the glances toward the ingredient before the verbal request (Real value)

Feature 4: Whether or not the ingredient was most recently glanced at (Boolean value)

We applied a support vector machine (SVM) (Cortes and Vapnik, 1995)—a type of supervised machine learning approach that is widely used for classification problems—to classify

the participants' gaze patterns into two categories, one for the intended ingredient (i.e., positive) and the other for the non-intended, competing ingredients (i.e., negative). In this work, we used Radial Basis Function (RBF) Kernels and the implementation of LIBSVM (Chang and Lin, 2011) for the analysis and evaluation reported below.

To evaluate the effectiveness of our model in classifying gaze patterns for user intentions, we conducted a 10-fold cross-validation using the 276 episodes of interaction. For each episode, we calculated a feature vector, including Features 1–4, for each ingredient that the customer looked toward before making a verbal request. To train the SVM, if an ingredient was the requested ingredient, the classification label was set to `1`; otherwise, it was set to `−1`. In the test phase, the trained SVM determined the classification for each ingredient glanced at. On average, the SVMs achieved 89.00% accuracy in classifying labels of customer intention. Feature selection analyses (Chen and Lin, 2006) revealed that Feature 3 was the most indicative in classifying intentions, followed by Feature 4, Feature 1, and then Feature 2.

3.3. Intention Prediction

The SVM classifier was further modified to predict the customers' intentions. The input to our SVM predictor was a stream of gaze fixations. As the interaction unfolded, we maintained a list of candidate ingredients, their corresponding feature vectors, and the estimated probabilities of the ingredient being the intended request, calculated using the method based on Wu et al. (2004). When a new gaze fixation on an ingredient occurred, we first checked whether or not the ingredient was in the candidate list. If the ingredient was already in the list, we updated its feature vector and estimated probability; otherwise, we added a new entry for the ingredient to the list.

A traditional SVM was used to classify an ingredient to be the potential request if the estimated probability was greater than 0.5. If more than one ingredient was classified as a potential request, the traditional SVM predictor picked the ingredient with the highest probability as the final prediction. If, however, none of the ingredients were classified as potential requests, the predictor made no prediction. The effectiveness of such a traditional SVM predictor was assessed via a 10-fold cross-validation using our 276 episodes. For this evaluation, a prediction was considered to be correct only when the prediction matched the actual request. Note that this intention prediction was different from the classification of gaze patterns reported in the previous section. The accuracy of intention prediction was assessed by whether or not the predicted ingredients matched the requested ones, whereas the accuracy of intention classification was based on comparisons of classified labels, including both positive and negative, with actual labels. The traditional SVM predictor on average reached 61.52% accuracy in predicting which ingredients the customer would pick. Further analysis revealed that 28.99% of the time the SVM predictor made no predictions. However, when it made predictions (i.e., 71.01% of the time), the SVM provided predictions at 86.43% accuracy. This accuracy could be interpreted as the confidence of the traditional SVM predictor in predicting intention when it had a positive classification.

We defined an anticipation window as the time period starting with the last change in the prediction and ending with the onset of the speech utterance (see **Figure 2** as an example). This anticipation window allowed us to understand how early the predictor could reach the correct predictions. For the traditional SVM predictor, the anticipation window for the correct predictions was on average 1420.57 ms before the actual verbal request, meaning that the predictor could anticipate the intended ingredient about 1.4 s in advance. The interaction

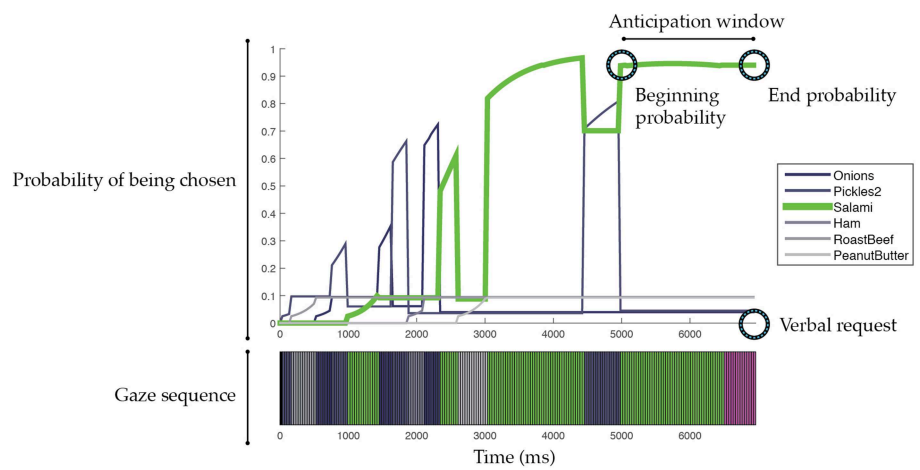


FIGURE 2 | Illustration of episodic prediction analysis. Each illustrated episode ends at the start of the verbal request. The top plot shows probabilities of glanced ingredients that may be chosen by a customer. Note that the plotted probability was with respect to each ingredient. By calculating the normalized probability across all ingredients, we can determine the likelihood of which ingredient will be chosen. The bottom plot shows the customer's gaze sequence. Ingredients are color coded. Purple indicates gazing toward the bread. Black indicates missing gaze data. An anticipation window is defined as the time period starting with the last change in the prediction and ending with the onset of the speech utterance. The beginning and end probabilities are the probabilities of the predicted ingredient at the beginning and end of the anticipation window.

duration before the verbal request for the episodes with correct predictions was on average 3802.56 ms ($SD = 1596.45$).

The predictive accuracy of the traditional SVM predictor was largely impaired by the frequency with which it made no predictions. To address this issue, we ensured that our SVM-based predictor always made a prediction, choosing the ingredient with the highest probability. A 10-fold cross-validation using the 276 episodes showed that our SVM-based predictor on average reached 76.36% predictive accuracy and could make those correct predictions 1831.27 ms ahead of their corresponding verbal requests (Interaction duration $M = 3802.56$, $SD = 1596.45$). **Table 1** summarizes these results. Moreover, we analyzed the probabilities of the chosen ingredients that were at the beginning and end of the anticipation window (see **Figure 2**). On average, the beginning and end probabilities for the correct predictions were 0.36 and 0.75, respectively, whereas the beginning and end probabilities for the incorrect predictions were 0.28 and 0.43, respectively. These probability parameters indicate the confidence of our SVM-based predictor in making a correct prediction. For example, when the probability of an ingredient is over 0.43, the ingredient is likely to be the intended choice. We note that this threshold (0.43) is lower than the threshold used by the traditional SVM (0.50). Similarly, if the probability of an ingredient is lower than 0.36, the ingredient is less likely to be the intended choice. These parameters allow the construction of a real-time intention predictor that anticipates the customers' choices on the fly.

In the next section, we provide examples and further analyses of when our SVM-based predictor made correct and incorrect predictions. These analyses revealed gaze patterns that may provide additional insight into understanding the customers' intentions.

3.4. Qualitative Analysis

To further understand how our intention predictor made correct and incorrect predictions in the collected interaction episodes, we plotted the probability of each glanced-at ingredient over time, aligned with the corresponding gaze sequence received from the gaze tracker, for each interaction episode (see **Figure 2** for an example). These plots facilitated a qualitative analyses of gaze patterns and further revealed patterns that were not captured in our designed features but may signify user intentions. In the following paragraphs, we present our analyses and discuss exemplary cases.

3.4.1. Correct predictions

Two categories—one dominant choice and the trending choice—emerged from the episodes with correct predictions (see examples in **Figure 3**).

TABLE 1 | Summary of our quantitative evaluation of the effectiveness of different intention prediction approaches.

	Predictive accuracy	Anticipation time
Chance	4.35–11.11%	N/A
Attention-based	65.22%	N/A
SVM-based	76.36%	1831 ms

3.4.1.1. One dominant choice

In this category, customers seemed to be focused toward one dominant ingredient, which was apparent in their gaze cues (**Figure 3**, Top). In particular, we found two types of gaze patterns. In the first, participants looked toward the intended ingredient for a prolonged time. In the second, they looked toward the intended ingredient multiple times in the course of their interaction. For both patterns, the intended ingredient received the majority of the gaze attention relative to other ingredients. This dominance allowed the predictor to give correct predictions.

3.4.1.2. Trending choice

In contrast to the previous category, there were situations in which customers did not seem to have a single ingredient in mind. In these situations, the customers exhibited a "shopping" behavior by looking toward multiple ingredients to decide which one to order. These situations usually involved the participants' visual attention being spread across multiple candidate ingredients. However, the customers generally looked toward the intended ingredient recurrently compared to other competing ingredients throughout the interaction. This recurrent pattern resulted in the intended ingredient becoming a trending choice, as illustrated in the bottom examples of **Figure 3**. The SVM-based predictor was observed to capture this pattern effectively.

3.4.2. Incorrect predictions

From the 10-fold evaluation of the SVM-based predictor, there were a total of 62 episodes resulting in incorrect predictions. In the following paragraphs, we describe the characteristics of four identified categories of these incorrect predictions.

3.4.2.1. No intended glances

Among the incorrect predictions, there were 23 episodes (37.10%) during which the customers did not glance at the intended ingredients (**Figure 4**, First row). There are three reasons that might explain these cases. First, the customers had made their decisions in previous episodes. For example, when they were glancing around to pick an ingredient, they may have also decided which ingredient to order next. Second, their intentions were not explicitly manifested through their gaze cues. Third, the gaze tracker did not capture the gaze of the intended ingredient (i.e., missing data). In each of these cases, the predictor could not make correct predictions as it did not have the necessary information about the intended ingredients.

3.4.2.2. Two competing choices

Sometimes, customers seemed to have two ingredients they were deciding between (**Figure 4**, Second row). In this case, their gaze cues were similarly distributed between the competing ingredients. Therefore, gaze cues alone were not adequate to anticipate the customers' intent. We speculate that the determinant factors in these situations were subtle and not well-captured via gaze cues. Therefore, the predictor was likely to make incorrect predictions in these situations.

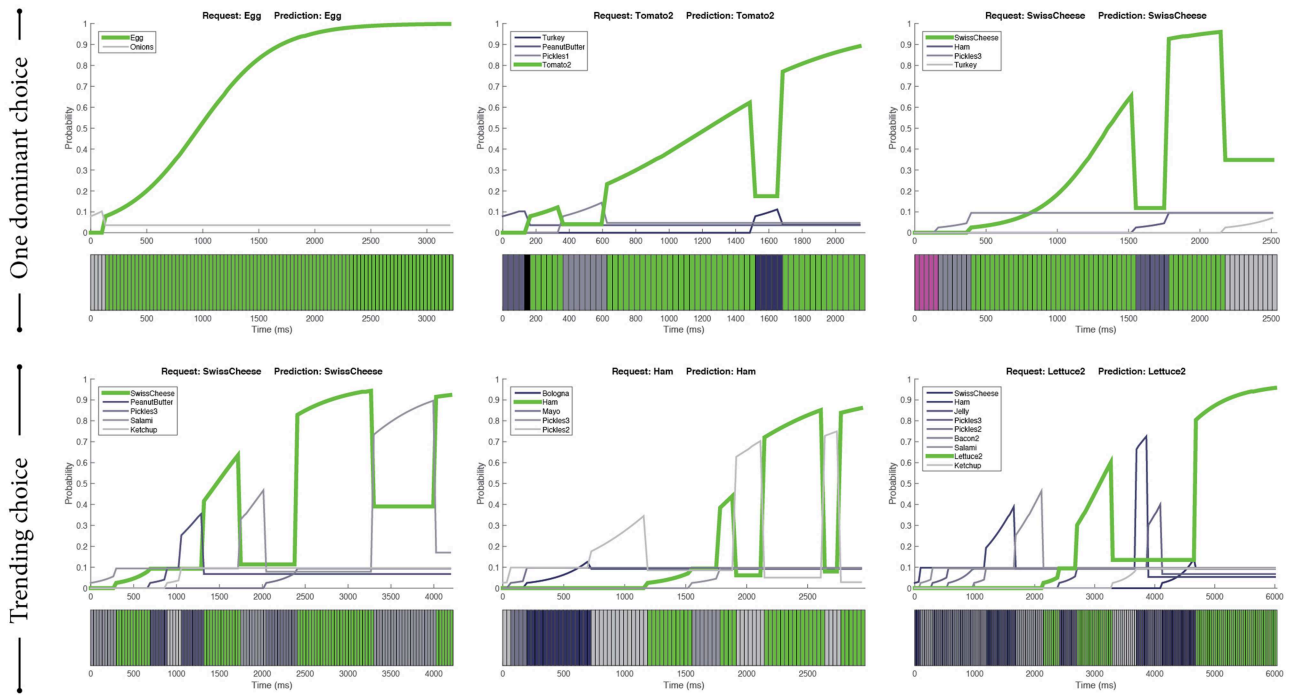


FIGURE 3 | Two main categories of correct predictions: one dominant choice (top) and the trending choice (bottom). Green indicates the ingredients predicted by our SVM-based predictor that were the same as the actual ingredients requested by the customers. Purple indicates gazing toward the bread and yellow indicates gazing toward the worker. Black indicates missing gaze data.

3.4.2.3. Multiple choices

Similar to the case of two competing choices, the customers sometimes decided among multiple candidate ingredients (**Figure 4**, Third row). As gaze cues were distributed across candidate ingredients, our predictor had difficulty in choosing the intended ingredient. Additional information, either from different behavioral modalities or new features of gaze cues, is necessary to distinguish the intended ingredient from the competing ones.

3.4.2.4. Favoring competing choices

In situations where the customers looked toward competing ingredients more frequently as compared to the intended ingredient, our predictor made incorrect predictions (see examples in **Figure 4**, Fourth row). One potential explanation for this type of gaze pattern is that the customers changed their decision after quick glances at the intended ingredients. For instance, as shown in the bottom examples of **Figure 4**, while the customers looked longer and multiple times at the red ingredient, they requested the blue ingredient with smaller gaze attention. Our features failed to capture such quick decisions, likely resulting in incorrect predictions.

3.4.3. Special patterns

In analyzing the efficacy of our SVM-based intention predictor, we observed some special, potentially informative gaze patterns that were not explicitly captured in our derived features emerge. We discuss these patterns in the following paragraphs.

3.4.3.1. Initiating joint attention

Initiating joint attention is the process of using behavioral cues to direct the other's attention to a shared artifact. One such behavioral instantiation involves alternating gaze cues—looking toward the intended ingredient, looking toward the worker, and then looking back at the intended ingredient (Mundy and Newell, 2007). We found such patterns of initiating joint attention in our data, as shown in the first row of **Figure 5**. This pattern usually emerged toward the end of the episode, serving as a signal to the worker that the intended ingredient had been chosen.

3.4.3.2. Confirmatory request

The inverse pattern of initiating joint attention is that of the customer looking toward the worker, toward the intended ingredient, and then back toward the worker. Conceptually, we can characterize this pattern as a confirmatory request, meaning that the customer sought the worker's attention, directed their attention, and checked if the intention was understood. From our data, this pattern of confirmatory request seemed to signify intention. As illustrated in the second row of **Figure 5**, the single ingredient between fixations at the worker was the intended ingredient.

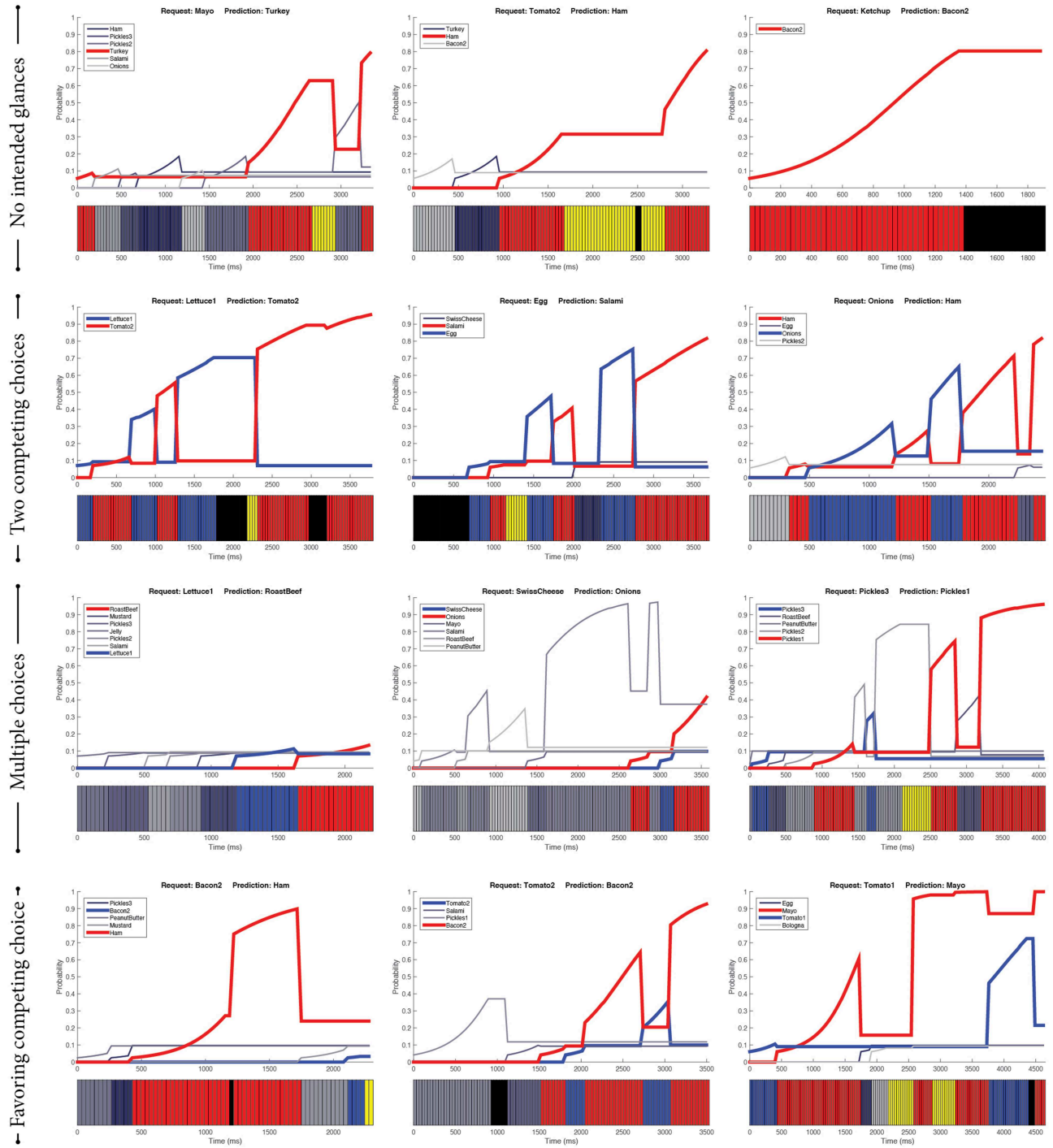


FIGURE 4 | Examples of incorrect predictions. Red indicates the prediction made by the SVM-based predictor, whereas blue indicates the actual ingredient requested by the customers. Purple indicates gazing toward the bread whereas yellow indicates gazing toward the worker. Black indicates missing gaze data.

3.4.3.3. Goal referencing

Another pattern that emerged from the data was visual references to the goal, which in our context was the bread where ingredients were moved. This type of reference was found in a variety of combinations. It could be found before, after, or in between choosing the intended ingredient. Examples are provided in the third row of **Figure 5**. There may be different meanings to these combinations. For instance, the customers might have checked which ingredients had been added to the sandwich and used that information to decide which ingredient to pick next.

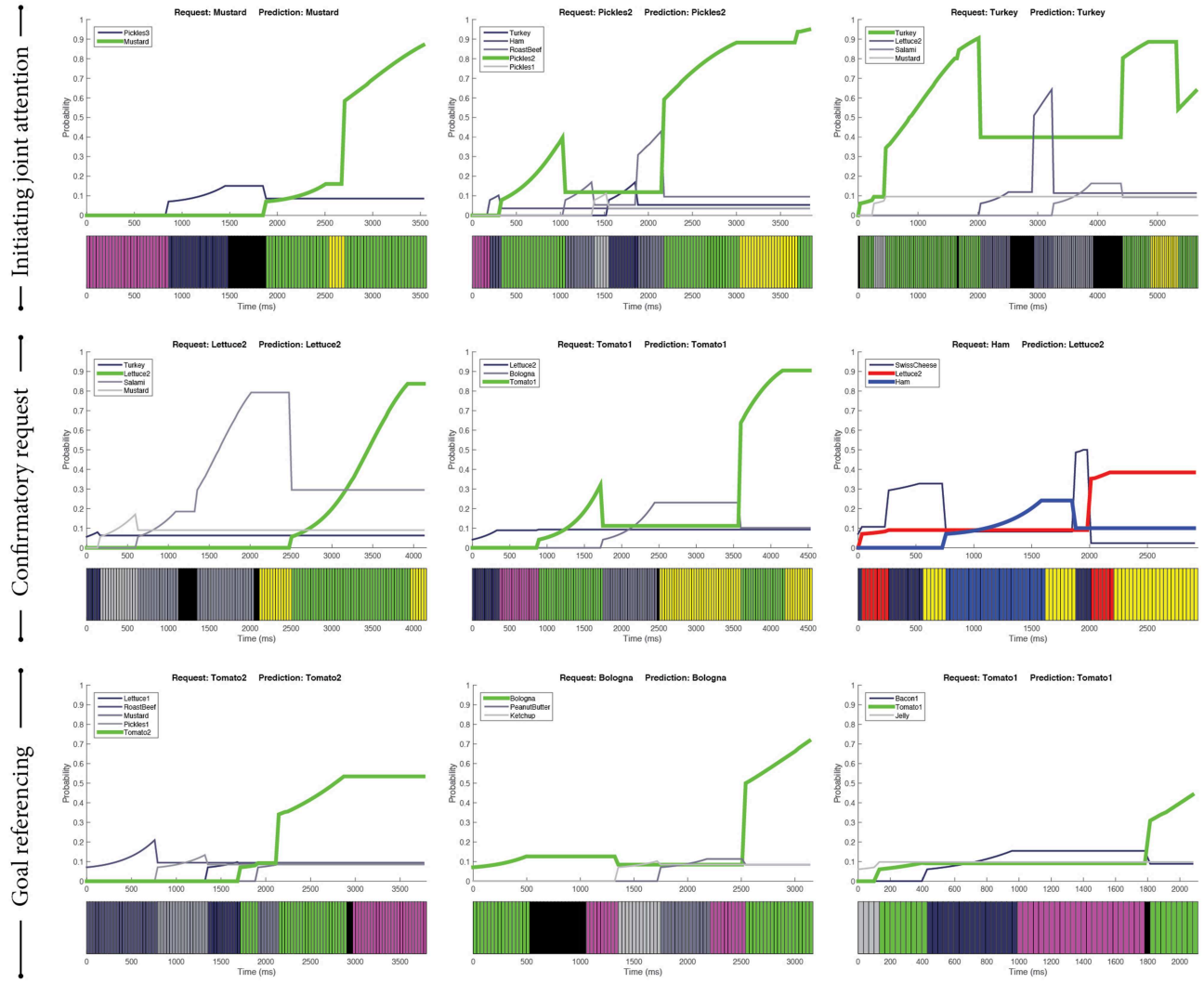


FIGURE 5 | Examples of special gaze patterns. Green indicates the ingredients predicted by our SVM-based predictor that were the same as the actual ingredients requested by the customers. Blue lines indicate the ingredients that the customers picked. Red lines are our predictions. Purple indicates gazing toward the bread, whereas yellow indicates gazing toward the worker. Black indicates missing gaze data.

4. Discussion

To quantitatively investigate the relationship between exhibited gaze cues and intentions, we contextualized our investigation in a sandwich-making scenario in which a worker made a sandwich using ingredients requested by a customer. We characterized intentions as the ingredients requested by the customers and hypothesized that the customers' gaze patterns would predict their choice of ingredients. We developed an SVM-based intention predictor using four features that aimed to represent characteristics of the customers' gaze patterns. The SVM-based predictor was demonstrated to outperform the basic attention-based predictor in predicting the customers' choices of ingredients. Moreover, the SVM-based predictor could make correct predictions approximately 1.8 s before the requests. Furthermore, we qualitatively analyzed the instances of correct and incorrect predictions made by the SVM-based predictor to better understand its performance in boundary cases. In this section, we discuss implications of our qualitative analyses, potential applications of our intention predictor, and limitations of the present work.

4.1. Implications of Qualitative Analyses

Our qualitative analyses (Section 3.4) provided not only insight into how the SVM-based predictor made correct and incorrect predictions, but they also revealed special patterns that may *signal* intentions via visual references to the other person and the goal. Signaling is an intentional strategy that people use to manifest actions and intentions in a way that is more predictable and comprehensible to interaction partners (Pezzulo et al., 2013). For example, parents exaggerate intonation in infant-directed speech (Kuhl et al., 1997). The use of signaling strategies facilitates the

formation of common ground. The special patterns *initiating joint attention* and *confirmatory request* involved interleaving gaze cues between the partner and the intended ingredient. These displays of interleaving gaze may serve as an intentional signaling strategy, highlighting the relevance of the intended ingredient. Similarly, the visual references to the goal, which is the bread in our scenario, may be signaling the intentional link between the bread and the intended ingredient, as shown in the pattern *goal referencing*.

The four features of gaze cues explored in this work were based on statistical measures of the customers' gaze sequences. While these features seemed to capture how the distribution of gaze cues may indicate intentions, they did not explicitly encode sequential structures from gaze sequences. However, sequential structures—such as gaze toward the target, then partner, and then the target again—may encapsulate particular semantic meanings, such as directing the partner's attention toward the target. The capability to recognize these sequential structures as those of *initiating joint attention*, *confirmatory request*, and *goal referencing*, could reveal the underlying meanings of gaze sequence and potentially improve the efficacy of the SVM-based predictor. For example, the last plot of the examples of *confirmatory request* showed that the intention predictor could benefit from recognizing the sequential human-target-human pattern. One way to recognize such sequential structures is through template matching, which has been explored to recognize communicative backchannels (Morency et al., 2010).

However, the special patterns, identified in Section 3.4.3, should be used with caution when predicting intentions. The last plot in **Figure 4** illustrated a contradictory example; even though there was a clear pattern of *confirmatory request*, it did not signify the intended ingredient. Further research is necessary to investigate how the incorporation of sequential structures into the predictive model may enhance predictive performance.

4.2. Applications

The capability to interpret others' intentions and anticipate actions is critical in performing joint actions (Sebanz and Knoblich, 2009; Huber et al., 2013). Prior research has explored how reading intention and performing anticipatory actions might benefit robots in providing assistance to their users, highlighting the importance of intention prediction in joint actions between humans and robots (Sakita et al., 2004; Hoffman and Breazeal, 2007). Building on prior research, this work provides empirical results showing the relationship between gaze cues and human intentions. It also presents an implementation of an intention predictor using SVMs. With the advancement of computing and sensing technologies, such as gaze tracking systems, we anticipate that an even more reliable intention predictor could be realized in the foreseeable future. Computer systems such as assistive robots and ubiquitous devices could utilize intention predictors to augment human capabilities in many applications. For example, robot co-workers could predict human workers' intentions by monitoring their gaze cues, enabling the robots to choose complementary tasks to increase productivity in manufacturing applications. Similarly, assistive robots could provide necessary assistance to people by interpreting their gaze patterns that signal intended help. In addition to applications involving physical interactions, recommendation systems could provide better recommendations to users by utilizing their gaze patterns. For instance, an online shopping website could dynamically recommend products to customers by tracking and interpreting their gaze patterns.

4.3. Limitations

The current work also has limitations that motivate future investigations. First, we employed SVMs for data analysis and modeling to quantify the potential relationship between gaze cues and intentions. Alternative approaches, such as decision trees and hidden Markov models (HMMs), may also be used to investigate such relationships and interaction dynamics. However, similar to most machine learning approaches that are sensitive to the data source, our results were subject to the interaction context and the collected data. For instance, the parameters of the predictive window (e.g., size) might be limited to our present context. Yet, in this work, we demonstrated that characteristics of gaze cues, especially duration and frequency, are a rich source for understanding human intentions. Furthermore, we used a toy set of sandwich items as our research apparatus. Participants working with the toy sandwich may have produced different gaze patterns then they would when working with real sandwich materials.

Second, we formulated the problem of intention prediction in the context of sandwich-making as the problem of using the customers' gaze patterns to predict their choices of ingredients. Intention is a complex construct that may not be simply represented as the requested ingredient. While our work focused solely on using gaze cues to predict customer intent, workers in this scenario may rely on additional features, including facial expressions and other cues from the customer, and other forms of contextual information, such as preferences expressed previously toward particular toppings or knowledge of what toppings might "go together." Disentangling the contributions of different features to observer performance in these predictions would significantly enrich our understanding of the process people follow to predict intent. However, our findings were in line with literature indicating that gaze cues manifest attention and lead intended actions (Butterworth, 1991; Land et al., 1999; Johansson et al., 2001). In addition, the sequences of gaze cues, as inputs to our predictive model, were obtained via a gaze tracker worn by the customers. Future research may consider acquiring the gaze sequences from the perspective of the worker. This approach may be beneficial in developing an autonomous robotic assistant (Ognibene and Demiris, 2013; Ognibene et al., 2013) that can leverage its onboard camera to obtain the different items human users gaze toward. Future work may also compare the performance of human observers and the types of errors they make to those of our machine learning model. Such a comparison may inform our selection of features or learning algorithms in building systems that recognize user intent.

5. Conclusion

Eye gaze is a rich source for interpreting a person's intentions. In this work, we developed a SVM-based approach to quantify how gaze cues may signify a person's intention. Using the data collected from a sandwich-making task, we demonstrated the effectiveness of our approach in a laboratory evaluation, where our predictor provided improved accuracy in making correct predictions of the customers' choices of ingredient (76%) compared to the attention-based approach (65%) that only relied on the most recently glanced-at ingredient. Moreover, our SVM-based approach provided correct predictions approximately 1.8 s before the requests, whereas the attention-based approach did not afford such intention anticipation. Analyses of the episodic interactions further revealed gaze patterns that suggested semantic meanings and that contributed to correct and incorrect predictions. These patterns informed the design of gaze features that offer a more complete picture of human intentions. Our findings provide insight into linking human intentions and gaze cues and offer implications for designing intention predictors for assistive systems that can provide anticipatory help to human users.

Acknowledgments

This work was supported by National Science Foundation awards 1149970 and 1426824. The dataset analyzed in this paper is also used in another submission (Andrist et al., 2015) to this Research Topic. The authors would like to thank Ross Luo and Jing Jing for their contributions to data collection and analysis.

References

Adams, R. B., and Kleck, R. E. (2005). Effects of direct and averted gaze on the perception of facially communicated emotion. *Emotion* 5:3. doi: 10.1037/1528-3542.5.1.3

Andrist, S., Collier, W., Gleicher, M., Mutlu, B., and Shaffer, D. (2015). Look together: analyzing gaze coordination with epistemic network analysis. *Front. Psychol.* 6:1016. doi: 10.3389/fpsyg.2015.01016

Astington, J. W. (1993). *The Child's Discovery of the Mind, Vol. 31*. Cambridge, MA: Harvard University Press.

Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. *Psychol. Sci.* 15, 415–419. doi: 10.1111/j.0956-7976.2004.00694.x

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., and Plumb, I. (2001). The reading the mind in the eyes test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *J. Child Psychol. Psychiatry* 42, 241–251. doi: 10.1111/1469-7610.00715

Blakemore, S.-J., and Decety, J. (2001). From the perception of action to the understanding of intention. *Nat. Rev. Neurosci.* 2, 561–567. doi: 10.1038/35086023

Bratman, M. (1987). *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Brennan, S. E., Chen, X., Dickinson, C. A., Neider, M. B., and Zelinsky, G. J. (2008). Coordinating cognition: the costs and benefits of shared gaze during collaborative search. *Cognition* 106, 1465–1477. doi: 10.1016/j.cognition.2007.05.012

Buccino, G., Binkofski, F., Fink, G. R., Fadiga, L., Fogassi, L., Gallese, V., et al. (2001). Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. *Eur. J. Neurosci.* 13, 400–404. doi: 10.1046/j.1460-9568.2001.01385.x

Butler, S. C., Caron, A. J., and Brooks, R. (2000). Infant understanding of the referential nature of looking. *J. Cogn. Dev.* 1, 359–377. doi: 10.1207/S15327647JCD0104_01

Butterworth, G. (1991). *The Ontogeny and Phylogeny of Joint Visual Attention*. Oxford, UK: Basil Blackwell.

Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2, 27:1–27:27. doi: 10.1145/1961189.1961199

Chen, Y.-W., and Lin, C.-J. (2006). "Combining SVMs with various feature selection strategies," in *Feature Extraction*, eds I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh (Berlin; Heidelberg: Springer), 315–324.

Clark, H. H. (1996). *Using Language, Vol. 1996*. Cambridge: Cambridge University Press.

Clark, H. H., and Brennan, S. E. (1991). Grounding in communication. *Perspect. Soc. Shared Cogn.* 13, 127–149.

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1007/BF00994018

d'Andrade, R. (1987). *A Folk Model of the Mind*. Cambridge, UK: Cambridge University Press.

Doshi, A., and Trivedi, M. M. (2009). On the roles of eye gaze and head dynamics in predicting driver's intent to change lanes. *IEEE Trans. Intell. Transport. Syst.* 10, 453–462. doi: 10.1109/TITS.2009.2026675

Feinfield, K. A., Lee, P. P., Flavell, E. R., Green, F. L., and Flavell, J. H. (1999). Young children's understanding of intention. *Cogn. Dev.* 14, 463–486.

Frischen, A., Bayliss, A. P., and Tipper, S. P. (2007). Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychol. Bull.* 133:694. doi: 10.1037/0033-2909.133.4.694

Griffin, Z. M. (2001). Gaze durations during speech reflect word selection and phonological encoding. *Cognition* 82, B1–B14. doi: 10.1016/S0010-0277(01)00138-X

Haji-Abolhassani, A., and Clark, J. J. (2014). An inverse Yarbus process: predicting observers task from eye movement patterns. *Vis. Res.* 103, 127–142. doi: 10.1016/j.visres.2014.08.014

Hayhoe, M., and Ballard, D. (2014). Modeling task control of eye movements. *Curr. Biol.* 24, R622–R628. doi: 10.1016/j.cub.2014.05.020

Heal, J. (2005). "Joint attention and understanding the mind," in *Joint Attention: Communication and Other Minds*, eds D. Bourget and D. Chalmers (Oxford, UK: Oxford University Press), 34–44.

Hoffman, G., and Breazeal, C. (2007). "Effects of anticipatory action on human-robot teamwork efficiency, fluency, and perception of team," in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction* (Arlington, VA: ACM), 1–8.

Huber, M., Kupferberg, A., Lenz, C., Knoll, A., Brandt, T., and Glasauer, S. (2013). Spatiotemporal movement planning and rapid adaptation for manual interaction. *PLoS ONE* 8:e64982. doi: 10.1371/journal.pone.0064982

Johansson, R. S., Westling, G., Bäckström, A., and Flanagan, J. R. (2001). Eye–hand coordination in object manipulation. *J. Neurosci.* 21, 6917–6932.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., et al. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science* 277, 684–686.

Land, M., Mennie, N., and Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception* 28, 1311–1328.

Leslie, A. M. (1987). Pretense and representation: the origins of "theory of mind." *Psychol. Rev.* 94, 412.

Malle, B. F., and Knobe, J. (1997). The folk concept of intentionality. *J. Exp. Soc. Psychol.* 33, 101–121.

Masur, E. F. (1983). Gestural development, dual-directional signaling, and the transition to words. *J. Psycholinguist. Res.* 12, 93–109.

McCall, J. C., Wipf, D. P., Trivedi, M. M., and Rao, B. D. (2007). Lane change intent analysis using robust operators and sparse bayesian learning. *IEEE Trans. Intell. Transport. Syst.* 8, 431–440. doi: 10.1109/TITS.2007.902640

Meltzoff, A. N. (1995). Understanding the intentions of others: re-enactment of intended acts by 18-month-old children. *Dev. Psychol.* 31, 838.

Meltzoff, A. N., and Brooks, R. (2001). ""Like Me" as a building block for understanding other minds: bodily acts, attention, and intention," in *Intentions and Intentionality: Foundations of Social Cognition*, eds B. F. Malle, L. J. Moses, and D. A. Baldwin (Cambridge, MA: MIT Press), 171–191.

Meyer, A. S., Sleiderink, A. M., and Levelt, W. J. (1998). Viewing and naming objects: eye movements during noun phrase production. *Cognition* 66, B25–B33.

Morales, M., Mundy, P., and Rojas, J. (1998). Following the direction of gaze and language development in 6-month-olds. *Infant Behav. Dev.* 21, 373–377.

Morency, L.-P., de Kok, I., and Gratch, J. (2010). A probabilistic multimodal approach for predicting listener backchannels. *Auton. Agent. Multi. Agent. Syst.* 20, 70–84. doi: 10.1007/s10458-009-9092-y

Morris, D., and Desebrock, G. (1977). *Manwatching: A Field Guide to Human Behaviour*. New York, NY: HN Abrams.

Mundy, P., and Newell, L. (2007). Attention, joint attention, and social cognition. *Curr. Dir. Psychol. Sci.* 16, 269–274. doi: 10.1111/j.1467-8721.2007.00518.x

Ognibene, D., Chinellato, E., Sarabia, M., and Demiris, Y. (2013). Contextual action recognition and target localization with an active allocation of attention on a humanoid robot. *Bioinspirat. Biomimet.* 8:035002. doi: 10.1088/1748-3182/8/3/035002

Ognibene, D., and Demiris, Y. (2013). "Towards active event recognition," in *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence* (Beijing: AAAI Press), 2495–2501.

Pezzulo, G., Donnarumma, F., and Dindo, H. (2013). Human sensorimotor communication: a theory of signaling in online social interactions. *PLoS ONE* 8:e79876. doi: 10.1371/journal.pone.0079876

Rizzolatti, G., and Craighero, L. (2004). The mirror-neuron system. *Annu. Rev. Neurosci.* 27, 169–192. doi: 10.1146/annurev.neuro.27.070203.144230

Sakita, K., Ogawara, K., Murakami, S., Kawamura, K., and Ikeuchi, K. (2004). "Flexible cooperation between human and robot by interpreting human intention from gaze information," in *Proceedings 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2004 (IROS 2004),* Vol. 1 (Sendai: IEEE), 846–851.

Sebanz, N., Bekkering, H., and Knoblich, G. (2006). Joint action: bodies and minds moving together. *Trends Cogn. Sci.* 10, 70–76. doi: 10.1016/j.tics.2005.12.009

Sebanz, N., and Knoblich, G. (2009). Prediction in joint action: what, when, and where. *Top. Cogn. Sci.* 1, 353–367. doi: 10.1111/j.1756-8765.2009.01024.x

Shibata, S., Tanaka, K., and Shimizu, A. (1995). "Experimental analysis of handing over," in *Proceedings 4th IEEE International Workshop on Robot and Human Communication, 1995, RO-MAN'95 TOKYO* (Tokyo: IEEE), 53–58.

Tomasello, M. (1995). "Joint attention as social cognition," in *Joint Attention: Its Origins and Role in Development*, eds C. Moore and P. J. Dunham (New York, NY: Psychology Press), 103–130.

White, S. (1989). Backchannels across cultures: a study of Americans and Japanese. *Lang. Soc.* 18, 59–76.

Wu, T.-F., Lin, C.-J., and Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *J. Mach. Learn. Res.* 5, 975–1005.

Yi, W., and Ballard, D. (2009). Recognizing behavior in hand-eye coordination patterns. *Int. J. Humanoid Robot.* 6, 337–359. doi: 10.1142/S0219843609001863

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

OPINION
published: 08 July 2015
doi: 10.3389/fpsyg.2015.00929

What do we think we are doing: principles of coupled self-regulation in human-robot interaction (HRI)

Idit Shalev[1]* *and Tal Oron-Gilad*[2]

[1] Department of Education and the Zlotowski Center of Neuroscience, Ben-Gurion University of the Negev, Beer-Sheva, Israel, [2] Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Beer-Sheva, Israel

Keywords: human-robot interaction (HRI), embodied cognition, self-regulation, goal pursuit, teamwork

The use of domestic service robots is becoming widespread. While in industrial settings robots are often used for specified tasks, the challenge in the case of robots put to domestic use is to afford human-robot collaboration in a variety of non-predefined and different daily tasks. Herein, we aim at identifying and understanding the conditions that will facilitate flexible collaboration between humans and robots. Past research of social and personality psychology was mainly focused on individual's self-regulation, defined as the ability to govern, or direct attention, resources, or action toward the realization of a particular goal (Higgins, 1989; Kruglanski et al., 2002). There is evidence that pursuing goals with the presence of others influences self-control (Fishbach and Trope, 2005), however only little is known on dyadic processes of self-regulation. Additionally, whereas research of goal pursuit in social psychology has mainly been associated with general processes of the structure and function of goals (Gollwitzer and Bargh, 1996; Carver and Scheier, 1998; Kruglanski et al., 2002; Fishbach and Ferguson, 2007; Custers and Aarts, 2010), human-robot interaction involves pragmatic interpersonal dilemmas such as how to coordinate human-robot activity and what knowledge should be shared between humans and robots over the course of action. To fill this gap, in what follows, we will define the unique characteristics of what we term as *human-robot coupled self-regulation*, which has the unique features of a dyadic asymmetric team aimed to increase the affordances of an individual in different activities. We will describe the unique characteristics of human-robot interaction and its special challenges toward goal pursuit.

OPEN ACCESS

Edited by:
Sebastian Loth, Universität Bielefeld, Germany

Reviewed by:
Serge Thill, University of Skövde, Sweden
Maria Koutsombogera, Institute for Language and Speech Processing, Greece
Julian Hough, Universität Bielefeld, Germany

***Correspondence:**
Idit Shalev, shalevid@bgu.ac.il

Specialty section:
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology

Received: *15 March 2015*
Accepted: *22 June 2015*
Published: *08 July 2015*

Citation:
Shalev I and Oron-Gilad T (2015) What do we think we are doing: principles of coupled self-regulation in human-robot interaction (HRI). Front. Psychol. 6:929. doi: 10.3389/fpsyg.2015.00929

Human and Robot are a Dyadic Instrumental Asymmetric Team

Our first assumption is that self-regulation of a human-robot couple could be conceptualized as a unique team configuration. A team is "a distinguishable set of two or more people who interact, dynamically, interdependently, and adaptively toward a common and valued goal/objective/mission, who have each been assigned specific roles or functions to perform, and who have a limited life-span of membership" (Salas et al., 1992, p. 4; Salas et al., 2010). Team members have differentiated responsibilities and roles (Cannon-Bowers et al., 1993). Therefore, essential for a team's successful performance is the understanding of the abilities and behaviors of its members that fit their experience and unique expertise for the task at hand.

Because humans and robots differ in their level of agency (the capacity to act and do) and their level of experience (the capacity to feel and sense), (Gray and Wegner, 2012), we argue that their contribution to the team is not symmetric. Based on the reasoning that genuine authorship of an action or situation may not always be clear (Dijksterhuis et al., 2008), we suggest that defined requirements of person, robot, and situation are essential to reduce the expectation gap.

Our perspective is that human-robot collaboration should be viewed in terms of functionality, to extend possibilities for the kinds of goals that humans want to pursue. These instrumental relations

between a person and her tool, used to increase the fit between person and environment, are termed affordances (Gibson, 1979). Following this view, we argue that robots can be perceived as self-regulatory tools to increase affordances across different situations (Koole and Veenstra, 2015). Our instrumental relational approach enables flexibility in tuning the robot's level of responsiveness and dominance in human-robot social contexts. For example, whereas the human member of the team holds a fixed ownership position, the robot's level of dominance could vary by user demands, or depending on the situation. To understand the usefulness of this principle, let us take for example 80 year old Mrs. Brown. She is physically fragile, but it is important for her to maintain an independent life style. This is why she has "Rupert," a multi-functional platform robot that serves as her aid. When she leaves the house she may want "Rupert" to lead and find the safest walking path to the store, thus she may set it to high dominance and responsiveness, in case she startles. At home, she may not desire high level of proactive care-taking and leave "Rupert" to be on call.

Concrete Level of Human-robot Negotiation

Our second assumption is that human-robot coupled self-regulation is based on concrete rather than abstract level of agreement. Carrying out human-robot joint actions demands continuous coordination on at least five elements: (1) who takes part; (2) what is the role of each member; (3) what is the joint goal; (4) how does each team member contribute to the timing and synchronization; and (5) where the actions take place (Clark, 2005). To address this, the robot should identify where the focus of attention of the human is, to what degree the attention of the human is focused on team actions, and how to convey feedback. Similarly, the human needs to calibrate expectations from the robot, i.e., be invested in the robot's immediate action or approval of action, and how to respond to the robot's requests (Alami et al., 2005).

Coupled self-regulation of goals requires agreement on goal setting and goal striving as two basic phases in goal pursuit (Gollwitzer and Oettingen, 2011). Whereas, robots may act automatically from initiation to completion of the task, humans' possible reflection on their performance may involve conscious awareness and create new representations of behavior, thus leading to communication gaps (Baumeister and Bargh, 2014). According to the action identification theory, a specific action can be verbally identified and interpreted from different levels of abstraction, ranging from low-level identities that specify how the action is performed, to high-level identities that signify why the action is performed. For instance, a person who "drinks water" can identify it as "holding a glass" (low level), or as "relieving thirst" (high level) (Vallacher and Wegner, 1987, 1989). This helps explain why different action identifications by human and robot may lead to dissimilar systems of goals and means of attainment (Kruglanski et al., 2002; Shah et al., 2002).

To address these challenges, we suggest the use of multiple human-robot forms of communication to pursue the joint goal. Lohan et al. (2014) proposed a distinction between two kinds of actions: path-oriented and manner-oriented, that can be communicated via two different linguistic utterance styles. Whereas, in path-oriented utterances the goal is stressed, in manner-oriented utterances, the means of motion are emphasized (e.g., Talmy, 1991). In our example, Mrs. Brown and "Rupert" carry a recliner to the porch (Path-"let's move the chair to the porch" or Manner-"I want to read my book on the porch"). Suddenly the phone rings and Mrs. Brown wants to go and answer ((Path-"let me go get the phone" or Manner-"I need to answer this call"). "Rupert" must understand that the goal has changed and pause.

Continuous and Various Communication Forms Over Goal Pursuit

Research indicates that professional and social interactions between team members can develop the team's social cognition (Klimoski and Mohammed, 1994). There is evidence that a team's fluent on-going communication regarding goal pursuit reduces the need for preexisting knowledge (Kozlowski and Bell, 2003). In social HRI, it is critical to generate many levels of interaction with the automation. Hence, the robot should always be present and aim to facilitate the goal, even if only to provide recommendations. In civil aviation, for example, communication is key especially if things turn out unexpectedly. In the Northwest 2009 incident in Minneapolis the automation had the capability, but was not designed to point out that the task was not performed as planned and that the pilots missed their destination. To borrow from our previous example, let us suppose Mrs. Brown wants to grab a pillow from the upper cabinet. The robot may not be able to reach so high, but it should continue to collaborate by providing feedback and advice; I cannot reach the uppermost cupboard (failure to complete task) but it is too dangerous for you to try to reach it on your own, if not urgent, perhaps we should call your son, or is there another pillow on a lower shelf?

Much of human communication over goal pursuit is based on social cues (e.g., gestures, and mimicry) that automatically generate social judgment and behavior (Chartrand and Bargh, 1999; van Baaren et al., 2003; Leander et al., 2010). Similarly, translation of social cues to social signals leads to inference of human intentions by robotic agents (Fiore et al., 2013). The relevance of automatic embodied cues for joint goal pursuit was demonstrated in human-human and human-robot synchronicity, suggesting that physical synchronicity is associated with experience of responsiveness and empathy (Sebanz and Knoblich, 2009; Cohen et al., 2010; Paladino et al., 2010; Boucher et al., 2012; Hoffman et al., 2014). Embodied communication is not only "used" by robots, but integrated in them to support both the recognition of the human's behavior and the generation of their behavior. Research of social signal processing and modeling multimodal communication, suggests that social and behavioral cues may be detectable from a machine, hence perceivable. Likewise, models of behavior are integrated in a way that a robot exhibits a more natural behavior, aiming at

a more successful interaction with the human (Pentland, 2007; Vinciarelli et al., 2012).

However, despite emerging findings from the field of embodied cognition on the potential of physical and social cues as an alternative route for communication, it was also claimed that embodied cognition cues can lead to different patterns of activation across different contexts (Loersch and Payne, 2011), thus prediction of behavior may be difficult (Shalev, 2015). A possible way to address this limitation is to use robots in fixed context, where interpretation to human's embodied signals is less ambiguous. For example Loth et al. (2013), have demonstrated that bar staff responded to a set of two non-verbal signals. Foster (2014), indicated that robotic sensors can similarly detect and respond to these signals.

Addressing the Human-robot Communication Gap over Goal Pursuit

Individuals frequently use embodied cues for functional self-regulatory purposes (Balcetis and Cole, 2009; Schnall et al., 2010; Bargh and Shalev, 2012; Shalev, 2014). However, using embodied cues as diagnostic inputs (Williams et al., 2009; Ackerman et al., 2010; Meier et al., 2012; Robinson and Fetterman, 2015; Winkielman et al., 2015) may lead to human-robot miscommunications. For example, human speakers expect co-located listeners to link visually perceivable objects and the verbally described references to them. Thus, humans may expect a co-located robot to have the same visual-verbal linking abilities (e.g., look at the green object on the right), thus developers must integrate the robot's visual system with natural language components to enable this flow of communication (Kopp, 2010; Cantrell et al., 2012; Vollmer et al., 2013).

Furthermore, there is also anecdotal evidence of human-human communication misunderstandings in complex scenes. For example orientation can be relative to egocentric, or exocentric (absolute or relative) locations. Soldiers for example, are taught to communicate via the exocentric coordinates of the compass rose. However, most humans tend to naturally orient relative to their egocentric perspective, which may be difficult for robots to depict. Interestingly, Cassenti et al. (2012) found that instructors used exocentric references to direct the robot and that it improved their performance relative to egocentric-only commands.

To address this communication gap, we argue that shared database, sensors and multiple types of displays and interaction means (e.g., physiological measures, eye tracking, voice, touch, text, button presses etc.) can enrich the robot's capacity of perception and expression. Similarly, to reduce expectation issues, technology can shape the way the user acts on the robot, how individuals understand what to expect from it, and how they can interact with a robot to refine their mutual understanding of the task at hand. Providing the relevant information about the current state of the robot, the progress of the task, and of the surrounding environment, can facilitate successful performance. Similarly, education efforts need to convey the ambiguity of ongoing human-robot communication, particularly the robot's physical and data-driven limitations, and to encourage problem solving and novelty seeking.

Acknowledgments

The research was partially supported by the Helmsley Charitable Trust through the Agricultural, Biological and Cognitive Robotics Center of Ben-Gurion University of the Negev. We thank the editor and reviewers for their helpful comments.

References

Ackerman, J. M., Nocera, C. C., and Bargh, J. A. (2010). Incidental haptic sensations influence social judgments and decisions. *Science* 328, 1712–1715. doi: 10.1126/science.1189993

Alami, R., Clodic, A., Montreuil, V., Sisbot, E. A., and Chatila, R. (2005). *Task Planning for Human-Robot Interaction*. Grenobel: JointsOc-EUSAI Conference.

Balcetis, E., and Cole, S. (2009). Body in mind: the role of embodied cognition in self-regulation. *Soc. Pers. Psychol. Commun*. 759–774. doi: 10.1111/j.1751-9004.2009.00197.x

Bargh, J. A., and Shalev, I. (2012). The substitutability of physical and social warmth in daily life. *Emotion* 12, 154–162. doi: 10.1037/a0023527

Baumeister, R. F., and Bargh, J. A. (2014). "Conscious and unconscious: toward an integrative understanding of human life and action," in *Dual Process Theories of the Social Mind*, eds J. Sherman, B. Gawronski, and Y. Trope (New York, NY: Guilford Press), 33–49.

Boucher, J. D., Pattacini, U., Lelong, A., Bailly, G., Elisei, F., Fagel, S. P. F., et al. (2012). I reach faster when i see you look: gaze effects in human-human and human-robot face-to-face cooperation. *Front. Neurorob.* 6:3. doi: 10.3389/fnbot.2012.00003

Cannon-Bowers, J. A., Salas, E., and Converse, S. A. (1993). "Shared mental models in expert team decision making," in *Current Issues in Individual and Group Decision Making*, ed N. J. Castellan (Mahwah, NJ: Erlbaum), 221–246.

Cantrell, R., Krause, E., Scheutz, M., Zillich, M., and Potapova, E. (2012). "Incremental referent grounding with NLP-biased visual search," in *Proceedings of AAAI 2012 Workshop on Grounding Language for Physical Systems*.

Carver, C. S., and Scheier, M. F. (1998). *On the Self-regulation of Behavior*. New York, NY: Cambridge University Press.

Cassenti, D. N., Kelley, T. D., Yagoda, R., and Avery, E. (2012). Improvements in robot navigation through operator speech preference. *Paladyn* 3, 102–111. doi: 10.2478/s13230-012-0100-6

Chartrand, T. L., and Bargh, J. A. (1999). The chameleon effect: the perception–behavior link and social interaction. *J. Pers. Soc. Psychol.* 76:893. doi: 10.1037/0022-3514.76.6.893

Clark, H. H. (2005). Coordinating with each other in a material world. *Dis. Stud.* 7, 507–525. doi: 10.1177/1461445605054404

Cohen, E., Ejsmond-Frey, R., Knight, N., and and Dunbar, R. I. M. (2010). Rowers' high: behavioural synchrony is correlated with elevated pain thresholds. *Biol. Lett.* 6, 106–108. doi: 10.1098/rsbl.2009.0670

Custers, R., and Aarts, H. (2010). The unconscious will: how the pursuit of goals operates outside of conscious awareness. *Science* 329, 47–50. doi: 10.1126/science.1188595

Dijksterhuis, A., Preston, J., Wegner, D. M., and Aarts, H. (2008). Effects of subliminal priming of self and God on self-attribution of authorship for events. *J. Exp. Soc. Psychol.* 44, 2–9. doi: 10.1016/j.jesp.2007.01.003

Fiore, S. M., Wiltshire, T. J., Lobato, E. J., Jentsch, F. G., Huang, W. H., and Axelrod, B. (2013). Toward understanding social cues and signals in human–robot interaction: effects of robot gaze and

proxemic behavior. *Front. Psychol.* 4:859. doi: 10.3389/fpsyg.2013.00859

Fishbach, A., and Ferguson, M. F. (2007). "The goal construct in social psychology," in *Social Psychology: Handbook of Basic Principles*, eds A. W. Kruglanski and T. E. Higgins (New York, NY: Guilford Press), 490–515.

Fishbach, A., and Trope, Y. (2005). The substitutability of external control and self-control in overcoming temptation. *J. Exp. Soc. Psychol.* 41, 256–270. doi: 10.1016/j.jesp.2004.07.002

Foster, M. E. (2014). "Validating attention classifiers for multi-party human-robot interaction," in *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction: Workshop on Attention Models in Robotics*. (Bielefeld: ACM Press).

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston, MA: Houghton Mifflin.

Gollwitzer, P. M., and Bargh, J. A. (eds.). (1996). *The Psychology of Action: Linking Cognition and Motivation to Behavior*. New York, NY: Guilford Press.

Gollwitzer, P. M., and Oettingen, G. (2011). "Planning promotes goal striving," in *Handbook of self-regulation: Research, Theory, and Applications, 2nd edn.*, eds K. D. Vohs and R. F. Baumeister (New York, NY: Guilford), 162–185.

Gray, K., and Wegner, D. M. (2012). Feeling robots and human zombies: mind perception and the uncanny valley. *Cognition* 125, 125–130. doi: 10.1016/j.cognition.2012.06.007

Higgins, E. T. (1989). Continuities and discontinuities in self-regulatory and self-evaluative processes: a developmental theory relating self and affect. *J. Pers.* 57, 407–444. doi: 10.1111/j.1467-6494.1989.tb00488.x

Hoffman, G., Birnbaum, G. E., Reis, H. T., Vanunu, K., and Sass, O. (2014). "Robot responsiveness to human disclosure affects social impression and appeal," in *Proceedings of the 9th ACM/IEEE International Conference on Human-Robot Interaction*.

Klimoski, R., and Mohammed, S. (1994). Team mental model: construct or metaphor? *J. Man.* 20, 403–437.

Koole, S. L., and Veenstra, L. (2015). Does emotion regulation occur only inside people's heads? Towards a situated cognition analysis of emotion-regulatory dynamics. *Psychol. Inq.* 26, 61–68. doi: 10.1080/1047840X.2015.964657

Kopp, S. (2010). Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors *Speech Commun.* 52, 587–597. doi: 10.1016/j.specom.2010.02.007

Kozlowski, S. W. J., and Bell, B. S. (2003). "Work groups and teams in organizations," in *Handbook of Psychology: Industrial and Organizational Psychology* eds W. C. Borman, D. R. Ilgen, R. J. Klimoski (London: Wiley), 333–375.

Kruglanski, A. W., Shah, J. Y., Fishbach, A., Friedman, R., Chun, W. Y., and Sleeth-Keppler, D. (2002). A theory of goal systems. *Adv. Exp. Soc. Psychol.* 34, 331–378. doi: 10.1016/s0065-2601(02)80008-9

Leander, N. P., Chartrand, T. L., and Wood, W. (2010). Mind your mannerisms: behavioral mimicry elicits stereotype conformity. *J. Exp. Soc. Psychol.* 47, 195–201. doi: 10.1016/j.jesp.2010.09.002

Loersch, C., and Payne, B. K. (2011). The situated inference model an integrative account of the effects of primes on perception, behavior, and motivation. *Pers. Psychol. Sci.* 6, 234–252. doi: 10.1177/1745691611406921

Lohan, K. S., Griffiths, S. S., Sciutti, A., Partmann, T. C., and Rohlfing, K. J. (2014). Co−development of manner and path concepts in language, action, and eye−gaze behavior. *Top. Cogn. Sci.* 6, 492–512. doi: 10.1111/tops.12098

Loth, S., Huth, K., and de Ruiter, J. P. (2013). Automatic detection of service initiation signals used in bars. *Front. Psychol.* 4:557. doi: 10.3389/fpsyg.2013.00557

Meier, B. P., Schnall, S., Schwarz, N., and Bargh, J. A. (2012). Embodiment in social psychology. *Top. Cogn. Sci.* 4, 705–716. doi: 10.1111/j.1756-8765.2012.01212.x

Paladino, M. P., Mazzurega, M., Pavani, F., and Schubert, T. W. (2010). Synchronous multisensory stimulation blurs self-other boundaries. *Psychol. Sci.* 21, 1202–1207. doi: 10.1177/0956797610379234

Pentland, A. A. (2007). Social signal processing. *IEEE Signal Process. Mag.* 24, 108–111. doi: 10.1109/MSP.2007.4286569

Robinson, M. D., and Fetterman, A. K. (2015). The embodiment of success and failure as forward versus backward movements. *PLoS ONE* 10:e0117285. doi: 10.1371/journal.pone.0117285

Salas, E., Cooke, N. J., and Gorman, J. C. (2010). The science of team performance: progress and the need for more. *Int. J. Hum. Comput. Stud.* 45, 75–104. doi: 10.1177/0018720810374614

Salas, E., Dickinson, T. L., Converse, S. A., and Tannenbaum, S. I. (1992). "Toward an understanding of team performance and training," in *Teams: Their Training and Performance*, eds R. J. Swezey and E. Salas (Norwood, NJ: Ablex), 3–29.

Schnall, S., Zadra, J. R., and Proffitt, D. R. (2010). Direct evidence for the economy of action: glucose and the perception of geographical slant. *Perception* 39, 464. doi: 10.1068/p6445

Sebanz, N., and Knoblich, G. (2009). Prediction in joint action: what, when, and where. *Top. Cogn. Sci.* 1, 353–367. doi: 10.1111/j.1756-8765.2009.01024.x

Shah, J. Y., Friedman, R., and Kruglanski, A. W. (2002). Forgetting all else: on the antecedents and consequences of goal shielding. *J. Pers. Soc. Psychol.* 83:1261. doi: 10.1037/0022-3514.83.6.1261

Shalev, I. (2014). Implicit energy loss: embodied dryness cues influence vitality and depletion. *J. Cons. Psychol.* 24, 260–270. doi: 10.1016/j.jcps.2013.09.011

Shalev, I. (2015). The architecture of embodied cue integration: insight from the "motivation as cognition" perspective. *Front. Psychol.* 6:658. doi: 10.3389/fpsyg.2015.00658

Talmy, L. (1991). "Path to realization: a typology of event conflation," in *Proceedings of the 17th Annual Meeting of the Berkeley Linguistics Society*, eds L. A. Sutton, C. Johnson and R. Shields (Berkeley, CA: Berkeley Linguistic Society), 480–519.

Vallacher, R. R., and Wegner, D. M. (1987). What do people think they're doing? Action identification and human behavior. *Psychol. Rev.* 94:3. doi: 10.1037/0033-295X.94.1.3

Vallacher, R. R., and Wegner, D. M. (1989). Levels of personal agency: individual variation in action identification. *J. Per. Soc. Psychol.* 57:660. doi: 10.1037/0022-3514.57.4.660

van Baaren, R. B., Maddux, W. W., Chartrand, T. L., DeBouter, C., and van Knippenberg, A. (2003). It takes two to mimic: behavioral consequences of self-construals. *J. Pers. Soc. Psychol.* 84, 1093–1102. doi: 10.1037/0022-3514.84.5.1093

Vinciarelli, M., Pantic, D., Heylen, C., Pelachaud, I., Poggi, F., D'ericco, M., et al. (2012). Bridging the gap between social animal and unsocial machine: a survey of social signal processing. *IEEE Trans. Affect. Comput.* 3, 69–87. doi: 10.1109/T-AFFC.2011.27

Vollmer, A.-L., Wrede, B., Rohlfing, K. J., and Cangelosi, A. (2013). Do beliefs about a robot's capabilities influence alignment to its actions? *Proc. EpiRob/ICDL* 2013, 1–6. doi: 10.1109/DevLrn.2013.6652521

Williams, L. E., Huang, J. Y., and Bargh, J. A. (2009). The scaffolded mind: higher mental processes are grounded in early experience of the physical world. *Eur. J. Soc. Psychol.* 39, 1257–1267. doi: 10.1002/ejsp.665

Winkielman, P., Niedenthal, P., Wielgosz, J., Eelen, J., and Kavanagh, L. C. (2015). "Embodiment of cognition and emotion," in *APA handbooks in psychology. APA handbook of personality and social psychology, Vol. 1. Attitudes and social cognition* eds M. Mikulincer, P. R. Shaver, E. Borgida and J. A. Bargh (Washington, DC: American Psychological Association), 151–175. doi: 10.1037/14341-004

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

PERSPECTIVE
published: 09 September 2015
doi: 10.3389/fpsyg.2015.01362

Investigating the ability to read others' intentions using humanoid robots

Alessandra Sciutti[1], Caterina Ansuini[1], Cristina Becchio[1,2] and Giulio Sandini[1]*

[1] Departments of Robotics, Brain and Cognitive Sciences, Istituto Italiano di Tecnologia, Genoa, Italy, [2] Department of Psychology, Centre for Cognitive Science, University of Torino, Torino, Italy

The ability to interact with other people hinges crucially on the possibility to anticipate how their actions would unfold. Recent evidence suggests that a similar skill may be grounded on the fact that we perform an action differently if different intentions lead it. Human observers can detect these differences and use them to predict the purpose leading the action. Although intention reading from movement observation is receiving a growing interest in research, the currently applied experimental paradigms have important limitations. Here, we describe a new approach to study intention understanding that takes advantage of robots, and especially of humanoid robots. We posit that this choice may overcome the drawbacks of previous methods, by guaranteeing the ideal trade-off between controllability and naturalness of the interactive scenario. Robots indeed can establish an interaction in a controlled manner, while sharing the same action space and exhibiting contingent behaviors. To conclude, we discuss the advantages of this research strategy and the aspects to be taken in consideration when attempting to define which human (and robot) motion features allow for intention reading during social interactive tasks.

Keywords: motor cognition, second-person interaction, contingency, kinematics, intention reading, human–robot interaction

OPEN ACCESS

Edited by:
Sebastian Loth, Universität Bielefeld, Germany

Reviewed by:
Jodi Forlizzi, Carnegie Mellon University, USA
Greet Van De Perre, Vrije Universiteit Brussel, Belgium

****Correspondence:***
Alessandra Sciutti, Departments of Robotics, Brain and Cognitive Sciences, Istituto Italiano di Tecnologia, Via Morego 30, Genoa 16163, Italy
alessandra.sciutti@iit.it

Specialty section:
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology

Received: *15 April 2015*
Accepted: *24 August 2015*
Published: *09 September 2015*

Citation:
Sciutti A, Ansuini C, Becchio C and Sandini G (2015) Investigating the ability to read others' intentions using humanoid robots. Front. Psychol. 6:1362. doi: 10.3389/fpsyg.2015.01362

Reading Intentions from Others' Movement

The ability to attend prospectively to others' actions is crucial to social life. Our everyday, commonsense capability to predict another person's behavior hinges crucially on judgments about that person's intentions, whether they act purposefully (with intent) or not, as well as judgments about the specific content of the intentions guiding others' actions – what they intend in undertaking a given action (Baldwin and Baird, 2001).

Humans rely on several sources to understand others' intention (**Figure 1**). For instance, by looking at the context of the surrounding environment we are often able to infer what is another person's intention. If a closed bottle of wine is on the table and a person reaches for a drawer, we guess that he is more probably looking for a bottle opener than for a fork. Under similar circumstances, the information provided by the context would allow an observer to constraint the number of possible inferences, thus facilitating the action prediction process (Kilner, 2011). But actions can also take place in contexts that do not provide sufficient information to anticipate others' intention. In such cases it has been demonstrated that others' gaze behavior may be a suitable cue to anticipate the intention to act (Castiello, 2003) as well as the specific goal of an action (Ambrosini et al., 2015). Moreover, there is a growing body of evidence indicating that, in

FIGURE 1 | An illustrative picture of human–robot interaction with the humanoid robot iCub. The mutual and spontaneous information exchange is mediated by context (i.e., the game on the touch screen that the two partners are playing) and by the agents' gazing behavior, but also by the intention information embedded in their movement properties. Copyright photo: Agnese Abrusci, Istituto Italiano di Tecnologia© IIT.

absence of gaze or contextual information, intentions can be inferred from body motion. But how is this possible?

How another agent moves can represent a cue to infer his intention because the way he moves is intrinsically related to his intention. In keeping with previous evidence (e.g., Marteniuk et al., 1987), recent studies have shown that in humans different motor intentions translate into different kinematics patterns (Ansuini et al., 2006, 2008; Sartori et al., 2011b). For instance, Ansuini et al. (2008) asked participants to reach for and grasp the very same object (i.e., a bottle) to accomplish one of four possible actions (i.e., pouring, displacing, throwing, or passing). Kinematic assessment revealed that when the bottle was grasped with the intent to pour, the fingers were shaped differently than in the other conditions. Further studies have extended these effects to the domain of social intention, reporting that not only the presence of a social vs. individual intention (Becchio et al., 2008b), but also the type of "social" intention (compete vs. cooperate) has an effect on action kinematics (Becchio et al., 2008a; see also Georgiou et al., 2007).

Recent evidence suggests that observers are sensitive to early differences in visual kinematics and can use them to discriminate between movements performed with different intentions (Vingerhoets et al., 2010; Manera et al., 2011; Sartori et al., 2011a; Stapel et al., 2012). For instance, Sartori et al. (2011a) tested whether observers use pre-contact kinematic information to anticipate the intention in grasping an object. To this end, they first analyzed the kinematics of reach-to-grasp movements performed with different intents: cooperate, compete against an opponent, or perform an individual action at slow or fast speed. Next, they presented participants with videos representative of each type of intention, in which neither the part of the movement after the grasping, nor the interacting partner, when present, were visible. The results revealed that observers were able to judge the agent's intent by simply observing the initial reach-to-grasp phase of the action.

The above findings suggest that intentions influence action planning so that different kinematic features are selected depending on the overarching intention. The observer is sensitive to this information and can use it to anticipate the unfolding of an action. Reading intention by observing movement therefore enables humans to anticipate others' actions, even when other sources of information are absent or ambiguous.

Research on the topic of understanding intention from movement has been traditionally the domain of psychology and neuroscience. However, there is growing interest in applying these ideas to computer vision, robotics, and human–robot interaction (e.g., Strabala et al., 2012; Shomin et al., 2014; Dragan et al., 2015). Unfortunately, the methodologies and paradigms currently used present important limitations. In the next sections, we will first briefly describe the methods traditionally applied to investigate this topic, and we will point out their potential shortcomings. Thereafter we will propose a new potential role for robots: before becoming anticipatory companions, robots could serve as suitable tools to overcome these limitations in research.

Barriers to Investigation of Intention-from-Movement Understanding

Reading intention from movement observation has been traditionally investigated with *video clips* used as stimuli. In these paradigms, for instance, temporally occluded goal-oriented actions are shown and the participant is asked to watch them and guess which is the actor's intention. This approach guarantees full control on the stimulation in all its aspects: timing,

information content, and perfect repeatability. Moreover, with video manipulation it is also possible to create behaviors that are impossible or unnatural, by modifying selectively relevant properties of the action. However, when looking at a video presentation, the subject is merely an observer, rather than a participant in the interaction. In other words, the use of videos eliminates some fundamental aspects of real collaborative scenarios, including the shared space of actions, the physical presence, the possibility to interact with the same objects and even potential physical contact between the two partners. Furthermore the video paradigm progresses in a fixed design and does not react to the action of the participant. It therefore precludes the possibility to build a realistic interactive exchange. Hence, the use of movie stimuli provides a fundamental way to investigate how others' actions and intentions are processed, but it should be used to complement other approaches that allow for actual interaction and contingent behavior.

More recently, the use of *virtual reality* systems has been proposed as a tentative solution to this problem. With this kind of settings it is possible to create virtual characters, or avatars, that respond contingently to participant's behavior (e.g., his gaze or his actions), while still maintaining the proper controllability of video stimuli. This type of methodology has strong potential, but it also has the limitation of detaching the participant from the real world. The resulting subject's behavior might then be affected not only by actions of the avatar, but also by being immersed in an environment that is not his everyday reality and which might not feature the same physical laws (e.g., gravity). Many aspects of our movements may derive from an optimization or a minimization of energy expenditure computed over life-long interaction with environmental constraints (e.g., Berret et al., 2008). Thus, removing the real environment from the equation could actually cause important changes in the performance of even simple interactions such as passing an object back and forth.

To summarize, the use of video stimuli allows full controllability, but it lacks of the possibility of contingent reaction and compromises the investigation of reading intention-from-motion in the context of a real interaction. On the other hand, virtual reality provides a certain degree of action contingency, but forces the participant to be immersed in a reality, that is different from his everyday experience. Thus, a new tool that goes beyond these limitations and allows an actual interaction with a high level of control is needed. In our opinion, the application of robots may meet these requirements. In the following, we propose a brief description of the main properties that would make robots, especially humanoids, a valuable instrument to investigate human ability to read intentions from others' movements.

Humanoid Robots as New Tool to Investigate Intention Understanding

Second-Person Interaction

As mentioned above, current paradigms investigating intention understanding are often based on a "spectator" approach to the phenomenon. However, social cognition differs in three important ways when we actively interact with others ('second-person' social cognition) compared to when we merely observe them ('third-person' social cognition; Schilbach and Timmermans, 2013). First, being involved in an interaction has an emotional component that is missing in a detached action observation setting (Schilbach and Timmermans, 2013). Second, it changes the perception of the environment, which is processed in terms of the range of possible actions of the two partners rather than those of the single participant (e.g., Richardson et al., 2007; Doerrfeld et al., 2012). Third, it is characterized by a higher flexibility, as the partners can adaptively change their actions during the interaction itself (e.g., Sartori et al., 2009). Robots provide the unique opportunity to investigate second-person social cognition, by engaging the participant in a face-to-face interaction without losing the controllability of the experiment or the shared environment. Although an experimenter or a human actor can be used as co-agent in a real interaction, the very fact that two people interacting influence each other in a complex way would easily result in behaviors that go beyond experimental control (see Streuber et al., 2011). Moreover, the automatic processes that constitute a great part of implicit communication (e.g., unintentional movements or gazing) are very difficult to restrain. As suggested by Bohil et al. (2011), "an enduring tension exists between ecological validity and experimental control" in psychological research. A robotic platform might provide a way out of this dilemma because it could sense the ongoing events and elaborate the incoming signals through its onboard sensors so to be able to react contingently to the behavior of the human partner, according to predefined rules.

Modularity of the Control

A further advantage of the use of robotic platforms relates to the possibility to isolate the contributions of specific cues that inform intention-from-movement understanding. When we observe other's actions, the incoming flow of sensory information provides multiple sources of evidence about the agent's goal, such as their gaze direction, arm trajectory, and hand pre-shape. The contribution of these factors in isolation is indicated by several empirical studies (e.g., Rotman et al., 2006; Manera et al., 2011). However, how these factors contribute together to mediate intention understanding remains unclear (Stapel et al., 2012; Furlanetto et al., 2013; Ambrosini et al., 2015). It is difficult in practice to separate and independently manipulate individual cues. For instance, the temporal dynamics of eye-hand coordination in a passing action or the relationship between the speed of a reaching movement and its accuracy are not independently planned by a human actor (see Ambrosini et al., 2015). Conversely, on a robot these aspects can be separated, distorted, or delayed, to assess the relative importance of each feature of the motion. For instance, we know that the unfolding of an action kinematics occurs within a specific temporal structure, e.g., the peak deceleration occurs at around 70–80% of a reach-to-grasp movement (Jeannerod, 1986). The robot allows the experimenter to selectively manipulate the time of peak deceleration to assess precisely which temporal deviations from human-like behavior could be tolerated by

an observer, without hindering the possibility to infer other's intentions.

Shared Environment

Robots are embodied agents, moving in our physical world, and therefore sharing the same physical space, and being subject to the same physical laws that influence our behavior. In contrast to virtual reality avatars, robots bring the controllability and contingency of the interaction into the real-world, where actual interaction usually occurs. Furthermore, robots with a humanoid shape have the advantage of being able to use the tools and objects that belong to a human environment and have been designed for human use. These properties make robots more adaptable to our common environments. In addition, the human shape and the way humans move are encoded by the brain differently with respect to any other kind of shape and motion (Rizzolatti and Sinigaglia, 2010). Consequently, humanoid platforms can probe some of the internal models naturally developed to interact with humans and allow studying exactly those basic mechanisms that make human–human interaction so efficient.

Necessary Robot Features to Investigate Human Ability to Read Intentions

When using a robot to investigate intention understanding in humans, some potential issues have to be considered. It could be objected, for instance, that the ability to anticipate others' intentions is strongly related to the properties of the human motor repertoire (Rizzolatti and Sinigaglia, 2010) and a robot does not exactly replicate the shape or the movements of human agent.

Although some researchers have succeeded in copying human appearance quite precisely (Nishio et al., 2007), human movement is indeed much harder to reproduce. This is due, for instance, to the materials and actuators with which robots are built, which are quite dissimilar from human elastic tissues and muscles, and to the complexity of human articulations. Still, entire research areas are devoted to build new robots that more closely resemble motor control and actuation of a human body (e.g., Kenshiro robot, Kozuki et al., 2013).

It is worth noticing that robotic platforms currently available offer interactive contexts in which robotic motion could be *sufficiently* similar to human motion. In this respect, investigation of reading intention-from-movement is particularly suitable for the use of humanoid robots, because it is traditionally focused on simple actions such as reaching to pass, grasping, transporting, or handing-over an object. This choice derives from the observation that most everyday collaborative behaviors are made of combinations of these simple acts. With this "vocabulary" as the focus of interest, it is possible to find existing robotic platforms that allow for human-like visuo-manual coordination, i.e., a control of gaze and manual actions that resembles that of a human (e.g., iCub, Metta et al., 2010, see **Figure 2**). Additionally, an approximate human-like shape, at least in the apparent humanoid structure of the robot body (e.g., torso, arm, hand, neck, head), might be required. This way humans can easily match their own bodily configuration with that of the robot and it is also simpler for experimenters to design robot behaviors approximating human motions both in end-effector and joint trajectories.

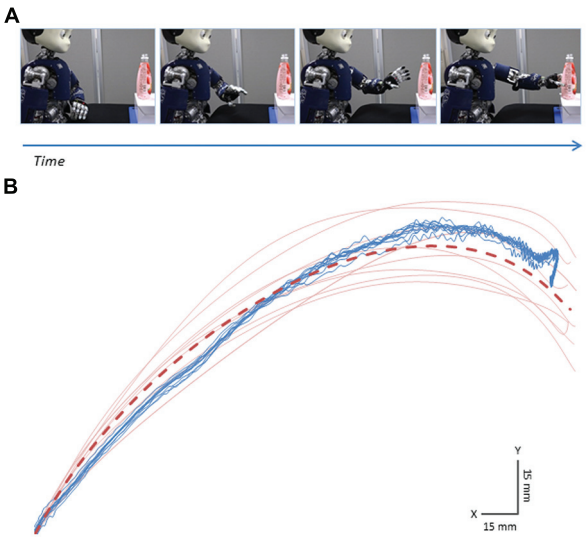


FIGURE 2 | It is possible to replicate simple movements with a humanoid robot that are *sufficiently* similar to those of a human actor. Here we show an example where the robot approximates a previously recorded human reaching. **(A)** Snapshots of the humanoid robot iCub reaching for a bottle with the aim to pour its content (i.e., pouring intention). **(B)** Sample trajectories of the palm of the hand on the horizontal (X) and vertical (Y) planes of the motion. Blue lines represent robot actions while red lines indicate human motions. Each line refers to a single movement. Data from ten trials are reported. It can be noticed that robot motion is highly repeatable and reflects quite accurately the average trajectory of the human action to be reproduced (dashed red line). Image by Oskar Palinko.

Since a robot is not an exact replica of a human, the doubts remain about whether a humanoid actually elicits in the human observers the same class of phenomena that are activated when they are observing a fellow human. A general answer to this question is not available yet (see Sciutti et al., 2012 for a review on the topic). However, there is some evidence suggesting that a humanoid robot exhibiting properly programmed motions can evoke the same automatic behavioral reactions as a human – at least in the context of the simple motions listed before.

One of these phenomena is the automatic anticipation of the action goal of another agent. Such prediction is associated to the activation of the observer's motor system (Elsner et al., 2013) and therefore does not occur when an object is self-propelling toward a goal position with the same predictable motion (Flanagan and Johansson, 2003). In an action observation task in which the humanoid robot iCub transported an object into a container, the observers exhibited a similar degree of automatic anticipation as for a human actor, suggesting that a comparable motor matching (and goal reading) occurred for both agents (Sciutti et al., 2013a). This result was replicated with another behavioral effect related to motor matching, namely automatic imitation (Bisio et al., 2014). When witnessing someone else performing an action,

humans spontaneously adapt their speed to that of their partner. It has been demonstrated that a similar unconscious adaptation occurs also after the observation of a humanoid robot action, but only if robot motion complies with the regularities of human biological motion. Additionally, humans process humanoid and human lifting actions in a similar manner. In line with this, it has been shown that observers are able to infer the weight of an unknown lifted object with the same accuracy both when looking at a human actor or at the iCub robot performing the lifting (Sciutti et al., 2013b, 2014). These results expand previous studies that showed that other behavioral phenomena associated to motor resonance (i.e., the activation of the observer's motor system during action perception) can generalize to humanoid robot observation, such as priming (Liepelt et al., 2010) and motor interference (Oztop et al., 2005).

Taken together, this evidence indicates that, as far as simple collaborative behaviors are concerned, humanoid robot actions are processed similarly to human actions and trigger a similar response in the human partners. Hence, using a humanoid robot as stimulus could give us insights not only about which mechanisms could facilitate human–robot interaction, but also about the laws subtending the dynamics of human–human interaction.

Conclusion

We predict that the use of robots as tools for investigating the phenomenon of reading intentions from movement observation will have a substantial impact not only on cognitive science research, but also from a technological standpoint. The tangible benefits for psychology and cognitive science of using humanoid robots to investigate intention reading consist in adding to the research the controllability of each single aspect of interaction (*modularity of control*), a property which is well beyond the possibilities of a human actor, while at the same time preserving a real reciprocity and involvement (*second-person interaction*), also in terms of space (*shared environment*). In turn, the possibility to have robots that move so as to seamlessly reveal their intents, would result in a more efficient, safe, and fluent human-robot collaboration. Indeed, by exploiting the same subtle kinematics signals that enable the timely and rich mutual understanding observed among humans, the implicit reading of robot intentions would happen naturally, with no need of specific training or instructions. Hence this line of research will allow us to build better, more interpretable robots and at the same time to deepen our understanding of the complex field of human–human interaction.

Acknowledgments

The research presented here has been supported by the CODEFROR project (FP7-PIRSES-2013-612555) – https://www.codefror.eu/. CA and CB were supported by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement n. 312919. Authors would like to thank Oskar Palinko for his help in figure and robot motion preparation and Matthew Tata for his help in proofreading the manuscript.

References

Ambrosini, E., Pezzulo, G., and Costantini, M. (2015). The eye in hand: predicting others' behavior by integrating multiple sources of information. *J. Neurophysiol.* 113, 2271–2279. doi: 10.1152/jn.00464.2014

Ansuini, C., Giosa, L., Turella, L., Altoè, G., and Castiello, U. (2008). An object for an action, the same object for other actions: effects on hand shaping. *Exp. Brain Res.* 185, 111–119. doi: 10.1007/s00221-007-1136-1134

Ansuini, C., Santello, M., Massaccesi, S., and Castiello, U. (2006). Effects of end-goal on hand shaping. *J. Neurophysiol.* 95, 2456–2465. doi: 10.1152/jn.01107.2005

Baldwin, D. A., and Baird, J. A. (2001). Discerning intentions in dynamic human action. *Trends Cogn. Sci.* 5, 171–178. doi: 10.1016/S1364-6613(00)01615-1616

Becchio, C., Sartori, L., Bulgheroni, M., and Castiello, U. (2008a). Both your intention and mine are reflected in the kinematics of my reach-to-grasp movement. *Cognition* 106, 894–912. doi: 10.1016/j.cognition.2007.05.004

Becchio, C., Sartori, L., Bulgheroni, M., and Castiello, U. (2008b). The case of Dr. Jekyll and Mr. Hyde: a kinematic study on social intention. *Conscious. Cogn.* 17, 557–564. doi: 10.1016/j.concog.2007.03.003

Berret, B., Darlot, C., Jean, F., Pozzo, T., Papaxanthis, C., and Gauthier, J. P. (2008). The inactivation principle: mathematical solutions minimizing the absolute work and biological implications for the planning of arm movements. *PLoS Comput. Biol.* 4:e1000194. doi: 10.1371/journal.pcbi.1000194

Bisio, A., Sciutti, A., Nori, F., Metta, G., Fadiga, L., Sandini, G., et al. (2014). Motor contagion during human-human and human-robot interaction. *PLoS ONE* 9:e106172. doi: 10.1371/journal.pone.0106172

Bohil, C. J., Alicea, B., and Biocca, F. A. (2011). Virtual reality in neuroscience research and therapy. *Nat. Rev. Neurosci.* 12, 752–762. doi: 10.1038/nrn3122

Castiello, U. (2003). Understanding other people's actions: intention and attention. *J. Exp. Psychol. Hum. Percept. Perform.* 29, 416–430. doi: 10.1037/0096-1523.29.2.416

Doerrfeld, A., Sebanz, N., and Shiffrar, M. (2012). Expecting to lift a box together makes the load look lighter. *Psychol. Res.* 76, 467–475. doi: 10.1007/s00426-011-0398-394

Dragan, A. D., Bauman, S., Forlizzi, J., and Srinivasa, S. S. (2015). Effects of robot motion on human-robot collaboration. *Proceedings* 15, 1921–1930. doi: 10.1145/2696454.2696473

Elsner, C., D'Ausilio, A., Gredebäck, G., Falck-Ytter, T., and Fadiga, L. (2013). The motor cortex is causally related to predictive eye movements during action observation. *Neuropsychologia* 51, 488–492. doi: 10.1016/j.neuropsychologia.2012.12.007

Flanagan, J. R., and Johansson, R. S. (2003). Action plans used in action observation. *Nature* 424, 769–771. doi: 10.1038/nature01861

Furlanetto, T., Cavallo, A., Manera, V., Tversky, B., and Becchio, C. (2013). Through your eyes: incongruence of gaze and action increases spontaneous perspective taking. *Front. Hum. Neurosci.* 7:455. doi: 10.3389/fnhum.2013.00455

Georgiou, I., Becchio, C., Glover, S., and Castiello, U. (2007). Different action patterns for cooperative and competitive behaviour. *Cognition* 102, 415–433. doi: 10.1016/j.cognition.2006.01.008

Jeannerod, M. (1986). The formation of finger grip during prehension. A cortically mediated visuomotor pattern. *Behav. Brain Res.* 19, 99–116. doi: 10.1016/0166-4328(86)90008-90002

Kilner, J. M. (2011). More than one pathway to action understanding. *Trends Cogn. Sci.* 15, 352–357. doi: 10.1016/j.tics.2011.06.005

Kozuki, T., Motegi, Y., Shirai, T., Asano, Y., Urata, J., Nakanishi, Y., et al. (2013). Design of upper limb by adhesion of muscles and bones - Detail human mimetic

musculoskeletal humanoid kenshiro. *IEEE Int. Conf. Intel. Rob. Syst.* 935–940. doi: 10.1109/IROS.2013.6696462

Liepelt, R., Prinz, W., and Brass, M. (2010). When do we simulate non-human agents? Dissociating communicative and non-communicative actions. *Cognition* 115, 426–434. doi: 10.1016/j.cognition.2010.03.003

Manera, V., Becchio, C., Cavallo, A., Sartori, L., and Castiello, U. (2011). Cooperation or competition? Discriminating between social intentions by observing prehensile movements. *Exp. Brain Res.* 211, 547–556. doi: 10.1007/s00221-011-2649-2644

Marteniuk, R. G., MacKenzie, C. L., Jeannerod, M., Athenes, S., and Dugas, C. (1987). Constraints on human arm movement trajectories. *Can. J. Psychol.* 41, 365–378. doi: 10.1037/h0084157

Metta, G., Natale, L., Nori, F., Sandini, G., Vernon, D., Fadiga, L., et al. (2010). The iCub humanoid robot: an open-systems platform for research in cognitive development. *Neural Netw.* 23, 1125–1134. doi: 10.1016/j.neunet.2010.08.010

Nishio, S., Ishiguro, H., and Hagita, N. (2007). "Geminoid: teleoperated android of an existing person," in *Humanoid Robots: New Developments*, ed. A. C. de Pina Filho (Vienna: I-Tech Education and Publishing), 343–352. doi: 10.5772/4876

Oztop, E., Franklin, D. W., and Chaminade, T. (2005). Human – humanoid Interaction: is a humanoid robot perceived as a human. *Int. J. Humanoid Robot.* 2, 537–559. doi: 10.1142/S0219843605000582

Richardson, M. J., Marsh, K. L., and Baron, R. M. (2007). Judging and actualizing intrapersonal and interpersonal affordances. *J. Exp. Psychol. Hum. Percept. Perform.* 33, 845–859. doi: 10.1037/0096-1523.33.4.845

Rizzolatti, G., and Sinigaglia, C. (2010). The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations. *Nat. Rev. Neurosci.* 11, 264–274. doi: 10.1038/nrn2805

Rotman, G., Troje, N. F., Johansson, R. S., and Flanagan, J. R. (2006). Eye movements when observing predictable and unpredictable actions. *J. Neurophysiol.* 96, 1358–1369. doi: 10.1152/jn.00227.2006

Sartori, L., Becchio, C., Bulgheroni, M., and Castiello, U. (2009). Modulation of the action control system by social intention: unexpected social requests override preplanned action. *J. Exp. Psychol. Hum. Percept. Perform.* 35, 1490–1500. doi: 10.1037/a0015777

Sartori, L., Becchio, C., and Castiello, U. (2011a). Cues to intention: the role of movement information. *Cognition* 119, 242–252. doi: 10.1016/j.cognition.2011.01.014

Sartori, L., Straulino, E., and Castiello, U. (2011b). How objects are grasped: the interplay between affordances and end-goals. *PLoS ONE* 6:e025203. doi: 10.1371/journal.pone.0025203

Schilbach, L., and Timmermans, B. (2013). Toward a second-person neuroscience. *Behav. Brain Sci.* 36, 393–414. doi: 10.1017/S0140525X12000660

Sciutti, A., Bisio, A., Nori, F., Metta, G., Fadiga, L., Pozzo, T., et al. (2012). Measuring human-robot interaction through motor resonance. *Int. J. Soc. Robot.* 4, 223–234. doi: 10.1007/s12369-012-0143-141

Sciutti, A., Bisio, A., Nori, F., Metta, G., Fadiga, L., and Sandini, G. (2013a). Robots can be perceived as goal-oriented agents. *Interact. Stud.* 14, 1–31. doi: 10.1075/is.14.3.02sci

Sciutti, A., Patanè, L., Nori, F., and Sandini, G. (2013b). "Do humans need learning to read humanoid lifting actions?," in *Proceedings of the Conference 2013 IEEE 3rd Joint International Conference on Development and Learning and Epigenetic Robotics, ICDL 2013 – Electronic*, Osaka. doi: 10.1109/DevLrn.2013.6652557

Sciutti, A., Patanè, L., Nori, F., and Sandini, G. (2014). Understanding object weight from human and humanoid lifting actions. *IEEE Trans. Auton. Ment. Dev.* 6, 80–92. doi: 10.1109/TAMD.2014.2312399

Shomin, M., Vaidya, B., Hollis, R., and Forlizzi, J. (2014). "Human-approaching trajectories for a person-sized balancing robot," in *Proceeding of the IEEE International Workshop on Advanced Robotics and Its Social Impacts*. Evanston, IL. doi: 10.1109/arso.2014.7020974

Stapel, J. C., Hunnius, S., and Bekkering, H. (2012). Online prediction of others' actions: the contribution of the target object, action context and movement kinematics. *Psychol. Res.* 76, 434–445. doi: 10.1007/s00426-012-0423-422

Strabala, K., Lee, M. K., Dragan, A., Forlizzi, J., and Srinivasa, S. S. (2012). Learning the communication of intent prior to physical collaboration. *Proc. IEEE Int. Work. Robot Hum. Interact. Commun.* 968–973. doi: 10.1109/ROMAN.2012.6343875

Streuber, S., de la Rosa, S., Knoblich, G., Sebanz, N., and Buelthoff, H. H. (2011). The effect of social context on the use of visual information. *Exp. Brain Res.* 214, 273–284. doi: 10.1007/s00221-011-2830-9

Vingerhoets, G., Honoré, P., Vandekerckhove, E., Nys, J., Vandemaele, P., and Achten, E. (2010). Multifocal intraparietal activation during discrimination of action intention in observed tool grasping. *Neuroscience* 169, 1158–1167. doi: 10.1016/j.neuroscience.2010.05.080

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 25 August 2015
doi: 10.3389/fpsyg.2015.01253

Robot Comedy Lab: experimenting with the social dynamics of live performance

*Kleomenis Katevas *, Patrick G. T. Healey and Matthew Tobias Harris*

Cognitive Science Research Group, School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK

The success of live comedy depends on a performer's ability to "work" an audience. Ethnographic studies suggest that this involves the co-ordinated use of subtle social signals such as body orientation, gesture, gaze by both performers and audience members. Robots provide a unique opportunity to test the effects of these signals experimentally. Using a life-size humanoid robot, programmed to perform a stand-up comedy routine, we manipulated the robot's patterns of gesture and gaze and examined their effects on the real-time responses of a live audience. The strength and type of responses were captured using SHORE™ computer vision analytics. The results highlight the complex, reciprocal social dynamics of performer and audience behavior. People respond more positively when the robot looks at them, negatively when it looks away and performative gestures also contribute to different patterns of audience response. This demonstrates how the responses of individual audience members depend on the specific interaction they're having with the performer. This work provides insights into how to design more effective, more socially engaging forms of robot interaction that can be used in a variety of service contexts.

Keywords: human robot interaction, affective computing, humanoid robots, live performance, social signals

OPEN ACCESS

Edited by:
Sebastian Loth, Universität Bielefeld, Germany

Reviewed by:
Candace L. Sidner, Worcester Polytechnic Institute, USA
Martina Mara, Ars Electronica Futurelab, Austria

****Correspondence:***
Kleomenis Katevas, Cognitive Science Research Group, School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End Road, London E1 4NS, UK
k.katevas@qmul.ac.uk

Specialty section:
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology

Received: *15 March 2015*
Accepted: *05 August 2015*
Published: *25 August 2015*

Citation:
Katevas K, Healey PGT and Harris MT (2015) Robot Comedy Lab: experimenting with the social dynamics of live performance. Front. Psychol. 6:1253. doi: 10.3389/fpsyg.2015.01253

1. Introduction

Not everyone knows how to tell a joke. A good joke depends as much on the quality of the delivery as it does on the quality of the material. Intonation, posture, gaze, gesture, expression, and timing all contribute to successful comic delivery. Moreover, effective delivery is not just a matter of what the speaker does, it also depends on the reciprocal dynamics of the speaker–listener interaction. The fluency of speakers' performance in conversation depends on the moment-to-moment responsiveness of their audience and, in turn, on the speakers' ability to concurrently accommodate and adjust to these responses while they are speaking (Goodwin, 1979; Bavelas et al., 2000). If addressees appear to be bored or distracted, speakers become disfluent. Conversely, an appropriately timed smile or raised eyebrow by an addressee provides useful feedback that speakers can use to adapt their message.

Our basic hypothesis is that these interactional dynamics should be just as important to the mass interaction involved in performing in front of a live comedy audience as they are to telling a joke to a friend. Ethnographic studies of stand-up comedy and street performance support this idea. Gardair (2013) demonstrated the pervasive relevance of interaction to the achievement of a successful street performance. Street performances are actively established and managed using

patterns of interaction similar to those used to establish and maintain conversational clusters or "F-formations" (Kendon, 1990). Street performers use variations of body position, orientation and gaze to manage engagement and define the performance space (Gardair et al., 2011). They invest considerable effort in orchestrating, explicitly eliciting and in some cases actively training the audiences' responses. This process appears to be key to the development of a collective sense of audience membership and often takes up more than 90% of the performance time. It also appears to play a critical role in obtaining money from the audience (Gardair, 2013).

Rutter (1997, 2000) argued that stand-up comedy is also defined by interaction: the performance is an interactive organization and delivery of material constantly informed by audience responses; for an audience, becoming involved in the developing flow of the act engenders not just an active and responsive manner but one where all can be held to account. We assume that it is these interactional processes that contribute to the distinction that performers make between "good" and "bad" audiences for the same performance. Furthermore, it is the same processes that underpin people's experience of moments of "crackle", "movement" and "lift", or "drop" and "drift" that are part of the practical language of performance (Healey et al., 2009).

Embodied robots provide a unique opportunity to experiment with these interactional processes by enabling the introduction of controlled manipulations directly into a live performance. Although robots have the disadvantage of eliciting responses that may be different from those of a human performer, they can hold the "content" of the routine constant (e.g., the prosody, semantics and syntax of the jokes) while selectively manipulating aspects of delivery (e.g., body orientation, gaze, and gesture). This strategy of using embodied robots as tools for human interaction experiments has precedents in work by MacDorman and Ishiguro (2006), Sidner et al. (2005), and Knight and Simmons (2013). These studies use robots to experiment with different aspects of overt robot behavior, including gaze and gesture, as a means of probing the detailed organization of social interaction. This enables direct comparisons of the effects of different behaviors on interaction and can provide a principled basis on which to design robots that can engage successfully with humans.

Previous work has also specifically made use of embodied robots to tell jokes. Hayashi et al. (2005) created a robot–robot dialogue system so that two robots could enact Japanese "Manzai" routines in front of an audience. Although the robot's movements were scripted, the timing of their jokes was sensitive to audience responses. Sjöbergh and Araki (2009) used the Robovie-i platform to show that the same jokes delivered by a robot are rated as funnier on average than when delivered in text form only. They also showed a larger effect of robot responses (positive or negative) on perceived funniness of jokes. Knight et al. (2011) used the Nao robot to explore how the choice of jokes from a larger repertoire could be customized according to the strength of audience responses. These studies effectively held the non-verbal delivery of each joke constant.

Here we use a robot to explore specific non-verbal elements of performer-audience interaction in comic delivery. In order to motivate the choice of experimental manipulations we briefly describe a pilot study of a stand-up comedy performance. Building on the observations from this study and previous ethnographic work on performer-audience interaction, we describe the "Comedy Parser" system we developed to support performative gaze and gestures in a commercial robot platform (Katevas et al., 2014). The impact of these manipulations was analyzed in a live performance experiment conducted over two nights at the Barbican Centre in London. As far as we are aware this is the first attempt to use a robot to probe the moment-by-moment, embodied aspects of how stand-up comedians "work" an audience.

2. Comic Observations

A pilot observational study was made using video data taken from a "Comedy Lab" hosted in the Performance Lab at Queen Mary University of London. The aim of this study was to extend previous ethnographic observations of performer-audience interaction in street performance (Gardair et al., 2011; Gardair, 2013) to the specific context of stand-up comedy. In particular, to get a more detailed sense of some of the elements of non-verbal delivery required from a robot to "read" and respond to an audience.

The Comedy Lab session featured live stand-up performances by two professional comedians: Tiernan Douieb (compère) and Stuart Goldsmith (main act), with 25 audience participants recruited through social media channels. Douieb provided a 5 min "warm-up" and then introduced Goldsmith who did a 15 min set (**Figure 1**). Full-HD audio-visual recordings were made using fixed cameras approximating the audience's view of the performer and the performer's view of the audience. All data collection and analysis was made with informed consent and approved by the Queen Mary University of London research ethics committee (Reference: QMREC1199b).

The video recordings of the performer and audience were imported into ELAN, a multimedia annotation tool (Wittenburg et al., 2006). A simple qualitative analysis of the performer's and audiences' use of non-verbal signals was made using multiple passes over the tape using ELAN to control speed of playback and to code significant events.

FIGURE 1 | Comedy Lab with Stuart Goldsmith, at Queen Mary University of London.

2.1. Comic Delivery

Patterns of performer-audience interaction are complex and a detailed analysis is outside the scope of the current paper, however several observations are relevant to the robot behaviors described below. First, the gaze of the performer was predominately either to the floor or to an audience individual. Second, the performer's gaze tended to shift at the end of every sentence, and sometimes between phrases, in a pattern similar to that observed in conversation where speakers use gaze to elicit responses from their addressees (Kendon, 1967; Goodwin, 1979). The performer would also focus their gaze on an audience participant accompanied by a pointing gesture and reference to the participant in the talk, making it clear that a specific audience member was being addressed.

Punchlines were typically distinguished by faster delivery and a short pause and change of gaze. In such cases, the change of gaze was always onto an audience member. In addition, the performer usually followed their punchlines with a smile and sometimes laughter. Resumption following a punchline appeared to be primarily contingent on the audience response. The duration of pause after the punchline was determined by whether and how quickly laughter ensued. If the laughter was significant, the performer would remain silent until the laughter started to subside—a pattern also noted for Japanese Manzei performances by Hayashi et al. (2005). Audience laughter was marked not only by facial displays but also large visible body movements of the head and upper body. Audience members directed their gaze mostly at the performer, occasionally to each other or the floor.

As with street performance, the comedians also occasionally used large gestures (**Figure 1**) designed to promote a stronger or more prolonged audience response similar to the applause elicitation gestures described by Gardair (2013). Another interesting shared feature with street performance is the stand-up comics' use of explicit commentaries on the character of audience responses as a way to generate more active engagement, e.g., complaining about isolated or weak laughter (Gardair et al., 2011; Gardair, 2013).

Drawing on these findings our experiment manipulated the robot's gaze at audience members and the production of specific performative gestures (see Section 3.3.2) and then assesses their impact on the audience responses.

3. Experimental Study

3.1. Study Overview and Predictions

"Comedy Lab: Human vs. Robot" was conceived as a performance experiment, carried out in an arts venue in front of a live audience (**Figure 2**). The basic rationale was to use a robotic performer to perform a predetermined comedy script while different aspects of the delivery were manipulated and live audience responses gathered for analysis.

Our first manipulation involves gaze: we dynamically allocated different audience members as gaze targets for the robot during the performance. Although it seems intuitive that people simply smile when they are happy, displays of positive affect are conditioned by social context. Following Bavelas et al. (1986) we assumed that even in the somewhat anonymized context of a live performance, the overt responses and facial expressions produced by audience members are communicative displays designed for specific recipients. This is a relatively strong assumption because we are proposing that the principal recipient of the audiences' facial displays in this context is the robot. This lead to the prediction that audience members should display more positive affect when they believe the robot is attending to them and less when they believe it is not. Note that this is independent of how funny they find the jokes themselves.

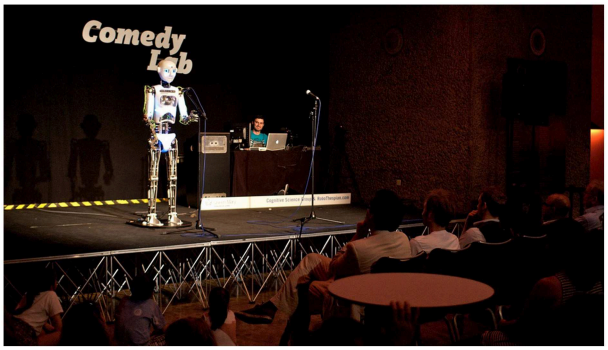


FIGURE 2 | Comedy Lab with RoboThespian^{TM} at the Barbican Centre in London.

Our second manipulation involves gesture: a number of special gestures were programmed as exceptions to the default "canned" movements delivered by the robot platform. Drawing on our pilot observational study (Section 2.1) and observational studies of street performances (Gardair, 2013), we opted to test the effects of four specific gestures that appeared to be designed to elicit positive audience responses (illustrated in Section 3.3.2). The first of these was a raised arm "welcome" gesture, the second an "emphasis" gesture, the third a pointing gesture, and the fourth an applause eliciting gesture. If these gestures are effective in promoting stronger engagement this should be evident in their effects on measures of positive affect in the audience responses.

We also tested our basic assumption that the experiment succeeds in creating a credible stand-up performance by assessing whether the jokes themselves, written by the compère Douieb, elicit positive responses. Although not central to the questions about delivery that are the main concern of this paper, this is an important issue for the validity of the study. It also provides a test of whether people responded to the specifics of the performance rather than adopting a generic positive (or sceptical) attitude due to the novelty of seeing a robot performing. Other studies showed robot performers may elicit similar or stronger positive responses than their human counterparts (Hayashi et al., 2005).

3.2. Materials and Methods

Before proceeding to describe the design of the study we first introduce the computer vision software and robot platform that were used in this study.

3.2.1. Measures of Audience Response

To obtain fine-grained real-time response measures and automatic measures of facial display and position, we used

sentiment analysis techniques developed in computer vision research. Fraunhofer SHORE™ (Sophisticated High-speed Object Recognition Engine) was selected for this purpose.

Provided with video imagery, SHORE™ detects faces within each frame and provides properties for each of them. In our tests we found SHORE™ able to detect audience faces when seated under low (but under our control) lighting conditions and filmed from the front, and able to do so in real-time. The properties the software produces for each identified face, and so makes available for experimental measures, are the following:

- Location of the face in the space.
- Position of the eyes, nose and mouth.
- Gender classification ("Male", "Female", or "Unknown").
- Age estimation in years.
- Facial expression recognition, expressed as percentages of "Happy", "Sad", "Angry", and "Surprised".
- Identify whether the eyes are open or closed.
- Identify how much the mouth is open.
- Detection of up to 60° of face rotation.

Most of the above features have been validated using external data sets (Ernst et al., 2009). The face detection has been validated using the CMU+MIT data sets and showed good accuracy relative to other classification methods (91.5% detection rate with a 1 in 10 miss rate). The gender classification has been validated using the BioID data set (94.3% recognition rate) as well as the Feret fafb data set (92.4% recognition rate). Finally, the happiness analyzer has been validated on the JAFFE data base (95.3% recognition rate). Note that none of these test datasets were used as training sets for the framework. Further information can be found on the Fraunhofer IIS website: http://iis.fraunhofer.de.

3.2.2. The Robot Platform

RoboThespian™ is a humanoid robot designed for interaction in public places created by Engineered Arts Ltd. (see **Figures 2**, **4**). Following a human body model, it consists of a robotic head, two arms with hands, the robot's torso as well as the two legs. The robotic head has two rectangular LCD screens for eyes as well as embedded LED lighting in the cheeks that allows it to make facial expressions as it talks. The mouth can only move vertically, a process that is automated and synchronized with the speech engine. The two arms and hands can move fast and fluently, while the torso's movement is relatively limited and slow. The robot cannot walk by itself as it only has passive leg movement. It also uses the Acapela Text-To-Speech engine from Acapela Group Babel Technologies SA, providing voice synthesis in customizable voices, as well as control over speed, timing, volume, and shape (sound pitch).

To provide the robot platform with the capabilities required for the experiment we built a "Comedy Parser" system that controls the robotic behavior including its interaction with the audience (Katevas et al., 2014). Provided with a specially marked-up script, it delivers the content while enacting the behaviors described in Section 2. It uses SHORE™ computer vision software to analyze the audience in real-time and identify each person's location in the space as well as to capture characteristics such as gender, age and moment-by-moment display of "happiness". The complete source-code, licensed under an MIT License, is available at https://github.com/minoskt/ComedyParser.

3.2.3. Procedure

Two performances were staged as part of the "Hack the Barbican" event at 6 p.m. on 7th and 8th of August 2013 at the Barbican Centre in London. The club stage that was used is freely accessible to the public. Each performance comprised the compère's warm up, and the (human) comedian's act followed by the robot act. The compère's warm-up lasted approximately 10 min, the comedian's lasted 13 min, and the robot's 8 min. Two professional stand-up comedians, Tiernan Douieb and Andrew O'Neill, were recruited for the compère and comedian roles, respectively. This format was used both to widen the appeal of the event and to help create a more convincing stand-up comedy context. Although the compère and comedian made normal stage entries and exits the robot cannot walk and therefore its position and the control desk were fixed throughout (see **Figure 2**). During the robot's performance, an experimenter monitored the control equipment, visible to the side at the rear of the stage.

Figure 3 shows the configuration of the staging, with the seat placement, the position of the performers, as well as the position of the two speakers, the tracking camera and the three directional microphones. The tracking camera was an inconspicuous Gig-E Vision camera positioned high at the back of the stage with a field of view that encompassed the seated area.

An audio-visual recording of each performance was obtained by placing an HD video camera toward the back of the audience area. SHORE™ software analyzed video imagery from the tracking camera and passed the output to the Comedy Parser. Comedy Parser also archived all dynamic aspects of the performance, in particular the robot's gaze and point.

Participants were informed that they were being captured on video for research purposes and all data capturing and handling

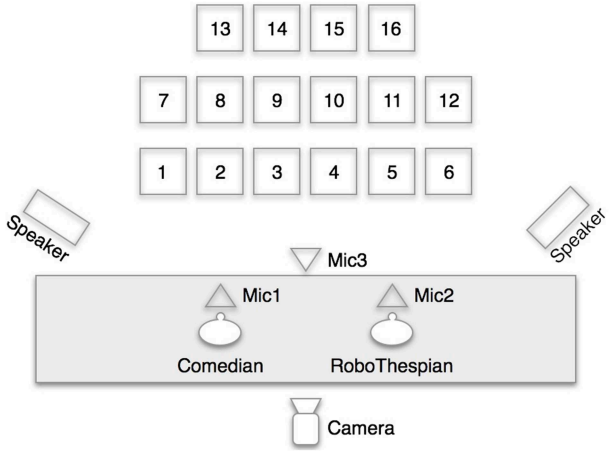


FIGURE 3 | Barbican Comedy Lab configuration.

procedures were audited by the Queen Mary University of London research ethics committee (Reference: QMREC1199b).

3.2.4. Participants

Audience participants were recruited by advertising "Comedy Lab" through social media channels of the two performers, the venue (The Barbican Centre, London), research group (Cognitive Science, Queen Mary University of London) and "Hack the Barbican". The following context was provided in the advert:

> What makes a good performance? By pitting stand-up comics Tiernan Douieb and Andrew O'Neill against a life size robot in a battle for laughs, researchers at Queen Mary University of London hope to find out more— and are inviting you along.
>
> A collaboration between the labs of Queen Mary's Cognitive Science Research Group, RoboThespian's creators Engineered Arts, and the open-access spaces of Hack The Barbican, the researchers are staging a stand-up gig where the headline act is a robot as a live experiment into performer-audience interaction.
>
> This research is part of work on audience interaction being pioneered by the Cognitive Science Group. It is looking at the ways in which performers and audiences interact with each other and how this affects the experience of "liveness". The experiment with RoboThespian™ is testing ideas about how comedians deliver their material to maximize comic effect.

Approximately 50 people attended each performance on each night. Data from SHORE™ were captured for 22 people for the first night (15 men and 7 women between the ages of 28 and 64 years, $M = 46.4$, $SD = 8.0$) and 19 for the second night (13 men and 6 women between the ages of 27 and 60 years, $M = 46.2$, $SD = 8.1$).

3.3. Results

The measures of "Happiness", "Anger", "Surprise", and "Sadness" produced by SHORE™ showed substantial inter-correlations. For example, in our data "Happiness" and "Sadness" are negatively correlated for: Pearson's $r_{(121)} = -0.484, p < 0.01$ (Note: $N = 121$ because there are three measures for each of 48 people corresponding to Before, During and After a punchline, discussed in more detail below); and so are "Happiness" and "Anger": Pearson's $r_{(121)} = -0.433, p < 0.01$. These correlations make these measures partially redundant and we therefore report results only for the "Happiness" measure in the following analysis.

Throughout we report computed probabilities for completeness but adopt a criterion level of $p < 0.05$ for inferences. We use Generalized Linear Mixed Model (GLMM) analyses to model the combined random effects, categorical and interval fixed effects and repeated measures involved in the audience responses measured in this study.

3.3.1. Punchlines

To test if audience members respond selectively to the jokes, their facial displays of "Happiness" were averaged over three "Response Phases": "Before", "During", and "After" defined as, 2 s before the punchline, the duration of the punchline delivery and 2 s after.

Average "Happiness" displayed by the audience was analyzed in a GLMM using a Linear Model. This treated Response Phase (Before/During/After) as a fixed factor and Audience Member nested within Night as random factors. It shows a main effect of Response Phase [$F_{(2,\ 120)} = 5.66, p < 0.01$]. Planned, pairwise comparisons show that people displayed more happiness after the punchlines than before them [$t_{(120)} = 3.32, p < 0.01$] or during them [$t_{(120)} = 2.67, p = 0.01$] but no difference in displayed happiness before and during the punchlines [$t_{(120)} = -0.86, p = 0.39$]. The estimated means and standard errors are summarized in **Table 1**. Fixed (B) coefficients provide estimates of effect size: After = 2.3, (95% CI lower = 0.6, upper = 4.0); Before = −0.59, (95% CI lower = −1.9, upper = 0.7). *During* is the reference category.

3.3.2. Gestures

During each performance, RoboThespian™ used four specific performative gestures. Due to timing issues, the first "welcome" gesture (**Figure 4A**) was not obvious to the participants as they were still applauding, welcoming RoboThespian™ on stage. Consequently this gesture is excluded from the analysis. The following three gestures are analyzed:

1. *Gesture B:* A reprise "I said hello" gesture that emphasizes the expected return of greetings suggested by Gardair's (2013) analysis of street performances (**Figure 4B**).
2. *Gesture C:* A pointing gesture while saying "you go first", inspired by our observational study of stand-up comedy (**Figure 4C**).
3. *Gesture D:* The applause elicitation gesture "Thank you, and good night," inspired by Gardair's (2013) analysis of street performances (**Figure 4D**).

As the gestures are qualitatively different in their effects we analyze them separately.

3.3.2.1. Gesture B

A GLMM Linear Model analysis of average displayed Happiness in response to Gesture B with Response Phase (Before vs. During vs. After) as a fixed factor and Night (1 vs. 2) and Audience Member as random factors showed no main effect of Response Phase [$F_{(2,\ 99)} = 1.63, p = 0.20$]. Planned pairwise comparisons also showed no difference between the different response phases {Estimated Means: Before = 37.0, During = 42.6, After = 41.4; Pairwise Comparisons: Before vs. During: $t_{(99)} = -1.71, p =$

TABLE 1 | Estimated means and standard errors for "Happiness" before, during and after punchlines.

Response phase	Estimated mean	Std. error
Before	44.2	3.17
During	44.8	3.13
After	47.1	3.12

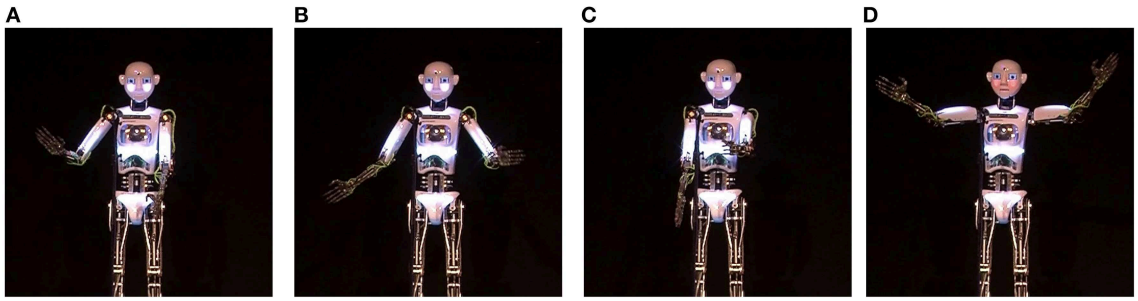


FIGURE 4 | Performative Gestures used during the live performance. (A) "Welcome" gesture, **(B)** Reprise "I said hello" gesture, **(C)** Pointing gesture, **(D)** Applause elicitation gesture.

TABLE 2 | Estimated means and standard errors for "Happiness" before, during and after execution of Gesture C.

Response phase	Estimated mean	Std. error
Before	42.3	4.3
During	52.2	4.1
After	51.1	4.0

TABLE 3 | Estimated means and standard errors for "Happiness" before, during and after robot gaze.

Response phase	Estimated mean	Std. error
Before	45.1	4.3
During	45.1	4.3
After	42.1	4.3

0.09, Before vs. After = $[t_{(99)} = -1.34, p = 0.18]$, During vs. After $[t_{(99)} = 0.34, p = 0.73]$}.

3.3.2.2. Gesture C

A GLMM analysis with the same factors as above showed a different pattern of responses. For Gesture C there was a significant main effect of Response Phase $[F_{(2, 106)} = 6.11, p < 0.01]$. The estimated means are provided in **Table 2**. Pairwise comparisons show that displayed happiness increased during and immediately after the production of Gesture C but were not reliably different while the gesture was produced and immediately after: Before vs. During: $t_{(106)} = -3.23, p < 0.01$, Before vs. After = $[t_{(106)} = -3.1, p < 0.01]$, During vs. After $[t_{(106)} = 0.44, p = 0.66]$. Fixed ($B$) coefficients: Before = −8.7, (95% CI lower = −14.4, upper = −3.14); During = 0.44, (95% CI lower = −3.8, upper = 6.0). After is redundant.

3.3.2.3. Gesture D

The parallel GLMM analysis for Gesture D shows no main effect of Response Phase: $F_{(2, 92)} = 2.13, p = 0.13$. Planned pairwise comparisons showed no reliable differences between the three response phases: Before vs. During $t_{(92)} = -1.11, p = 0.27$; Before vs. After $t_{(92)} = 0.44, p = 0.66$ During vs. Before $t_{(92)} = 1.12, p = 0.27$.

Overall, only Gesture C produced a reliable positive response. The three different Gestures are, of course, designed to achieve different effects. Gesture B is designed primarily to prompt applause and cheering. Gesture C works to underline the point of a joke and responses seem to be closely tied to the timing of the gesture delivery. For Gesture D the aim is to elicit applause. Unfortunately we do not have robust quantitative measures of these different responses.

3.3.3. Gaze

A total of 14 unique individuals in the audience were randomly fixated a total of 153 times by the robot over the two nights. Three people were fixated only once and are excluded from the analysis.

The effect of Gaze on displayed "Happiness" is analyzed in a GLMM linear model with Audience Member as a random factor and Gaze Phase (2 s Before, During and 2 s After) a fixed factors. The robot's fixation points were not exact so the distance in pixels between a participant's actual location in the video and the fixation point of the robot is included as a co-variate. This analysis shows a main effect of Gaze Phase $[F_{(2, 238)} = 14.5, p < 0.01]$ and a main effect of Distance $[F_{(1, 242)} = 5.19, p < 0.05]$. The estimated means are provided in **Table 3**. Planned pairwise comparisons of Gaze Phase show no difference in the fixated person's displayed "Happiness" before and during the fixation [Before vs. During $t_{(238)} = 0.02, p = 0.99$] but a significant drop afterwards [After vs. Before $t_{(238)} = -4.68, p < 0.01$, After vs. During $t_{(238)} = -4.96, p < 0.01$]. Fixed ($B$) coefficients for Gaze Phase: Before = 5.8, During = 5.8, After = 3.2.

The fixed coefficient (B) for the distance co-variate of −0.12 additionally showed that the further an audience member's face was from the center of the robot's fixation point the lower the estimated facial display of "Happiness".

3.3.4. Human vs. Robot

Although direct comparison of the human and robot performers was not part of the original study design (and is in some respects problematic see Section 4 below) it is interesting to compare the overall "Happiness" response evoked by the compère, comedian and robot.

A *post-hoc* GLMM linear model analysis of average percentage happiness of each audience member on each night with Performer (Compère vs. Human Comedian vs. Robot) as a fixed

factor and Audience Member and Night as a random factors shows a main effect of Performer [$F_{(2,\ 227)} = 9.37, p < 0.01$]. Planned pairwise comparisons show people responded more positively to the human comedian than the compére [$t_{(227)} = 4.33, p < 0.01$] but no other comparisons were significant [Compère vs. Robot $t_{(227)} = 1.85, p < 0.1$; Comedian vs. Robot $t_{(227)} = 1.37, p < 0.2$]. As **Table 4** shows, people's responses to the Robot were essentially intermediate between the two human performers.

4. Discussion

At the broadest level these results demonstrate the viability of using embodied robots to study the social dynamics of live performance. The ability to make controlled, fine-grained manipulations of gaze and gesture while holding other aspects of performance constant creates exciting possibilities for future research that go beyond what is possible using human confederates; people are simply unable to selectively control their own performances to the same degree as a robot (Kuhlen and Brennan, 2013). Balanced against this are the issues that arise from the fact that the performer is plainly a robot.

Anecdotally, our observation and personal discussions with people afterwards suggested that audience members on the two nights of Comedy Lab found the jokes generally amusing despite the restricted prosody and cadence of the robot platform's speech synthesis. However, audience responses might have been biased by the novelty of the situation. For example, Hayashi et al. (2005) provided evidence that people are more sympathetic to a robot comedian than a human comedian, although in this work a live robot performance was compared with a recorded human performance. Audience bias might also run in the opposite direction; our audience was explicitly recruited for a robot vs. human "Comedy Lab" and contained some journalists and people with a technical interest in robotics. Consequently, they might be atypical of a stand-up comedy audience and more interested in the technical than the comic material. We note that both comedians said informally that they found the audience harder to engage than a typical stand-up comedy club.

The present study does not provide data that enables us to assess audience bias directly. The finding that people responded as positively to the robot as they did to the human stand-up suggests that any potential bias was limited. However, we note that this comparison is confounded by differences in, amongst other things, staging, materials and delivery. We avoid this problem here by focusing our analysis on the comparison of audience responses to the robot before, during and after the specific manipulated behaviors. This allows us to broadly discount the potential influence of people's generic dispositions toward robot performers; positive or negative.

TABLE 4 | Estimated means and standard errors for "Happiness" response to each performer.

Response phase	Estimated mean	Std. error
Compére	38.2	4.1
Comedian	45.6	4.0
Robot	42.6	4.1

Importantly, the results showed that audience responses are closely co-ordinated with the delivery of the punchlines and robot gaze. Specifically, displays of positive affect peaked just after the punchlines but declined after gaze. This showed that people were selectively responsive to both the content and delivery. Audience members appeared to be particularly sensitive both to whether the robot was looking at them and to the specific angle of the robot's gaze. The more closely they were fixated by the robot, the more positive affect they displayed. The robot was treated as a social agent that successfully elicited social response patterns typical of human interactions. This finding supports our hypothesis that performers use gaze as a means of eliciting audience responses (Kendon, 1967). It is also consistent with prior work that has noted the importance of social gaze in storytelling performances by embodied robots (Mutlu et al., 2006) and humans (Goodwin, 1979).

The pattern of results for the manipulated gestures is less clear. Only the pointing gesture (see **Figure 4C**) had a statistically significant effect. There may be several reasons for this. It might be due to a lack of measures appropriate to each gesture, e.g., the emphasis gesture may have caused a louder or more emphatic response that would not necessarily show up in the measures of facial affect. It might be due to problems in the execution of the gestures that made them difficult to interpret or it might be that the gestures simply do not function in the way we expected.

Overall, the results demonstrate a fine-grained link between specific aspects of delivery and specific audience responses. This is consistent with the general hypothesis that part of what underpins the experience of live performance is the social dynamics of audience-performer interactions (Rutter, 2000; Gardair et al., 2011; Gardair, 2013). As noted in the introduction, laughter, smiles and other displays of affect are themselves performances designed for audiences including our conversational partners (Kraut and Johnston, 1979; Bavelas et al., 1986; Fernández-Dols and Ruiz-Belda, 1995). The data presented here show how this use of displays of affect extends to live comedy audiences and, in particular, to specific moments of engagement between performers and individual audience members. Like the observational studies described in the introduction it provides evidence that performers modulate audience responses not only through large performative gestures but also through the use of fine-grained mechanisms such as eye contact. Moreover, it shows that these mechanisms lead to different patterns of response for different audience members. Understanding specific moment-by-moment processes that underpin these interactional dynamics is key to developing more compelling live experiences and more engaging robots.

This exploratory study needs to be replicated and extended. A larger repertoire of gestures and other non-verbal signals needs to be tested together with a richer set of response measures. An obvious limitation here is the possible influence of the specific audience and context. Testing alternative patterns of delivery across a wider range of audiences would help to establish how general these patterns are. Greater realism could be achieved

by using motion capture sequences from a human comedian to drive the robot. These sequences could provide the basis for more naturalistic manipulations of different non-verbal elements of performance and would also support more credible robot-human comparison.

The Comedy Parser platform (Katevas et al., 2014) demonstrates how robot performances can use the computational vision capabilities provided by systems like SHORE™ to make the details of delivery contingent on how each individual in an audience is responding in real-time. We note that this goes beyond what a human comic can do. This approach can be extended to other modalities such as automatic, real-time audio processing to sense oral responses, applause and more subtle cues such as collective inbreaths or rustling paper. There is also potential for experimenting with speech rhythm and intonation. Alternative speech engines provide some interesting capabilities. For example, CereVoice is capable of changing of the voice's "mood" into "happy", "calm", or "joke" (Aylett and Pidcock, 2007).

5. Conclusion

This paper demonstrates how humanoid robots can be used to probe the complex social signals that contribute to the experience of live performance. Using qualitative, ethnographic work as a starting point we can generate specific hypotheses about the use of social signals in performance and use a robot to operationalize and test them. This can provide a principled basis on which to give humanoid robots the capabilities needed to interpret and respond to the social dynamics of massed audiences.

Moreover, this paper provides insight into the nature of live performance. We showed that audiences have to be treated as heterogeneous, with individual responses differentiated in part by the interaction they are having with the performer. Equally, performances should be further understood in terms of these interactions. Successful performance manages the dynamics of these interactions to the performer's- and audiences'-benefit.

Author Contributions

This work is based on research carried out for an advanced placement project by KK under the supervision of PH. Concept: PH, KK, MH. Observational study: MH, KK. Experimental design: PH. Engineering of robot platform: KK. Data collection and processing: KK, MH. Statistical analysis and interpretation: PH. Manuscript: PH, KK, MH.

Acknowledgments

This work was funded by EPSRC (EP/G03723X/1) through the Media and Arts Technology Program, an RCUK Center for Doctoral Training. The live performances were made possible by "Hack the Barbican" who accepted "Comedy Lab: Human vs. Robot" into their program, and by Engineered Arts who provided RoboThespian™ for the performances. The staff of the Barbican Centre assisted—admirably—in staging these performances. The authors extend a personal thanks to Tiernan Douieb, a professional comedian excited to experiment with comedy in ways perhaps he hadn't foreseen. Tiernan advised us throughout, wrote the script for the robot's routine, and hosted each night's performance.

References

Aylett, M. P., and Pidcock, C. J. (2007). "The cerevoice characterful speech synthesiser sdk," in *IVA '07: Proceedings of the 7th International Conference on Intelligent Virtual Agents* (Paris: Springer-Verlag).

Bavelas, J. B., Black, A., Lemery, C. R., and Mullett, J. (1986). "I show how you feel": motor mimicry as a communicative act. *J. Pers. Soc. Psychol.* 50:322. doi: 10.1037/0022-3514.50.2.322

Bavelas, J. B., Coates, L., and Johnson, T. (2000). Listeners as co-narrators. *J. Pers. Soc. Psychol.* 79:941. doi: 10.1037/0022-3514.79.6.941

Ernst, A., Ruf, T., and Küblbeck, C. (2009). "A modular framework to detect and analyze faces for audience measurement systems," in *Proceedings of the 2nd Workshop on Pervasive Advertising*, eds J. Müller, P. Holleis, A. Schmidt, and M. May (Lübeck: GI Jahrestagung), 75–87.

Fernández-Dols, J.-M., and Ruiz-Belda, M.-A. (1995). Are smiles a sign of happiness? gold medal winners at the olympic games. *J. Pers. Soc. Psychol.* 69:1113. doi: 10.1037/0022-3514.69.6.1113

Gardair, C. (2013). *Audience Interaction in Street Performances*. Ph.D. thesis, Queen Mary University of London.

Gardair, C., Healey, P. G., and Welton, M. (2011). "Performing places," in *Proceedings of the 8th ACM Conference on Creativity and Cognition, C&C '11* (New York, NY: ACM), 51–60.

Goodwin, C. (1979). "The interactive construction of a sentence in natural conversation," in *Everyday Language: Studies in Ethnomethodology*, ed Charles Goodwin (New York, NY: Irvington Publishers) 97–121.

Hayashi, K., Kanda, T., Miyashita, T., Ishiguro, H., and Hagita, N. (2005). "Robot manzai - robots' conversation as a passive social medium," in *Humanoid Robots, 2005 5th IEEE-RAS International Conference on* (Tsukuba), 456–462.

Healey, P. G., Frauenberger, C., Oxley, R., Schober, M., and Welton, M. (2009). "Engaging audiences," in *CHI 2009 Workshop: Crowd Computer Interaction* (Boston, MA).

Katevas, K., Healey, P. G. T., and Harris, M. T. (2014). "Robot stand-up: engineering a comic performance," in *Proceedings of the Workshop on Humanoid Robots and Creativity at the IEEE-RAS International Conference on Humanoid Robots Humanoids* (Madrid).

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychol.* 26, 22–63. doi: 10.1016/0001-6918(67)90005-4

Kendon, A. (1990). *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Vol. 7. Cambridge, UK: CUP Archive.

Knight, H., Satkin, S., Ramakrishna, V., and Divvala, S. (2011). "A savvy robot standup comic: online learning through audience tracking," in *Workshop Paper (TEI'10)* (Funchal).

Knight, H., and Simmons, R. (2013). "Estimating human interest and attention via gaze analysis," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on* (Karlsruhe: IEEE), 4350–4355.

Kraut, R. E., and Johnston, R. E. (1979). Social and emotional messages of smiling: an ethological approach. *J. Pers. Soc. Psychol.* 37:1539. doi: 10.1037/0022-3514.37.9.1539

Kuhlen, A. K., and Brennan, S. E. (2013). Language in dialogue: when confederates might be hazardous to your data. *Psychon. Bull. Rev.* 20, 54–72. doi: 10.3758/s13423-012-0341-8

MacDorman, K. F., and Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interact. Stud.* 7, 297–337. doi: 10.1075/is.7.3.03mac

Mutlu, B., Forlizzi, J., and Hodgins, J. (2006). "A storytelling robot: modeling and evaluation of human-like gaze behavior," in *Humanoid Robots, 2006 6th IEEE-RAS International Conference on* (Genoa), 518–523.

Rutter, J. (1997). *Stand-up as Interaction: Performance and Audience in Comedy Venues*. Ph.D. thesis, University of Salford.

Rutter, J. (2000). The stand-up introduction sequence: comparing comedy compères. *J. Pragmat.* 32, 463–483. doi: 10.1016/S0378-2166(99)00059-4

Sidner, C. L., Lee, C., Kidd, C. D., Lesh, N., and Rich, C. (2005). Explorations in engagement for humans and robots. *Artif. Intell.* 166, 140–164. doi: 10.1016/j.artint.2005.03.005

Sjöbergh, J., and Araki, K. (2009). "Robots make things funnier," in *New Frontiers in Artificial Intelligence, Vol. 5447, of Lecture Notes in Computer Science*, eds H. Hattori, T. Kawamura, T. Id, M. Yokoo, and Y. Murakami (Berlin; Heidelberg: Springer), 306–313.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). "Elan: a professional framework for multimodality research," in *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)* (Genoa).

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 20 October 2015
doi: 10.3389/fpsyg.2015.01578

Comprehension and engagement in survey interviews with virtual agents

Frederick G. Conrad[1,2]*, *Michael F. Schober*[3], *Matt Jans*[4], *Rachel A. Orlowski*[5], *Daniel Nielsen*[6] *and Rachel Levenstein*[7]

[1] Michigan Program in Survey Methodology, Institute for Social Research, University of Michigan, Ann Arbor, MI, USA, [2] Joint Program in Survey Methodology, University of Maryland, College Park, MD, USA, [3] Department of Psychology, New School for Social Research, New York, NY, USA, [4] Center for Health Policy Research, University of California at Los Angeles, Los Angeles, CA, USA, [5] Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI, USA, [6] Department of Biostatistics, Center for Cancer Biostatistics, University of Michigan Medical School, Ann Arbor, MI, USA, [7] University of Chicago Consortium on Chicago School Research, Urban Education Institute, University of Chicago, Chicago, IL, USA

This study investigates how an onscreen virtual agent's dialog capability and facial animation affect survey respondents' comprehension and engagement in "face-to-face" interviews, using questions from US government surveys whose results have far-reaching impact on national policies. In the study, 73 laboratory participants were randomly assigned to respond in one of four interviewing conditions, in which the virtual agent had either high or low dialog capability (implemented through Wizard of Oz) and high or low facial animation, based on motion capture from a human interviewer. Respondents, whose faces were visible to the Wizard (and videorecorded) during the interviews, answered 12 questions about housing, employment, and purchases on the basis of fictional scenarios designed to allow measurement of comprehension accuracy, defined as the fit between responses and US government definitions. Respondents answered more accurately with the high-dialog-capability agents, requesting clarification more often particularly for ambiguous scenarios; and they generally treated the high-dialog-capability interviewers more socially, looking at the interviewer more and judging high-dialog-capability agents as more personal and less distant. Greater interviewer facial animation did not affect response accuracy, but it led to more displays of engagement—acknowledgments (verbal and visual) and smiles—and to the virtual interviewer's being rated as less natural. The pattern of results suggests that a virtual agent's dialog capability and facial animation differently affect survey respondents' experience of interviews, behavioral displays, and comprehension, and thus the accuracy of their responses. The pattern of results also suggests design considerations for building survey interviewing agents, which may differ depending on the kinds of survey questions (sensitive or not) that are asked.

Keywords: virtual agent, survey interviewing, social signals, comprehension, dialog capability, facial animation

OPEN ACCESS

Edited by:
Sebastian Loth, Universität Bielefeld, Germany

Reviewed by:
Jonathan Gratch, University of Southern California, USA
Verena Nitsch, Universität der Bundeswehr München, Germany

***Correspondence:**
Frederick G. Conrad
fconrad@umich.edu

Specialty section:
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology

Received: *08 March 2015*
Accepted: *29 September 2015*
Published: *20 October 2015*

Citation:
Conrad FG, Schober MF, Jans M, Orlowski RA, Nielsen D and Levenstein R (2015) Comprehension and engagement in survey interviews with virtual agents. Front. Psychol. 6:1578. doi: 10.3389/fpsyg.2015.01578

INTRODUCTION

An important source of knowledge about society is what people report in survey interviews that produce the data for official (government) statistics, e.g., population estimates on employment, health and crime. Data from such survey interviews, which provide essential input for policy decisions, are administered on a very large scale; for example, more than 60,000 US households

per month are recruited to participate in the Current Population Survey, from which the US unemployment rate is calculated, and for the European Social Survey (ESS) in 2012, 54,600 standardized face-to-face interviews were carried out in 29 countries (Ferrin and Kriesi, 2014). Results from these interviews can have far-reaching consequences: even small changes in reported US unemployment rates, for example, can affect world financial markets, and results from the ESS make "a major contribution to the creation of effective social and economic policies in Europe" (Geoghegan-Quinn, 2012). So understanding what leads to accurate responses, and to participants' willingness to engage in such surveys, is societally important (Schober and Conrad, 2015).

Although survey interviews have traditionally been administered by humans either face-to-face or on the telephone, the landscape is changing: surveys are increasingly "self-administered" (that is, administered by automated systems, as in online surveys in a web browser possibly on a mobile device; Mavletova and Couper, 2014), and new human and automated modes are being explored (Conrad and Schober, 2008), e.g., videomediated interviewing (Anderson, 2008), text message surveys (Schober et al., 2015), and speech dialog system surveys (Bloom, 2008; Johnston et al., 2013). Exploring new ways of administering surveys is sensible given declining survey response rates and the growing expenses of carrying out human-administered interviews (see, e.g., Groves, 2011; Keeter, 2012; Massey and Tourangeau, 2013), but the task is complex: new interviewing methods will only be adopted if they lead to high quality data (accurate responses, and response and completion rates comparable to or better than those in other modes) and to respondents satisfied with their experience.

One new interviewing technology that has been proposed to promote high quality data—as measured by disclosure of sensitive information and (presumably more) honest responding—uses animated virtual humans to ask questions and capture responses (Lucas et al., 2014; see also DeVault et al., 2014; Gratch et al., 2014). The promise is that virtual interviewers can promote rapport and engagement with participants while simultaneously providing a feeling of safety and anonymity that is much more difficult to achieve with a human interviewer, and at the same time allowing users to display (and even learn to improve) the social cues they display in interaction with humans (Baur et al., 2013). And some of the findings are promising: Lucas et al. (2014) found that people in a semi-structured clinical health screening interview disclosed more sensitive information in open-ended responses to a virtual interviewer they believed was automated than to one that was clearly operated by a human. von der Pütten et al. (2011) found that a more talkative interviewing agent led students to reveal more personal information and to produce more words in answering some open-ended questions on love and relationships.

The evidence on how virtual interviewers might affect responses in surveys that produce social science and government data, on the other hand, is less promising with respect to disclosure. The one study thus far (Lind et al., 2013) focused on responses to questions about sensitive and potentially embarrassing topics (alcohol and drug use, sexual behavior) and questions about personal behaviors (exercise, religious attendance); such questions can lead at least some respondents to answer in ways that present themselves in a more positive light in survey interviews where human interviewers ask the questions compared to when a computer presents textual or spoken questions (Tourangeau and Smith, 1996; Turner et al., 1998; Kreuter et al., 2008). The finding was that automation did increase disclosure relative to a human interviewer, but only with the audio-only (no facial representation) interface; there were few if any differences in responses to the virtual interviewers relative to a human interviewer (Lind et al., 2013).

Here we explore how virtual interviewers affect answers to the kinds of questions about facts and behaviors (e.g., "How many bedrooms are there in your house?" "Last week did you do any work for pay?") that are especially common in survey interviews that produce official statistics and that, in most cases, are not particularly threatening or embarrassing to answer. Because these questions generally concern non-sensitive, mundane topics, we are not focused on how virtual human interviewers might affect disclosure. Instead, we explore how and whether virtual human interviewers promote conscientious task performance—accurate survey responding, which depends on comprehending the questions in the way the survey designers intended—and respondent engagement in these particular kinds of interviews. In our experiment we varied two features (among the many other potentially manipulable features of a virtual survey interviewer, see Lind et al., 2013)—the interviewer's *dialog capability* and *facial animation*—and explored whether they have independent or compound effects.

Background

The kinds of survey interviews we examine here have particular features that distinguish them from other kinds of interaction (Schaeffer, 1991; Houtkoop-Steenstra, 2000; Schober and Conrad, 2002), as well as from other kinds of interviews. The survey interview is an interactive situation in which (usually) the interviewer, as a representative of the survey designers (researchers), initiates the dialog and "drives" the interaction according to a script (Suchman and Jordan, 1990), asking the respondent questions (that usually specify the answer categories) about her opinions and behaviors.

This kind of standardized wording and administration procedure is intended to make responses comparable across interviews. In the most strictly standardized interviews, interviewers are required to ask questions exactly as scripted and use only "neutral probes" like "Let me repeat the question" or "Whatever it means to you" if respondents say anything that isn't an acceptable answer (e.g., something other than a response option included in the question), so as to ensure that all respondents receive the same stimulus and to avoid the possibility that interviewers will bias responses (Fowler and Mangione, 1990). This can lead to perverse interactions in which interviewers thwart respondents' efforts to understand what they are being asked by refusing to provide the clarification that respondents seek (Suchman and Jordan, 1990), and in which interviewers violate ordinary norms of conversation by failing

to "ground" the meaning of utterances they themselves have produced (Schober and Conrad, 2002).

Analyses of these kinds of survey interviews demonstrate that respondents can misinterpret ordinary expressions in questions (like "bedroom" and "work for pay")—that is, interpret them differently than the survey designers intend—much more often than one might think (Conrad and Schober, 2000), because the mapping or "fit" between their circumstances and the question concepts may not be straightforward (consider someone whose room originally designed as a den is being used as a bedroom, or whose freelance work included pay-in-kind). This is particularly a problem when interviews are strictly standardized; in more collaborative or "conversational" interviews, where interviewers and respondents work together to make sure respondents understand questions as intended (e.g., Schober and Conrad, 1997; Conrad and Schober, 2000), respondents generally interpret questions much more accurately. The best response accuracy, overall, seems to result not only when respondents request clarification if they believe they need it ("What do you mean by work for pay exactly?"), but when interviewers can also volunteer clarification when they believe respondents need it (Schober et al., 2004).

When designing a virtual interviewer for these kinds of surveys, a key consideration is, therefore, which features will best help respondents understand the questions as they are intended. Based on what is known about respondent comprehension in human-administered interviews, a virtual interviewer that can clarify question meaning when explicitly asked to do so and when it determines the respondent would better understand the question if its meaning were clarified—what we will call here a virtual interviewer with greater *dialog capability*—should, in principle, lead to more accurate comprehension. Whether this is actually the case with a virtual interviewer has not been demonstrated. Evidence from other automated implementations of survey interviews suggests that it could be the case, but it is not a foregone conclusion that it will be. For example, respondents' accuracy in a text-based web survey (Conrad et al., 2007) and in a (wizarded) spoken dialog survey system (Ehlen et al., 2007) improves when the system can provide clarification after a long period of inactivity or silence, but it does not improve in conditions where the only way to obtain clarification is to explicitly request it.

Whether high dialog capability interviewing systems with a facial representation will similarly promote comprehension is unclear. The addition of a face to the interface could make respondents even more reluctant to request clarification about ordinary words like "bedroom" and "job," as they sometimes seem to be with human interviewers (Schober et al., 2004). Or, on the other hand, it could make them think the automated interviewer has greater agency and capabilities, and is thus better positioned to engage in clarification dialog. Because users' attributions about animated agents are likely to vary depending on the characteristics of the face—both static and dynamic (e.g., McDonnell et al., 2012; Piwek et al., 2014)—one might expect that survey response accuracy could be affected by how an animated virtual interviewer is visually implemented: survey respondents may evaluate the agent's competence and its likelihood of being able to provide useful clarification as greater when it behaves in a more human-like way. That is, they might assume that a more human-like face on a virtual interviewer means that the interviewer will comprehend requests for clarification better, and that the interviewer may better perceive the *respondent's* paralinguistic and facial displays of need for clarification (Schober et al., 2012).

Hypotheses

In the study reported here, we test the following hypotheses about how a virtual survey interviewer's dialog capability and facial characteristics affect respondents' comprehension (as measured by the accuracy of their answers—our primary measure of task success). We also test how these factors affect respondents' social engagement with the interviewer, as measured by their behavioral displays as well as their subjective assessments of the interviewer. The facial characteristic that our hypotheses focus on is motion or *facial animation*: whether the face moves in a more or less human-like way, that is, with more or fewer channels of motion. We examine facial animation because this strikes us an attribute that is particularly likely to affect respondents' interpretation of a virtual interviewer's humanness; this is consistent with evidence in other task contexts that users interpret an embodied agent's intentions based more on audio and animation than on the render style of the character (McDonnell et al., 2012).

Hypotheses about Comprehension

Hypothesis 1: *Dialog capability and comprehension.* A virtual interviewer with greater dialog capability will improve respondents' comprehension of survey questions, particularly when the fit between terms in the survey questions and the circumstances respondents are answering about is not straightforward.

This hypothesis will be supported to the extent that respondents treat a virtual interviewer with high dialog capability as better able than a low-dialog-capability virtual interviewer to interpret (1) their explicit requests for clarification and (2) indirect evidence of comprehension difficulty, both spoken and visual. If dialog capability affects comprehension in this way, its effect should be measurable both by response accuracy and by the number of requests for clarification. The basic mechanism is that more clarification should correct more misconceptions and resolve more ambiguities; the effect of dialog capability should be most evident when comprehension problems of this sort are most frequent, i.e., when the virtual interviewer asks questions about concepts that correspond in an ambiguous way to respondents' circumstances or whose definitions run counter to respondents' intuitions. We manipulate this experimentally in the study reported here.

The evidence to date that evaluates the effect of virtual survey interviewers on the quality of responses does not provide evidence about whether dialog capability works the same way or to the same extent with human and virtual interviewers. For example, while the Lind et al. (2013) study concerned survey interviews, the authors did not design the virtual interviewers to provide clarification; moreover the interaction was not entirely spoken: the interviewing agents asked questions orally but

respondents answered by clicking or typing. If clarification does not work the same way—if respondents don't solicit or interpret clarification in the same way—with virtual interviewers in a spoken dialog interview as they do with human interviewers, the hypothesis will not be supported. This could occur if, for example, respondents do not treat the virtual interviewer as conversationally competent—which might be affected by the interviewer's facial animation.

> *Hypothesis 2*: *Facial animation and comprehension.* A virtual interviewer with more facial animation will improve respondents' comprehension of survey questions.

This hypothesis will be supported if survey respondents attend better or try harder at the survey response task when an interviewer seems more human-like, which can result from a virtual agent's increased motion (Hyde et al., 2013; Piwek et al., 2014). The evidence is that perceiving another person's facial motion *can* improve at least some kinds of task success. For perceptual tasks, for example, people tend to be better at detecting a speaker's identity when presented with a moving than a static face (see Xiao et al., 2014, for a review), and they can comprehend speech even in noisy conditions better with facial (especially mouth) motion cues than without (Alexanderson and Beskow, 2014). In avatar-mediated communication, participants are better able to detect truth and deception when an avatar has realistic eye motion (Steptoe et al., 2010).

On the other hand, in a survey interview setting where the measure was disclosure of sensitive information rather than comprehension accuracy, Lind et al. (2013) found less disclosure to a high-motion virtual interviewer than to a low-motion interviewer for some survey questions, and no difference in disclosure for others. To the extent that these disclosure findings are relevant to comprehension and response accuracy for non-sensitive survey questions, increased facial motion in a virtual interviewer may not improve survey task performance, and this hypothesis will not be supported.

> *Hypothesis 3: Interactive effects of facial animation and dialog capability on comprehension.* A virtual interviewer with more facial animation may improve respondents' comprehension of survey questions particularly when the interviewer has greater dialog capability. To put it another way, a virtual interviewer's dialog capability may improve comprehension particularly when the interviewer's facial animation is consistent with greater dialog competence.

If a virtual interviewer's greater facial animation suggests that it has greater dialog competence, respondents may be particularly more likely to seek clarification (explicitly request it) or to provide indirect evidence of their need for clarification (paralinguistic or facial), and thereby comprehend and answer more accurately, than if an interviewer has less facial animation. If so, this would predict an interaction: greater clarification-seeking or evidence of need for clarification, and thus improved response accuracy, with a high-animation agent in a high-dialog-capability condition.

On the other hand, greater facial animation could lead to unrealistic expectations that the agent's dialog competence is fully human, which could subsequently conflict with the agent's actual abilities; in this case, greater facial animation could, paradoxically, lead to poorer comprehension if the respondent relies solely on the interviewer to diagnose need for clarification. One could also imagine other interactive effects: an interviewer with low facial animation might lead users to *under*estimate the dialog capability of high-dialog-capability agents, and thus request clarification or produce indirect evidence of need for clarification less often than would be optimal.

Although hypotheses about interactive effects of a virtual interviewer's dialog capability and facial animation have not been tested before, the plausibility of such effects is strengthened by the finding that survey respondents in face-to-face interviews produce more paralinguistic displays of need for clarification (speech disfluencies) and avert their gaze more often for unreliable answers in high-dialog-capability (conversational) than low-dialog-capability (strictly standardized) interviews (Schober et al., 2012). Of course, human interviewers have high facial animation in the sense we are exploring here, unless their facial mobility is impaired from neurological illness or cosmetic interventions, and yet when they conduct standardized interviews they are required to restrict their ordinary dialog capability; so a mismatch between facial animation and dialog capability is not unusual in human-administered survey interviews. On the other hand, if comprehension in surveys depends mostly on the conceptual content conveyed by dialog, the interviewer's facial animation will not interact with dialog capability in affecting respondents' comprehension.

Hypotheses about Engagement

Independent of comprehension or clarification-seeking behavior, a virtual interviewer's dialog capability and facial animation could have independent or interactive effects on survey respondents' engagement with the interview, as evidenced by their social behaviors during the interaction (e.g., time spent looking at the virtual interviewer, nods and verbal acknowledgments, and smiles) and by how they experience the interview subjectively.

Respondents' engagement in survey interviews—their involvement, attentiveness, and conscientiousness—is critical for obtaining accurate data. But respondents can be less engaged in the interview task than would be desirable, perhaps because most do not ask to be interviewed (the researchers invite them via an interviewer). In conventional survey modes, evidence of respondents' lack of engagement can be seen in their terminating an interview before it is completed (see Peytchev, 2009 for a discussion of breakoffs in online questionnaires) and in their least-effort "satisficing" as they answer questions, for example selecting the same response option again and again in a battery of questions (e.g., Chang and Krosnick, 2010). Our focus here is on respondents' behavioral displays of engagement during the course of a virtual interview—their gaze, their spoken and visual acknowledgments, and their smiles—and their reported post-interview assessments of their interview experience.

With this focus, we test the following hypotheses:

> *Hypothesis 4: Dialog capability and engagement.* A virtual interviewer whose interaction is more like everyday conversation—who can clarify the questions—will engage respondents more than a virtual interviewer with low dialog capability.

One might expect that when survey respondents interact with a virtual interviewer with more human-like capabilities they will behave more as they do in ordinary conversation: they will look at their interlocutor more, acknowledge their understanding more (nod, produce backchannels like "okay"), display social cues (smile), and rate the interaction as more positive. To our knowledge this has not been examined directly, but accounts of frustration experienced by respondents whose standardized interviewers are prevented from providing clarification (e.g., Suchman and Jordan, 1990) are consistent with this hypothesis.

> *Hypothesis 5: Facial animation and engagement.* A virtual interviewer whose facial movement is more human-like will engage respondents more than a virtual interviewer with low facial animation.

From other domains of interaction with virtual agents, the evidence is that people judge agents with more (bodily) motion as more acceptable and human (Piwek et al., 2014), and that realistic characters that move more are judged more positively (Hyde et al., 2013). The benefits of more human-like behavior may well extend to the survey context: Conrad et al. (2013) demonstrated that people invited to participate in (human-administered) telephone survey interviews were more likely to agree to participate when the interviewers spoke less robotically (with more disfluencies) during the invitation interaction. And Foucault Welles and Miller (2013) demonstrated that respondents in face-to-face (human-administered) survey interviews reported feeling greater rapport (which is presumably related to their feelings of engagement) when *interviewers* nodded and smiled more, and when they gazed at respondents' faces less.

> *Hypothesis 6: Interactive effects of facial animation and dialog capability on engagement.* A virtual interviewer with more facial animation may increase respondents' engagement particularly when the interviewer has greater dialog capability.

Any effects of dialog capability and facial animation on respondents' display of social cues or assessment of the interviewer could be independent, or they could interact. The same range of possible interaction effects exists for measures of engagement as for comprehension. The combination of low dialog capability and low facial animation could lead to particularly unengaging or alienating interaction. High facial animation could lead to unrealistic expectations about an interviewer's dialog capability, which when thwarted could lead respondents to be *less* engaged with the interviewer. Low facial animation could lead to underestimation of a high dialog capability interviewer's competence, which could lead respondents to attend less fully to or disengage with the interviewer.

MATERIALS AND METHODS

Our strategy in this study was to bring participants to our laboratory to respond to 12 questions about housing, work and purchases taken from US government surveys, which they answered on the basis of scenarios describing fictional circumstances. This allowed us to directly assess the accuracy of their responses—which also measures the extent to which their comprehension of the terms in the survey questions fits what the official definitions of those terms would require. Participants (respondents) were randomly assigned to be interviewed by a (Wizard-of-Oz) interviewing agent with either high or low facial animation (many channels/points of motion vs. few) and high or low dialog capability (conducting interviews in either a collaborative or strictly standardized style). For each respondent, half the fictional scenarios were designed to map onto the survey questions in a straightforward way and half in a complicated way. Thus, the experimental design was $2 \times 2 \times 2$.

Although having respondents answer about fictional scenarios as opposed to about their own lives reduces ecological validity, it has the advantage of allowing direct assessment of accuracy of comprehension during the interviews. In other studies with human interviewers we have used post-interview self-administered questionnaires (Suessbrick et al., 2000; Schober et al., 2012) and human-administered re-interviews (Conrad and Schober, 2000; Suessbrick et al., 2000) as alternate (less direct) methods for assessing comprehension and survey response accuracy, under the logic that response change when respondents are provided with a standard definition of a survey term is likely to reflect the correction of a misinterpretation in the original interview; the findings in those studies are highly consistent with the findings produced when responses are based on fictional scenarios, and so in the current study we use fictional scenarios. The questions and scenarios in the current study are the same as those used in previous laboratory studies of telephone interviews (Schober and Conrad, 1997; Schober et al., 2004) and of online text- and speech-based interviewing systems (Conrad et al., 2007; Ehlen et al., 2007). Although the participant sample and time frame for this experiment make a comparison with those studies not entirely parallel, they provide relevant context for evaluating respondents' performance with a virtual interviewer in the current study.

Experiment Materials

Survey Questions

The 12 survey questions were adapted to apply to the fictional scenarios that respondents would be answering about, rather than about the respondent's own circumstances: four questions about employment from the US Current Population Survey (e.g., "Last week, did Chris do any work for pay?" filling in the name of the fictional character Chris in the question "Last week, did you do any work for pay?"), four questions about purchases from the Current Point of Purchase Survey (e.g., "Has Kelly purchased or had expenses for household furniture?"), and four questions about housing from the Consumer Price Index Housing Survey (e.g., "How many bedrooms are there in this house?"). Each question had a corresponding official definition for its key concepts developed by the sponsoring agency. For example, for the question "Has Kelly purchased or had expenses for household furniture," the official definition of household furniture is this:

> Tables, chairs, footstools, sofas, china cabinets, utility carts, bars, room dividers, bookcases, desks, beds, mattresses, box springs, chests of drawers, night tables, wardrobes, and unfinished furniture. Do not include TV, radio, and other sound equipment, lamps and lighting fixtures, outdoor furniture, infants' furniture, or appliances (US Bureau of the Census and US Bureau of Labor Statistics, 1993).

(Supplementary Table 1 includes all questions and the official definitions relevant to each question).

The questions were ordered for the experiment to correspond with the order in which they appeared in the survey from which they were drawn, and counterbalanced across domains for different respondents to make sure that any effects observed in the experiment could not be attributed to the order in which the virtual interviewer asked about the different domains. So one respondent would answer purchase questions followed by housing questions followed by employment questions, another would answer housing questions followed by employment questions followed by purchase questions, etc.

Respondent Scenarios

Fictional scenarios on the basis of which respondents were to answer the questions were assembled into paper packets, with one page per scenario. In actual surveys respondents most often answer based on their recall and self-assessment; using scenarios is more similar to situations when respondents answer by consulting their personal records, and, more importantly, allows us to isolate and focus on comprehension—there is no autobiographical recall involved when respondents answer based on scenarios. For factual questions about respondents' behaviors or circumstances, the outcome of each exchange—an answer to a survey question—is either accurate or not (e.g., the respondent either has or has not done any work for pay in the last week). In principle this could be independently assessed if researchers were to have independent evidence about the respondent's circumstances (e.g., trustworthy records from the respondent's place of employment), but of course, in many cases (e.g., for many personal behaviors and for respondents' opinions) there is no independently verifiable evidence about the accuracy of responses.

The scenarios, which were not seen by the interviewing Wizard during the interview, consisted of work descriptions, purchase receipts, and floor plans. Two alternate scenarios were created for each question, one describing situations that mapped onto questions and the corresponding official definitions in a straightforward way ("straightforward mappings") and one describing situations that mapped onto questions and official definitions in a complicated way ("complicated mappings")—for which respondents might well need clarification in order to answer the question in a way that fit the definition. For example, for the question about household furniture, the straightforward scenario was a receipt for the purchase of an end table. The complicated scenario was a receipt for the purchase of a floor lamp. The official definition—which was not part of the materials given to the respondents, but could only be presented orally by a high-dialog-capability virtual interviewer—clarified that for the purposes of this survey, a floor lamp is not to be counted as a household furniture purchase, and thus the answer to this question should be "no." (The answer for the straightforward scenario should be "yes," as an end table counts as a furniture purchase).

The selection of these scenarios thus allowed direct evaluation of whether the respondent had comprehended the question in a way that fit the official definitions. A respondent who answers "yes" to the household furniture question with a floor lamp receipt, or "no" with an end table receipt, is not interpreting the question as the survey designers intended; these responses can be classified as incorrect.

Scenario packets were assembled for each respondent that included half (6) straightforward and half (6) complicated scenarios, with two straightforward and two complicated scenarios per domain (employment, purchases, housing). The orderings of mappings were counterbalanced across respondents, such that the particular combination of straightforward and complicated mappings for one respondent was the complement of the combination for another. Across all respondents, both straightforward and complicated scenarios were presented equally often and in different orders, both so that the interviewing Wizard could not anticipate which scenario a particular respondent was encountering and so that any effects observed in the experiment could not be attributed to a particular sequence of mappings.

Additional Interviewer Utterances

In addition to the survey questions and the full definitions of relevant terms in the questions, all other allowable interviewer utterances in low and high dialog capability interviews were scripted. These included several introductions of the interview (e.g., "Hello, my name is Derek and today I will be asking you a few questions about housing, jobs and purchases."), pre-interview practice material, neutral probes (e.g., "Is that a yes or a no?"), partial definitions (just the text that resolves the ambiguity in the corresponding complicated scenario), clarification offers ("It sounds like you're having some trouble. Can I give you a definition that might help?"), utterances to manage the dialog (e.g., "Yes," "No," "Please wait one moment"), and utterances to run the experimental session ("Please turn to the next page of your packet"; "I am going to ask the research assistant to help you. Just a minute please"). Supplementary Table 2 lists the full set of additional scripted utterances.

Developing the Virtual Interviewers

The virtual interviewers for the four experimental conditions were created using famous3D's ProFACE video software (version 2.5) to make variants of a single 3D model of a head. We first video- and audio-recorded a human interviewer (a male graduate student in survey methodology who spoke American English) administering all survey questions, prompts, clarifications, and additional interviewer utterances, with 21 green and blue dots affixed to his face to capture 21 different motion channels (forehead, outer and inner brows, furrow, upper eyelids, region below the eyes, cheeks, right and left sides of nose, right and left lower lips, chin, etc.). With the ProFace software we captured his facial motion and mapped it to a face template, which could then be projected onto one of ProFace's existing models (Derek,

in our case; see **Figure 1**) either using all motion channels (for the high facial animation conditions) or a subset (for the low facial animation conditions). All audio files used in the low dialog capability conditions were also used in the high dialog capability conditions; there were, of course, extra speech files (and accompanying video) for high dialog capability conditions (e.g., offers of clarification).

Note that because all four virtual interviewers were based on the same head model, the interviewer's base level of visual realism or naturalism, which can affect how users judge and respond to virtual agents in other task contexts (e.g., Baylor and Kim, 2004; MacDorman and Ishiguro, 2006; Gong, 2008; MacDorman et al., 2009), was the same across all four conditions. In a job interview training task, Baur et al. (2013) found that interviewees criticized their interviewer as not looking realistic enough; our interviewer has a level of realism that reflects the photographic origins of the model, and is more realistic than the more cartoon-like survey interviewer in Lind et al. (2013), but there is simply not enough evidence in survey tasks about the optimal levels of realism for a virtual survey interviewer.

Also note that because the interviewers differed behaviorally on more than one feature, any effects on respondents must be attributed to bundles of behavioral features rather than individual features.

Facial Animation

Table 1 summarizes the major features of motion in the high and low facial animation interviewers. For the low facial animation conditions, seven motion channels were projected onto the Derek model: chin, left and right lower lips, left and right corners of mouth, and left and right peaks of lip. The low animation interviewer head and face do not move, the eyes do not blink, and the mouth does not change shape as the interviewer speaks—it just opens and closes.

For high facial animation conditions, in addition to the 21 channels of captured motion the interviewer's head and face move (applying ProFace's jitter function) at all times (even while waiting for responses, to give the appearance of listening), and his eyes blink. The interviewer's mouth forms appropriate shapes for the sounds he is producing; to improve the correspondence between the interviewer's mouth movements and speech, additional keyframes were added by hand beyond the captured motion at a fine level of granularity, with particular combinations of motions for different consonants and vowel sounds in the recordings, based on the judgments of an animator. Finally, stationary shoulders were added to make the head movements look more realistic.

See Supplemental Data for video examples of low and high facial animation introductions to the interview (Videos 1, 2) and for low and high facial animation variants of Purchases Question 3 (Videos 3, 4).

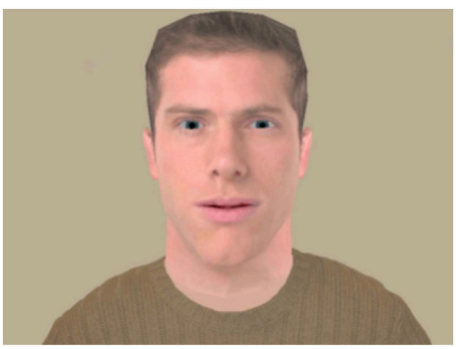


FIGURE 1 | The Derek model that formed the basis of the four virtual interviewers.

Dialog Capability

Table 2 summarizes the major features of dialog capability in the high and low dialog capability interviewers. These were implemented by an experimenter behind the scenes (the Wizard) following protocols for which interviewer files (questions, neutral probes, definitions, etc.) were to be played to respondents in which sequence and in response to which respondent behaviors (see Wizard protocols below). In all cases the virtual interviewers presented the same questions, and (from the

TABLE 1 | Facial animation features of virtual interviewers.

	Low facial animation	High facial animation
Head moves	No	Yes, even when "listening"
Face moves	Only mouth	Yes
Eyes move	No	Yes
Eyes blink	No	Yes
Mouth movement	Only opens and closes during speech, but does not change shape	Mouth forms appropriate shapes for sounds being produced

TABLE 2 | Dialog capability features of virtual interviewers.

	Low dialog capability	High dialog capability
Reads question as worded	Yes	Yes
Understands spoken answers	Yes	Yes
Repeats question if asked	Yes	Yes
Understands explicit requests for clarification	Yes	Yes
Provides clarification when explicitly requested	No: presents neutral probe (e.g., "Whatever it means to you"; "Let me repeat the question")	Yes: reads definition
Offers clarification when it seems needed (based on respondent's verbal and visual behavior)	No	Yes

respondents' perspective) they could comprehend and register spoken answers.

The low dialog capability protocol was to administer a strictly standardized interview, as implemented in previous studies in this line of research (e.g., Schober and Conrad, 1997; Schober et al., 2004). The virtual interviewer presented the questions exactly as worded and could repeat questions if asked, but if a respondent explicitly requested clarification the interviewer would only provide a neutral probe (of the Wizard's choosing, just as in human-administered standardized interviews; see Video 5 in Supplementary Materials for an example).

The high dialog capability protocol was to administer "conversational" interviews, again as in Schober and Conrad (1997). After reading the question exactly as worded, the (wizarded) interviewer (a male graduate student) did whatever he thought was needed to make sure that the respondent had interpreted the question as intended—to "ground" the meaning of terms in survey questions, to use Clark and colleagues' term (e.g., Clark and Wilkes-Gibbs, 1986; Clark and Schaefer, 1987; Clark, 1996). In other words, the interviewer's task was to make sure that the respondent's interpretation fit the official definition. This included not only providing the full official definition if the respondent explicitly requested it but also offering clarification if the interviewer (Wizard) got the sense that the respondent might need it (see Video 6 in Supplementary Materials for an example). Given the nature of the video files and wizarding protocols, this implementation of conversational interviewing is not as fully flexible as human interviewers can provide, because our virtual interviewers could not provide fully tailored partial definitions or improvise unscripted dialog, but it is on the most flexible end of the continuum (see Schober et al., 2004).

Pre-study: Verifying Distinctiveness of Virtual Interviewers

In order to increase our confidence that we had successfully manipulated what we hoped to in creating the virtual interviewer videos, we collected ratings of all 130 video clips in the experiment, both low and high facial animation versions. The clips included all questions, probes, definitions, and introductions to be used by both the low and high dialog capability virtual interviewers. Thirteen raters (11 female, two male; mean age 28.8, ranging from 24 to 34; all with bachelors' degrees, six graduate students in survey methodology) each rated 65 high- and low-animation video clips in one of two group viewing sessions. For each clip, each rater judged the virtual interviewer on a ten point scale for warmth ("How warm was Derek, with 0 being Not At All Warm and 10 being Very Warm?"), naturalness ("How natural was Derek, with 0 being Not At All Natural and 10 being Very Natural?"), and similarity to an actual interviewer ("To what degree did Derek seem like an actual interviewer, with 0 being Not At All Like An Interviewer and 10 being Very Much Like An Interviewer?").

The ratings confirmed that the high facial animation virtual interviewers were, in the aggregate, perceived to be reliably warmer [4.58 vs. 2.78 on the 10-point scale, $F_{(1, 12)} = 28.56$, $p < 0.001$, $\eta^2 = 0.704$], more natural [5.23 vs. 2.95 on the 10-point scale, $F_{(1, 12)} = 36.24$, $p < 0.001$, $\eta^2 = 0.751$], and more like a human interviewer [6.24 vs. 4.36 on the 10-point scale, $F_{(1, 12)} = 21.35$, $p = 0.001$, $\eta^2 = 0.640$] than the low realism versions. The same pattern was observed for most individual clips, though not all. Although none of the ratings reached the top of the 10 point scale, these strongly reliable differences suggested to us that these implementations of virtual interviewers would be suitable for the experiment.

Wizarding Protocols

The virtual interviewers were controlled by mapping each video file to a key on the computer keyboard using ArKaos VJ software. This allowed the Wizard to present the next relevant file to the respondent by pressing a key, according to the relevant protocol for high or low dialog capability interviewing (see **Table 3** for the Wizard's decision rules). Using the VJ software allowed seamless

TABLE 3 | Wizard's decision rules.

Low dialog capability	High dialog capability
Give respondent 3 min to familiarize him/herself with packet, and ignore respondent if he/she says he/she is ready	Give respondent 3 min to familiarize him/herself with packet, but begin interview if respondent says he/she is ready
Wait 10 s between transition and question clip, despite respondent behavior	Wait for respondent to look at virtual interviewer before presenting next question clip
Do not modify presentation of clips based on respondent's gaze or attention	Stop presenting a clip if respondent stops looking at virtual interviewer
Send research assistant to help respondent if in trouble	Use virtual interviewer to assist respondent if in trouble. If not successful send research assistant
If respondent seems hesitant or confused, do nothing	If respondent seems hesitant or confused, then offer help
If respondent asks for help, then present neutral probe	If respondent asks for help not related to scenario, then present neutral probe
	If respondent asks for help pertaining to scenario, then present entire definition
	If respondent asks for help with specific mention of key concept, then present partial definition
If respondent interrupts virtual interviewer, then finish presenting clip. Wait for respondent to repeat him/herself	If respondent interrupts virtual interviewer, then present waiting clip and address respondent's concern immediately

presentation of the video clips, so that the virtual interviewer appeared to the respondent to be acting on its own. The Wizard sat in a control room with a one-way mirror and live video feed view of the respondent. The control computers were set up so that the Wizard could view the respondent from a frontal overhead position and could also see the video file of the virtual interviewer as it was playing for the respondent.

The use of a Wizard allowed us to implement the high and low dialog capability virtual interviewers without programming a full survey dialog system with speech recognition and dialog management, which was beyond the scope of the current study [In other projects we have implemented a standardized survey spoken dialog system for mobile devices (Johnston et al., 2013) and experimented with an automated telephone system that implements conversational interviewing, including modeling respondents' paralinguistic displays of need for clarification (Ehlen et al., 2007)]. Because the same Wizard manipulated the virtual interviewers in this study across all conditions, his detection of and judgments of the meaning of respondents' facial and bodily displays and verbal behavior were likely to be consistent in the different conditions. This means that across the high and low facial animation conditions, the timing of turn transitions (the point at which speakers and listeners trade roles in conversation), which has been shown to affect perceptions of (in particular rapport with) virtual humans (Huang et al., 2011), were deployed based on the same human Wizard judgments, appropriately for either the high or low dialog conditions. Thus, although by necessity the Wizard needed to be informed about respondents' experimental conditions (so that he could deploy the appropriate video files), the particular linguistic and interactive intuitions that the Wizard brought to the experiment did not differ across the conditions.

Post-interview Measures

After completing the interview, respondents filled out an online questionnaire in which they reported their subjective experience interacting with the virtual interviewer on a number of dimensions (e.g., "How much did you enjoy interacting with Derek?", "Would you say that Derek acted more like a computer or a person?", "How often did Derek seem to act on his own?"). They also provided information about their technological experience ("How often, on average, do you use a computer?") and their demographic and linguistic background (e.g., "Is English your native language?"). The full questionnaire is presented in Supplementary Table 3.

Participants

Seventy-five participants (respondents) were recruited from the local site of the Craig's List online forum (https://annarbor.craigslist.org/) ($n = 51$) and through word of mouth ($n = 21$); for three respondents we do not have records about how they heard about the study.

Respondents, who were paid $35 for participating, were each randomly assigned to an experimental condition, except for two who were recruited specifically to replace two respondents (one in each high-dialog-capability condition) who expressed suspicion that the virtual interviewer was wizarded (the replaced and replacement respondents were all recruited through Craig's List). This led to a final data set with 18 respondents in three of the four conditions and 19 in the high-dialog-capability-high-facial animation condition.

In the final data set, the composition of the four groups did not differ reliably in age ($F < 1$), nor in recruitment source (p-values for all $\mathbf{X}^2 > 0.15$.) The respondents ranged in age from 18 to 67 years (mean = 36.8); 38 were female and 35 were male. 56.2% of respondents reported being White, 20.5% Black, 16.4% Asian or Pacific Islander, and 5.5% reported being members of other groups. 37.4% of respondents reported their highest level of education as less than a bachelor's degree, 42.5% as a bachelor's degree, and 19.2% as a graduate or professional degree. As a group they were highly computer literate, with 84.9% reporting using a computer 5–7 days per week. 89% reported that English was their native language.

All procedures that respondents followed, and all materials that were presented to them, were reviewed and approved by the University of Michigan IRB-HSBS (Institutional Review Board—Health Sciences and Behavioral Sciences).

Procedure

Each respondent was escorted to a first room where he or she signed consent forms and was handed the packet of experimental scenarios on the basis of which he or she would be answering survey questions. A research assistant instructed respondents using the following script:

> In this study, you will be asked 12 questions about fictional purchases, housing, and jobs. This interview is not like typical interviews. We will not be asking you about your own experiences but about the information contained in scenarios in this packet, so we can assess the accuracy of your responses. On each page there is one scenario, which corresponds to one question. You should answer each question based only on information in the corresponding scenario. Each scenario is independent of each other, so you should *not* use information from the previous page to answer a subsequent question. Some of the scenarios are dated; consider the date in the packet to be current, rather than responding based on today's date. You will receive additional information about this procedure once the interview begins. Let's enter the room now to start the interview.

Respondents were then led to a second room, which contained two mounted cameras, a chair, a table, a computer, a monitor displaying the virtual interviewer, a microphone on the table, and (in the high-dialog-capability conditions) a non-functioning web camera trained on the respondent to increase the plausibility that the virtual interviewer could sense the respondent. The room was free of other distractions. If a respondent asked about any of the equipment, the research assistant answered by saying, "I will be happy to answer your questions after the interview." The research assistant then pointed at the monitor with the virtual interviewer and gave the following instructions:

> You are going to be interviewed by Derek. Derek will speak to you, and you should respond aloud. Please look at Derek when he's speaking to you. Okay?
>
> When I leave the room, Derek will introduce himself and give you the opportunity to familiarize yourself with the scenario. Please use all the available time to fully acquaint yourself with the entire packet. You may also want to review each scenario before answering its respective question.
>
> This is a new way to conduct interviews and, therefore, might be a little rough around the edges. Please bear with us if there are any problems. Let me know if you experience any difficulty with the equipment. I am leaving now, but please feel free to knock on the door if you need my help. The interview will begin as soon as I leave the room. Any questions?

In the high-dialog-capability conditions, the research assistant presented the following additional instructions:

> Please look at Derek when you are ready for the next question. Derek can hear and see you.
>
> Sometimes, survey questions use ordinary words in unexpected ways. To be sure you understand the question, you may need to ask Derek to clarify particular words so please ask for clarification if you are *at all* unsure about what they mean. In fact, you may need to get clarification from Derek in order to answer accurately. Unlike what happens in some survey interviews, Derek *is* able to help you when you indicate you need help. So you should be sure to ask Derek for clarification if you are at all unsure about what a word means.

This description of the respondent's role in conversational interviews parallels the additional instructions in Schober and Conrad (1997).

The research assistant then left the room and the interview proceeded, starting with a first training question and scenario to familiarize the respondent with the task. The research assistant, who monitored the video and audio of the interview along with the Wizard, was available to enter the room if there were technical difficulties or if the respondent gave evidence of not having understood the instructions (e.g., about turning the page in their scenario packet for each next survey question).

After the interview, the research assistant escorted respondents to another lab room, where they filled out the on-line post-experiment questionnaire. Finally, they were asked whether they felt they were indeed interacting with a computer (to give them the opportunity to voice any suspicions that the virtual interviewer was wizarded), debriefed about the actual Wizard-of-Oz experiment setup, and paid for their participation.

The reported analyses are based on the 73 respondents who gave no evidence in the experiment debriefing of suspecting that the virtual interviewer was wizarded. From transcripts of the interviews, we know that no participant ever expressed any suspicion or asked any questions about how the virtual interviewer worked during the interview.

RESULTS

Comprehension

To test our Hypotheses 1–3 about comprehension, we first focus on response accuracy and then on respondents' and virtual interviewers' clarification behaviors. We adopt conventional thresholds for alpha, with levels of $p < 0.05$ as statistically significant (reliable) and $0.05 < p < 0.10$ as marginal.

Response Accuracy

Respondents' comprehension was measured by observing, for each response, whether it matched what the official definition of the survey term would require.

As **Figure 2** shows, Hypothesis 1 was supported: virtual interviewers with high dialog capability led to significantly greater response accuracy (74.3%) than virtual interviewers with low dialog capability (60.2%), $F_{(1, 69)} = 21.69$, $p < 0.001$, $\eta^2 = 239$. This was entirely driven by the effect of dialog capability on response accuracy for complicated mapping scenarios (50.9% for high dialog capability and 25.9% for low dialog capability interviewers); in contrast, for straightforward mappings there was no effect of interviewer dialog capability on response accuracy (response accuracy was uniformly high in all conditions), as demonstrated by the interaction of mapping by dialog capability $F_{(1, 69)} = 15.38$, $p < 0.001$, $\eta^2 = 0.182$.

Figure 2 also shows that, contrary to Hypothesis 2, there was no evidence that the virtual interviewer's facial animation affected response accuracy, $F_{(1, 69)} = 0.15$, $p = 0.70$, $\eta^2 = 0.002$. To further investigate whether there really was no effect of facial animation on response accuracy, we computed a $Bayes_{10}$ factor (using the JASP, 2015 package) comparing the fit of the data under the null hypothesis (no effect of facial animation) and the alternative (see Jarosz and Wiley, 2014 for an account of the underlying logic). An estimated $Bayes_{10}$ factor (alternative/null) of 0.193 suggested that the data were 5.18:1 in favor of the null hypothesis, that is, 5.18 times more likely to occur under a model without including an effect of facial animation, rather than a model with it (in comparison, an estimated Bayes factor [alternative/null] for dialog capability is 2.092 in favor of the alternative hypothesis).

Contrary to Hypothesis 3, and further supporting the interpretation that the virtual interviewer's dialog capability was entirely responsible for response accuracy, is the finding that the interaction between response accuracy and facial animation was not significant, $F_{(1, 69)} = 0.006$, $p = 0.94$, $\eta^2 = 0.000$; the $Bayes_{10}$ factor for the interaction between dialog capability and facial animation is 0.386, suggesting that the data are 2.59:1 in favor of the null hypothesis.

Clarification Behaviors

So that we could examine direct and indirect requests for clarification and their relationship with respondents' comprehension, complete transcripts of the survey question-answer sequences in each interview were created and coded. A coding scheme for all interviewer and respondent moves (see Supplementary Table 4) was adapted from our previous studies with human interviewers (Schober et al., 2004) that included

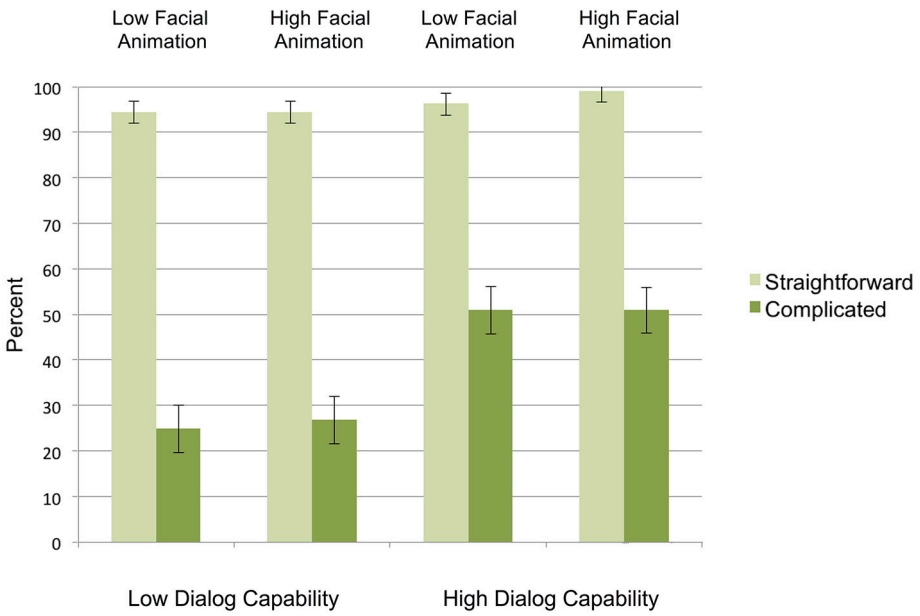


FIGURE 2 | Response accuracy (percentage of survey responses that matched what the official definition would require) for straightforward and complicated scenarios (error bars represent SE's).

codes for the behaviors we expected to differ between high- and low-dialog-capability interviews (e.g., offering clarification, providing definitions, providing neutral probes). In order to verify reliability of the coding, the majority of the question-answer sequences (86.6%) were coded again by a different coder; agreement between these two sets of codes was measured with Cohen's kappa, which at 0.988 was "almost perfect" by Everitt and Haye's (1992, p. 50) characterization.

Consistent with Hypothesis 1, respondents only ever requested or received clarification in the high dialog capability conditions, and not at all in the low dialog capability conditions. This makes sense because of course any requests with a low-dialog-capability virtual interviewer would be met with a neutral probe (e.g., "Let me repeat the question" or "whatever it means to you") rather than substantive clarification (e.g., "In this survey we do not include floor lamps as furniture").

Also consistent with Hypothesis 1 (see **Table 4**), respondents with the high-dialog-capability virtual interviewers explicitly requested clarification more often—nearly twice as often—for complicated scenarios than for straightforward scenarios, and they correspondingly received clarification more than twice as often for complicated scenarios. The virtual interviewer also was more likely to comment on the respondent's need for clarification for complicated scenarios. Compared to explicitly requesting clarification, respondents indirectly indicated that they were having comprehension trouble (e.g., "I don't know whether to count that or not") far less frequently, and they did not do this at different rates for different scenario types.

Contrary to Hypothesis 2 (see **Table 4**), there was no evidence that respondents in the high dialog capability conditions explicitly requested clarification any more often when the virtual interviewer had high than low facial animation, nor did they reject clarification or receive definitions any more often.

Even though there was no evidence that the virtual interviewer's facial animation affected respondents' requests for clarification, respondents with high animation virtual interviewers did have different clarification dialog experiences in a few other ways. Respondents with the high animation virtual interviewer were marginally more likely to be presented with a comment about their confusion ("It sounds like you're having some trouble") than respondents with the low animation virtual interviewer. This is potentially consistent with Hypothesis 2, to the extent that respondents' non-verbal or paralinguistic evidence of confusion (beyond explicit or implicit verbal requests for clarification) differed enough between high and low animation virtual interviewers so as to affect the Wizard's presentation of such comments. On the other hand, Hypothesis 2 seems clearly contradicted by the less interpretable finding that respondents with a high facial animation virtual interviewer were reliably *less* likely to be offered unsolicited clarification. This would make sense if we saw other evidence that respondents requested clarification or provided evidence of confusion more with the low facial animation interviewer, but that is not what we observe. In any case, although we see little evidence for Hypothesis 2, the fact that clarification dialog can proceed differently when the interviewer has high or low facial animation suggests that the impact of facial animation on clarification dialog deserves further exploration.

Analyses of potential interactive effects of the interviewer's dialog capability and facial animation on respondents' requests for clarification and receiving clarification are not significant.

TABLE 4 | Percentage of question-answer sequences in which clarification and related speech occurred (SE's in parentheses).

	Scenario mapping		Effect	Facial animation		Effect
	Straight forward	Complicated		Low	High	
Respondent explicit requests for clarification ("What do you mean by 'furniture'?")	**18.1 (3.7)**	**35.2 (5.2)**	**$F_{(1, 35)} = 20.74$, $p < 0.001$, $\eta^2 = 0.372$**	29.2 (5.9)	24.1 (5.8)	$F_{(1, 35)} = 0.37$, $p = 0.55$, $\eta^2 = 0.011$
Respondent implicit requests for clarification ("I don't know whether to count that or not")	6.3 (1.8)	4.4 (1.8)	$F_{(1, 35)} = 0.88$, $p = 0.354$, $\eta^2 = 0.025$	4.6 (2.1)	6.2 (2.1)	$F_{(1, 35)} = 0.27$, $p = 0.605$, $\eta^2 = 0.008$
Virtual interviewer comments on respondent's confusion ("It sounds like you're having some trouble.")	**3.6 (1.2)**	**8.9 (1.7)**	**$F_{(1, 35)} = 8.08$, $p = 0.007$, $\eta^2 = 0.188$,**	**4.2 (1.6)**	**8.4 (1.6)**	**$F_{(1, 35)} = 3.42$, $p = 0.073$, $\eta^2 = 0.089$**
Virtual interviewer offers clarification ("Can I help you?")	25.8 (3.4)	25.2 (3.2)	$F_{(1, 35)} = 0.022$, $p = 0.882$, $\eta^2 = 0.001$	**31.5 (3.8)**	**19.6 (3.7)**	**$F_{(1, 35)} = 4.98$, $p = 0.032$, $\eta^2 = 0.124$**
Respondent rejects offer	5.3 (1.6)	3.2 (1.1)	$F_{(1, 35)} = 1.36$, $p = 0.251$, $\eta^2 = 0.037$,	5.1 (3.4)	3.4 (1.4)	$F_{(1, 35)} = 0.67$, $p = 0.42$, $\eta^2 = 0.019$
Virtual interviewer presents definition	**16.3 (3.5)**	**36.6 (4.5)**	**$F_{(1, 35)} = 26.55$, $p < 0.001$, $\eta^2 = 0.431$**	29.6 (5.1)	23.3 (5.0)	$F_{(1, 35)} = 0.80$, $p = 0.38$, $\eta^2 = 0.022$

Statistically reliable and marginal differences are in bold face.

Consistent with the response accuracy evidence, Hypothesis 3 is not supported by evidence from clarification behavior.

Respondents' Engagement

To test our Hypotheses 4–6 about respondents' engagement, we first focus on respondents' gaze at the virtual interviewers, and then on their acknowledgment behaviors, smiles, and subjective assessments of the virtual interviewer.

Gaze at the Virtual Interviewer

From the video recordings of respondents' faces, we used Sequence Viewer (http://www.sequenceviewer.nl/) to code whether respondents were looking at the screen (i.e., at the virtual interviewer), at their paper packet, or elsewhere at every moment in each interview (from the research assistants' observations of video monitors during the pre-interview training sessions, we knew that respondents had all looked at the virtual interviewer for several minutes before the survey interview, as instructed). Respondents looked almost exclusively at their scenario packet and the virtual interviewer; they looked elsewhere in the room so rarely (less than 1% of the time) as to be negligible (see **Figure 3**).

Consistent with Hypothesis 4, respondents spent a greater proportion of the interview time looking at the high-dialog-capability virtual interviewers (29.8% of the time) than the low-dialog-capability virtual interviewers (21.1%), $F_{(1, 69)} = 6.73$, $p = 0.012$, $\eta^2 = 0.089$. In order to further explore this phenomenon (that is, to further understand how respondents' engagement as measured by gaze connected with clarification dialog), we examined respondents' gaze at the virtual interviewers for complicated and straightforward scenarios, because it was only in complicated scenarios that clarification dialog ever occurred. As **Figure 3** shows, respondents looked slightly but reliably *less* at the virtual interviewer (and more at their scenario packets) when the mappings between questions and scenarios were complicated (24.7% of the time) rather than straightforward (26.3% of the time), $F_{(1, 69)} = 4.20$, $p < 0.05$, $\eta^2 = 0.057$. This overall difference resulted particularly from the low-dialog-capability conditions (19.2% of the time for complicated scenarios and 23.1% for straightforward) rather than the high-dialog-capability conditions, where there was no difference in the proportions of time spent looking at the virtual interviewer based on scenario mappings (30.1 vs. 29.6%), interaction $F_{(1, 69)} = 7.16$, $p < 0.01$, $\eta^2 = 0.094$. Our interpretation is that in the low-dialog-capability conditions respondents were left to their own devices to figure out the right answer to the survey question, and so the only available useful information, if the virtual interviewer would not provide clarification, could come from examining the scenarios more closely. In the high dialog capability conditions, engagement with the virtual interviewer through gaze was greater and not related to the content of the scenarios[1].

[1]This pattern of findings rules out an interpretation that greater gaze at the high-dialog-capability virtual interviewer resulted simply from respondents looking at the interviewer more at turn transitions, which is a well-documented phenomenon

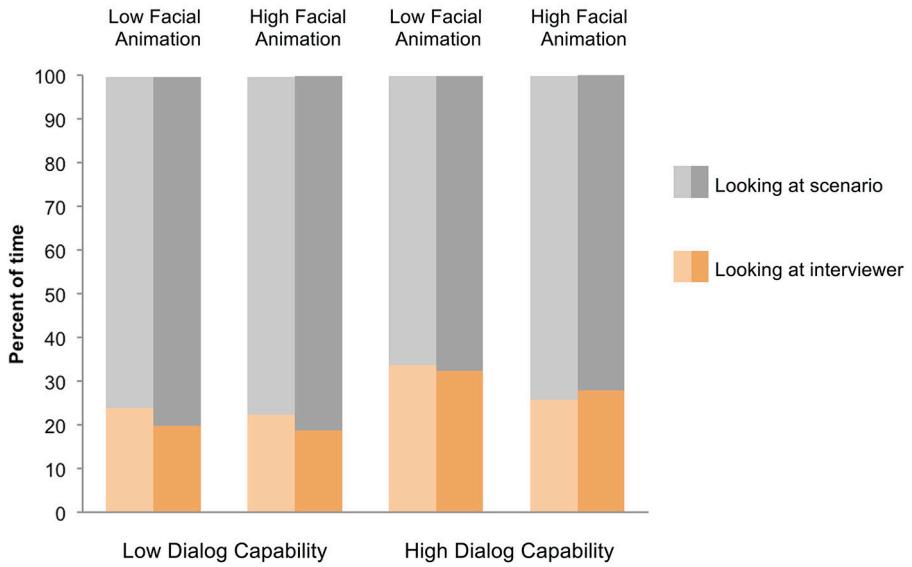


FIGURE 3 | Percentage of time that respondents looked at the virtual interviewer and the scenario packet, on average, across the four conditions, broken down by whether they were answering questions that mapped onto the scenario in a straightforward (lighter shades) or complicated (darker shades) way. Gaze elsewhere (not at the virtual interviewer or scenario packet) was so rare (less than 1% of the time in all conditions) that it is not plotted.

Contrary to Hypothesis 5, there is not sufficient evidence that respondents looked more at the virtual interviewers with high facial animation than those with low facial animation, $F_{(1, 69)} = 1.22$, $p = 0.27$, $\eta^2 = 0.017$. An estimated $Bayes_{10}$ factor (alternative/null) of 0.669 suggested that the data were 1.49:1 in favor of the null hypothesis, that is, 1.49 times more likely to occur under a model that does not include an effect of facial animation, rather than a model that does include it.

Contrary to Hypothesis 6, the virtual interviewer's facial animation did not interact with its dialog capability in affecting respondents' gaze behavior, $F_{(1, 69)} = 0.50$, $p = 0.48$, $\eta^2 = 0.017$. An estimated $Bayes_{10}$ factor (alternative/null) of 1.771 does not rule out the possibility that the data may favor Hypothesis 6, but it seems unlikely.

There are at least two possible explanations for this pattern of results—that gaze increased with high-dialog-capability but not high-facial-animation interviewers—given that our experimental conditions varied on more than one feature. One is that respondents with a high-dialog-capability virtual interviewer found the content of the interviewer's contributions (e.g., clarification dialog) compelling and human-like enough to spend a greater proportion of their time looking at the interviewer. Another is that respondents with a high-dialog-capability virtual interviewer fully trusted what they were told about the interviewer's perceptual capacity in the experiment instructions: that the high-dialog-capability interviewer could perceive their facial expression and gaze. The fact that respondents in the high-dialog-capability conditions were explicitly instructed to look at the interviewer when ready for the next question makes disentangling this more difficult, but we note that the increase in looking time at the high-dialog-capability interviewer is *proportional*, and occurs along with a substantial increase in interview duration; high-dialog-capability interviews took 7.26 min on average (SE 0.36 min) compared with low-dialog-capability interviews (5.53 min, SE 0.37 min), $F_{(1, 69)} = 11.23$, $p = 0.001$, $\eta^2 = 0.140$. So the increase in looking time seems to us unlikely to result only from looking at the interviewer during transitions between survey questions, which would need to be quite long (a full minute of the interview, or a full 5 s at each question transition) to account for the effect.

Although respondents in this experiment did look at their paper packets a substantial proportion of the time during the interview (which means that at those moments they could only have been listening to—not watching—the virtual interviewer), we consider the proportions of time looking at the virtual interviewer observed here to be sufficient to allow us to detect potential effects of the virtual interviewer's facial animation even in the conditions with less looking time. The fact that we did observe significant differences in multiple measures based on facial animation corroborates this judgment.

Acknowledgment Behaviors

In face-to-face interactions interlocutors can acknowledge each other's utterances verbally and visually: they can use back channel utterances (e.g., "okay," "all right," "got it," "thank you"; Yngve, 1970) and they can nod, shake their heads, shrug their shoulders, raise their eyebrows, etc., in order to communicate continued attention and possible understanding (Allwood et al., 1992; McClave, 2000). Verbal and visual acknowledgments can be seen as part of an integrated multimodal system

in human dialog (e.g., Kendon, 1967). If this were the case, then there should be more looking at the interviewer for complicated than straightforward scenarios with the high dialog capability interviewers, because complicated scenarios involved more transitions (because of more clarification).

(Carter and Adolphs, 2008) that displays engagement in an interaction.

To examine acknowledgments in our virtual interviews, we counted respondents' verbal back channel utterances from the interactional moves we had coded (see Supplementary Table 4). We also coded head movements (nods, head shakes, other head movements like tilts), and other body or facial movements (like shoulder shrugs and eyebrow raising), using Sequence Viewer, based on the video recordings of the interviews. Just as reliability was measured for the interactional move coding (86.6% of question-answer sequences double-coded, see Section Comprehension), it was measured for these behaviors as well. Each of the individual behaviors was relatively rare in our sample, but coders' level of agreement was high: for head movements the coders' judgments agreed 92.5% of the time, and for other body movements they agreed 94.1% of the time [Cohen's kappas for these reliabilities were low, at 0.32 and 0.27, but as Viera and Garrett (2005) demonstrate, kappa can easily be a misleading index of agreement when the occurrence of what is coded is rare].

In our tests of Hypotheses 4–6, we first looked at verbal backchannels alone, head movements alone, and particular body and facial movements alone. Because backchannels and particular head movements and particular body and facial movements occurred rarely enough that there was a risk that we would miss patterns relevant to our hypotheses given our sample size, we also aggregated across nods, head shakes, other head movements, and other body and facial movements.

For Hypothesis 4 (effects of interviewer's dialog capability on respondent engagement), we see only suggestive evidence in support of it. Respondents did not produce many backchannels (and many produced none), but they produced marginally more of them with the high dialog capability agents (0.32 per interview) than with the low-dialog-capability agents (0.18 per interview), $F_{(1, 69)} = 2.82$, $p = 0.098$, $\eta^2 = 0.039$. Analyses of all facial and bodily movements do not show any significant effects.

The evidence for Hypothesis 5 (effects of interviewer facial animation on respondent engagement) is also suggestive. Respondents were marginally more likely to produce one of these movements when the virtual interviewer had high facial animation (averaging 0.13 occurrences per speaking turn) than when virtual interviewer had low facial animation (0.08 occurrences per speaking turn, $F_{(1, 69)} = 3.21$, $p = 0.078$, $\eta^2 = 0.039$. But support for Hypothesis 5 becomes stronger if we also include verbal back channel utterances, taking Carter and Adolphs' (2008) multimodal view of acknowledgment behavior. As **Figure 4** shows, respondents were nearly twice as likely to display our aggregated acknowledgment behaviors (visual and verbal) when the virtual interviewer had high facial animation (at a rate of 0.18 occurrences per speaking turn) than when the virtual interview had low facial animation (0.11 occurrences per speaking turn), $F_{(1, 69)} = 4.29$, $p < 0.05$, $\eta^2 = 0.059$.

Hypothesis 6 predicted an interaction of the form that respondents would produce disproportionately more engagement behaviors with high-dialog-capability high-facial-animation virtual interviewers, and proportionately fewer with low-dialog-capability low-facial-animation interviewers. We see partial evidence in support of this hypothesis in one significant interaction of interviewer dialog capability and facial animation with respect to nods, $F_{(1, 69)} = 5.81$, $p = 0.019$, $\eta^2 = 0.078$. Partially consistent with Hypothesis 6, respondents nodded least with the low-dialog capability low-facial-animation interviewer (0.05 times per interview), but (unexpectedly) most with the high-dialog-capability *low*-facial animation-interviewer (0.26 times per interview). There were no other significant interaction effects.

Smiles

Another measure of respondents' engagement with the virtual interviewers is their frequency of smiling.

We thus coded respondents' smiles in order to compute smile frequency and duration. The coder (one of the authors) had been certified in the Facial Action Coding System (FACS; Ekman and Friesen, 1978). We determined coding reliability for all the question-answer sequences for a subsample of 20% of the respondents, equally distributed in the four experimental conditions, as independently coded by a second coder (four respondents had to be excluded from this analysis because the resolution of the video was not sufficient for this level of facial coding). Coders' level of agreement on smile frequency was high (92.1%), with a Cohen's kappa of 0.66. Coders' judgments on smile duration were also highly correlated, $r = 0.835$, $p < 0.0001$ (considering all sequences) and $r = 0.762$, $p < 0.0001$ (considering only those sequences in which at least one smile was found by at least one coder).

Regarding Hypothesis 4, there were no reliable effects of the interviewer's dialog capability on smiles.

Regarding Hypothesis 5, respondents interacting with a high facial animation virtual interviewer smiled marginally more often (2.25 times over the course of their interview, SE 0.55) than respondents interacting with a low facial animation virtual interviewer (0.78 times, SE 0.55), $F_{(1, 68)} = 3.62$, $p = 0.061$, $\eta^2 = 0.050$. Respondents interacting with a high facial animation virtual interviewer also smiled marginally longer (11.5 s over the course of the interview, SE 3.1) than respondents interacting with a low facial animation virtual interviewer (3.0 s, SE 3.1), $F_{(1, 68)} = 3.67$, $p = 0.060$, $\eta^2 = 0.051$.

Regarding Hypothesis 6, there were no significant interactive effects of virtual interviewers' dialog capability and facial animation on respondents' smiles.

Respondents' Self-reported Subjective Experience

A final set of measures of respondents' engagements was their responses to the post-experiment questionnaire in which they reported how they felt about and evaluated the virtual interviewers.

Table 5 presents the average ratings as well as ANOVA statistics for tests of Hypotheses 4–6[2]. Note that all of these ratings are lower than one would expect if the respondents

[2]We report parametric statistical analyses on these rating scale data so as to straightforwardly test our hypotheses about our two experimental factors and potential interactions. This approach is supported by Norman's (2010) arguments and evidence about the robustness of parametric analyses for interval data. But this does not mean that we are claiming that our participants treated the distances between intervals on our scales as equal, which is, of course, unknowable (Jamieson, 2004); we simply are claiming that higher ratings are higher.

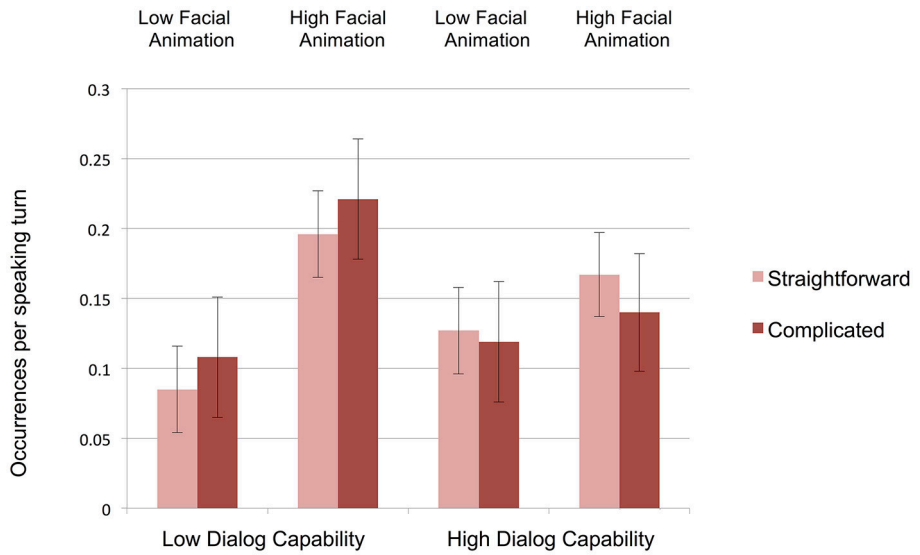


FIGURE 4 | Respondents' rates of aggregated acknowledgment behaviors (verbal back channels, nods, head shakes, other head movements, and other body and facial movements) per speaking turn (error bars represent SE's).

evaluated the virtual interviewer as being very human-like. But given the constraints of a standardized interviewing situation, it is also plausible that human interviewers who implemented these interviews would not be rated as particularly autonomous, personal, close, or sensitive, and they might also be rated as more robotic than human (the term "robotic" is sometimes used to caricature the behavior of rigidly standardized interviewers, for example in survey invitations, see Conrad et al., 2013).

As detailed in **Table 5**, Hypothesis 4 is supported on several fronts. Respondents with an interviewer high in dialog capability reported enjoying the interview more, and they rated the interviewer as more autonomous, more personal, less distant, and more sensitive than respondents with an interviewer low in dialog capability[3]. They also rated the interviewer as less like a computer. Unexpectedly, respondents with high dialog capability interviewers reported a greater decrease in comfort across the interview than respondents with the low dialog capability interviewers.

In contrast to the predictions of Hypothesis 5, there were significant effects of facial animation suggesting that interviewers with *low* facial animation were in some ways preferred. Respondents with low facial animation interviewers reported marginally greater comfort with the interviewer at the start of the session, and they rated the interviewer as marginally more natural and as reliably more autonomous (acting on his own), than did respondents with high facial animation interviewers.

[3]We interpret the finding that autonomy was rated as lower for low-dialog-capability agents as reflecting respondents' assessment of the virtual interviewer's ability to reason and think ("act on his own"), as opposed to reflecting a judgment that the virtual interviewer had a human operator rather than being stand-alone software. While we can't, of course, rule out this possibility, the fact that the patterns of ratings are consistent on more items beyond the question about autonomy supports this view.

The pattern uncovered in tests of Hypothesis 6 is consistent with that found for acknowledgments. Respondents with low facial animation interviewers were more likely (albeit marginally) to rate the interviewer as autonomous when the interviewer had high dialog capability (see **Table 5**). These same respondents were also particularly more likely to rate the low facial animation interviewer as more personal, as less distant (closer), and as marginally more like a person than a computer. In other words, respondents found the interviewer to be particularly autonomous and personal when he looked more robotic (displayed less facial movement) but could converse like a human. The fact that the mean ratings in this condition (low facial animation/high dialog capability) stand out from the others, along with the (marginal) interaction effects, suggests that part of what is driving the main effects of dialog capability and facial animation on these items are the perceptions of this subgroup.

DISCUSSION

Summary

The findings reported here document that two important elements of human face-to-face interaction—dialog capability and facial movement—implemented in virtual survey interviewers differently affect respondents' comprehension and the nature of their engagement with the virtual interviewer. As tested in Hypotheses 1 and 4, respondents who interacted with a virtual interviewer with greater dialog capability (that is, which could help respondents interpret the questions as intended) provided more accurate answers and took more responsibility for their comprehension, requesting clarification more often. They looked at high-dialog-capability interviewers more, they produced marginally more backchannel responses, and they reported enjoying the interview more and finding

TABLE 5 | Respondents' subjective ratings of the virtual interviewers, presented in the order the ratings were elicited (SE's in parentheses).

	Response options	Low dialog capability		High dialog capability		Test of hypothesis 4: effect of dialog capability	Test of hypothesis 5: effect of facial animation	Test of hypothesis 6: interaction of dialog capability and facial animation
		Low facial animation	High facial animation	Low facial animation	High facial animation			
How comfortable were you with Derek at the start of the session?	1: Not at all comfortable 5: Very comfortable	3.83 (0.27)	3.22 (0.27)	3.33 (0.27)	3.00 (0.26)	$F_{(1, 69)} = 1.83$, $p = 0.180$, $\eta^2 = 0.026$	**$F_{(1, 69)} = 3.13$, $p = 0.081$, $\eta^2 = 0.043$**	$F_{(1, 69)} = 0.27$, $p = 0.604$, $\eta^2 = 0.004$
As the interview progressed, did your comfort with Derek increase, decrease, or stay the same?	1: Decrease 2: Stay the same 3: Increase	1.78 (0.15)	1.78 (0.15)	1.56 (0.15)	1.21 (0.15)	**$F_{(1, 69)} = 7.05$, $p = 0.010$, $\eta^2 = 0.093$**	$F_{(1, 69)} = 1.35$, $p = 0.250$, $\eta^2 = 0.019$	$F_{(1, 69)} = 1.35$, $p = 0.250$, $\eta^2 = 0.019$
How natural was the interaction with Derek?	1: Not at all natural 5: Very natural	3.22 (0.27)	2.61 (0.27)	3.17 (0.27)	2.74 (0.27)	$F_{(1, 69)} = 0.02$, $p = 0.898$, $\eta^2 = 0.000$	**$F_{(1, 69)} = 3.65$, $p = 0.06$, $\eta^2 = 0.050$**	$F_{(1, 69)} = 0.11$, $p = 0.741$, $\eta^2 = 0.002$
How often did Derek seem to act on his own?	1: Never 5: All the time	2.56 (0.25)	2.44 (0.25)	3.83 (0.25)	2.74 (0.25)	**$F_{(1, 69)} = 9.82$, $p = 0.003$, $\eta^2 = 0.124$**	**$F_{(1, 69)} = 5.80$, $p = 0.019$, $\eta^2 = 0.078$**	**$F_{(1, 69)} = 3.86$, $p = 0.053$, $\eta^2 = 0.053$**
Would you say that Derek acted more like a computer or a person?	1: Just like a computer 2: As much like a computer as a person 3: Just like a person	1.44 (0.152)	1.61 (0.15)	2.11 (0.15)	1.68 (0.15)	**$F_{(1, 69)} = 6.02$, $p = 0.017$, $\eta^2 = 0.080$**	$F_{(1, 69)} = 0.75$, $p = 0.391$, $\eta^2 = 0.011$	**$F_{(1, 69)} = 3.88$, $p = 0.053$, $\eta^2 = 0.053$**
How much did you enjoy interacting with Derek?	1: Did not enjoy at all 5: Thoroughly enjoyed	3.06 (0.22)	3.22 (0.22)	3.83 (0.22)	3.37 (0.22)	**$F_{(1, 69)} = 4.41$, $p = 0.039$, $\eta^2 = 0.060$**	$F_{(1, 69)} = 0.46$, $p = 0.500$, $\eta^2 = 0.007$	$F_{(1, 69)} = 2.06$, $p = 0.156$, $\eta^2 = 0.029$
How frustrating was it to be interviewed by Derek?	1: Not at all frustrating 5: Very frustrating	1.88 (0.24)	2.11 (0.23)	1.50 (0.23)	1.84 (0.22)	$F_{(1, 68)} = 2.01$, $p = 0.161$, $\eta^2 = 0.029$	$F_{(1, 68)} = 1.54$, $p = 0.219$, $\eta^2 = 0.022$	$F_{(1, 68)} = 0.06$, $p = 0.806$, $\eta^2 = 0.001$
I felt that Derek was...	1: Impersonal 5: Personal	1.94 (0.23)	2.35 (0.24)	3.28 (0.23)	2.68 (0.23)	**$F_{(1, 68)} = 12.72$, $p = 0.001$, $\eta^2 = 0.158$**	$F_{(1, 68)} = 0.16$, $p = 0.693$, $\eta^2 = 0.002$	**$F_{(1, 68)} = 4.61$, $p = 0.035$, $\eta^2 = 0.063$**
I felt that Derek was...	1: Distant 5: Close	2.17 (0.24)	2.77 (0.25)	3.19 (0.26)	2.79 (0.24)	**$F_{(1, 66)} = 4.55$, $p = 0.037$, $\eta^2 = 0.064$**	$F_{(1, 66)} = 0.17$, $p = 0.685$, $\eta^2 = 0.003$	**$F_{(1, 66)} = 4.13$, $p = 0.046$, $\eta^2 = 0.059$**
I felt that Derek was...	1: Inexpressive 5: Expressive	2.56 (0.29)	2.78 (0.29)	3.00 (0.29)	2.84 (0.28)	$F_{(1, 68)} = 0.79$, $p = 0.377$, $\eta^2 = 0.011$	$F_{(1, 68)} = 0.01$, $p = 0.911$, $\eta^2 = 0.000$	$F_{(1, 68)} = 0.44$, $p = 0.509$, $\eta^2 = 0.006$
I felt that Derek was...	1: Insensitive 5: Sensitive	2.56 (0.22)	2.94 (0.22)	3.22 (0.22)	3.32 (0.22)	**$F_{(1, 69)} = 5.59$, $p = 0.021$, $\eta^2 = 0.075$**	$F_{(1, 69)} = 1.21$, $p = 0.275$, $\eta^2 = 0.017$	$F_{(1, 69)} = 0.45$, $p = 0.503$, $\eta^2 = 0.007$

Statistically significant and marginal effects are in bold.

the interviewer to be more personal and less distant. As tested in Hypotheses 2 and 5, respondents who interacted with a virtual interviewer with more facial animation displayed more evidence of engagement—more verbal back channels and visual acknowledgments of the interviewer's utterances, and marginally more smiles. They also reported *less* comfort with the high facial animation interviewers and rated these interviewers as less natural. In testing Hypotheses 3 and 6, we observed that respondents (unexpectedly) nodded more and rated the virtual interviewer as more personal and less distant if it had high dialog capability and *low* facial animation.

The current findings extend work on people's reactions and behaviors when they talk with interviewing agents, for example telling stories to an agent that exhibits listening behavior (e.g., Gratch et al., 2006; von der Pütten et al., 2010), answering open-ended questions asked by a peer (e.g., Bailenson et al., 2006) or answering open-ended questions asked by a deception-detecting kiosk agent (Nunamaker et al., 2011), to the task of a survey interview for social measurement that uses closed categories as response options and that is designed to make statistical estimates of a population. The findings also extend work on disclosure of sensitive information in a survey interview with a virtual interviewer (Lind et al., 2013) to an interview with non-sensitive questions that have verifiably correct and incorrect answers, and in which accurate comprehension of the terms in the questions is critical. Because of the nature of this survey task, our measures focus on aspects of the interaction and of respondents' behavior (e.g., response accuracy, smiles, acknowledgments) that have not been the focus in previous studies, where users' nuanced interpretation of what the virtual interviewer is asking is less essential.

While it is unclear where exactly our survey task fits into a taxonomy of tasks for which virtual humans have been designed, what is clear is that for this task the two features we experimentally manipulated have quite distinct effects. We assume this is because they engage different channels of communication (the exchange of spoken vs. visual information) and manifest themselves over different time scales—a virtual agent's facial animation is visible to users as soon as any talking starts, while evidence of the agent's dialog capability unfolds more incrementally over time as the interviewer does or does not respond to the user's need for clarification. We hypothesize that our findings should generalize to other interactive tasks with virtual agents that share the central features of the current task: a need for grounding interpretation of terms in an agent's utterances and a need for the user to be sufficiently engaged to complete a task that someone else has initiated (Schober et al., 2003).

While our experimental design allows us to see effects of what we manipulated, it does not allow us to disentangle the relative contributions of the bundled features that comprise the different agents. Of course, our agents' particular features could have been implemented differently (e.g., the agents could have had different vocal or visual attributes, or been unnamed or have had different names), and it is unknown how our findings would generalize to different implementations. Our experimental design also does not allow inference about potential (and intriguing) causal connections between our different measures. For example, we do not know whether respondents' attributions about the high-dialog-capability interviewer *result from* or *cause* or are *independent of* their improved comprehension: did respondents answer more accurately with a high dialog capability virtual interviewer because they enjoyed the interview more and found the interviewer more perceptive and responsive? Or did they enjoy the interview more because they were confident that they had comprehended the questions as intended? Did respondents smile more often and longer with a high facial animation virtual interviewer because they felt more engaged, as one might expect given Krämer et al.'s (2013) finding that users who were engaged in small talk with a virtual agent smiled more when the virtual agent smiled more? Or, alternatively (and consistent with our respondents' reports of less comfort), did they smile more because their smiles reflected distress or discomfort (e.g., Ansfield, 2007)? The fact that respondents' subjective experience of a virtual survey interviewer—their level of comfort, their enjoyment, how natural they feel the interaction to be—can be correlated with their disclosure of sensitive information (Lind et al., 2013) makes it plausible that users' affective reactions could be causally connected with their comprehension and behavioral displays even with non-sensitive survey questions of the sort asked here, but the current data only allow speculation.

Designing Virtual Survey Interviewers

Animating virtual interviewing agents that could be used in, for example, a web survey with textual response is becoming increasingly straightforward with off-the-shelf tools. Instantiating dialog capability and speech recognition is a greater challenge, but the constrained nature of the survey interview task (a finite set of possible turns that can occur, standardized wording of questions, closed response options with limited vocabulary that a speech recognition system can handle, definitions of key terms already existing) can make implementing clarification dialog in a textual or speech interviewing system more plausible than in more open-ended or free-form conversational domains (Johnston et al., 2013; Schober et al., 2015).

Given the many possible ways to instantiate a virtual interviewer—a range of possible expressivity, sensing capabilities and responsiveness to respondents' signals, and a range of more and less human-like facial motion and detail—we propose the following design considerations for building virtual interviewers for actual surveys that produce population estimates:

- *Designing to maximize participation:* Potential respondents are likely to vary in whether they will consent to interact with a virtual interviewer, for example, in an online survey. Perhaps the greatest deterrent is uncanniness (e.g., MacDorman et al., 2009). The fact that participants in the current study reported that the virtual interviewers with more facial animation made them less comfortable and were less natural than virtual interviewers with less facial movement could result from people's finding the increased realism of high facial animation to be eerie, and this might reduce participation in virtual interviews by some sample members. But for others, this might

not affect participation; in the Lind et al. (2013) study with a more cartoon-like interviewer, different respondents had completely opposite affective reactions from each other to the very same interviewing agent, and this correlated with their willingness to disclose sensitive information.

- *Designing to maximize completion:* Although in this study we did not include an interviewing condition without a facial representation, the increased engagement (more acknowledgments and smiles) that we observed with the high facial animation interviewers could translate to increased completion of questionnaires compared to self-administered online questionnaires without any virtual interviewer. Engagement could promote completion if respondents apply social norms from face-to-face interaction in which it would be rude to break off a conversation midstream, or because a moving talking face simply makes the task more interesting. To investigate this, one would need to carry out a study outside the laboratory (e.g., online) with naturalistic incentives (rather than our laboratory method with payment).
- *Designing to maximize comprehension:* As we have proposed for human interviewers (Schober and Conrad, 1997; Conrad and Schober, 2000), enabling virtual survey interviewers to engage in clarification dialog is likely to improve respondents' understanding of questions and thus the quality of the data collected in the survey. There are a number of ways to instantiate clarification dialog in a virtual interviewer, from providing scripted (spoken or even textual) definitions only when respondents request them to diagnosing the potential need for clarification based on respondents' disfluencies and gaze aversion (e.g., Ehlen et al., 2007; Schober et al., 2012). The findings in the current study suggest that system-initiated clarification is likely to be important for maximizing comprehension.
- *Designing the interviewer's appearance and voice:* It is essentially impossible to design a virtual human interviewer without creating the perception of some demographic characteristics. If the virtual interviewer communicates by speaking, its speech will inevitably have attributes such as dialect, a pitch range, prosody, and vocal quality. How the current findings, which are based on one 3D head model with particular visual and linguistic attributes, will generalize to virtual interviewers with other visual and linguistic attributes, will be a key design question: how a virtual interviewer's visual attributes (skin shade, eye color, hair style, facial features, clothing, hair covering, etc.) or speech style (accent, vocabulary, pronunciation) will affect respondents' judgments about the interviewer's perceived "social identity" (gender, race, social class, education, religious affiliation) and potentially respondents' answers to questions on some interview topics. It is well known that demographic characteristics of human interviewers can (undesirably) affect the distribution of responses (e.g., Hatchett and Schuman, 1975) even in telephone interviews where only voice attributes are available (e.g., Cotter et al., 1982; Finkel et al., 1991). There is preliminary evidence that this kind of interviewer effect may also appear with virtual interviewers (Conrad et al., 2011), and that gender and nationality attributions can occur for embodied agents more generally (Eyssel and Hegel, 2012; Eyssel and Kuchenbrandt, 2012).
- *Designing for different types of survey questions*: The current research suggests that virtual interviewers implemented with high dialog capability may promote accurate answers to factual questions about mundane topics for which complicated mappings are possible. However, it has been shown (Lind et al., 2013) that when virtual interviewers ask questions about sensitive topics, respondents seem to answer most questions less truthfully (disclose less sensitive information) than when the same questions are spoken by a disembodied (audio) interviewer. If a survey investigates both non-sensitive and sensitive topics, one could imagine implementing the virtual interviewer for only the non-sensitive questions. To our knowledge this has never been attempted; much is unknown about how the intermittent display of a virtual interviewer might affect respondents' affective responses and whether removing an interviewer—after being present—could convincingly create a sense of privacy.
- *Giving respondents a choice of interviewer?* One potential advantage of implementing virtual survey interviewers is that one could let *respondents* choose an interviewer with the attributes (appearance, speech style) that they prefer, which is not a possibility with human interviewers. It is entirely unknown which attributes respondents would most want to be able to choose, whether providing choices will increase respondents' engagement and data quality, or how choosing an interviewer that makes respondents most comfortable might affect their effort in producing accurate responses.

Considering factors such as these, as well as those raised by Cassell and Miller (2008), will be essential if virtual survey interviewing systems are to be effective. The need for accurate survey data will continue; the question will be what kinds of interviewers and interviewing systems will best promote accurate data and respondent engagement in new technological environments (Schober and Conrad, 2008), and what role embodied interviewing agents might best play.

ACKNOWLEDGMENTS

Many thanks to Andy Bailin, Xiao Chen, Brian Lewis, Laura Lind, Miyuki Nishimura, Catherine Militello, Christian Prinz, and David Vannette for their assistance. This research was supported by National Science Foundation grants SES-0551294 to FC and MS and SES-0454832 to FC, and by faculty research funds to MS from the New School for Social Research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2015.01578

REFERENCES

Alexanderson, S., and Beskow, J. (2014). Animated Lombard speech: motion capture, facial animation and visual intelligibility of speech produced in adverse conditions. *Comput. Speech Lang.* 28, 607–618. doi: 10.1016/j.csl.2013.02.005

Allwood, J., Nivre, J., and Ahlsen, E. (1992). On the semantics and pragmatics of linguistic feedback. *J. Semantics* 9, 1–26. doi: 10.1093/jos/9.1.1

Anderson, A. H. (2008). "Video-mediated interactions and surveys," in *Envisioning the Survey Interview of the Future,* eds F. G. Conrad and M. F. Schober (Hoboken, NJ: John Wiley & Sons, Inc), 95–118. doi: 10.1002/9780470183373.ch5

Ansfield, M. E. (2007). Smiling when distressed: when a smile is a frown turned upside down. *Pers. Soc. Psychol. Bull.* 33, 763–775. doi: 10.1177/0146167206297398

Bailenson, J. N., Yee, N., Merget, D., and Schroeder, R. (2006). The effect of behavioral realism and form realism of real-time avatar faces on verbal disclosure, nonverbal disclosure, emotion recognition, and copresence in dyadic interaction. *Presence* 15, 359–372. doi: 10.1162/pres.15.4.359

Baur, T., Damian, I., Gebhard, P., Porayska-Pomsta, K., and André, E. (2013). "A job interview simulation: social cue-based interaction with a virtual character," in *2013 International Conference on Social Computing (SocialCom)* (Alexandria, VA: IEEE), 220–227. doi: 10.1109/SocialCom.2013.39

Baylor, A. L., and Kim, Y. (2004). "Pedagogical agent design: the impact of agent realism, gender, ethnicity, and instructional role," in *Intelligent Tutoring Systems, Proceedings, Lecture Notes in Computer Science,* Vol. 3220 (Berlin; Heidelberg: Springer), 592–603. doi: 10.1007/978-3-540-30139-4_56

Bloom, J. (2008). "The speech IVR as a survey interviewing methodology," in *Envisioning the Survey Interview of the Future,* eds F. G. Conrad and M. F. Schober (Hoboken, NJ: John Wiley & Sons, Inc), 119–136. doi: 10.1002/9780470183373.ch6

Carter, R., and Adolphs, S. (2008). Linking the verbal and visual: new directions for corpus linguistics. *Lang. Comput.* 64, 275–291. doi: 10.1163/9789401205474_019

Cassell, J., and Miller, P. (2008). "Is it self-administration if the computer gives you encouraging looks?" in *Envisioning the Survey Interview of the Future*, eds F. G. Conrad and M. F. Schober (Hoboken, NJ: John Wiley & Sons),161–178.

Chang, L., and Krosnick, J. A. (2010). Comparing oral interviewing with self-administered computerized questionnaires: an experiment. *Public Opin. Q.* 74, 154–167. doi: 10.1093/poq/nfp090

Clark, H. H. (1996). *Using Language*. Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9780511620539

Clark, H. H., and Schaefer, E. F. (1987). Collaborating on contributions to conversations. *Lang. Cogn. Process.* 2, 19–41. doi: 10.1080/01690968708406350

Clark, H. H., and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition* 22, 1–39. doi: 10.1016/0010-0277(86)90010-7

Conrad, F. G., Broome, J. S., Benkí, J. R., Kreuter, F., Groves, R. M., Vannette, D., et al. (2013). Interviewer speech and the success of survey invitations. *J. R. Stat. Soc.* 176, 191–210. doi: 10.1111/j.1467-985X.2012.01064.x

Conrad, F. G., and Schober, M. F. (2000). Clarifying question meaning in a household telephone survey. *Public Opin. Q.* 64, 1–28. doi: 10.1086/316757

Conrad, F. G., and Schober, M. F. (eds.). (2008). *Envisioning the Survey Interview of the Future,* Vol. 542 (Hoboken, NJ: John Wiley & Sons). doi: 10.1002/9780470183373

Conrad, F. G., Schober, M. F., and Coiner, T. (2007). Bringing features of dialogue to web surveys. *Appl. Cogn. Psychol.* 21, 165–187. doi: 10.1002/acp.1335

Conrad, F. G., Schober, M. F., and Nielsen, D. (2011). "Race-of-virtual-interviewer effects," in *The 66th Annual Conference of the American Association for Public Opinion Research* (Phoenix, AZ).

Cotter, P. R., Cohen, J., and Coulter, P. B. (1982). Race-of-interviewer effects in telephone interviews. *Public Opin. Q.* 46, 278–284. doi: 10.1086/268719

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., et al. (2014). "SimSensei Kiosk: a virtual human interviewer for healthcare decision support," in *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems* (Paris: International Foundation for Autonomous Agents and Multiagent Systems), 1061–1068.

Ehlen, P., Schober, M. F., and Conrad, F. G. (2007). Modeling speech disfluency to predict conceptual misalignment in speech survey interfaces. *Discourse Process.* 44, 245–265. doi: 10.1080/01638530701600839

Ekman, P., and Friesen, W. V. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Palo Alto, CA: Consulting Psychologists Press.

Everitt, B. S., and Haye, D. F. (1992). *Talking About Statistics: A Psychologist's Guide to Data Analysis.* New York, NY: Halsted Press.

Eyssel, F., and Hegel, F. (2012). (S)he's got the look: gender stereotyping of robots. *J. Appl. Soc. Psychol.* 42, 2213–2230. doi: 10.1111/j.1559-1816.2012.00937.x

Eyssel, F., and Kuchenbrandt, D. (2012). Social categorization of social robots: anthropomorphism as a function of robot group membership. *Br. J. Soc. Psychol.* 51, 724–731. doi: 10.1111/j.2044-8309.2011.02082.x

Ferrin, M., and Kriesi, H. (2014). *Europeans' Understandings and Evaluations of Democracy: Topline Results from Round 6 of the European Social Survey. ESS Topline Results Series*, *(4)*. Available online at: http://www.europeansocialsurvey.org/permalink/800ea36f-3a8d-11e4-95d4-005056b8065f.pdf

Finkel, S. E., Guterbock, T. M., and Borg, M. J. (1991). Race-of-interviewer effects in a preelection poll Virginia 1989. *Public Opin. Q.* 55, 313–330. doi: 10.1086/269264

Foucault Welles, B., and Miller, P. V. (2013). "Nonverbal correlates of survey rapport: an analysis of interviewer behavior," in *The Cannell Workshop at the 2013 Meeting of the American Association for Public Opinion Research (AAPOR)* (Boston, MA).

Fowler, F. J., and Mangione, T. W. (1990). *Standardized Survey Interviewing: Minimizing Interviewer-Related Error.* Newbury Park, CA: SAGE Publications, Inc.

Geoghegan-Quinn, M. (2012). *Opening Address to the ESS International Conference – 'CrossNational Evidence from European Social Survey: Exploring Public Attitudes, Informing Public Policy in Europe'*, Nicosia. Available online at: http://www.europeansocialsurvey.org/docs/findings/ESS1_5_select_findings.pdf [Accessed Nov 23, 2012].

Gong, L. (2008). How social is social responses to computers? The function of the degree of anthropomorphism in computer representations. *Comput. Hum. Behav.* 24, 1494–1509. doi: 10.1016/j.chb.2007.05.007

Gratch, J., Lucas, G. M., King, A. A., and Morency, L. P. (2014). "It's only a computer: the impact of human-agent interaction in clinical interviews," in *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems* (New York, NY: International Foundation for Autonomous Agents and Multiagent Systems), 85–92.

Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R. J., et al. (2006). "Virtual rapport," in *Proceedings of the 6th International Conference Intelligent Virtual Agents (IVA 2006),* eds J. Gratch, M. Young, R. Aylett, D. Ballin, and P. Olivier (Marina del Rey, CA: Springer Verlag Berlin Heidelberg), 14–27. doi: 10.1007/11821830_2

Groves, R. M. (2011). Three eras of survey research. *Public Opin. Q.* 75, 861–871. doi: 10.1093/poq/nfr057

Hatchett, S., and Schuman, H. (1975). White respondents and race-of-interviewer effects. *Public Opin. Q.* 39, 523–528. doi: 10.1086/268249

Houtkoop-Steenstra, H. (2000). *Interaction and the Standardized Survey Interview: The Living Questionnaire*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511489457

Huang, L., Morency, L. P., and Gratch, J. (2011). "Virtual rapport 2.0," in *Proceedings of the 10th International Conference Intelligent Virtual Agents (IVA 2011),* eds H. Högni Vilhjálmsson, S. Kopp, S. Marsella, and K. R. Thórisson (Marina del Rey, CA: Springer Verlag Berlin Heidelberg), 68–79.

Hyde, J., Carter, E. J., Kiesler, S., and Hodgins, J. K. (2013). "Perceptual effects of damped and exaggerated facial motion in animated characters," in *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG13)* (Shanghai: IEEE). doi: 10.1109/FG.2013.6553775

Jamieson, S. (2004). Likert scales: how to (ab)use them. *Med. Educ. 38*, 1217–1218. doi: 10.1111/j.1365-2929.2004.02012.x

Jarosz, A. F., and Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes Factors. *J. Probl. Solving* 7, 2. doi: 10.7771/1932-6246.1167

JASP. (2015). Available online at: http://jasp-stats.org/

Johnston, M., Ehlen, P., Conrad, F. G., Schober, M. F., Antoun, C., Fail, S., et al. (2013). "Spoken dialog systems for automated survey interviewing," in *Proceedings of the 14th Annual SIGDIAL Meeting on Discourse and Dialogue (SIGDIAL 2013)* (Metz), 329–333.

Keeter, S. (2012). Presidential address: survey research, its new frontiers, and democracy. *Public Opin. Q.* 76, 600–608. doi: 10.1093/poq/nfs044

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychol.* 26, 22–63. doi: 10.1016/0001-6918(67)90005-4

Krämer, N., Kopp, S., Becker-Asano, C., and Sommer, N. (2013). Smile and the world will smile with you—The effects of a virtual agent's smile on users' evaluation and behavior. *Int. J. Hum. Comput. Stud.* 71, 335–349. doi: 10.1016/j.ijhcs.2012.09.006

Kreuter, F., Presser, S., and Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and Web surveys: the effects of mode and question sensitivity. *Public Opin. Q.* 72, 847–865. doi: 10.1093/poq/nfn063

Lind, L. H., Schober, M. F., Conrad, F. G., and Reichert, H. (2013). Why do survey respondents disclose more when computers ask the questions? *Public Opin. Q.* 77, 888–935. doi: 10.1093/poq/nft038

Lucas, G. M., Gratch, J., King, A., and Morency, L. P. (2014). It's only a computer: virtual humans increase willingness to disclose. *Comput. Hum. Behav.* 37, 94–100. doi: 10.1016/j.chb.2014.04.043

MacDorman, K. F., Green, R. D., Ho, C.-C., and Koch, C. T. (2009). Too real for comfort: uncanny responses to computer generated faces. *Comput. Hum. Behav.* 25, 695–710. doi: 10.1016/j.chb.2008.12.026

MacDorman, K. F., and Ishiguro, H. (2006). The uncanny advantage of using androids in social and cognitive science research. *Interact. Stud.* 7, 297–337. doi: 10.1075/is.7.3.03mac

Massey, D. S., and Tourangeau, R. (2013). Introduction: new challenges to social measurement. *Ann. Am. Acad. Pol. Soc. Sci.* 645, 6–22. doi: 10.1177/0002716212463314

Mavletova, A., and Couper, M. P. (2014). Mobile web survey design: scrolling versus paging, SMS versus email invitations. *J. Surv. Stat. Methodol.* 2, 498–518. doi: 10.1093/jssam/smu015

McClave, E. Z. (2000). Linguistic functions of head movements in the context of speech. *J. Pragmat.* 37, 855–878. doi: 10.1016/S0378-2166(99)00079-X

McDonnell, R., Breidt, M., and Bülthoff, H. H. (2012). Render me real? Investigating the effect of render style on the perception of animated virtual humans. *ACM Trans. Graph.* 31, 91. doi: 10.1145/2185520.2185587

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Adv. Health Sci. Educ. Theory Pract.* 15, 625–632. doi: 10.1007/s10459-010-9222-y

Nunamaker, J. F., Derrick, D. C., Elkins, A. C., Burgoon, J. K., and Patton, M. W. (2011). Embodied conversational agent-based kiosk for automated interviewing. *J. Manage. Inf. Syst.* 28, 17–48. doi: 10.2753/MIS0742-1222280102

Peytchev, A. (2009). Survey breakoff. *Public Opin. Q.* 73, 74–97. doi: 10.1093/poq/nfp014

Piwek, L., McKay, L. S., and Pollick, F. E. (2014). Empirical evaluation of the uncanny valley hypothesis fails to confirm the predicted effect of motion. *Cognition* 130, 271–277. doi: 10.1016/j.cognition.2013.11.001

Schaeffer, N. C. (1991). "Conversation with a purpose – or conversation? Interaction in the standardized interview," in *Survey Measurement and Process Quality*, eds P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman (New York, NY: John Wiley), 367–391.

Schober, M. F., and Conrad, F. G. (1997). Does conversational interviewing reduce survey measurement error? *Public Opin. Q.* 61, 576–602.

Schober, M. F., and Conrad, F. G. (2002). "A collaborative view of standardized survey interviews," in *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview,* eds D. Maynard, H. Houtkoop-Steenstra, N. C. Schaeffer, and J. van der Zouwen (New York, NY: John Wiley & Sons), 67–94.

Schober, M. F., and Conrad, F. G. (2008). "Survey interviews and new communication technologies," in *Envisioning the Survey Interview of the Future,* eds F. G. Conrad and M. F. Schober (Hoboken, NJ: John Wiley & Sons), 1–30. doi: 10.1002/9780470183373.ch1

Schober, M. F., and Conrad, F. G. (2015). Improving social measurement by understanding interaction in survey interviews. *Policy Insights Behav. Brain Sci.* 2, 211–219. doi: 10.1177/2372732215601112

Schober, M. F., Conrad, F. G., Antoun, C., Ehlen, P., Fail, S., Hupp, A. L., et al. (2015). Precision and disclosure in text and voice interviews on smartphones. *PLoS ONE* 10:e0128337. doi: 10.1371/journal.pone.0128337

Schober, M. F., Conrad, F. G., Dijkstra, W., and Ongena, Y. P. (2012). Disfluencies and gaze aversion in unreliable responses to survey questions. *J. Off. Stat.* 28, 555–582. Available online at: http://www.jos.nu/Articles/abstract.asp?article=284555

Schober, M. F., Conrad, F. G., Ehlen, P., and Fricker, S. S. (2003). "How web surveys differ from other kinds of user interfaces," in *Proceedings of the American Statistical Association, Section on Survey Research Methods* (Alexandria, VA: American Statistical Association). Available online at: http://www.amstat.org/sections/srms/proceedings/y2003/files/jsm2003-000336.pdf

Schober, M. F., Conrad, F. G., and Fricker, S. S. (2004). Misunderstanding standardized language in research interviews. *Appl. Cogn. Psychol.* 18, 169–188. doi: 10.1002/acp.955

Steptoe, W., Steed, A., Rovira, A., and Rae, J. (2010). "Lie tracking: social presence, truth and deception in avatar-mediated telecommunication," in *CHI'10: Proceedings of the 28th International Conference on Human Factors in Computing Systems* (Vancouver, BC), 1039–1048.

Suchman, L., and Jordan, B. (1990). Interactional troubles in face-to-face survey interviews. *J. Am. Stat. Assoc.* 85, 232–253. doi: 10.1080/01621459.1990.10475331

Suessbrick, A. L., Schober, M. F., and Conrad, F. G. (2000). "Different respondents interpret ordinary questions quite differently," in *Proceedings of the American Statistical Association, Section on Survey Research Methods* (Alexandria, VA: American Statistical Association). Available online at: https://www.amstat.org/sections/srms/proceedings/papers/2000_155.pdf

Tourangeau, R., and Smith, T. W. (1996). Asking sensitive questions: the impact of data collection mode, question format, and question context. *Public Opin. Q.* 60, 275–304. doi: 10.1086/297751

Turner, C. F., Ku, L., Rogers, S. M., Lindberg, L. D., Pleck, J. H., and Sonenstein, F. L. (1998). Adolescent sexual behavior, drug use, and violence: increased reporting with computer survey technology. *Science* 280, 867–873. doi: 10.1126/science.280.5365.867

US Bureau of the Census and US Bureau of Labor Statistics. (1993). *Point of Purchase Survey 1993: Checklist A: Five-Year to One-Week Recall, Vol. 3.* Washington, DC: US Department of Commerce, Bureau of the Census.

Viera, A. J., and Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Fam. Med.* 37, 360–363. Available online at: http://www.stfm.org/FamilyMedicine/Vol37Issue5/Viera360

von der Pütten, A., Hoffmann, L., Klatt, J., and Krämer, N. C. (2011). "Quid pro quo? Reciprocal self-disclosure and communicative accommodation towards a virtual interviewer," in *Proceedings of the 10th International Conference Intelligent Virtual Agents (IVA 2011),* eds H. Högni Vilhjálmsson, S. Kopp, S. Marsella, and K. R. Thórisson (Marina del Rey, CA: Springer Verlag Berlin Heidelberg), 183–194.

von der Pütten, A. M., Krämer, N. C., Gratch, J., and Kang, S. H. (2010). "It doesn't matter what you are!" Explaining social effects of agents and avatars. *Comput. Hum. Behav.* 26, 1641–1650. doi: 10.1016/j.chb.2010.06.012

Xiao, N. G., Perrotta, S., Quinn, P. C., Wang, Z., Sun, Y. H. P., and Lee, K. (2014). On the facilitative effects of face motion on face recognition and its development. *Front. Psychol.* 5:633. doi: 10.3389/fpsyg.2014.00633

Yngve, V. (1970). "On getting a word in edgewise," in *Papers from the 6th Regional Meeting, Chicago Linguistic Society* (Chicago, IL: Chicago Linguistic Society).

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 08 July 2015
doi: 10.3389/fpsyg.2015.00931

Systematic analysis of video data from different human–robot interaction studies: a categorization of social signals during error situations

*Manuel Giuliani *, Nicole Mirnig, Gerald Stollnberger, Susanne Stadler, Roland Buchner and Manfred Tscheligi*

Department of Computer Sciences, Center for Human-Computer Interaction, University of Salzburg, Salzburg, Austria

OPEN ACCESS

Edited by:
Snehlata Jaswal, Indian Institute of Technology Jodhpur, India

Reviewed by:
Harry J. Witchel, University of Sussex, UK
Fady Alnajjar, Brain Science Institute, RIKEN, Japan

***Correspondence:**
Manuel Giuliani, Department of Computer Sciences, Center for Human-Computer Interaction, University of Salzburg, Sigmund-Haffner-Gasse 18, 5020 Salzburg, Austria
manuel.giuliani@sbg.ac.at

Specialty section:
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology

Received: *14 March 2015*
Accepted: *22 June 2015*
Published: *08 July 2015*

Citation:
Giuliani M, Mirnig N, Stollnberger G, Stadler S, Buchner R and Tscheligi M (2015) Systematic analysis of video data from different human–robot interaction studies: a categorization of social signals during error situations. Front. Psychol. 6:931. doi: 10.3389/fpsyg.2015.00931

Human–robot interactions are often affected by error situations that are caused by either the robot or the human. Therefore, robots would profit from the ability to recognize when error situations occur. We investigated the verbal and non-verbal social signals that humans show when error situations occur in human–robot interaction experiments. For that, we analyzed 201 videos of five human–robot interaction user studies with varying tasks from four independent projects. The analysis shows that there are two types of error situations: social norm violations and technical failures. Social norm violations are situations in which the robot does not adhere to the underlying social script of the interaction. Technical failures are caused by technical shortcomings of the robot. The results of the video analysis show that the study participants use many head movements and very few gestures, but they often smile, when in an error situation with the robot. Another result is that the participants sometimes stop moving at the beginning of error situations. We also found that the participants talked more in the case of social norm violations and less during technical failures. Finally, the participants use fewer non-verbal social signals (for example smiling, nodding, and head shaking), when they are interacting with the robot alone and no experimenter or other human is present. The results suggest that participants do not see the robot as a social interaction partner with comparable communication skills. Our findings have implications for builders and evaluators of human–robot interaction systems. The builders need to consider including modules for recognition and classification of head movements to the robot input channels. The evaluators need to make sure that the presence of an experimenter does not skew the results of their user studies.

Keywords: social signals, error situation, social norm violation, technical failure, human–robot interaction, video analysis

1. Introduction

The interaction between humans and robots is often affected by problems that occur during such interactions. Human users interact with robots based on their mental models, expectations, and previous experiences. When problems occur, users are often confused. Their expectations are violated and they do not know how to react. In the worst case, such problems can even result in a termination of the interaction (Scheutz et al., 2011). The users, most likely, attribute the error to the robot, and these problems are in fact often caused by the robot. Some examples for occurring problems are insufficient or defective sensor data, errors due to misinterpretation of sensor data from the robot's reasoning module, and general implementation errors (Goodrich and Schultz, 2007). In some cases, however, an interruption of the interaction may also be caused by the human, for example, if the human interaction partner wants to perform a task that is not within the capability of the robot. Irrespective of the origin, the human interaction partner gets confused and the continuation of the interaction is at stake.

Ample evidence for the occurrence of the above described problems can be found in the data of human–robot interaction (HRI) user studies, in which humans directly interact with robots. The matter of interest in these studies is often an envisioned flawless interaction. Therefore, data of problematic interactions may get discarded from further analysis or the problem itself is not part of the analysis. We argue that these data potentially bear valuable insights and ideas for improving future HRI. We are interested in the following questions: what are the social signals that humans display in the event of these errors and what kind of error situations do arise in human–robot interactions?

1.1. Social Signals

The term *social signal* is used to describe verbal and non-verbal signals that humans use in a conversation to communicate their intentions. Vinciarelli et al. (2009) argue that the ability to recognize social signals is crucial to mastering social intelligence. In their view, the recognition of social signals will be the next step toward a more natural human-computer and human–robot interaction. Ekman and Friesen (1969) define five classes of human non-verbal behavior. *Emblems* are gestures that have a meaning for members of a group, class, or culture, e.g., the thumbs up sign that means positive agreement in many western countries. *Illustrators* are gestures or movements that are directly tied to speech and are used to illustrate what has been said verbally, e.g., humans forming a triangle with their fingers while speaking about a triangular-shaped object. *Affect displays* are signals used to convey an emotional state, often by facial expressions or body posture. *Regulators* are signals used to steer the conversation with a conversation partner, e.g., to regulate turn taking. Finally, *adaptors* are actions used on objects in the environment or on oneself, e.g., lip biting or brushing back hair. The social signals that we detected in our video analysis are mostly affect displays, regulators, and adaptors (see Section 3). For annotating social signals, we are not following Ekman's taxonomy. Instead, we separate the signals into the body parts that the participants in the HRI studies used to express the signal, which makes it easier to annotate combinations of social signals (see Section 2.4).

In recent years, more and more researchers worked on the automatic recognition of social signals, an area that is called *social signal processing*. Vinciarelli et al. (2012) give an overview of the field. They classify social signals with a similar taxonomy that we are using in the annotation scheme in Section 2.4. According to Vinciarelli et al. (2009), human social signals come either from physical appearance, gesture and posture, face and eyes behavior, space and environment, or vocal behavior. The authors also review early work from social signal processing. In human–robot interaction, social signal processing also receives more attention by researchers from different areas. Jang et al. (2013) present a video analysis that is similar to our work. In the analysis, they annotated recordings of six one-on-one teacher–student learning sessions, in order to find the social signals with which students signal their engagement in the interaction. The goal of this work is to implement an engagement classifier for a robot teacher. Tseng et al. (2014) present a robot that automatically recognizes the spatial patterns of human groups by analysing their non-verbal social signals in order to appropriately approach the group and offer services.

A second area of interest to HRI, is the generation of social signals by robots. Bohus and Horvitz (2014) presented a direction-giving robot that forecasts when the user wants to conclude the conversation. This robot uses hesitations (e.g., the robot says "so…") when it is not certain about the user state in order to get more time to compute a correct forecast and also to convey the uncertainty of the robot. Bohus and Horvitz did not report an improvement in disengagement forecasts for their robot which used hesitations. This might have been due to the conservative strategy they were using in their study, which was tuned to avoid false disengagements. Sato and Takeuchi (2014) researched how the eye gaze behavior of a robot can be used to control the turn taking in non-verbal human–robot interactions. In their study, three humans played a game with a robot that was programmed to look at the other players during the game. The study shows that the robot's gaze can influence who will be the next speaker in the conversation. In another eye gaze generation study, Stanton and Stevens (2014) found that robot gaze positively influences the trust of experiment participants who had to give answers to difficult questions in a game, but negatively influences trust when answering easy questions. However, robot gaze positively influences task performance for easy questions, but negatively influences task performance for difficult questions. Stanton and Stevens discuss that robot gaze might put pressure on the experiment participants. Carter et al. (2014) presented a study, in which participants repeatedly threw a ball to a humanoid robot that attempted to catch the ball. In one of the study conditions, when the robot did not catch the ball, it generated social signals, e.g., it shrugged its shoulders. The study results show that participants smile more when the robot displays social signals and rate the robot as more engaging, responsive, and human-like.

1.2. Error Situations

In the videos of the experiments that we annotated for this work, we found two different kinds of error situations. On one

hand, there were situations in which unusual robot behavior led to a violation of a social norm; on the other hand, there were error situations because of technical failures of the robot. In this section, we will review related work on both of these areas to define our notion of the term *error situation*.

We follow the definition of Sunstein (1996) that *social norms* are "social attitudes of approval and disapproval, specifying what ought to be done and what ought not to be done" (Sunstein, 1996, p. 914). Human interaction is defined by social norms. For example, they define how one should ask for directions on the street or how you should behave in a bar. Schank and Abelson (1977) showed that everyday social interactions have an underlying *social script*, a definition of interaction steps to which humans usually obey. The order of these interaction steps is guided by social signals. Loth et al. (2013) found that customers use two combined non-verbal social signals to signal bartenders that they would like to order a drink: they position themselves directly at the bar counter and look at a member of staff. We define a *violation of a social norm* as a deviation from the social script or the usage of the wrong social signals. For example, in our videos there are instances in which the robot executed unexpected actions in the interaction (e.g., asking for directions several times although the human already gave correct instructions and the robot acknowledged the instructions) or showed unusual social signals (e.g., not looking directly at the person it is talking to).

The second class of error situations in our experiment videos arises from *technical failures* of the robot. Interestingly, we can resort to definitions of technical failures of humans interacting with machines, in order to classify these errors, since all robots we observed are autonomous agents. Rasmussen (1982) defines two kinds of human errors: *execution failures* happen when a person carries out an appropriate action, but carries it out incorrectly, and *planning failures* happen when a person correctly carries out an action, but the action is inappropriate. To transfer these definitions to autonomous robots and to make the definitions clearer, consider the following two examples. The robot makes an execution failure, when it picks up an object, but loses it while grasping it; the robot has a planning failure, when the decision mechanism of the robot decides to ask the human for directions, although it already did so and the human correctly gave the information. Execution failures are also called slips or lapses, while planning failures are mistakes[1].

To summarize these two definitions (further described in Section 3), we found two types of error situations in the videos we annotated. The robots either violated social norms by executing interaction steps at the wrong time or by showing unusual social signals, or they had obvious technical failures. It is interesting to note that social norm violations often arise of planning failures by the robot, while technical failures are usually execution failures.

Social neuroscientists have studied error situations and how they are perceived by humans. Forbes and Grafman (2013) define social neuroscience as "The systematic examination of how social psychological phenomena can be informed by neuroscience methodologies, and how our understanding of neural function can be informed by social psychological research" (Forbes and Grafman, 2013, p. 1). In recent years, several neuroscientists conducted studies to research the neural correlations when humans observe error situations.

Berthoz et al. (2002) conducted a study to find the neural systems that support processing of intentional and unintentional social norm violations. They used event-related functional magnetic resonance imaging (fMRI) to compare the neural responses of humans listening to stories describing either normal behavior, embarrassing anecdotes, or social norm violations. Berthoz et al. found that the neural systems involved in processing social norm violations are the same as systems involved in representing mental states of others and in responding to aversive emotional expressions. The authors conclude that the findings have implications for understanding the pathology of patients who exhibit social behavioral problems associated with the identified neural systems.

de Bruijn et al. (2011) conducted a study to research whether humans represent the task of a co-actor during error monitoring in joint action. The authors showed through measurement of electroencephalogram (EEG) signals and behavioral data that the study participants show increased amplitudes on the response-locked error-related negativity, an event-related brain potential that is generated after an erroneous response (Falkenstein et al., 1990), and longer reaction times following own errors in a social go/no-go task. The findings show that people incorporate the tasks of others into their own error monitoring and adjust their own behavior during joint action.

Radke et al. (2011) investigated brain activities in humans when monitoring errors that only influenced themselves or also had implications for others. They found in an fMRI study that monitoring errors that have implication for others activates the medial prefontal cortex, a part of the mentalizing system. The authors conclude from the results that this for example explains symptoms of patients with obsessive-compulsive disorder, who have fears that their own actions will harm others.

Ridderinkhof et al. (2004) conducted a meta-analysis of primate and human studies as well as of human functional neuroimaging literature. The analysis showed that the detection of unfavorable outcomes, response errors, response conflict, and decision uncertainty enhances brain activity in an extensive part of the posterior medial frontal cortex. This indicates that performance monitoring, including error monitoring, is associated with this brain region. Koban et al. (2013) recently showed in an event-related fMRI study that error monitoring is integrated with the representation of pain of others. The results of their study show that the same brain regions are involved in error monitoring and empathy for pain and that the brain activity in these regions is enhanced when the pain of the other person is caused by oneself.

In this paper, we perform a systematic analysis of video data from different HRI user studies. The first goal for this analysis is to identify those situations in interactions between humans and robots that lead to problems and create error situations. Such problems include long dialogue pauses, repetitions in the dialogue, misunderstandings, and even a complete abruption of the interaction. In the next step, we categorize the detected error

[1] A good depiction of human error types can be found at http://www.skybrary.aero/index.php/Human_Error_Types.

situations into problems resulting from social norm violations and problems that occur due to a technical error. Based on this categorization, we analyse the social signals that humans produce during the problematic situation, in order to map situations and social signals. We distinguish verbal and non-verbal social signals, e.g., speech, gaze, head orientation, and body posture.

For the analysis, we use video data from a variety of HRI user studies. The videos were taken from different projects, providing us with a wide spectrum of robots, robot tasks, and experimental settings. The JAMES project (Joint Action for Multimodal Embodied Social Systems[2]) used a stationary bartender robot with social skills, presenting humans with socially appropriate interaction; the JAST project (Joint-Action Science and Technology[3]) used a stationary robot that cooperates with a human in an assembly task; the IURO project (Interactive Urban Robot[4]) used a mobile, wheeled robot that autonomously navigates through densely crowded inner-city environments and actively asks information from pedestrians; and the RPBD project (Robot Programming by Demonstration) used a NAO robot to research kinesthetic robot teaching in an industrial environment.

In our data analysis, we peruse three goals: (1) provide a ranked categorization of social signals including their frequency of occurrence; (2) develop a mapping between error types and social signals in order to understand if there are particular social signals that are typically evoked either by social norm violation or technical failure; and (3) explore the influence of independent variables (e.g., presence of experimenter during the interaction, single vs. group interaction) on the display of social signals.

2. Methods and Materials

Figure 1 shows the work flow of the method that we applied in this paper. First, we executed five HRI user studies[5], from which we collected a video corpus of 201 interactions. We then annotated the videos in two steps. We introduce the HRI user studies in Section 2.1. Following that, we give an overview of the video corpus in Section 2.2 and information on the study participants in Section 2.3. Finally, we describe the annotation process in more detail in Section 2.4.

[2]http://www.james-project.eu.

[3]http://www6.in.tum.de/Main/ResearchJast.

[4]http://www.iuro-project.eu/.

[5]To be clear: we did not carry out the user studies specifically for this paper. We are revisiting the results from prior studies and analyse them from a different viewpoint in this work.

2.1. Human–Robot Interaction Studies

Our video analysis of social signals in error situations is based on videos from five human–robot interaction studies. These studies were carried out as part of the projects JAMES, JAST, IURO, and RPBD. Each of the studies had a different task for human and robot, except for the two JAMES studies. This enables us to study social signals in the context of a variety of tasks with robots that have different appearances. We have three different humanoid robots (**Figure 2**), from which one robot is stationary and two robots are mobile. Furthermore, we have four different scenarios, the bartender scenario from JAMES, the joint assembly scenario from JAST, the direction asking scenario from IURO, and the robot teaching scenario from RPBD.

All user studies have in common that the robots were able to understand and produce speech, and that they had visual perception modules for person tracking. The studies were carried out either in Germany or Austria. A majority of the spoken interactions were in German, for the rest human and robot spoke English. We received ethical approval for all of the studies. All study participants signed an informed consent and gave us permission to use the videos taken from the studies for further analysis. The JAMES studies complied to the ethics standards of fortiss (2012, 2013). The JAST study complied to the ethics standards of the Technical University of Munich (2010). The IURO study complied to the Ethics standards of the University of Salzburg (2015). The RPBD study complied to the Ethics standards of the University of Salzburg (2014). For more details on each of the studies, please refer to the publications that we cite for each study in the respective section.

In the following sections we shortly introduce all four projects and describe the user studies from which we used videos. **Figure 2** shows images of all four robots.

2.1.1. JAMES, Stationary Robot Bartender

The goal of the JAMES project was to implement successful joint action that is based on social interaction. The task of the JAMES robot was that of a bartender. It had to take drink orders from customers and to hand out the correct drinks to the person who ordered it. **Figure 2A** shows the robot interacting with a customer. The bartender robot consisted of two industrial robot arms with humanoid hands, mounted in a position to resemble human arms. Furthermore, the robot has an animatronic talking head, the iCat (van Breemen, 2005), which is capable of producing lip-synchronized speech as well as

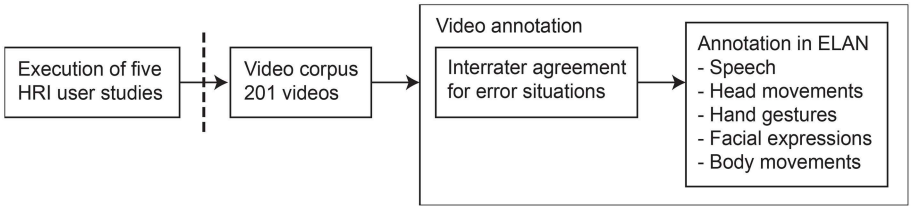


FIGURE 1 | Chronological steps for the work we carried out in this paper. First, we executed five human–robot interaction user studies; second, we collected the videos from these studies to form a video corpus; third, we annotated the videos in a two-step process.

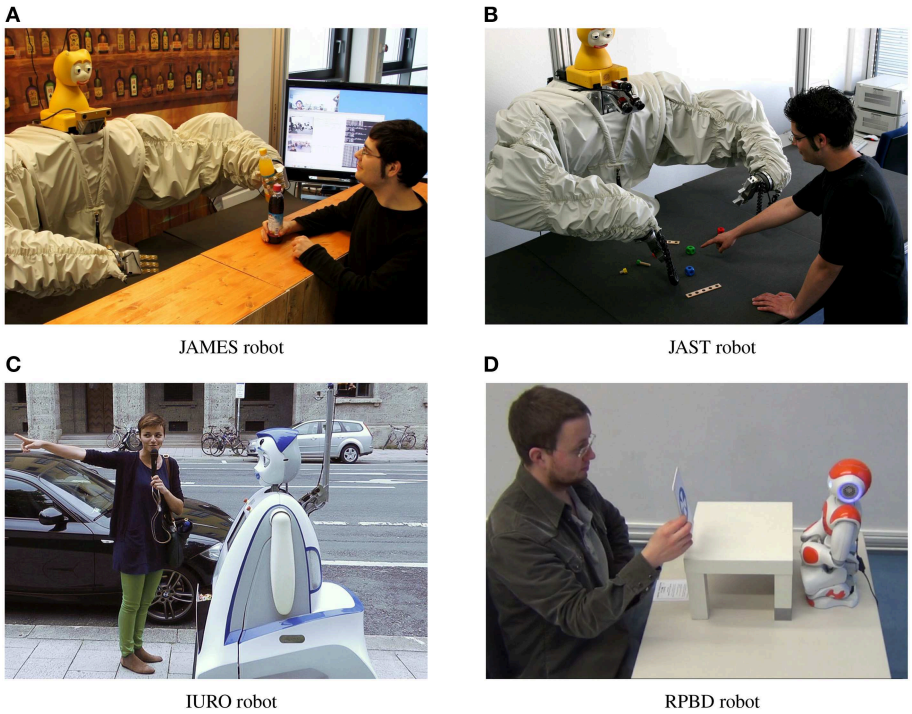


FIGURE 2 | The four robots used in the human–robot interaction studies. Pictures show interactions from the studies. **(A)** JAMES robot, **(B)** JAST robot, **(C)** IURO robot, and **(D)** RPBD robot

expressing basic facial expressions such as smiling and frowning. The robot was mounted behind a bar, which could be reached by the robot arms. Furthermore, the robot was able to hand over bottles to its customers.

The videos that we are using in this work are from two user studies that were executed with the robot. Foster et al. (2012) researched how the behavior of the robot has to change when interacting with either single customers or groups of customers. Giuliani et al. (2013) compared how user groups perceive the robot when it shows only task-based actions or when it also uses social actions. Both user studies used the same instructions for the study participants, they were simply asked to walk up to the robot and to order a drink. An experimenter was visible at all times during both JAMES studies.

2.1.2. JAST, Stationary Robot with Assembly Task

The goal of the JAST project was to develop jointly-acting autonomous systems that communicate and work intelligently on mutual tasks in dynamic unstructured environments. The task of the JAST robot was to assemble target objects from a wooden toy construction set together with a human partner. **Figure 2B** shows the robot. It is the same robot system that was used in the JAMES project. The robot had a table in front of it on which the different assembly parts were laid out. It was able to recognize the objects and to hand them over to the human.

The videos we are using in this work are from the user study reported by Giuliani et al. (2010). The task of the study participants and the robot was to jointly construct two target objects. In the experiment, the authors compared two different strategies for generating referring expressions to objects, a traditional strategy that always generated the same expression for the objects and an adaptive strategy that made use of the situated context knowledge of the robot. The participants worked together with the robot in one-on-one interactions. During the study, participants were not able to see the experimenter, who was sitting behind a poster wall.

2.1.3. IURO, Mobile Robot Asking for Directions

The goal of the IURO project was to develop a robot that navigates and interacts in densely populated, unknown human-centred environments and retrieves information from human partners in order to navigate to a given goal. The IURO robot was developed to autonomously navigate in an unstructured public-space environment and proactively approach pedestrians to retrieve directions. **Figure 2C** shows the robot interacting with a pedestrian in the city center of Munich, Germany. The IURO robot was designed with anthropomorphic, but not entirely humanoid appearance. A humanoid head is combined with a functionally designed body. The head is able to produce lip-synchronized speech and express basic facial expressions (Ekman, 1992). Additionally, the robot has two arms, but no hands (to avoid wrong expectations, since the robot is not able to grasp objects). A pointing device for indicating directions is mounted above the robot head.

The videos we used for annotating social signals in error situations were taken from the field trial of the IURO robot in which the final set-up of the robot was validated. To ensure the final robot version running at the best possible set-up, the

robot platform was subject to manifold evaluation on different interaction aspects at different points in the project. For a detailed overview on the evaluation set-ups, timeline, and results which led to the final robot prototype, refer to Weiss et al. (2015). The IURO robot interacted with single users and groups of users. During the interactions, the experimenters were mostly, but not always, visible to the participants.

2.1.4. RPBD, Mobile Robot with Kinesthetic Teaching

The videos of the last user study that we are evaluating for error situations in this work, were taken from a master's thesis. The goal of this master's thesis was to determine user acceptance factors of robots with different appearances. Specifically, the thesis researches how kinesthetic teaching with an anthropomorphic robot in an industrial context is perceived by users with different backgrounds (programmers vs. naïve users).

The videos we are using in this work are from the user study reported by Stadler et al. (2014). The authors implemented a kinesthetic teaching approach on the humanoid robot platform NAO. The robot was able to record and replay a behavior—a pick-and-place task—taught by the participants. During the experiment, human and robot had direct contact via kinesthetic teaching. Furthermore, the robot was able to recognize and produce speech, and had visual object recognition based on landmark and color detection. **Figure 2D** shows an experiment participant in interaction with the robot. All study participants interacted alone with the robot, but an experimenter was visible to them at all times.

2.2. Video Corpus

Our video corpus consists of 201 videos, from which 129 are from the two JAMES user studies, 34 are from the JAST user study, 27 are from the IURO user study, and 11 are from the RPBD user study. We chose only those videos from all user studies that show at least one error situation in which the robot either violated a social norm or had a technical failure. Overall, the videos show 272 individual interactions between a single user or a group of users. The difference between numbers of videos and numbers of interactions is because the videos of the JAST and IURO studies show more than one interaction. The interactions between the study participants and the robots are on average 108.467 s long (standard deviation 47.927 s). During the interactions, 578 error situations occurred in total.

2.3. Participants

The videos feature 137 unique study participants, who interacted individually or in groups with the robots. Ninety-four participants were male, 43 were female. Although all experiments took place either in Germany or in Austria, the robot spoke in German with 86 participants, and English with the other 51 participants.

2.4. Annotation

For data analysis, we annotated our video corpus using the video coding tool ELAN (Wittenburg et al., 2006). **Figure 3** shows an example of an ELAN annotation of a video of one of the JAMES user studies using our annotation format.

For annotating the videos, we followed a two-step process. In the first step, we annotated all passages in the videos in which an error in the interaction occurred. For example, when the robot did not understand what a participant had said. We labeled these instances as *error situation*. Since not every error situation is easy to recognize, we coded the error situations in each video file by two independent raters. Afterwards, we calculated the percentage of overlap for the annotated error situations between the two raters. For videos which had less than 75% coding agreement, the two coders looked at the data material again, discussed the differences and reached a consensus on the error situations.

There were two main reasons why coding differed between the two raters. On one hand, one coder annotated the data from a more technical perspective, while the other coder considered the material from a more social viewpoint. For example, if the robot says that it did not understand the study participant it could either mean that the speech recognition module failed (technical perspective) or it could be considered as socially appropriate (social perspective: people sometimes inquire when they do not understand an utterance). From a technical perspective, the utterance would likely be coded as an error situation, while from a social perspective it might be considered as socially acceptable and not an error situation. On the other hand, for correctly identified error situations, the coders did not always agree on when exactly the error situation begins or ends. For example, in case of the bartender robot, one coder started annotating the situation as soon as the robot hand moved toward the bottles, whereas the other coder only started after it was clear that the robot would actually grasp the wrong bottle. At the end, the annotators agreed that all codings should be done from the viewpoint of the study participant, which means that in the example the error situation would start from where the participant can see that the robot will grasp the wrong bottle.

In the second coding step, we annotated the actions the robot performed during the error situations, together with the social signals the study participants exhibited at the same time. For annotation of the social signals we used the following five classes, that we chose in order to be able also to annotate social signals that occur in parallel:

- *Speech*: verbal utterances by the participants, including *task-related sentences* for the task given in the user study, *questions* that the participants ask to the robot, a group member or the experimenter, and *statements* the participants make.
- *Head movements*: instances where the participant *looks* to the robot, a group member, or the experimenter. Head movements can also be *nodding*, *shaking*, and *tilting* the head.
- *Hand gestures*: movements that participants make with their hands, including *pointing* gestures, *emblems*, instances where the participants *manipulate an object*, or when they *touch* themselves on the body or in the face.
- *Facial expressions*: expressions as for example *smiling* that can be observed on the participants' faces. These also include signals like *rising the eyebrows* or making a *grimace*.
- *Body movements*: all movements that the participants make with their whole body, including *leaning* toward or away from

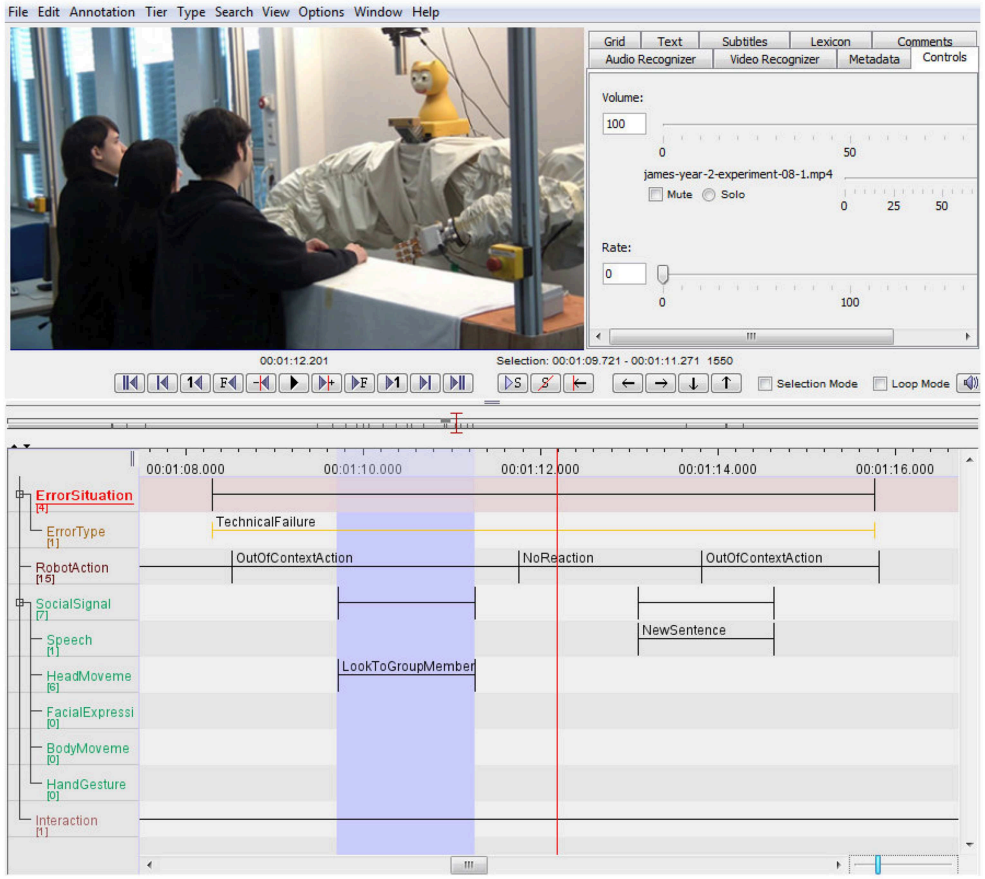


FIGURE 3 | Screenshot of an error situation in the ELAN annotation tool.

the robot, *moving* toward or away from the robot, and *changes in body posture*.

We annotated the social signals that occur during the error situations in our videos according to these five classes. In the next section, we present the results of these annotations.

3. Results

Table 1 shows an overview of all annotated verbal and non-verbal social signals that occurred during the error situations in our study videos. In the category **head movements** we found that participants often look back and forth between robot and experimenter or a group member if present. Depending on the study task, they also look back and forth between the robot and objects in front of them. The participants also sometimes nod, shake, or tilt their head. We annotated 947 items in the **speech** category. We subdivided the speech utterances into task-related sentences, sentences that the study participants said to the robot to move the given task forward, statements that participants made to comment on the situation to either the robot or another human, questions that participants asked to the robot or a human, audible laughter, and other utterances, for example attempts to speak or hesitation sounds. One participant whistled at the robot to get its attention. In the category **facial expressions**, we found that participants often smiled in error situations. Sometimes the participants grimaced, for example when they showed a concerned look or pouted. Quite often, the participants raised their eyebrows. When interacting with the robot, we found that participants mostly stand still and do not show much **body movement**. For the majority of body movements, participants leaned toward or away from the robot, less often they completely stepped away from the robot or changed their posture. In comparison to other social signals, we found only a few **hand gestures**. Participants often touched themselves in the face or put their hands on the hips. If they had an object in reach, they manipulated that object. Pointing gestures and iconic gestures were quite rare, we annotated only 1 thumbs up gesture and 9 persons, who waved at the robot. Other hand gestures include for example drumming with the fingers on a surface, raising one or both hands, and making a fist with the hand.

Next, we performed a statistical analysis to compare the differences in shown social signals for three dependent variables: *social norm violation* vs. *technical failure*; *experimenter visible* vs. *experimenter not visible*; and *group interaction* vs. *single*

TABLE 1 | Counts for all annotated social signals in the categories head movements, speech, facial expressions, body movements, and hand gestures.

Head movements	**1279**	**Speech**	**947**	**Facial expressions**	**484**
Look at robot	434	Task-related sentence	487	Smile	314
Look at experimenter	230	Statement	170	Grimace	124
Look into a direction	230	Question	111	Raise eyebrows	46
Look at group member	151	Laugh	98		
Tilt head	83	Correction	20		
Look to object	72	Other	61		
Nod	40				
Shake head	39				
Body movements	**272**	**Hand gestures**	**248**		
Lean	191	Touch own body	45		
Move	33	Manipulate object	61		
Other	48	Pointing	21		
		Emblem	10		
		Other	111		

user interaction. For that, we first performed an analysis of our data and found that all variables are not normally distributed. Therefore, we chose to compare the data populations with a Wilcoxon–Mann–Whitney test. Furthermore, we extracted the data for each error situation individually from the annotations in order to be able to group them by the dependent variables. The error situations had an average duration of 18.314 s (standard deviation 20.861 s). These numbers indicate that many of the error situations are either quite short or last very long. From the 578 annotated error situations, 427 are social norm violations and 151 are technical failures, in 420 error situations the experimenter was visible and in 158 situations the experimenter was not visible, and we annotated 257 group interactions and 321 single user interactions.

Table 2 shows the result of the Wilcoxon–Mann–Whitney test for the dependent variable *social norm violation* vs. *technical failure*. We only show the social signals for which we found statistically significant results. The results show that study participants more often smile and laugh audibly during technical failures than during social norm violations. The participants more often look back and forth between the robot head and objects in front of them during social norm violations. In contrast to that, they look more often to the experimenter during technical failures. The other statistically significant differences we found fall into the range of verbal social signals. During social norm violations, the participants in general speak more, they say task-related sentences to the robot and also repeat these sentences more often than during technical failures. However, during technical failures, the participants comment the situation more often and make statements to group members.

Table 3 shows the result of the Wilcoxon–Mann–Whitney test for the dependent variable *experimenter visible* vs. *experimenter not visible*. We only present the social signals for which we found statistically significant results. The results show that the study participants display much more non-verbal social signals when the experimenter is visible, for example tilting the head, making hesitation sounds, smiling, laughing audibly, nodding, and leaning back. Overall, the participants also talk more when the experimenter is visible, they say more task-related sentences to the robot, make more statements, and ask more questions. In contrast to that, the participants more often look back and forth between robot hand, robot head, and objects in front of them when the experimenter is not visible.

Finally, **Table 4** shows the result of the Wilcoxon–Mann–Whitney test for the dependent variable *group interaction* vs. *single user interaction*. Similar to the variable experimenter visible/not visible, we found that the participants show much more non-verbal signals, when interacting in groups with the robot. They laugh audibly, smile, and tilt their heads more often, when in an error situation. The participants look more often to the experimenter or a group member, when in a group interaction, but they look more often back and forth between the robot and an object, when they interact alone with the robot. We also found that the participants say more task-related sentences and make more statements commenting the situation when they are in a group. However, the participants ask more questions to the experimenter when they are in a single interaction with the robot.

After the presentation of the results, we now discuss their meaning and the implications for HRI in the following section.

4. Discussion

While annotating, we found that the difference in change of behavior during error situations and non-error situations is sometimes visible even in single user study instances. For example, during one of the studies of the JAMES project (Giuliani et al., 2013), one of the study participants had already ordered a drink and watched a group member ordering his drink. The bartender robot did not understand the other group member

TABLE 2 | Social signals shown during social norm violations and technical failures.

Social signal	Social norm violation	Technical failure	Wilcoxon–Mann–Whitney	
	Mean (std)	Mean (std)	*p*-value	W
Laughter	0.10 (0.42)	**0.33 (1.02)**	<0.0001	28346.0
Smile	0.46 (0.97)	**0.69 (1.03)**	0.0004	26997.5
Look to robot head	**0.53 (1.39)**	0.35 (1.64)	0.0009	36378.5
Look to object	**0.17 (0.73)**	0.03 (0.18)	0.0308	33934.0
Look to experimenter	0.27 (0.71)	**0.54 (1.21)**	0.0124	29161.5
Say task-related sent.	**0.37 (0.70)**	0.24 (1.00)	0.0012	36552.5
Repeat task-related sent.	**0.36 (1.12)**	0.26 (0.98)	0.0026	35808.5
Make statement to group	0.04 (0.27)	**0.11 (0.41)**	0.0079	30524.5

We annotated 427 social norm violations and 151 technical failures. We present for each social signal the mean number of occurrences per interaction and the standard deviation in parenthesis. Higher results are marked in bold.

TABLE 3 | Social signals shown when the experimenter is visible during an interaction or not.

Social signal	Exp. visible	Exp. not visible	Wilcoxon–Mann–Whitney	
	Mean (*SD*)	Mean (*SD*)	*p*-Value	*W*
Tilt head	**0.16 (0.52)**	0.05 (0.29)	0.0048	30636.0
Make sound	0.08 (0.50)	**0.09 (0.33)**	0.0468	34547.5
Smile	**0.56 (1.00)**	0.41 (0.95)	0.0342	30063.5
Laughter	**0.22 (0.74)**	0.01 (0.08)	< 0.0001	28720.0
Raise eyebrows	0.06 (0.29)	**0.14 (0.43)**	0.0038	35322.0
Nod	**0.08 (0.32)**	0.02 (0.14)	0.0345	31748.5
Lean back	**0.07 (0.28)**	0.01 (0.08)	0.0038	31255.5
Attempt to take object	0.00 (0.00)	**0.07 (0.80)**	0.0212	33600.0
Look to experimenter	**0.43 (1.00)**	0.08 (0.32)	<0.0001	27085.5
Look to group	**0.35 (0.91)**	0.00 (0.00)	< 0.0001	26860.0
Look to object	0.02 (0.15)	**0.44 (1.15)**	< 0.0001	39812.5
Look to robot hand	0.14 (0.88)	**0.63 (1.76)**	< 0.0001	38557.5
Look to robot head	0.19 (1.00)	**1.27 (2.07)**	< 0.0001	46870.5
Say task-related sent.	0.32 (0.85)	**0.39 (0.63)**	0.0159	36427.5
Make statement to robot	**0.09 (0.51)**	0.01 (0.11)	0.0397	31933.0
Make statement to group	**0.09 (0.36)**	0.00 (0.00)	0.0011	31047.0
Make statement	**0.13 (0.47)**	0.03 (0.16)	0.0044	30840.0
Question to experimenter	**0.11 (0.43)**	0.01 (0.11)	0.0016	30748.0
Question to group	**0.04 (0.26)**	0.00 (0.00)	0.0255	32153.0
Question to robot	**0.05 (0.36)**	0.00 (0.00)	0.0255	32153.0

The experimenter was visible in 420 and not visible in 158 situations. We present for each social signal the mean number of occurrences per interaction and the standard deviation in parenthesis. Higher results are marked in bold.

and repeatedly kept asking for the order. Because of this, the participant repeatedly had to smile and even sometimes laughed audibly. He furthermore kept looking back and forth between the other group member and the robot. This behavior changed completely as soon as the experimenter resolved the situation by declaring that there was an error with the system. Following this statement, the experiment participant did not smile any more and kept looking to the experimenter, although the robot kept asking for the order. This instance clearly shows how the behavior of the participant changed in seconds when coming from a social norm violation to a technical failure.

The counts of social signals in our annotations, that we show in **Table 1**, reveal three interesting results: firstly, there were many examples for error situations in which participants kept looking back and forth between robot and a group member, or robot and experimenter, or robot and objects in front of them. This is an indicator that the experiment participants are quite literally "looking" for a solution to resolve the error situation. The recognition and analysis of head movements are typically not part of the input modalities of human–robot interaction systems. Our results suggest that developers of input modalities for HRI systems should also look into expanding into this direction. Secondly, we found that participants do not use many hand

TABLE 4 | Social signals shown during single user and group interactions.

Social signal	Group	Single user	Wilcoxon–Mann–Whitney	
	Mean (*SD*)	Mean (*SD*)	*p*-Values	*W*
Tilt head	**0.20 (0.57)**	0.07 (0.37)	0.0002	45010.5
Smile	**0.71 (1.14)**	0.37 (0.82)	<0.0001	48209.0
Laughter	**0.30 (0.86)**	0.05 (0.35)	< 0.0001	47453.5
Change posture	0.02 (0.15)	**0.08 (0.36)**	0.0337	39755.0
Look to experimenter	**0.44 (1.03)**	0.26 (0.73)	0.0030	45377.0
Look to group	**0.57 (1.11)**	0.00 (0.00)	<0.0001	54088.5
Look to object	0.03 (0.20)	**0.21 (0.83)**	0.0002	37957.5
Look to robot head	0.18 (1.19)	**0.73 (1.60)**	<0.0001	31272.5
Say task-related sent.	**0.46 (1.03)**	0.24 (0.51)	0.0075	45263.5
Rephrase task-related sent.	**0.22 (0.94)**	0.04 (0.28)	0.0006	43924.5
Make statement to group	**0.14 (0.45)**	0.00 (0.00)	< 0.0001	45582.0
Make statement to robot	**0.14 (0.64)**	0.01 (0.11)	0.0002	43800.0
Make statement	0.06 (0.38)	**0.13 (0.44)**	0.0068	38765.5
Attempt to speak	**0.05 (0.23)**	0.01 (0.11)	0.0226	42502.0
Question to experimenter	0.05 (0.25)	**0.12 (0.44)**	0.0456	39535.5
Question to group	**0.06 (0.32)**	0.00 (0.00)	<0.0001	43335.0

We annotated 257 group interactions and 321 single user interactions. We present for each social signal the mean number of occurrences per interaction and the standard deviation in parenthesis. Higher results are marked in bold.

gestures during error situations. Furthermore, the majority of hand gestures do not fall into the categories that typically are studied in gesture communication. We found only a few pointing gestures and emblems, which questions the importance of these gesture categories for human–robot interaction. Also, we argue that the hand gestures that fall into the categories *touch own body*, *manipulate object*, and *other*, are not used by the participants to communicate their intentions. Thirdly, the participants often smiled during error situations, more often when they experienced technical failures and less often during a social norm violation. Work by Hoque et al. (2012) shows that humans smile in frustrating situations. They recorded the faces of participants who filled out a web form that was designed to elicit frustration. 90% of the participants smiled in these frustrating situations. We have no subjective data that could tell us whether the participants experienced frustration during the error situations with our robots. However, our video analysis indicates that they were frustrated, even more in the case of technical failures than when experiencing a social norm violation.

We often observed in the experiment videos that the participants kept standing still without moving at the beginning of an error situation. In psychology literature, this is referred to as "freezing." It is known that humans stop moving in certain situations. For example, Witchel et al. (2014) showed that the absence of non-instrumental movements can be a sign for the engagement of humans with media. It is also known that humans, as well as animals, freeze as a response to fear or stress (Hagenaars et al., 2014). We argue that the participants in our videos shortly freeze as response to the stress induced by the error situation and the presence of the experimenter. In future work, we will analyse how often, how long, and in which situations the participants kept standing still in our studies.

Our statistical evaluation of the error situations ordered by situation type in **Table 2** has to be interpreted with the tasks of the annotated user studies in mind. Of course, the users say more task-related sentences during social norm violations, because they want to solve the given task during the study. For example, many of the task-related sentences are said when the robot does not understand the participant so that the participant has to repeat the sentence. This indicates that speech is the most influential channel to resolve an error situation. However, it is interesting to see that the participants significantly talk less during technical failures.

We believe that the study participants are able to recognize if an error situation is purely technical and, therefore, stop saying task-related sentences to the robot. As we mentioned above, our results also show that the participants smile more often during technical failures, which may be elicited by the frustration they experience (Hoque et al., 2012). The participants also look more often to the experimenter. This suggests that they are looking for guidance from an authority figure during a situation they did not experience before (Smith et al., 2014).

Finally, the results of the statistical analysis in **Tables 3**, **4** in our opinion contain the most interesting result of our analysis. The participants show far more non-verbal signals, when they are in a group and/or can see the experimenter. This suggests that participants do not see the robot as an interaction partner that can interpret the same signals as a human. This is also supported by the fact that the participants make more statements about the error situation when they interact in a group with the robot or when the experimenter is visible. Of course, we also have to mention that some of the results for the conditions *experimenter visible* vs. *experimenter not visible* and *group interaction* vs. *single user interaction* are not surprising. For example, the participants

look less often to the experimenter when he/she is not visible, although they still attempt to look at him/her. This serves as a good test for the validity of our annotations.

5. Conclusion

Our video analysis of social signals in error situations during human–robot interaction experiments shows three main results. (1) The participants use head movements as a social signal to indicate when an error situation occurs. The participants do not use many hand gestures during these situations. Furthermore, the participants often smile during error situations, which could be an indication for experienced frustration. (2) The participants try to resolve social norm violations through speech. They can recognize technical failures of the robot, but they look for guidance by the experimenter in these situations. (3) The participants see the robot as an interaction partner that cannot interpret non-verbal social signals, such as smiling, laughing, nodding, shaking and tilting the head.

These findings have implications for the design and evaluation of HRI systems. *HRI system builders* should consider implementing modules for the automatic detection and interpretation of head movements, especially as an indicator for user engagement or confusion. This modality is not often used as an input channel for robots, but would be fairly easy to implement with modern sensors and image processing technology. It is known that humans use body posture to communicate their intentions (Bull, 1987; Clark, 2003). There is, however, not much work on the interpretation of head movements in particular. The importance of head movements is also supported by research from the cognitive sciences and neuropsychology that shows that head movements play a vital role in recognizing faces (O'Toole et al., 2002), especially for patients with congenital prosopagnosia, a condition that makes it difficult for an individual to recognize someone from his or her face (Longmore and Tree, 2013).

Evaluators of HRI systems should not discard the data of study trials in which errors occurred, because this data can contain valuable information, as our results show. Our analysis design also shows that the analysis of data from different HRI studies is possible and produces valuable results, when the study data can be coded in abstract categories. Furthermore, when designing the evaluation, one needs to thoroughly consider whether the experimenter or other humans should be present during the study or not, especially when measuring the social signals of study participants toward the robot. Our data clearly shows that the presence of other humans during an HRI study influences the social signals that the participants show. This is also supported by research in psychology, which has shown that study participants change their behavior when they are aware of being recorded (Laurier and Philo, 2006).

In future work, we plan to analyse parts of our video corpus in more detail. Specifically, we will execute a linguistic analysis of the task-related sentences, statements, and questions that the study participants said during the experiment. Furthermore, we will analyse the temporal connection between robot actions and the onset of the reactions of the participants during the error situations. As mentioned in the discussion, we will measure, how often and how long the participants freeze when they experience error situations. Finally, we plan to implement an automatic head movement analysis that interprets the head movements of humans, which is based on the findings of our analysis.

Acknowledgments

We would like to thank all project partners that were involved in preparing and running the user studies. This work was supported by the projects JAMES (Grant No. 270435), JAST (Grant No. 003747), IURO (Grant No. 248314), and ReMeDi (Grant No. 610902) funded by the European Commission through the 6th and 7th Framework Programme. We gratefully acknowledge the financial support by the Austrian Federal Ministry of Science, Research and Economy and the National Foundation for Research, Technology and Development in the Christian Doppler Laboratory for "Contextual Interfaces."

References

Berthoz, S., Armony, J., Blair, R., and Dolan, R. (2002). An fMRI study of intentional and unintentional (embarrassing) violations of social norms. *Brain* 125, 1696–1708. doi: 10.1093/brain/awf190

Bohus, D., and Horvitz, E. (2014). "Managing human-robot engagement with forecasts and ...um... hesitations," in *Proceedings of the 16th International Conference on Multimodal Interaction* (Istanbul: ACM), 2–9.

Bull, P. E. (1987). "Posture and gesture," in *International Series in Experimental Social Psychology* 16 (Oxford: Pergamon Press).

Carter, E. J., Mistry, M. N., Carr, G. P. K., Kelly, B. A., and Hodgins, J. K. (2014). "Playing catch with robots: incorporating social gestures into physical interactions," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication, 2014 RO-MAN* (Edinburgh: IEEE), 231–236.

Clark, H. H. (2003). "Pointing and placing," in *Pointing: Where Language, Culture, and Cognition Meet*, ed S. Kita (Psychology Press), 243–268.

de Bruijn, E. R., Miedl, S. F., and Bekkering, H. (2011). How a co-actor's task affects monitoring of own errors: evidence from a social event-related potential study. *Exp. Brain Res.* 211, 397–404. doi: 10.1007/s00221-011-2615-1

Ekman, P. (1992). An argument for basic emotions. *Cogn. Emot.* 6, 169–200. doi: 10.1080/02699939208411068

Ekman, P., and Friesen, W. V. (1969). The repertoire of nonverbal behavior: categories, origins, usage, and coding. *Semiotica* 1, 49–98. doi: 10.1515/semi.1969.1.1.49

Falkenstein, M., Hohnsbein, J., Hoormann, J., and Blanke, L. (1990). Effects of errors in choice reaction tasks on the erp under focused and divided attention. *Psychophysiol. Brain Res.* 1, 192–195.

Forbes, C. E., and Grafman, J. (2013). Social neuroscience: the second phase. *Front. Hum. Neurosci.* 7:20. doi: 10.3389/fnhum.2013.00020

Foster, M. E., Gaschler, A., Giuliani, M., Isard, A., Pateraki, M., and Petrick, R. P. A. (2012). "Two people walk into a bar: dynamic multi-party social interaction with a robot agent," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction (ICMI 2012)* (Santa Monica, CA).

Giuliani, M., Foster, M. E., Isard, A., Matheson, C., Oberlander, J., and Knoll, A. (2010). "Situated reference in a hybrid human-robot interaction system," in *Proceedings of the 6th International Natural Language Generation Conference (INLG 2010)* (Dublin).

Giuliani, M., Petrick, R. P., Foster, M. E., Gaschler, A., Isard, A., Pateraki, M., et al. (2013). "Comparing task-based and socially intelligent behaviour in a robot bartender," in *Proceedings of the 15th International Conference on Multimodal Interfaces (ICMI 2013)* (Sydney, NSW).

Goodrich, M. A., and Schultz, A. C. (2007). Human-robot interaction: a survey. *Found. Trends Hum. Comput. Interact.* 1, 203–275. doi: 10.1561/1100000005

Hagenaars, M. A., Oitzl, M., and Roelofs, K. (2014). Updating freeze: aligning animal and human research. *Neurosci. Biobehav. Rev.* 47, 165–176. doi: 10.1016/j.neubiorev.2014.07.021

Hoque, M. E., McDuff, D. J., and Picard, R. W. (2012). Exploring temporal patterns in classifying frustrated and delighted smiles. *IEEE Trans. Affect. Comput.* 3, 323–334. doi: 10.1109/T-AFFC.2012.11

Jang, M., Lee, D.-H., Kim, J., and Cho, Y. (2013). "Identifying principal social signals in private student-teacher interactions for robot-enhanced education," in *RO-MAN, 2013 IEEE* (Gyeongju), 621–626.

Koban, L., Corradi-Dell'Acqua, C., and Vuilleumier, P. (2013). Integration of error agency and representation of others' pain in the anterior insula. *J. Cogn. Neurosci.* 25, 258–272. doi: 10.1162/jocn-a-00324

Laurier, E., and Philo, C. (2006). "Natural problems of naturalistic video data," in *Video Analysis. Methodology and Methods. Qualitative Audiovisual Data Analysis in Sociology*, eds H. Knoblauch, B. Schnettler, J. Raab, and H.-G. Soeffner (Frankfurt am Main: Peter Lang).

Longmore, C. A., and Tree, J. J. (2013). Motion as a cue to face recognition: evidence from congenital prosopagnosia. *Neuropsychologia* 51, 864–875. doi: 10.1016/j.neuropsychologia.2013.01.022

Loth, S., Huth, K., and De Ruiter, J. P. (2013). Automatic detection of service initiation signals used in bars. *Front. Psychol.* 4:557. doi: 10.3389/fpsyg.2013.0055

O'Toole, A. J., Roark, D. A., and Abdi, H. (2002). Recognizing moving faces: a psychological and neural synthesis. *Trends Cogn. Sci.* 6, 261–266. doi: 10.1016/S1364-6613(02)01908-3

Radke, S., de Lange, F. P., Ullsperger, M., and De Bruijn, E. (2011). Mistakes that affect others: an fMRI study on processing of own errors in a social context. *Exp. Brain Res.* 211, 405–413. doi: 10.1007/s00221-011-2677-0

Rasmussen, J. (1982). Human errors. a taxonomy for describing human malfunction in industrial installations. *J. Occup. Accid.* 4, 311–333. doi: 10.1016/0376-6349(82)90041-4

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., and Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science* 306, 443–447. doi: 10.1126/science.1100301

Sato, R., and Takeuchi, Y. (2014). "Coordinating turn-taking and talking in multi-party conversations by controlling Robot's eye-gaze," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication, 2014 RO-MAN* (Edinburgh: IEEE), 280–285.

Schank, R., and Abelson, R. (1977). *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures, Vol. 2*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Scheutz, M., Cantrell, R., and Schermerhorn, P. (2011). Toward humanlike task-based dialogue processing for human robot interaction. *AI Mag.* 32, 77–84. doi: 10.1609/aimag.v32i4.2381

Smith, E. R., Mackie, D. M., and Claypool, H. M. (2014). *Social Psychology*. New York, NY: Psychology Press.

Stadler, S., Weiss, A., and Tscheligi, M. (2014). "I trained this robot: the impact of pre-experience and execution behavior on robot teachers," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication, 2014 RO-MAN* (Edinburgh: IEEE), 1030–1036.

Stanton, C., and Stevens, C. J. (2014). "Robot pressure: the impact of robot eye gaze and lifelike bodily movements upon decision-making and trust," in *Social Robotics* (Springer International Publishing), 330–339.

Sunstein, C. R. (1996). Social norms and social roles. *Columbia Law Rev.* 96, 903–968.

Tseng, S.-H., Hsu, Y.-H., Chiang, Y.-S., Wu, T.-Y., and Fu, L.-C. (2014). "Multi-human spatial social pattern understanding for a multi-modal robot through nonverbal social signals," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication, 2014 RO-MAN* (Edinburgh: IEEE), 531–536.

van Breemen, A. J. N. (2005). "iCat: experimenting with animabotics," in *Proceedings of AISB 2005 Creative Robotics Symposium* (Hatfield).

Vinciarelli, A., Pantic, M., and Bourlard, H. (2009). Social signal processing: survey of an emerging domain. *Image Vis. Comput.* 27, 1743–1759. doi: 10.1016/j.imavis.2008.11.007

Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., et al. (2012). Bridging the gap between social animal and unsocial machine: a survey of social signal processing. *IEEE Trans. Affect. Comput.* 3, 69–87. doi: 10.1109/T-AFFC.2011.27

Weiss, A., Mirnig, N., Bruckenberger, U., Strasser, E., Tscheligi, M., Kühnlenz, B., et al. (2015). The interactive urban robot: user-centered development and final field trial of a direction requesting robot. *Paladyn J. Behav. Rob.* 6, 42–56. doi: 10.1515/pjbr-2015-0005

Witchel, H. J., Westling, C. E., Tee, J., Healy, A., Needham, R., and Chockalingam, N. (2014). What does not happen: quantifying embodied engagement using nimi and self-adaptors. *Participations* 11, 304–331.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). "ELAN: a professional framework for multimodality research," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)* (Genoa).

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 21 July 2015
doi: 10.3389/fpsyg.2015.01016

Look together: analyzing gaze coordination with epistemic network analysis

Sean Andrist[1], Wesley Collier[2], Michael Gleicher[1], Bilge Mutlu[1] and David Shaffer[2]*

[1] Department of Computer Sciences, University of Wisconsin–Madison, Madison, WI, USA, [2] Department of Educational Psychology, University of Wisconsin–Madison, Madison, WI, USA

When conversing and collaborating in everyday situations, people naturally and interactively align their behaviors with each other across various communication channels, including speech, gesture, posture, and gaze. Having access to a partner's referential gaze behavior has been shown to be particularly important in achieving collaborative outcomes, but the process in which people's gaze behaviors unfold over the course of an interaction and become tightly coordinated is not well understood. In this paper, we present work to develop a deeper and more nuanced understanding of coordinated referential gaze in collaborating dyads. We recruited 13 dyads to participate in a collaborative sandwich-making task and used dual mobile eye tracking to synchronously record each participant's gaze behavior. We used a relatively new analysis technique—epistemic network analysis—to jointly model the gaze behaviors of both conversational participants. In this analysis, network nodes represent gaze targets for each participant, and edge strengths convey the likelihood of simultaneous gaze to the connected target nodes during a given time-slice. We divided collaborative task sequences into discrete phases to examine how the networks of shared gaze evolved over longer time windows. We conducted three separate analyses of the data to reveal (1) properties and patterns of how gaze coordination unfolds throughout an interaction sequence, (2) optimal time lags of gaze alignment within a dyad at different phases of the interaction, and (3) differences in gaze coordination patterns for interaction sequences that lead to breakdowns and repairs. In addition to contributing to the growing body of knowledge on the coordination of gaze behaviors in joint activities, this work has implications for the design of future technologies that engage in situated interactions with human users.

Keywords: referential gaze, epistemic network analysis, conversational repair, social signals, gaze tracking

OPEN ACCESS

Edited by:
Sebastian Loth, Universität Bielefeld, Germany

Reviewed by:
Nathan Kirchner, University of Technology Sydney, Australia
Thies Pfeiffer, Bielefeld University, Germany

****Correspondence:***
Sean Andrist, Department of Computer Science, University of Wisconsin–Madison, 1210 W. Dayton St., Madison, WI 53706, USA
sandrist@cs.wisc.edu

Specialty section:
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology

Received: *22 March 2015*
Accepted: *06 July 2015*
Published: *21 July 2015*

Citation:
Andrist S, Collier W, Gleicher M, Mutlu B and Shaffer D (2015) Look together: analyzing gaze coordination with epistemic network analysis. Front. Psychol. 6:1016. doi: 10.3389/fpsyg.2015.01016

1. Introduction

The key to successful communication is *coordination*, which in conversations enables participants to manage speaking turns (Sacks et al., 1974) and to draw each other's attention toward objects of mutual interest using actions such as pointing, placing, gesturing, and gazing (Clark, 2003; Clark and Krych, 2004). Through the course of an interaction, interlocutors mimic each other's syntactic structures (Branigan et al., 2000) and accents (Giles et al., 1991), and their bodies even begin to sway in synchrony (Condon and Osgton, 1971; Shockley et al., 2003). These acts of coordination

are critical to ensuring that *joint activities*, including conversation and collaboration, flow easily and intelligibly (Clark, 1996; Garrod and Pickering, 2004).

Of particular importance to successful interaction is the coordination of gaze and attention across a shared visual space (Clark and Brennan, 1991; Schober, 1993; Clark, 1996; Brown-Schmidt et al., 2005). *Gaze coordination* has been succinctly defined as a coupling of gaze patterns (Richardson et al., 2009). This coupling does not result from interlocutors explicitly aiming to synchronize their gaze movements, but instead gaze patterns become aligned over time due the need for coordination in joint activities. Mechanisms of gaze coordination, including mutual gaze and joint attention, serve as primary instruments of prelinguistic learning between infants and caregivers (Baldwin, 1995) and play a crucial role throughout life in coordinating conversations (Bavelas et al., 2002). Beyond coordination, gaze contributes to a larger number of important processes in everyday human interaction, including conveying attitudes and social roles (Argyle and Cook, 1976).

Although a large number of studies over the past several decades has investigated gaze behavior and the crucial role it plays in communication, how tightly coordinated gaze behaviors unfold over the course of an interaction is not well understood. For example, previous work has examined the timings of when people look toward referents—objects to which they or their interlocutors verbally refer (Tanenhaus et al., 1995; Griffin, 2004; Meyer et al., 2004). However, these investigations are generally one-sided, looking at each person's gaze in isolation, and do not capture the intricate coordinative patterns in which partners' referential gaze behaviors interact. Previous work has also investigated gaze alignment, exploring the extent to which conversational partners gaze toward the same targets at various time offsets (Richardson and Dale, 2005; Bard et al., 2009). However, existing research still lacks a more nuanced description of how gaze alignment changes over the different phases of the interaction.

In this paper, we present work to develop a deeper understanding of coordinated referential gaze in collaborating dyads. We are particularly interested in how the gaze behaviors of two collaborating participants unfold throughout a *reference-action sequence* in which one participant makes a verbal reference to an object in the shared workspace that the other participant is expected to act upon in some way. We collected data from 13 dyads outfitted with mobile eye-tracking glasses in a sandwich-making task; one participant (the instructor) made verbal references to visible ingredients they would like added to their sandwich while the other participant (the worker) was responsible for assembling those ingredients into the final sandwich (**Figure 2**). We chose this task to represent collaborative interactions that contain a large number of reference-action sequences. Because these behavior sequences are common and frequent across many kinds of interactions, we believe that the results of the analyses discussed in this work will generalize beyond the specific sandwich-making task to any interactions that involve reference-action sequences.

Due to the highly dynamic and interdependent nature of the data we collected, we utilized a relatively new analysis technique—*epistemic network analysis (ENA)*—to analyze and visualize the gaze targets of both participants as a complex and dynamic network of relationships. Our overall analysis was shaped by three research questions: (1) How do a collaborating dyad's gaze behaviors *unfold* over the course of a reference-action sequence? (2) How does the *alignment* of gaze behaviors shift throughout the different phases of a reference-action sequence? (3) How do coordinated gaze behaviors differ in sequences which include breakdowns and/or *repairs*?

To answer these three research questions, we conducted three separate analyses of the dyadic gaze data using ENA. In the first analysis, we used ENA to characterize different phases of a reference-action sequence, discovering clear differences in gaze behavior at each phase. This analysis also revealed a consistent pattern of gaze behavior that progresses in an orderly and predictable fashion throughout a reference-action sequence. In the second analysis, we explored the progression of gaze alignment between the interacting participants throughout a reference-action sequence. In general, we discovered a common rise and fall in the amount of aligned gaze throughout a sequence, as well as a back and forth pattern of which participant's gaze "led" the other's. In the third analysis, we explored the difference in gaze behaviors arising during sequences with repairs—verbal clarifications made in response to confusion or requests for clarification—vs. sequences without such repairs. ENA revealed detectably different patterns of gaze behavior for these two types of sequences, even at very early phases of the sequences before any verbal repair occurs.

In the next section, we review the relevant background on shared gaze in collaborative interactions. We also review cross-recurrence analysis, a common analytical tool used in prior work to analyze two-party gaze behaviors, in order to motivate our introduction of a newer approach. In the following section, we present network analysis, specifically epistemic network analysis (ENA), as an alternative to cross-recurrence analysis with a number of desirable properties for studying shared gaze in dyads. We then describe the data collection in the sandwich-making task, followed by the three analyses conducted in ENA. We conclude the paper with a discussion of the patterns of coordinated gaze uncovered in our analyses and their implications for interactive technologies and future research within this space.

2. Background

Previous research has revealed a significant amount of detail about the eye movements of speakers and listeners in isolation. In general, people look toward the things they are speaking about (Griffin, 2004; Meyer et al., 2004), toward the things they hear verbally referenced (Tanenhaus et al., 1995), and toward the things they anticipate will soon be referenced (Altmann and Kamide, 2004). For example, when speakers are asked to describe a simple scene, they fixate the objects in the order in which they mention them and roughly 800–1000 ms before naming them (Meyer et al., 1998; Griffin and Bock, 2000). Although fixation times are heavily modulated by context, research suggests that listeners will fixate an object roughly 500–1000 ms after

the onset of the spoken reference, which includes the 100–200 ms needed to plan and execute an eye movement (Fischer, 1998). When listeners view a scene containing referents for what they are hearing, their eye movements show that they can recognize a word before hearing all of it (Allopenna et al., 1998) and use visual information to disambiguate syntactic structures (Tanenhaus et al., 1995).

When collaborating over a shared workspace, conversational partners use each others' gaze to indicate attention toward and understanding of verbal references to objects in the shared environment (Gergle and Clark, 2011). Partners show increased shared gaze toward referents while they speak about those objects (Bard et al., 2009). Referencing is often a multimodal process, with objects being evoked through a speaker's actions, movement, or other pragmatic contextual cues such as gestures or head nods (Gergle and Clark, 2011). Speakers often under-specify their referents, relying on the listener to seek clarification if more information is needed to uniquely identify a particular referent (Campana et al., 2001). Previous research has shown that speakers look toward their addressees in order to check their understanding of references to new entities (Nakano et al., 2003) and that addressees rely on the speaker's gaze as a cue for disambiguating references, often before the reference could be disambiguated linguistically (Hanna and Brennan, 2007). This use of gaze has the effect of minimizing the joint effort of the participants in an interaction by reducing the time speakers must spend specifying referents.

Most previous research on gaze in interaction makes a simplifying assumption of *pseudounilaterality*—the implicit assumption that a behavioral variable is unilaterally determined by the actions of the participant expressing that behavior (Duncan et al., 1984). This assumption results in erroneously interpreting data on a participant's actions as representing the unilateral conduct of that participant, overlooking the partner's contribution to those data. A primary cause of pseudounilaterality is the use of simple-rate variables—generated by counting or by timing the occurrence of an action during an interaction and dividing that number by some broader count or timing. These variables do not contain information on the sequences in which actions occur in interaction.

Mobile dual eye-tracking is a relatively recent approach to capturing gaze behaviors that allows researchers to overcome problems of pseudounilaterality and develop more nuanced and ecologically valid accounts of how interlocutors coordinate their gaze during natural, situated conversations (Clark and Gergle, 2011). They have provided great opportunities for researchers to better understand the role of gaze as a coordination mechanism in conversation. Dual eye-tracking methods can be used to better understand the role gaze plays as a conversational resource during reference—how people specify the person, object, or entity that they are talking about (Clark and Gergle, 2012).

Cross-recurrence analysis is a commonly used technique for analyzing gaze data captured from participant dyads, as it permits the visualization and quantification of recurrent patterns of states between two time series, such as the gaze patterns of two conversational participants (Zbilut et al., 1998) (**Figure 1**). This analysis approach can reveal the temporal dynamics of a dataset without making assumptions about its statistical nature. The horizontal and vertical axes of a cross-recurrence plot specify the gaze of each of the two partners in interaction. Each diagonal on the plot (lower-left to upper-right) corresponds to an alignment of the participants' gaze with a particular time lag between them. A point is plotted on the diagonal whenever the gaze is *recurrent*—their eyes are fixating at the same object at the given time. The longest diagonal, from bottom-left to top-right of the plot, represents the gaze alignment at a lag of 0. Diagonals above and below that line represent alignments with positive and

FIGURE 1 | Cross-recurrence plots adapted from work by Richardson and Dale (2005). Horizontal and vertical axes specify the gaze of a speaker and a listener. Diagonal slices (lower-left to upper-right) correspond to an alignment of the participants' gaze with a particular time lag between them. A point is plotted on the diagonal whenever the gaze is recurrent. These plots visually compare a "good" listener (well aligned with the speaker's gaze) to a "bad" listener (not as well aligned). They also show the poor alignment of random gaze with a speaker's gaze.

negative offsets, shifting one of the participants' time-series gaze data in relation to the other participant.

Previous research utilizing cross-recurrence analysis has successfully expanded knowledge on gaze coordination. For example, research has shown that a listener's eye movements most closely match a speaker's eye movements at a delay of 2 s (Richardson and Dale, 2005) (**Figure 1**). In fact, the more closely a listener's eye movements are coupled with a speaker's, the better the listener does on a comprehension test. These results were later extended to find that eye movement coupling is sensitive to the knowledge that participants bring to their conversations (Richardson et al., 2007). The presence of the visual scene and beliefs about its perception by others also influence language use and gaze coordination in remote collaborations (Richardson et al., 2009). Gaze is not always well aligned; when speakers' referring expressions ignore listeners' needs, dyads show poor coordination of visual attention (Bard et al., 2009). Dyads whose members more effectively produce referring expressions better coordinate their attention better and in a way linked to the elaboration of the referring expressions.

Although cross-recurrence analysis has yielded some success in studying gaze coordination, it is best suited for examining data from short time windows and one pair at a time. Cross-recurrence plots do not support aggregating data from numerous dyads over long time spans in order to abstract away individual differences and discover generalizable patterns of interaction. These plots can also be difficult to interpret visually and lack the sophistication to represent the complex, dynamic relationships that characterize coordinated gaze over a shared physical workspace. In the next section, we present a particular instantiation of network analysis—epistemic network analysis—as an alternative analytical tool that overcomes these issues.

3. Epistemic Network Analysis

Studying gaze coordination and the temporal unfolding of collaborative gaze behaviors is difficult due to the highly dynamic and interdependent nature of the data. In order to explore this type of data, we were inspired to use an approach that is similar to social network analysis, which provides a robust set of analytical tools to represent networks of relationships, including complex and dynamic relationships (Wasserman, 1994; Brandes and Erlebach, 2005). However, social network analysis was developed to investigate relationships between people rather than relationships within discourse, gaze behaviors, or other indicators of cognitive processes.

Epistemic network analysis (ENA) is a relatively new analysis technique that is based in part on social network analytic models. ENA extends social network analysis by focusing on the patterns of relations among discourse elements, including the things people say and do. ENA networks are characterized by a relatively small number of nodes in contrast with the very large networks that techniques from social network analysis were designed to analyze, which often have hundreds, thousands, or even millions of nodes. In ENA networks, the weights of the connections between nodes (i.e., the association structures between elements) are particularly important, as are the dynamic changes in the weights and in the relative weighting of the links between different nodes.

ENA was designed to highlight connections among "actors," e.g., people, ideas, concepts, events, and behaviors, in a system. It was originally developed to measure relationships between elements of professional expertise by quantifying the co-occurrences of those elements in discourse and has been used for that purpose in a number of contexts (Rupp et al., 2009; Shaffer et al., 2009; Rupp et al., 2010; Orrill and Shaffer, 2012). However, ENA is a promising method to effectively analyze datasets that capture the co-occurrence of any behaviors or actions in social interactions over time.

The data within ENA are represented in a dynamic network model that quantifies changes in the strength and composition of *epistemic frames* over time. An epistemic frame is composed of individual frame elements, f_i, where i represents a particular coded element in a specified window of time. For our purposes, "coded elements" of the epistemic frame are annotated *gaze targets* for each participant in the interaction, and these elements are represented as nodes in a network. For any dyad, p, in any given reference-action sequence, s, each segment of interaction discourse, $D^{p,s}$, provides evidence of which epistemic frame elements (gaze targets) were active (being gazed toward). For this work, each segment of interaction represents 50 ms of time in the interaction.

Each segment of coded data is represented as a vector of `1` or `0`s representing the presence or absence, respectively, of each of the codes. Links, or relations, between epistemic frame elements are defined as co-occurrences of codes within the same segment. To calculate these links, each coded vector is converted into an adjacency matrix, $A^{p,s}$, for dyad p. For our purposes, co-occurrence of two codes is equivalent to the recurrence of gaze to the gaze targets represented by the codes. For any two gaze codes, the strength of their association in a network is computed based on the frequency of their co-occurrence in the data.

$$A^{p,s}_{i,j} = 1 \text{ if } f_i \text{ and } f_j \text{ are both in } D^{p,s}$$

Each coded segment's adjacency matrix, $A^{p,s}_{i,j}$ is then converted into an adjacency vector and summed into a single cumulative adjacency vector for each dyad p for each unit of analysis.

$$U^{p,s} = \sum A^{p,s}$$

For each dyad, p, and each reference-action sequence, s, the cumulative adjacency vector, $U^{p,s}$, is used to define the location of the segments in a high dimensional vector space defined by the intersections of each of the codes. Cumulative adjacency vectors are then normalized to a unit hypersphere to control for the variation in vector length, representing frequencies of co-occurring code pairs, by dividing each value by the square root of the sum of squares of the vector.

$$nU^{p,s} = U^{p,s} / \sqrt{\sum (U^{p,s})^2}$$

A singular value decomposition (SVD) is then performed to explore the structure of the code co-occurrences in the dataset.

The normalized cumulative adjacency vectors are first projected into a high dimensional space such that similar patterns of co-occurrences between coded elements would be positioned proximately. The SVD analysis then decomposes the structure of the data in this high dimensional space into a set of uncorrelated components, fewer in number than the number of dimensions that still account for as much of the variance in the data as possible, such that each accumulated adjacency vector, i, has a set of coordinates, P_i, on the reduced set of dimensions. The resulting networks are then visualized by locating the original frame elements, i.e., the network nodes, using an optimization routine that minimizes

$$\sum (P_i - C_i)^2$$

where P_i is the projection of the point under SVD, and C_i is the centroid of the network graph under the node positioning being tested. This operation produces a distribution of nodes in the network graph determined by the loading vectors that contain them in the space of adjacency vectors. Links are then constructed between the positioned network nodes according to the adjacency matrix.

The mean network for a group of networks can be calculated by computing the mean values of each edge weight in the networks. We can also conduct t-tests between groups of networks to determine if one group's networks (group A) are statistically different from a second group's networks (group B). The t-test operates on the distribution of the centroids of each group on one dimension. For example, we can determine if group A is statistically different from group B on the x-axis by calculating the means of each group's centroid projected to the x-axis and then conducting a t-test with a standard alpha level of 0.05.

4. Method

In order to gain a better understanding of how gaze coordination unfolds over reference-action sequences in dyadic collaborations, we conducted a data collection study in which pairs of participants engaged in a collaborative sandwich-making task. In this section, we present the collection of the data, followed by a number of analyses and visualizations conducted on that data using ENA.

4.1. Data Collection

We recruited 13 previously unacquainted dyads of participants from the University of Wisconsin–Madison campus. This data collection study was approved by the Education and Social/Behavioral Science Institutional Review Board (IRB) of the University of Wisconsin–Madison and all participants granted their written informed consent at the beginning of the study procedure. Participants sat across from each other at a table on which were laid out a number of potential sandwich ingredients and two slices of bread (**Figure 2**). One participant was assigned the role of *instructor*, and the other was assigned the role of *worker*. The instructor acted as a customer at a deli counter, using verbal instructions to tell the worker what ingredients they wanted on their sandwich, and the worker carried out the actions of moving the desired ingredients to the bread.

Each dyad carried out the sandwich-making task twice so that each participant would have a turn as both instructor and worker, resulting in 26 dyadic interactions. The experimenter told the instructor to request any 15 ingredients for their sandwich from among 23 ingredients laid out on the table. The choice of ingredients was left to the instructor; no list was provided by the experimenter. The instructor was asked to only request a single ingredient at a time and to refrain from pointing to or touching the ingredients directly. Upon completion of the first sandwich, an experimenter entered the study room to reset the ingredients back to their original locations on the table, and the participants switched roles for the second sandwich.

During the study, both participants wore mobile eye-tracking glasses developed by SMI[1]. These eye-trackers perform binocular dark-pupil tracking with a sampling rate of 30 Hz and gaze

[1]http://www.smivision.com/en/gaze-and-eye-tracking-systems.

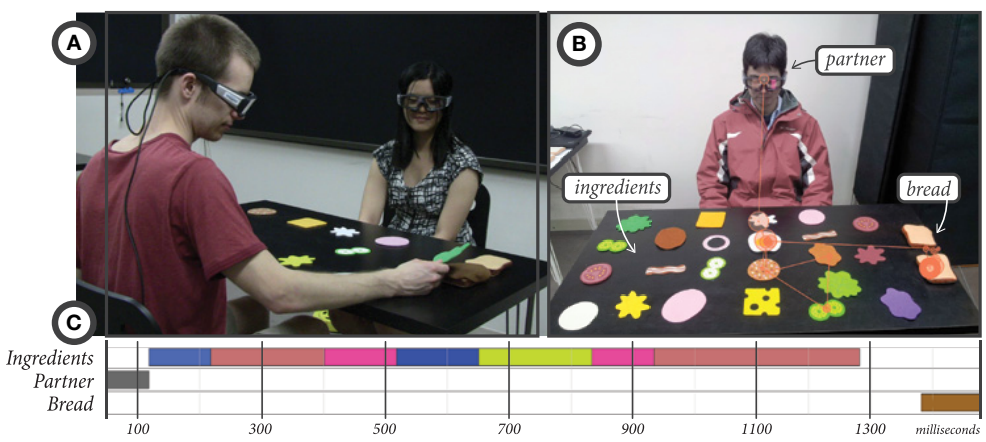


FIGURE 2 | (A) The setup of the data collection experiment in the sandwich-making task. **(B)** A view from one participant's eye-tracking glasses, showing their scan path throughout a reference-action sequence. **(C)** A timeline view of the gaze fixations to ingredients, the partner, and the bread shown in the scan path in **(B)**.

position accuracy of 0.5°. Each set of glasses contains a forward-facing high-definition camera that was used to record both audio and video (24 fps). The gaze trackers were time-synchronized with each other so that the gaze data from both participants could be correlated.

Following data collection, the proprietary BeGaze software created by SMI was used to automatically segment the gaze data into fixations—periods of time when the eyes were at rest on a single target—and saccades—periods of time when the eyes were engaged in rapid movement. Fixation identification minimizes the complexity of eye-tracking data while retaining its most essential characteristics for the purposes of understanding cognitive and visual processing behavior (Salvucci and Goldberg, 2000). BeGaze uses a dispersion-based (spatial) algorithm to compute fixations, emphasizing the spread distance of fixation points under the assumption that fixation points generally occur near one another. Eye fixations and saccades are computed in relation to a forward-facing camera located in the bridge of the eye-tracking glasses worn by the user. Thus, these fixations and saccades are defined within the coordinate frame of the user's head, and user head movements do not interfere with the detection of eye movements.

Gaze fixations are characterized by their duration and coordinates within the forward-facing camera view. Area-of-interest (AOI) analysis, which maps fixations to labeled target areas (AOIs) is a common method for adding semantic information to raw gaze fixations (Salvucci and Goldberg, 2000). In this work, all fixations were manually labeled for the target of the fixation. These labeled AOIs serve as the input data for ENA, rather than the raw gaze fixations. Possible target AOIs included the sandwich ingredients, the slices of bread, and the conversational partner's face and body. Around 80% of gaze fixations were mapped to these AOIs (79.47% for instructors, 81.65% for workers), and the remainder of gaze fixations were found to be directed elsewhere in space (e.g., to a spot on the table without a sandwich ingredient). Speech was also transcribed for each participant. Instructor requests for specific objects were tagged with the ID of the referenced object, and worker speech was labeled when it was either confirming a request or asking for clarification.

To make successful reference utterances, the speaker needs some form of feedback from the addressee. Despite the best efforts of speakers, there will inevitably be instances of breakdowns—misunderstandings or miscommunication—that can either impede ongoing progress of the interaction or lead to breakdowns in the future (Zahn, 1984). To correct breakdowns, humans engage in *repair*, a process that allows speakers to correct misunderstandings and helps ensure that the listener has the correct understanding of the relayed information (Zahn, 1984; Hirst et al., 1994). In the current data collection, if an instructor provided extra clarification for an initially inadequate reference, possibly prompted by the worker's request for clarification, that sequence was marked as containing a *repair*.

Following data collection, each interaction was divided into a set of reference-action sequences, such as a verbal request for bacon followed by the action of moving the bacon to the bread. Each sequence was further divided into five discrete phases: *pre-reference*, the time before any verbal reference has been made; *reference*, the time during the verbal request for a specific sandwich ingredient; *post-reference*, the time directly after the verbal reference and up until the worker's action; *action*, the time during the worker's action of moving the ingredient to the target bread; and *post-action*, the time immediately following this action.

We note that these phases are defined according to verbal and physical actions, not according to gaze behaviors, which are analyzed within each of these phases. The pre-reference phase (average length = 1.90 s) ends at the onset of the verbal reference. The reference phase (average length = 1.32 s) ends with the end of the utterance of the verbal reference. The end of the post-reference phase (average length = 0.78 s) is marked by the start of the physical action, which involves picking up the referent, particularly the moment it is first touched. The action phase (average length = 1.68 s) ends with the end of the physical action, which involves moving the ingredient to the bread and is marked by the moment it is let go. Finally, the end of any feedback provided by the instructor or the beginning of some preparatory utterance for the next reference, e.g., "so, uh, next I'll have...," marked the end of the post-action phase (average length = 0.81 s).

4.2. Analysis

As a first step of our analysis, we calculated common descriptive statistics for the gaze data. Unsurprisingly, we found very little mutual gaze during the reference-action sequences (0.92%) and a fairly large amount of simultaneous shared gaze toward the same target (31.16%). Instructors produced their verbal reference utterance on average 1.31 s after first fixating on it, although they made on average 1.93 fixations to the reference object before verbalizing it. Workers fixated on the reference object on average 1.65s after the verbal reference. Previous research has found that referential gaze in speech typically precedes the corresponding linguistic reference by approximately 800–1000 ms, and people look at what they hear after about 2000 ms (Meyer et al., 1998; Griffin and Bock, 2000). Our data seems to yield statistics close to these findings, and the slightly longer time offset between the gaze fixation and verbal reference among instructors may be due to occasionally having to search for an object, rather than having one already in mind at the beginning of the interaction.

4.2.1. Analysis 1

We analyzed the entirety of our collected data using ENA (**Figure 3**). For our first analysis, we considered each dyad ($n = 26$; 13 dyads × two interactions each) and phase ($n = 5$; pre-reference, reference, post-reference, action, or post-action) as the units of analysis. Each point in the central plot of **Figure 3** represents the centroid of a network for a single dyad's interaction in one of the five phases, collapsed across all reference-action sequences that occurred in the interaction. Solid squares represent the centroid of the mean network for all dyads in each of the five phases. These mean network centroids are surrounded by squares representing the confidence interval along both dimensions. A clear separation between each of the five phases can be observed, indicating that the patterns of gaze

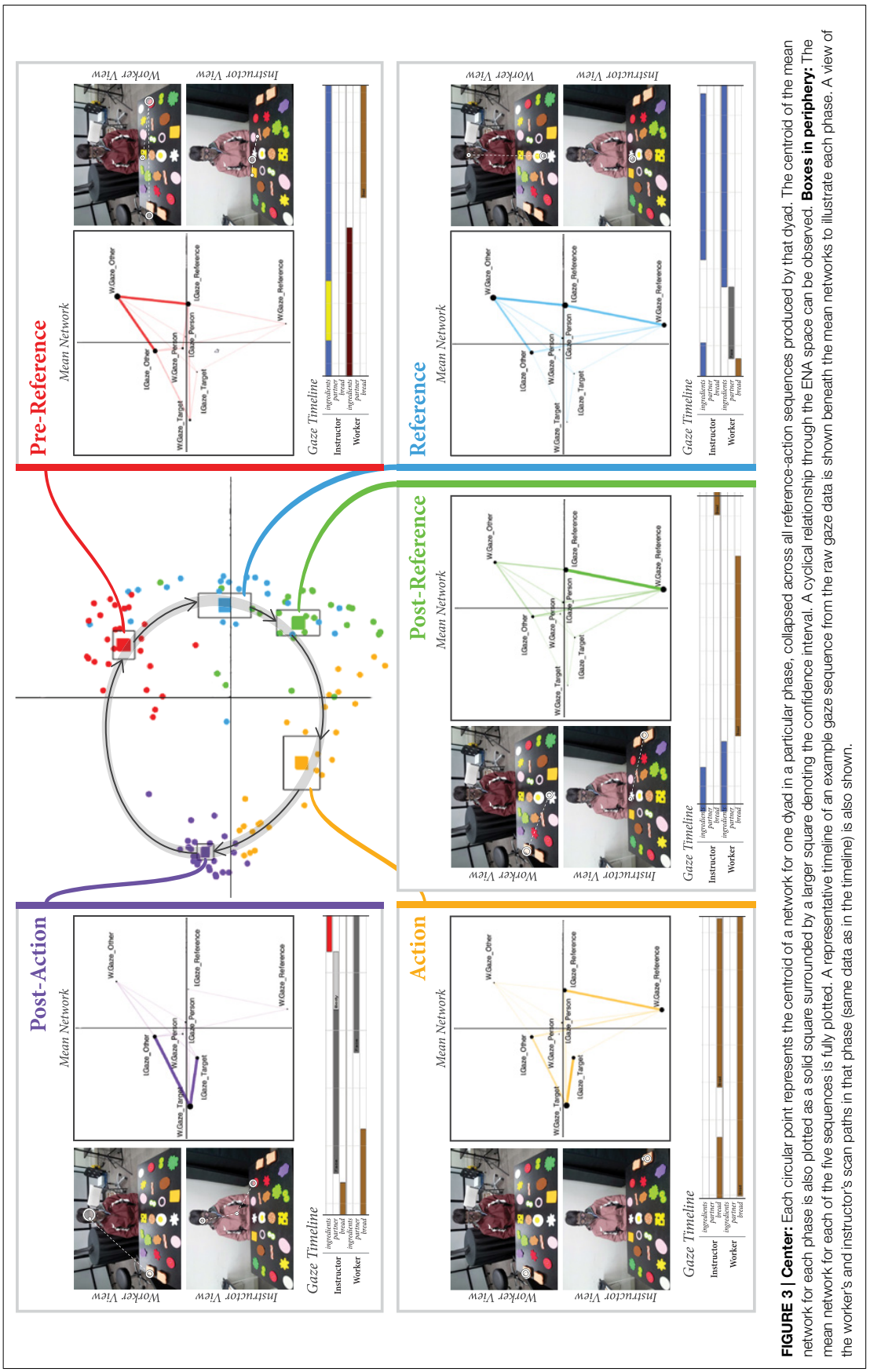


FIGURE 3 | Center: Each circular point represents the centroid of a network for one dyad in a particular phase, collapsed across all reference-action sequences produced by that dyad. The centroid of the mean network for each phase is also plotted as a solid square surrounded by a larger square denoting the confidence interval. A cyclical relationship through the ENA space can be observed. **Boxes in periphery:** The mean network for each of the five sequences is fully plotted. A representative timeline of an example gaze sequence from the raw gaze data is shown beneath the mean networks to illustrate each phase. A view of the worker's and instructor's scan paths in that phase (same data as in the timeline) is also shown.

coordination are significantly different in each of the five phases. We can also observe a clear cyclical pattern through the two-dimensional ENA space as we progress through each of the five phases in the reference-action sequence.

Figure 3 also plots the full mean networks for each of the five phases. As mentioned previously, nodes represent gaze targets, and edge weights represent the relative amount of recurrent gaze to those targets. There are four gaze target nodes for each participant: (1) the reference object for the sequence, (2) the interaction partner, (3) the action target (the bread to which ingredients are moved), and (4) all other objects. In these networks, edges only connect instructor and worker gaze target nodes, as simultaneous gaze within one person toward different targets is not possible. The naming conventions and meanings of all network nodes are explained in **Table 1**.

By examining the placement of nodes in the mean networks, we can develop an intuitive sense of the meaning of each axis in ENA space. As can be observed in the mean networks shown in **Figure 3**, ENA keeps the node positions identical across all plots for a given analysis. Nodes placed at extreme edges of the space, far from the center, are the most informative for intuitively labeling axes. In this respect, three nodes stand out: *W*.*Gaze_Other*, *W*.*Gaze_Reference*, and *W*.*Gaze_Target*. We can therefore recognize that networks with centroids located high on the y-axis are most characterized by strong connections to *W*.*Gaze_Other*. In other words, these networks include more worker gaze toward non-referents. In general, moving from high to low along the y-axis seems to indicate a shift from worker gaze toward non-referents to worker gaze toward the referent. Similarly, moving from right to left along the x-axis seems to indicate a shift from worker gaze toward sandwich ingredients (referents or non-referents) to worker gaze toward the target bread.

In each of the mean networks plotted in **Figure 3** for each of the five phases, the key differences to note are the shifting edge strengths between nodes. In the pre-reference phase, we can observe that the network—which has a centroid high along the y-axis in the central plot of **Figure 3**—has particularly strong connections between *W*.*Gaze_Other* and *I*.*Gaze_Other* and between *W*.*Gaze_Other* and *I*.*Gaze_Reference*.

TABLE 1 | ENA network node names and meanings.

Analysis 1, 2, 3	*I.Gaze_Reference*	Instructor gazing at reference ingredient
	I.Gaze_Other	Instructor gazing at non-reference ingredient
	I.Gaze_Target	Instructor gazing at target bread
	I.Gaze_Person	Instructor gazing at the worker
Analysis 1, 3	*W.Gaze_Reference*	Worker gazing at reference ingredient
	W.Gaze_Other	Worker gazing at non-reference ingredient
	W.Gaze_Target	Worker gazing at target bread
	W.Gaze_Person	Worker gazing at the instructor
Analysis 2	*W.Same*	Worker gazing at same object as instructor
	W.Different	Worker gazing at different object than instructor

Naming convention and meanings of all network nodes used throughout the different analyses.

These connections tell us that the pre-reference phase is characterized mostly by the worker looking toward non-referents while the instructor scans the objects, including the object that they will verbally indicate as the referent in the next phase of the sequence. In the reference phase, we can observe a growing connection between *W*.*Gaze_Reference* and *I*.*Gaze_Reference*, pulling the network centroids lower along the y-axis. In the post-reference phase, this connection is now strongest, and connections with *W*.*Gaze_Other* (the worker gazing to non-referents) have become much weaker, pulling these network centroids yet lower along the y-axis.

In the action phase, a strong connection between *W*.*Gaze_Target* and *I*.*Gaze_Target* appears, signaling simultaneous gaze toward the target, which, in this case, is the bread toward which the selected sandwich ingredient is being moved, pulling the network centroids left along the x-axis. Finally, the post-action phase retains the strong connection between *W*.*Gaze_Target* and *I*.*Gaze_Target*, with a new strong connection between *W*.*Gaze_Target* and *I*.*Gaze_Other*, indicating that the instructor has started to scan other objects in anticipation of the next reference-action sequence while the worker finishes gazing toward the target.

Our first analysis gives us an overall picture of the unfolding gaze patterns in dyadic collaborations throughout a reference-action sequence. We found the clear separation of shared gaze networks between each of the five phases in the reference-action sequence and the orderly cyclical pattern throughout the two-dimensional ENA space to be particularly striking. We highlight that, although the phases themselves are defined in terms of the temporal location of the reference speech and movement action, ENA is acting only upon the gaze data. Thus, patterns of shared gaze are uniquely different across the different phases of the sequence, e.g., before a verbal reference, during the reference, immediately after that reference, and so on. Furthermore, these patterns change and mutate in an orderly way through the abstract space defined by ENA. Theoretically, a mapping from the gaze networks back to the phases can be built. Given a segment of gaze, the phase of the reference-action sequence it came from could be predicted by computing the ENA network for that segment and plotting it in this space.

To validate and demonstrate the promise of the ENA analysis for prediction, we carried out a simple test that involved computing the ENA network as described above, but leaving out data from one of the 13 dyads, which resulted in an ENA space very similar to that shown in **Figure 3**. From the left-out dyad, 200 ms and 1000 ms segments of gaze data were randomly selected. Each of these segments were then modeled as adjacency vectors and projected into the ENA space constructed from data from the other 12 dyads. The predicted phase for each of the projected segments was labeled according to the nearest centroid of phase segments in the ENA space. **Table 2** illustrates the results from this analysis in the form of a confusion matrix. Rows are the actual phase that each segment of data is from, and columns are the predicted phase. As can be seen in the table, prediction appears to be fairly accurate except for some confusion in the shorter phases of *reference* and *action*. In realistic

TABLE 2 | Predicting phase from segments of gaze data.

		Predicted phase (200 ms segments)					Predicted phase (1000 ms segments)				
		Pre-reference	Reference	Post-reference	Action	Post-action	Pre-reference	Reference	Post-reference	Action	post-action
Actual phase	Pre-reference	117	3	10	60	16	31	3	1	2	4
	Reference	50	5	76	38	3	10	2	18	4	0
	Post-reference	6	0	31	6	0	0	2	7	0	0
	Action	7	1	46	52	54	2	0	10	7	13
	Post-action	7	0	0	33	61	0	0	0	2	18

To demonstrate how ENA analysis can be used for prediction, segments of gaze data were projected into the ENA space, and their phase was predicted according to the nearest centroid of phase networks. Rows are the actual phase that each segment of data is from, and columns are the predicted phase. Cells are colored in a gradient from dark green to white according to the quantity of segments in each cell. Prediction appears to be fairly accurate except for some confusion in the shorter phases of reference and action.

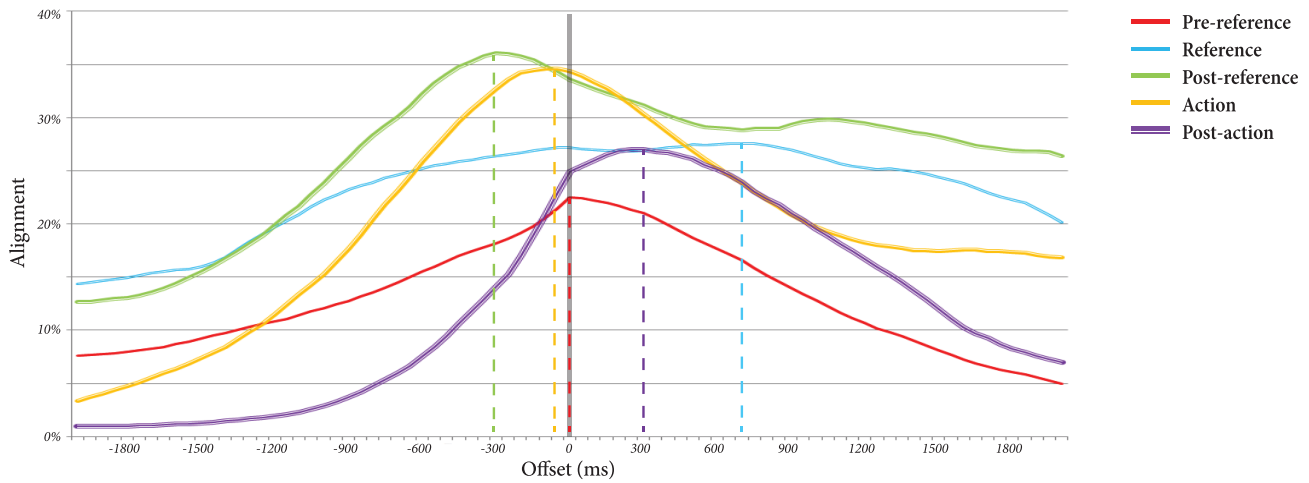


FIGURE 4 | Percentage of gaze alignment between the instructor and worker at each of the five phases, plotted at offset lags from −2 to 2 s. Positive lags indicate instructor lead, while negative lags put the worker ahead of the instructor.

prediction scenarios, prediction accuracy can be improved by using more sophisticated methods than the one employed here for demonstrative purposes, such as dynamically updating phase predictions as segments of gaze data are collected over time or assigning confidence weights to predictions based on their distance from phase centroids.

4.2.2. Analysis 2

In the second analysis, we were interested in finding the optimal lag of gaze alignment within each of the five phases. In other words, which participant's gaze leads that of the other, and by how much, in each phase? For this analysis, two new ENA codes were created: *same*, which is active if the worker and instructor are gazing at the same target (person, reference, target, or other), and *different*, which is active otherwise. For each phase of the reference-action sequence, across all dyads, we shifted the instructor's gaze from −2000 to 2000 ms in 50 ms increments and computed the value for each of the new codes. To find the optimal overlap, we divided the sum of the *same* code by the total number of increments in order to find a measure of "alignment" at each time lag. These alignments for each of the five phases are plotted in **Figure 4**.

The peak of the line graph for each of the five phases represents the optimal time lag at that phase. These lags, as well as the amount of gaze alignment that occurs at those lags, are summarized in **Table 3**. Positive lags put the instructor ahead of the worker, indicating that the instructor is "driving" the gaze patterns, while negative lags indicate that the worker is driving the gaze patterns. As can be observed, the pre-reference phase is characterized by neither participant driving the gaze patterns ($t = 0$ s) and a relatively low amount of gaze alignment (*alignment* = 22.5%). However, during the reference phase, the instructor starts to lead the gaze patterns ($t = 700$ ms), and the alignment increases (*alignment* = 27.6%). In the post-reference phase, the worker begins leading ($t = -300$ ms), and the dyad is most aligned (*alignment* = 36.1%). The action phase involves a slight lead by the worker ($t = -50$ ms) and slight drop in alignment (*alignment* = 34.6%). In the post-action phase, the instructor is once again leading ($t = 300$ ms), and the alignment has dropped further (*alignment* = 27.0%).

We next shifted the gaze streams in each phase of the reference-action sequence by that phase's optimal time lag (**Table 3**) and conducted an analysis in ENA by modeling from the instructor's perspective (**Figure 5**). Four nodes represent the

possible gaze targets for the instructor as before, but there are only two nodes for the worker: *W.Same*, signifying whether the worker is looking at the same target as the instructor, and *W.Different*, indicating a different target than the instructor.

By examining the placement of nodes in the mean networks shown in **Figure 5**, we can again develop an intuitive sense of the meaning of each axis in this new ENA space. Along the x-axis, we can observe *I.Gaze_Reference* far to the left and *I.Gaze_Target* far to the right, indicating as a progression from referent-directed gaze to target-directed gaze in this dimension, as the phases move from left to right along the x-axis.

For the y-axis, *I.Gaze_Person* is the lowest node, but the mean networks throughout the five phases in **Figure 5** show only a few strong connections with *I.Gaze_Person*, indicating that the instructor's gaze is not directed toward the worker. Instead, connections with *W.Same* get stronger as the phases move from *pre-reference* to *reference* to *post-reference* and then weaker again as they move to *action* and *post-action*. Strong connections with *W.Same* pull the network centroids lower along the y-axis in the central plot of **Figure 5**, suggesting an interpretation that this axis signifies "alignment." We can observe a rise and fall of alignment in the phases as their corresponding networks fall and rise respectively along the y-axis. This observation matches what we see in **Table 3** where the alignment percentages rise and fall throughout the five phases.

TABLE 3 | Optimal lag and alignment percentage.

	Pre-reference	Reference	Post-reference	Action	Post-action
Optimal Lag (ms)	0	700	−300	−50	300
Alignment (%)	22.5	27.6	36.1	34.6	27.0

Optimal time lags identified in Analysis 2 and the percentage of alignment at each offset.

4.2.3. Analysis 3

In our third and final analysis, we were interested in the differences between phases of reference-action sequences that

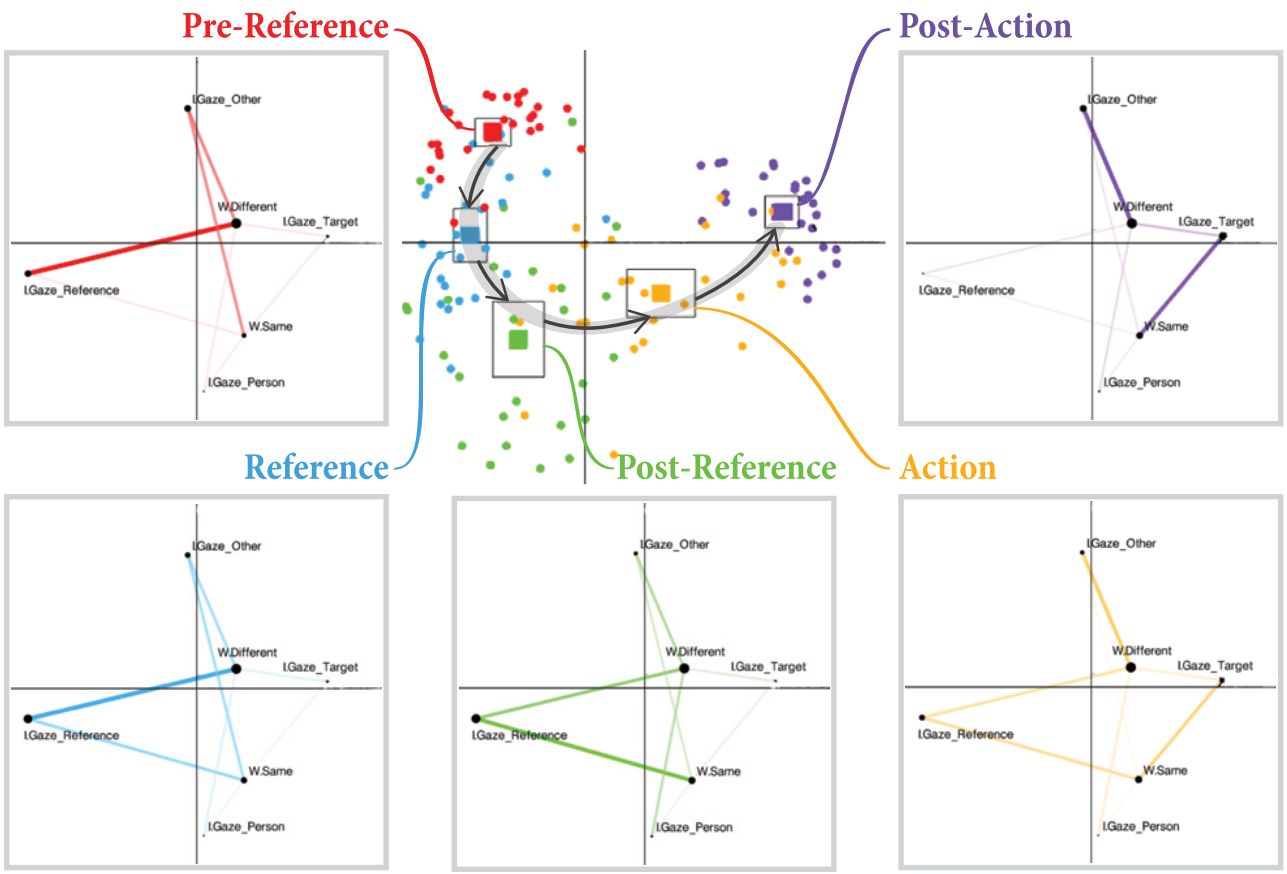


FIGURE 5 | Centroids and mean networks from the ENA that used gaze data from each phase that was shifted by the optimal lag for that phase. The data is modeled from the perspective of the instructor. Four nodes represent the possible gaze targets for the instructor as before, but there are only two nodes for the worker, signifying whether the worker is looking at the same target or a different target. *W_Different* and *W_Same* are largely vertically separated. Networks that are low on the y-axis have strong connections to *W_Same*, while networks high on the axis have strong connections to *W_Different*. Thus, the y-axis can be interpreted as signifying "alignment," and we can observe a rise and fall of alignment in the phases as their corresponding networks fall and rise respectively in the ENA space.

included a repair—in which the instructor had to provide a clarification to their first verbal reference, possibly at the explicit verbal request of the worker—from phases that did not include such repairs. The purpose of this analysis was to answer the following questions. Do the patterns of coordinated gaze in ENA look different during typical sequences vs. those involving repair? More importantly, can the gaze patterns from early phases (pre-reference, reference, and post-reference) be used to predict breakdowns later in the sequence, e.g., before the worker or the instructor offers repair or during repair?

For this analysis, we included "repair" ($n = 2$; repair or no-repair) as another unit of analysis in addition to the "dyad" and "phase" units we had before. As can be observed in **Figure 6**, gaze networks are significantly different between repair and no-repair along the y-axis for each of the first three phases in the reference-action sequence. The centroids of the mean networks (solid squares) for these phases are separated along the y-axis, and there is little vertical overlap in their confidence intervals. These phases, which occur before or during any possible repair, are thus potentially distinguishable along this dimension.

For the pre-reference phase, networks with repair are significantly higher on the y-axis than networks without repair, ($mean_{no-repair} = -0.46$, $mean_{repair} = -0.36$, $t = -2.17$, $p = 0.036$, Cohen's $d = -0.25$). Based on an inspection of the mean networks on the left side of **Figure 6**, this difference appears to be mostly due to the stronger connection between *I.Gaze_Reference* and *W.Gaze_Target* in the sequences with repair, which pulls the network centroids higher along the y-axis. This connection denotes a situation in which the worker is looking toward the target bread while the instructor is looking toward the referent. Here, the worker may still be cognitively engaged in the previous reference-action sequence, i.e., still looking toward the bread after moving the previous reference object there, while the instructor is already preparing their reference utterance for the current reference-action sequence, leading to an eventual breakdown in the interaction.

On the other hand, networks with repair are lower on the y-axis than networks without repair for the reference ($mean_{no-repair} = 0.057$, $mean_{repair} = -0.15$, $t = 2.12$, $p = 0.04$, Cohen's $d = 0.37$) and post-reference ($mean_{no-repair} = 0.42$, $mean_{repair} = 0.18$, $t = 2.79$, $p = 0.008$, Cohen's $d = 0.45$) phases. These differences appear to be mostly due to stronger connections with *W.Gaze_Other* (situated very low on the y-axis) in the sequences with repairs, as shown in **Figure 6**. In other words, the worker is gazing more toward non-referents in these sequences. Also, the networks coming from sequences without repairs appear to have stronger connections between *I.Gaze_Reference* and *W.Gaze_Reference*, pulling these networks higher along the y-axis. This observation implies that, when both the instructor and worker are fixated on the reference object, repairs are less likely to happen.

This analysis revealed that the pattern of coordinated gaze identified in Analysis 1 shows both similarities and differences during sequences involving a repair. More interestingly, the gaze behaviors from phases early in the sequence, particularly the pre-reference and reference phases, are visibly different when a repair occurs later in the sequence than when a repair does not occur later in the sequence. Thus, the need for repair can theoretically be anticipated in advance by observing the pattern of gaze behaviors early in a reference-action sequence.

5. Discussion

The overall goal of our analyses was to develop a more detailed and nuanced understanding of coordinated referential gaze patterns arising in physical dyadic collaborations. In particular, we sought answers to three research questions: (1) How do a collaborating dyad's gaze behaviors *unfold* over the course of a reference-action sequence? (2) How does the *alignment* of gaze behaviors shift throughout the different phases of a reference-action sequence? (3) How do coordinated gaze behaviors differ in sequences that include breakdowns and/or *repairs*? Due to the highly complex, dynamic, and interdependent nature of coordinated two-party gaze behavior, we turned to a relatively new analysis technique in order to explore these questions. Epistemic network analysis is ideally suited for analyzing datasets that capture the co-occurrence of social cues, including the gaze behaviors of multiple participants.

Each of the three analyses we conducted revealed important properties and patterns of coordinated referential gaze behavior in relation to the three research questions. In the first analysis, ENA was able to characterize and separate the five phases of a reference-action sequence (pre-reference, reference, post-reference, action, and post-action). We observed clear and significant differences in shared gaze behavior across these phases. This analysis also revealed a consistent cyclical pattern of gaze behavior that progresses in an orderly and predictable fashion through the two-dimensional abstract space created by ENA. An important implication of this analysis is that the tracked gaze of a collaborating dyad could be used *in situ* to track their progression through a reference-action sequence. By continuously applying ENA to segments of shared gaze behavior, these segments could potentially be classified according to their location within the ENA space as visualized in **Figure 3**.

In the second analysis, we explored the degree of alignment between the gaze behaviors of interacting participants throughout a reference-action sequence. We discovered a general rise and fall in alignment throughout a sequence, as well as a back and forth pattern of which participant was leading the interaction in terms of their gaze behavior. The worker's gaze follows the instructor's gaze during the beginning and end of the sequence when the instructor is leading the interaction by producing the verbal reference or preparing for the next sequence. In contrast, the instructor's gaze follows the worker's gaze during the middle of the sequence (post-reference and action phases) when the instructor appears to monitor the worker's behaviors as the worker attempts to fixate on the reference object and act on it appropriately.

In the third analysis, we explored the differences in gaze behavior between sequences with and without repairs. ENA revealed similar, but characteristically different, patterns of gaze behavior for these two types of sequences. An important implication of this analysis is that, by tracking the shared gaze of a collaborative dyad, repairs can potentially be anticipated

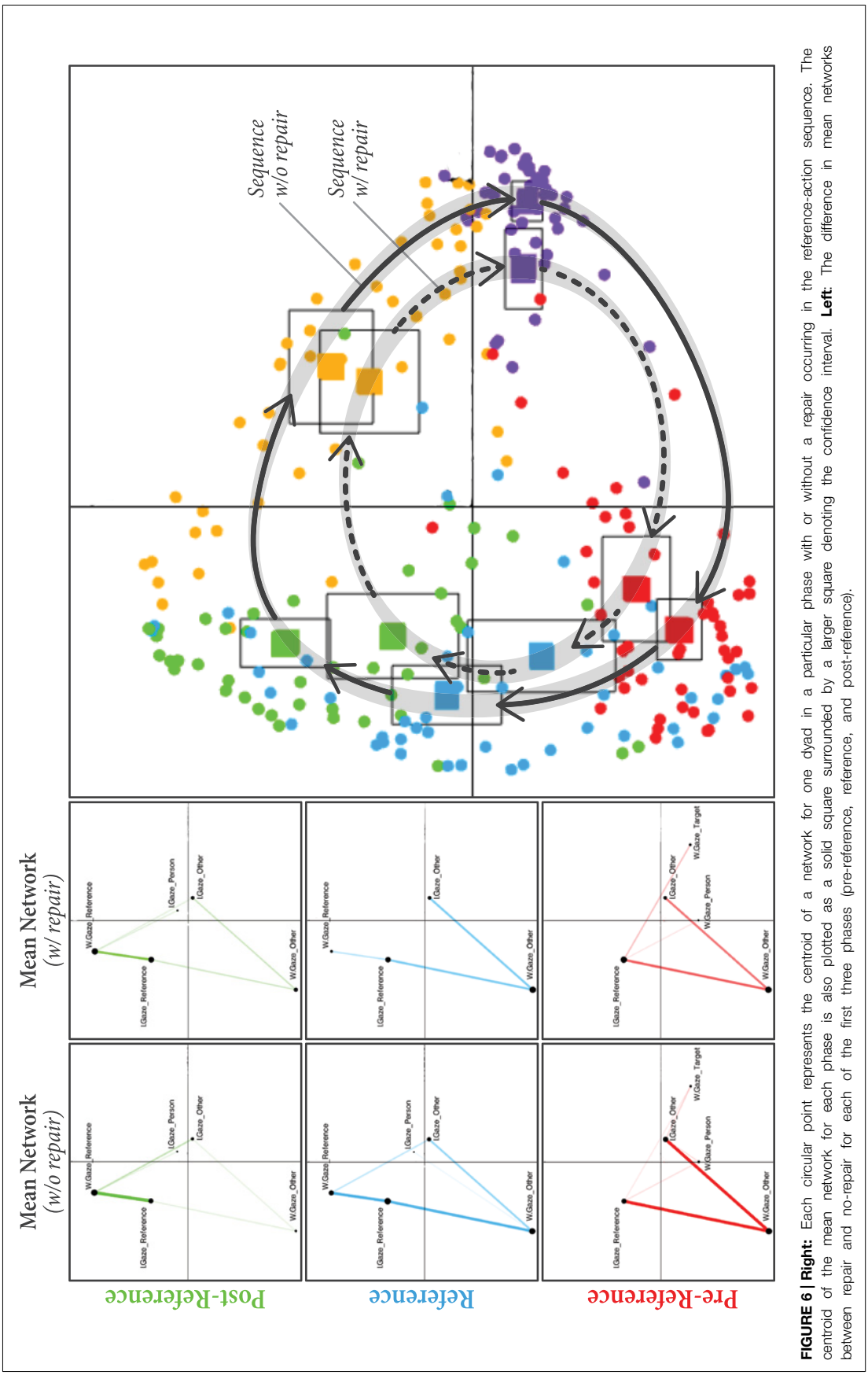


FIGURE 6 | Right: Each circular point represents the centroid of a network for one dyad in a particular phase with or without a repair occurring in the reference-action sequence. The centroid of the mean network for each phase is also plotted as a solid square surrounded by a larger square denoting the confidence interval. **Left:** The difference in mean networks between repair and no-repair for each of the first three phases (pre-reference, reference, and post-reference).

well in advance of their realization. By detecting when the sequence has entered the repair cycle, steps could be taken to quickly resolve any ambiguity or errors and move the interaction back to the non-repair cycle characterizing successful interactions.

There are a number of potential applications that could benefit from the properties and patterns of coordinated gaze discovered in this work. In particular, embodied artificial agents—including social robots and virtual characters—could utilize this knowledge to better align their gaze with human interlocutors and improve coordination in collaborative interactions. This application would require a shift from the *descriptive* analyses that we carried out in the current work to the development of *synthesizing* models that generate coordinative gaze behaviors. By synthesizing gaze behaviors appropriately in coordination with the detected gaze of a human interlocutor, the agent could attempt to produce gaze behaviors that follow the same cyclical pattern of natural humanlike gaze coordination as observed in Analysis 1.

The analyses presented in this paper yield insights that could be directly used to build computational models that generate appropriate gaze cues seen in natural conversations. For example, one such computational model could take the form of a state machine where *states* are represented by possible gaze targets (reference object, target object, conversational partner, etc.), and *transitions* in this state machine would be triggered either probabilistically (e.g., a high probability of gazing toward the referent during the reference phase) or directly by events (e.g., gazing toward the target object in reaction to the conversational partner's gaze toward it). These probabilities and event triggers would be updated from phase to phase according to the cyclical pattern of phases involved in a reference-action sequence as discovered in Analysis 1.

Analyses 2 and 3 similarly have specific implications for modeling and generating gaze behaviors for embodied artificial agents. Analysis 2 sheds light on the role of gaze in "mixed initiative" conversations (Novick et al., 1996). Specifically, the analysis suggests that the agent should shift between leading with its gaze (producing gaze behaviors to which the user is expected to respond) and following the user's gaze (gazing in response to the detected gaze behaviors of the user), as the interaction progresses through the phases of a reference-action sequence. Similarly, following the results of Analysis 3, an agent could recognize misunderstandings by the user before a repair is explicitly and verbally requested, potentially resulting in a more seamless interaction. Furthermore, the agent could make efforts to entirely avoid the patterns of gaze behavior that are characteristic of sequences involving disruptive breakdowns and repairs.

5.1. Future Work

The current work contributes to a growing body of knowledge on the coordination of gaze behaviors in joint activities and points toward a number of opportunities for more exploration within this space. For example, future work may explore other types of interactions, such as conversational or competitive interactions. Another avenue of future research is exploring the tangible implications of observed differences in gaze coordination for the overall success of the interaction. These differences could take the form of deviations from the observed cyclical pattern of Analysis 1 or from the alignments of Analysis 2. For example, there may be differences in participants' comprehension or task success, as was found in cross-recurrence analyses by Richardson and Dale (2005). Future work should also seek to uncover the ways in which gaze coordination can break down, and how breakdowns manifest themselves in ENA beyond our basic consideration of repairs in Analysis 3. This work may include the development of verbal and nonverbal strategies for bringing the interaction back on track when a diversion in the desired pattern of gaze coordination is observed.

Future work should also further investigate the temporal aspects of the gaze behaviors observed in reference-action sequences. The current work divides a reference-action sequence into an ordered sequence of five phases, but the gaze fixations within these phases are aggregated, and the low-level ordering of fixations is lost. While scanpath analysis is commonly used for analyzing temporal characteristics of gaze, scanpaths that result from this analysis only represent the gaze behaviors of individuals. Our analysis attempted to extract generalizable patterns of gaze behavior by aggregating data across multiple dyads and abstracting away the variability in gaze that results from individual differences and changing contextual factors. However, future work with ENA has the potential to extend our findings by retaining the information on the order of gaze fixations by moving from the bi-directional network graphs used in the current work to uni-directional network graphs and splitting each network node into a "sending" node and a "receiving" node. In this representation, a connection from, e.g., a partner-fixation (sending) node to a target-fixation (receiving) node would indicate a gaze fixation toward the target *after* a gaze fixation toward a person.

6. Conclusions

In this paper, we presented work to develop a deeper understanding of coordinated referential gaze in collaborating dyads. The behavioral context for our analyses was the *reference-action sequence*, a pattern of interaction in which one member of the dyad makes a verbal reference to an object in the shared workspace that the other member is expected to act upon in some way. We chose a dyadic sandwich-making task to study collaborative interactions that contain a large number of such sequences. A series of analyses of data collected in this task revealed how gaze coordination unfolded throughout an interaction sequence, how the gaze behaviors of individuals aligned at different phases of the interaction, and what gaze patterns indicated breakdowns and repairs in the interaction. We argue that our characterization of these patterns will generalize beyond this specific task to any interactions that involve reference-action sequences, as these sequences are commonly observed across many kinds of interactions. In addition to contributing to the growing body of knowledge on the coordination of gaze behaviors in joint activities, this work offers

a number of design implications for technologies that engage in dyadic interactions with people.

We used epistemic network analysis for the investigation presented in this paper and demonstrated the promise of ENA as a general tool that could be used for analyses that target not only gaze, but also gestures, language use, facial expressions, cognitive states, and so on. The use of this powerful analytical tool in different settings and in explorations of a variety of social behaviors can significantly expand our knowledge on the nuances of the coordination that naturally arises in successful joint human activities. Additionally, these explorations will enable us to design future technologies that utilize the newfound knowledge in order to more effectively coordinate and collaborate with human users.

Acknowledgments

The dataset analyzed in this paper is also used in another article by Huang et al. (2015) in this Research Topic. This work was funded in part by the MacArthur Foundation and by the National Science Foundation through grants DRL-0918409, DRL-0946372, DRL-1247262, DRL-1418288, DUE-0919347, DUE-1225885, EEC-1232656, EEC-1340402, REC-0347000, IIS-1208632, and IIS-1149970. The opinions, findings, and conclusions do not reflect the views of the funding agencies, cooperating institutions, or other individuals. We would like to thank Golnaz Arastoopour, Chien-Ming Huang, and Ross Luo for their contributions to this work.

References

Allopenna, P. D., Magnuson, J. S., and Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping models. *J. Mem. Lang.* 38, 419–439. doi: 10.1006/jmla.1997.2558

Altmann, G. T., and Kamide, Y. (2004). "Now you see it, now you don't: mediating the mapping between language and the visual world," in *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*, eds J. M. Henderson and F. Ferreira (New York, NY: Psychology Press), 347–386.

Argyle, M., and Cook, M. (1976). *Gaze and Mutual Gaze*. Cambridge, UK: Cambridge University Press.

Baldwin, D. A. (1995). "Understanding the link between joint attention and language," in *Joint Attention: Its Origins and Role in Development*, eds C. Moore and P. J. Dunham (Hillsdale, NJ: Erlbaum), 131–158.

Bard, E. G., Hill, R., and Arai, M. (2009). "Referring and gaze alignment: accessibility is alive and well in situated dialogue," in *Proceedings of CogSci 2009* (Amsterdam: Cognitive Science Society), 1246–1251.

Bavelas, J. B., Coates, L., and Johnson, T. (2002). Listener responses as a collaborative process: the role of gaze. *J. Commun.* 52, 566–580. doi: 10.1111/j.1460-2466.2002.tb02562.x

Brandes, U., and Erlebach, T. (eds.) (2005). "Network analysis: methodological foundations," in *Theoretical Computer Science and General Issues* (Springer-Verlag Berlin Heidelberg: Springer Science & Business Media), 472. doi: 10.1007/b106453

Branigan, H. P., Pickering, M. J., and Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition* 75, B13–B25. doi: 10.1016/S0010-0277(99)00081-5

Brown-Schmidt, S., Campana, E., and Tanenhaus, M. K. (2005). "Real-time reference resolution by naïve participants during a task-based unscripted conversation," in *Approaches to Studying World-Situated Language Use: Bridging the Language-as-Product and Language-as-Action Traditions*, eds J. C. Trueswell and M. K. Tanenhaus (Cambridge, MA: MIT Press), 153–171.

Campana, E., Baldridge, J., Dowding, J., Hockey, B. A., Remington, R. W., and Stone, L. S. (2001). "Using eye movements to determine referents in a spoken dialogue system," in *Proceedings of the 2001 Workshop on Perceptive User Interfaces* (New York, NY: ACM), 1–5.

Clark, A. T., and Gergle, D. (2011). "Mobile dual eye-tracking methods: challenges and opportunities," in *Proceedings of the International Workshop on Dual Eye Tracking (DUET 2011)* (Aarhus).

Clark, A. T., and Gergle, D. (2012). "Know what I'm talking about? Dual eye-tracking in multimodal reference resolution," in *Proceedings of CSCW 2012 Workshop on Dual Eye Tracking* (Seattle, WA), 1–8.

Clark, H. H. (1996). *Using Language*. Cambridge, UK: Cambridge University Press.

Clark, H. H. (2003). "Pointing and placing," in *Pointing: Where Language, Culture, and Cognition Meet*, ed S. Kita (Hillsdale, NJ: Erlbaum), 243–268.

Clark, H. H., and Brennan, S. E. (1991). Grounding in communication. *Perspect. Soc. Shared Cogn.* 13, 127–149. doi: 10.1037/10096-006

Clark, H. H., and Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *J. Mem. Lang.* 50, 62–81. doi: 10.1016/j.jml.2003.08.004

Condon, W. S., and Osgton, W. D. (1971). "Speech and body motion synchrony of the speaker-hearer," in *The Perception of Language*, eds D. H. Horton and J. J. Jenkins (Academic Press), 150–184.

Duncan, S., Kanki, B. G., Mokros, H., and Fiske, D. W. (1984). Pseudounilaterality, simple-rate variables, and other ills to which interaction research is heir. *J. Pers. Soc. Psychol.* 46, 1335–1348. doi: 10.1037/0022-3514.46.6.1335

Fischer, B. (1998). "Attention in saccades," in *Visual Attention*, ed R. D. Wright (New York, NY: Oxford University Press), 289–305.

Garrod, S., and Pickering, M. J. (2004). Why is conversation so easy? *Trends Cogn. Sci.* 8, 8–11. doi: 10.1016/j.tics.2003.10.016

Gergle, D., and Clark, A. T. (2011). "See what I'm saying?: using dyadic mobile eye tracking to study collaborative reference," in *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work* (New York, NY: ACM), 435–444.

Giles, H., Coupland, N., and Coupland, J. (eds.). (1991). "Accommodation theory: communication, context, and consequence," in *Contexts of Accommodation* (Cambridge: Cambridge University Press), 1–68. doi: 10.1017/CBO9780511663673.001

Griffin, Z. M. (2004). The eyes are right when the mouth is wrong. *Psychol. Sci.* 15, 814–821. doi: 10.1111/j.0956-7976.2004.00761.x

Griffin, Z. M., and Bock, K. (2000). What the eyes say about speaking. *Psychol. Sci.* 11, 274–279. doi: 10.1111/1467-9280.00255

Hanna, J. E., and Brennan, S. E. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *J. Mem. Lang.* 57, 596–615. doi: 10.1016/j.jml.2007.01.008

Hirst, G., McRoy, S., Heeman, P., Edmonds, P., and Horton, D. (1994). Repairing conversational misunderstandings and non-understandings. *Speech Commun.* 15, 213–229. doi: 10.1016/0167-6393(94)90073-6

Huang, C.-M., Andrist, S., Sauppé, A., and Mutlu, B. (2015). Using gaze patterns to predict task intent in collaboration. *Front. Psychol.* 6:1049. doi: 10.3389/fpsyg.2015.01049

Meyer, A., van der Meulen, F., and Brooks, A. (2004). Eye movements during speech planning: talking about present and remembered objects. *Vis. Cogn.* 11, 553–576. doi: 10.1080/13506280344000248

Meyer, A. S., Sleiderink, A. M., and Levelt, W. J. (1998). Viewing and naming objects: eye movements during noun phrase production. *Cognition* 66, B25–B33. doi: 10.1016/S0010-0277(98)00009-2

Nakano, Y. I., Reinstein, G., Stocky, T., and Cassell, J. (2003). "Towards a model of face-to-face grounding," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics* (Stroudsburg, PA: Association for Computational Linguistics), 553–561.

Novick, D. G., Hansen, B., and Ward, K. (1996). "Coordinating turn-taking with gaze," in *Proceedings of the Fourth International Conference on Spoken Language (ICSLP 96)* (Philadelphia, PA: IEEE), 1888–1891.

Orrill, C. H., and Shaffer, D. W. (2012). "Exploring connectedness: applying ENA to teacher knowledge," in *International Conference of the Learning Sciences (ICLS)* (Sydney, NSW).

Richardson, D. C., and Dale, R. (2005). Looking to understand: the coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cogn. Sci.* 29, 1045–1060. doi: 10.1207/s15516709cog0000/29

Richardson, D. C., Dale, R., and Kirkham, N. Z. (2007). The art of conversation is coordination: common ground and the coupling of eye movements during dialogue. *Psychol. Sci.* 18, 407–413. doi: 10.1111/j.1467-9280.2007.01914.x

Richardson, D. C., Dale, R., and Tomlinson, J. M. (2009). Conversation, gaze coordination, and beliefs about visual context. *Cogn. Sci.* 33, 1468–1482. doi: 10.1111/j.1551-6709.2009.01057.x

Rupp, A. A., Choi, Y., Gushta, M., Mislevy, R., Bagley, E., Nash, P., et al. (2009). "Modeling learning progressions in epistemic games with epistemic network analysis: principles for data analysis and generation," in *Proceedings from the Learning Progressions in Science Conference* (Iowa City, IA), 24–26.

Rupp, A. A., Gushta, M., Mislevy, R. J., and Shaffer, D. W. (2010). Evidence-centered design of epistemic games: measurement principles for complex learning environments. *J. Technol. Learn. Assess.* 8, 1–48. Available online at: http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1623

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language* 50, 696–735. doi: 10.1353/lan.1974.0010

Salvucci, D. D., and Goldberg, J. H. (2000). "Identifying fixations and saccades in eye-tracking protocols," in *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications* (New York, NY: ACM), 71–78.

Schober, M. F. (1993). Spatial perspective-taking in conversation. *Cognition* 47, 1–24. doi: 10.1353/lan.1974.0010

Shaffer, D. W., Hatfield, D., Svarovsky, G. N., Nash, P., Nulty, A., Bagley, E., et al. (2009). Epistemic network analysis: a prototype for 21st-century assessment of learning. *Int. J. Learn. Media* 1, 33–53. doi: 10.1162/ijlm.2009.0013

Shockley, K., Santana, M.-V., and Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *J. Exp. Psychol. Hum. Percept. Perform.* 29, 326–332. doi: 10.1037/0096-1523.29.2.326

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science* 268, 1632–1634. doi: 10.1126/science.7777863

Wasserman, S. (1994). *Social Network Analysis: Methods and Applications*, Vol. 8. Cambridge, UK: Cambridge University Press.

Zahn, C. J. (1984). A reexamination of conversational repair. *Commun. Monogr.* 51, 56–66. doi: 10.1080/03637758409390183

Zbilut, J. P., Giuliani, A., and Webber, C. L. (1998). Detecting deterministic signals in exceptionally noisy environments using cross-recurrence quantification. *Phys. Lett.* 246, 122–128. doi: 10.1016/S0375-9601(98)00457-5

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in
PSYCHOLOGY

TECHNOLOGY REPORT ARTICLE
published: 10 December 2014
doi: 10.3389/fpsyg.2014.01437

Social signal processing for studying parent–infant interaction

Marie Avril[1], Chloë Leclère[1,2,3], Sylvie Viaux[1,2], Stéphane Michelet[1], Catherine Achard[1], Sylvain Missonnier[3], Miri Keren[4], David Cohen[1,2]* and Mohamed Chetouani[1]

[1] CNRS, Institut des Systèmes Intelligents et de Robotiques, UMR 7222, Université Pierre et Marie Curie, Paris, France
[2] Department of Child and Adolescent Psychiatry, Pitié-Salpêtrière Hospital, Paris, France
[3] Laboratoire de Psychologie Clinique et Psychopathologie, Psychanalyse, Paris René Descartes University, Boulogne, France
[4] Department of Psychiatry, Infant Mental Health Unit, Geha Hospital, Tel Aviv University, Tel Aviv, Israel

Edited by:
Sebastian Loth, Universität Bielefeld, Germany

Reviewed by:
Pablo Gomez Esteban, Vrije Universiteit Brussel, Belgium
Guang Chen, Technische Universität München, Germany

***Correspondence:**
David Cohen, Department of Child and Adolescent Psychiatry, Pitié-Salpêtrière Hospital, 47-83 Boulevard de l'Hôpital, 75651 Paris, Cedex 13, France
e-mail: david.cohen@psl.aphp.fr

Studying early interactions is a core issue of infant development and psychopathology. Automatic social signal processing theoretically offers the possibility to extract and analyze communication by taking an integrative perspective, considering the multimodal nature and dynamics of behaviors (including synchrony). This paper proposes an explorative method to acquire and extract relevant social signals from a naturalistic early parent–infant interaction. An experimental setup is proposed based on both clinical and technical requirements. We extracted various cues from body postures and speech productions of partners using the IMI2S (Interaction, Multimodal Integration, and Social Signal) Framework. Preliminary clinical and computational results are reported for two dyads (one pathological in a situation of severe emotional neglect and one normal control) as an illustration of our cross-disciplinary protocol. The results from both clinical and computational analyzes highlight similar differences: the pathological dyad shows dyssynchronic interaction led by the infant whereas the control dyad shows synchronic interaction and a smooth interactive dialog. The results suggest that the current method might be promising for future studies.

Keywords: early parent–infant interaction, feature extraction, multimodal computational analysis, RGB-D sensor, synchrony, social signal processing

INTRODUCTION

Parent–child interactions are crucial for learning, later psychological traits, and psychopathology (Cohen, 2012). In many species, including mammals, parent–child interactions are based on close relationships that are characterized by (i) infant dependency on caregivers and (ii) a specific communication dynamic associated with a caregiver's adaptation and infant maturation. However, this type of study is complex, requiring the perception and integration of multimodal social signals. Combining several approaches within a multidisciplinary perspective at the intersection of social signal processing, computational neuroscience, developmental psychology, and child psychiatry may efficiently investigate the meaning of social signals during early parent–child interaction (Meltzoff et al., 2009). Exploring normal and pathological interactions during this early period of life has many implications including the possibility of understanding what the baby partner cannot explicitly express due to immaturity.

The Syned-Psy project (Synchrony, Early Development and Psychopathology, http://synedpsy.isir.upmc.fr/) aims to improve the synergy among three fields: child psychiatry, developmental psychology and social signal processing. The idea is to understand the clinical relevance of synchronic and dyssynchronic dyadic interactions and to develop automatic algorithmic tools to detect these phenomena in natural settings. Originally conceptualized and studied by developmental psychologists, the concept of synchrony is now relevant to many different research fields including social signal processing, robotics and machine learning. According to its conceptual framework, synchrony can be defined in many ways (Leclère et al., 2014). Delaherche et al. (2012) recently proposed that in most cases, one should distinguish between *what* is assessed (i.e., modalities such as body movement, gaze, smile, and emotion) and *how* the temporal link between partners' different modalities of interaction are assessed (i.e., speed, simultaneity, and smoothness). In the rest of the manuscript, we will follow this definition of synchrony.

The aim of this work was to characterize synchrony/dyssynchrony in parent–infant interactions occurring in situations of severe emotional neglect and to select interaction metrics that may be used in future clinical trials. To do this, we proposed to automatically detect and analyze behaviors. These behaviors are selected by considering clinical and technical requirements. Furthermore, the objective of our approach was to explore the capacity of new technological devices and tools to understand early parent–child interactions.

RELATED WORK IN PSYCHOLOGY

The quality of the parent–child relationship impacts children's social, emotional and cognitive development (Harrist and Waugh, 2002; Saint-Georges et al., 2013). Describing parent–child behavioral interactions is not a simple task because there are multiple modalities of interaction to explore. First, the interactive partnership between an infant and caregiver (usually called a "dyad") has to be defined and explored as a single unit. Second, given that

the relationship between an infant and their caregiver is bidirectional in nature, the dyad should be thought of as a dynamically interacting system (Sameroff, 2009). Third, given the dynamic relationship between an infant and their caregiver, a specific interest in the flow characterizing the exchange of information during infant-caregiver interactions has emerged (Weisman et al., 2012, 2013), leading to the study of rhythm (Berry et al., 1974; Condon, 1986; Stern, 2009), reciprocity (Lebovici, 1985; Bråten, 1998), and synchrony (Feldman, 2007). The recent discovery of both biological correlates of behaviorally synchronic phenomena (Dumas et al., 2010) and statistical learning (Kuhl, 2003; Saffran, 2003) has validated the crucial value of studying synchrony during child development (Feldman, 2007; Cohen, 2012). It appears that synchrony should be regarded as a social signal *per se* as it has been shown to be valid in both normal and pathological populations. Better mother–child synchrony is associated with familiarity (vs. unknown partner), a healthy mother (vs. pathological mother), typical development (vs. psychopathological development), and a more positive child outcome (Leclère et al., 2014).

In the field of human interactions, interactional synchrony can be defined as "the dynamic and reciprocal adaptation of the temporal structure of behaviors between interactive partners" (Delaherche et al., 2012). Here, behaviors include verbal and non-verbal communicative and emotional behaviors (e.g., gestures, postures, facial displays, vocalizations, and gazes). Synchronous interactions entail coordination between partners and intermodality. Caregivers and their children are able to respond to each other using different modalities starting from birth (Vandenberg, 2006; Hart, 2010). Thus, synchrony differs from mirroring or the chameleon effect. Instead, synchrony describes the intricate "dance" that occurs during short, intense, playful interactions; it builds on familiarity with the partner's behavioral repertoire and interaction rhythms, and it depicts the underlying temporal structure of highly aroused moments of interpersonal exchange that are clearly separated from the stream of daily life (Beebe and Lachmann, 1988; Tronick and Cohn, 1989; Fogel et al., 1992; Bråten, 1998; Stern, 2009). Therefore, synchrony has been measured in many different ways due to its broad range of theoretical applicability. The most common terms referring to synchrony are mutuality, reciprocity, rhythmicity, harmonious interaction, turn-taking and shared affect; all terms are used to characterize the mother–child dyad. Three main types of assessment methods for studying synchrony have emerged: (1) global interaction scales with dyadic items; (2) specific synchrony scales; and (3) micro-coded time-series analyzes (for a detailed review, see Leclère et al., 2014).

RELATED WORK IN COMPUTATIONAL PROCESSING

Many studies have been conducted (Gatica-Perez, 2009) to assess social interactions using automatic and computational methods, including automatic extraction of non-verbal cues and/or models of the multimodal nature of interaction. These studies have been performed in various contextual applications including role recognition (Salamin et al., 2009), partner coordination during interaction (Hung and Gatica-Perez, 2010), automatic analysis of meeting (Campbell, 2009; Vinciarelli et al., 2009), studying interactive virtual agents (Prepin and Pelachaud, 2011), and understanding of early development (Meltzoff et al., 2009). In the health domain, these applications include recognition or classification of psychopathological states (Cohn, 2010), psychotherapeutic alliance (Ramseyer and Tschacher, 2011), classification of autistic dimensions (Demouy et al., 2011) or the recognition of early expression of autism (Cohen et al., 2013).

Signals that have been investigated during social interactions are specific because they are not semantic in nature and often occur without consciousness. They include amplitude, frequency and duration for the non-verbal signals such as fillers, backchannels or gestures. Vinciarelli et al. (2009) distinguish five categories of cues: (1) physical appearance; (2) gesture and posture; (3) gaze and facial behaviors and mimics; (4) vocal cues; and (5) behavior related to the space and environment. Regarding audio signals, some cues have been better studied such as pitch, intensity and vocal quality (Batliner et al., 2011), intonation (Ringeval et al., 2011), rhythm (Hogan, 2011), motherese (Saint-Georges et al., 2013), and perceived emotion (Schuller et al., 2010). Regarding video signals, cues usually investigated include the quantity of body movements (Altmann, 2011; Ramseyer and Tschacher, 2011; Paxton and Dale, 2014) or facial movements (Carletta et al., 2006), the study of hand movements (Marcos-Ramiro et al., 2013; Ramanathan et al., 2013) or finger movements (Dong et al., 2013), the study of gaze (Sanchez-Cortes et al., 2013), and data with a higher level of annotation including smiling (Rehg et al., 2013), facial expressions (Bilakhia et al., 2013), posture (Feese et al., 2012) or the emotional body language (McColl and Nejat, 2014). In the era of RGB-D sensors (e.g., Kinect), online extraction of the skeleton is now available and has enabled the study of action recognition based on the joint architecture of the human body (Aggarwal and Xia, 2014; Chan-Hon-Tong et al., 2014). As a consequence, new body movement cues have been proposed based on the position of articulated arms, the trunk, head, and legs (Caridakis and Karpouzis, 2011; Yun et al., 2012; Anzalone et al., 2014a).

Some cues have been extracted to assess social characteristics and interaction at the level of the dyad (Yun et al., 2012; Ramanathan et al., 2013). Several studies (Campbell, 2009; Delaherche et al., 2012; Bilakhia et al., 2013; Rolf and Asada, 2014) have considered the multimodal nature of social signals and simultaneously studied several modalities. Various authors have used different metrics and modeling techniques to study synchrony (Delaherche et al., 2012), including correlation (Altmann, 2011), recurrent analysis (Varni et al., 2009), regression models (Bilakhia et al., 2013), quantity of mutual information (Rolf and Asada, 2014), or influence models (Dong et al., 2013).

PAPER CONTRIBUTION AND ORGANIZATION

The aim of this paper is to describe our methodology and to test its feasibility. Here, we present a pilot study in which we extracted and analyzed behavioral features in two case reports, one pathological situation of severe emotional neglect and one normal control, to study the feasibility and the coherence of the method. From an experimental point of view, the particularity of this work is to employ a computational setup in a clinical setting, where both needs and constraints had to be completed. The acquisition application had to preserve a natural free-play interaction

between pathological dyads and be usable by a non-expert. All the interactive scenarios and the applications have been designed in collaboration with psychologists. This collaboration has continued with the selection of relevant behavioral features from the raw data and their interpretation. The rest of the paper is organized as follows: in section 2, we present the method used to set up a computational system in a clinical setting and how we analyzed data acquired during the interactions. In section 3, results of clinical and computational analysis are presented for two representative dyads, and in section 4, the method and results are discussed.

MATERIALS AND METHODS

In this section, we focus on the integration of a computational setup in a clinical study and how the data recorded during the interaction can be treated. From a clinical point of view, the protocol aims to offer an optimal acquisition of parent–infant interactions and to preserve the natural interaction. The method of acquiring data must be as minimally intrusive as possible. From a technical perspective, the acquisition must be sufficiently efficient and robust to be able to collect significant and exploitable data for off-line processing.

CLINICAL PROTOCOL

The current protocol is part of a clinical study conducted in a French perinatal ambulatory unit "Unité Petite Enfance et Parentalité Vivaldi" of the Pitié-Salpêtrière University Hospital. The main objective of the study, named "*ESPOIR Bébé Famille*," is to evaluate the relevance of an early intensive intervention program for dyads in severe child neglect (CN) situations. CN is the persistent failure of the caregiver to meet the child's basic physical and/or psychological needs, resulting in interaction disorders (Glaser, 2002) and serious impairment of the child's development with short and long term negative impacts on the child's cognitive, socio emotional, behavioral and psychological development and emotional regulation (Rees, 2008). Thus, a severe neglectful situation presents interaction difficulties and dyssynchrony.

The inclusion criteria were as follows: (1) Dyads consisted of mothers (or fathers) with their children whose age varied between 12 and 36 months. At 12 months, the interactive pattern of the dyad is already built, and data extraction is facilitated because the child is able to sit in a small chair. The oldest age accepted was 36 months because that is the age limit for the parent child health care in this unit. (2) Mothers (or fathers) have been referred to the unit by social services or court petitions due to CN. (3) Clinical confirmation of CN is based on a child psychiatrist's assessment using the PIRGAS scale (Parent–Infant Relationship Global Assessment Scale, Axe II of DC 0-3 R), a clinical intensive scale of parent–child interaction quality. A control group of dyads with normal development and without interactional difficulty was also recruited.

The clinical evaluation of these dyads included interviews, questionnaires and filmed play sessions used for clinical annotations. Specifically, to assess synchrony, we used the coding interactive behavior (CIB), which is one of the most often used and validated global interaction scales (for a review of clinical instruments see Leclère et al., 2014). The CIB includes 43 codes rated on a 5-point Likert scale, divided into parent, child and dyadic codes. Codes were averaged into eight composites that were theoretically derived, concerned with diverse aspects of early parent–infant relationships and showed acceptable to high levels of internal consistency (Feldman et al., 1996; Keren et al., 2001). The French version has been validated and offers the same factorial distribution (Viaux-Savelon et al., 2014). The composites and items used in the present study are presented in **Table 1**.

The proposed computational system has been used in the filmed play sessions where parents and infants have a natural interaction. Play session are composed of three stages to capture the dyad behaviors in different contexts: (1) Free interactive play (4 min): parent and infant are invited to play together with toys

Table 1 | CIB relative items according to the eight composite subscores.

Composites	Relatives items
Parental sensitivity	Acknowledging; imitating; elaborating; parent gaze; positive affect; vocal appropriateness, clarity; appropriate range of affect; resourcefulness; praising; affectionate touch; supportive presence; infant led interaction
Parent intrusiveness	Forcing-physical manipulation; overriding, intrusiveness; parent negative affect, anger; parent anxiety; criticizing; parent-led interaction
Parent limit setting	Consistency of style; resourcefulness; appropriate structure, limit setting
Child compliance	Compliance to parent; reliance on parent for help; on-task persistence
Child withdrawal	Child negative emotionality, fussy; withdrawal; labile affect; avoidance of parent
Child engagement	Joint attention; child positive affect; affection to parent; alertness; fatigue; vocalizations, verbal output; initiation; competent use of the environment; creative-symbolic play; infant-led interaction
Dyadic joint negative state	Parent negative affect, anger; hostility; child negative emotionality, fussy; withdrawal, labile affect; fatigue; constriction; tension
Dyadic reciprocity	Parent gaze; positive affect; praising; affectionate touch; joint attention; child positive affect; vocalization, verbal output; initiation; dyadic reciprocity; adaptation-regulation; fluency

as usual. The goal is to create an interaction that is as natural as possible; the only directive given is "play as if you were at home." (2) Directed game (2 min): a complex game is given to the child (a puzzle for example) to encourage the parent to help them. With a difficult game, the purpose is to determine how the child will solicit the parent and how the parent will respond. In addition, this situation will incite the parent to intervene spontaneously during the game. (3) Free play while the parent is occupied (2 min): a questionnaire is given to the parent while the child is playing with toys. In this final situation, the aim is to observe how the child solicits the parent and how the parent shares their attention between the task and their infant.

Play sessions take place in a consultation room, controlled by a psychologist, where the parent and infant are invited to sit around a small table to play. Although a face-to-face disposition facilitates interactions, it complicates the data acquisition. Thus, the parent and infant are placed at 90° to one another around the table. To collect information from the interaction, two synchronized RGB-D sensors are placed in front of each participant and connected to a computer. This will run an acquisition application to record scene data. Additionally, a camera is used to film the scene for the clinical evaluation. **Figure 1A** shows the hardware setup in the consultation room.

Given our aim to run a study in a clinical setting, the acquisition application has to be easy to use and robust. Indeed, dyads with emotional neglect present interaction difficulties and thus the play sessions are subject to variations due to the child's (e.g., standing on a chair, looking for other toys) and parent's behavior (e.g., difficulty in controlling their child, wearing a large coat, hiding their face). Moreover, the psychologist has to leave the room each time the play session takes place. To reach these needs: (i) the hardware system is hidden to offer the most natural environment possible and avoid interest and distraction from the participants. (ii) The psychologist had to prepare the parent for the presence of a camera that is sometimes problematic. (iii) The hardware and the acquisition application were computed to be easily setup.

ACQUISITION APPLICATION

To respond to all of the technical and clinical constraints cited above, an acquisition application has been implemented with a

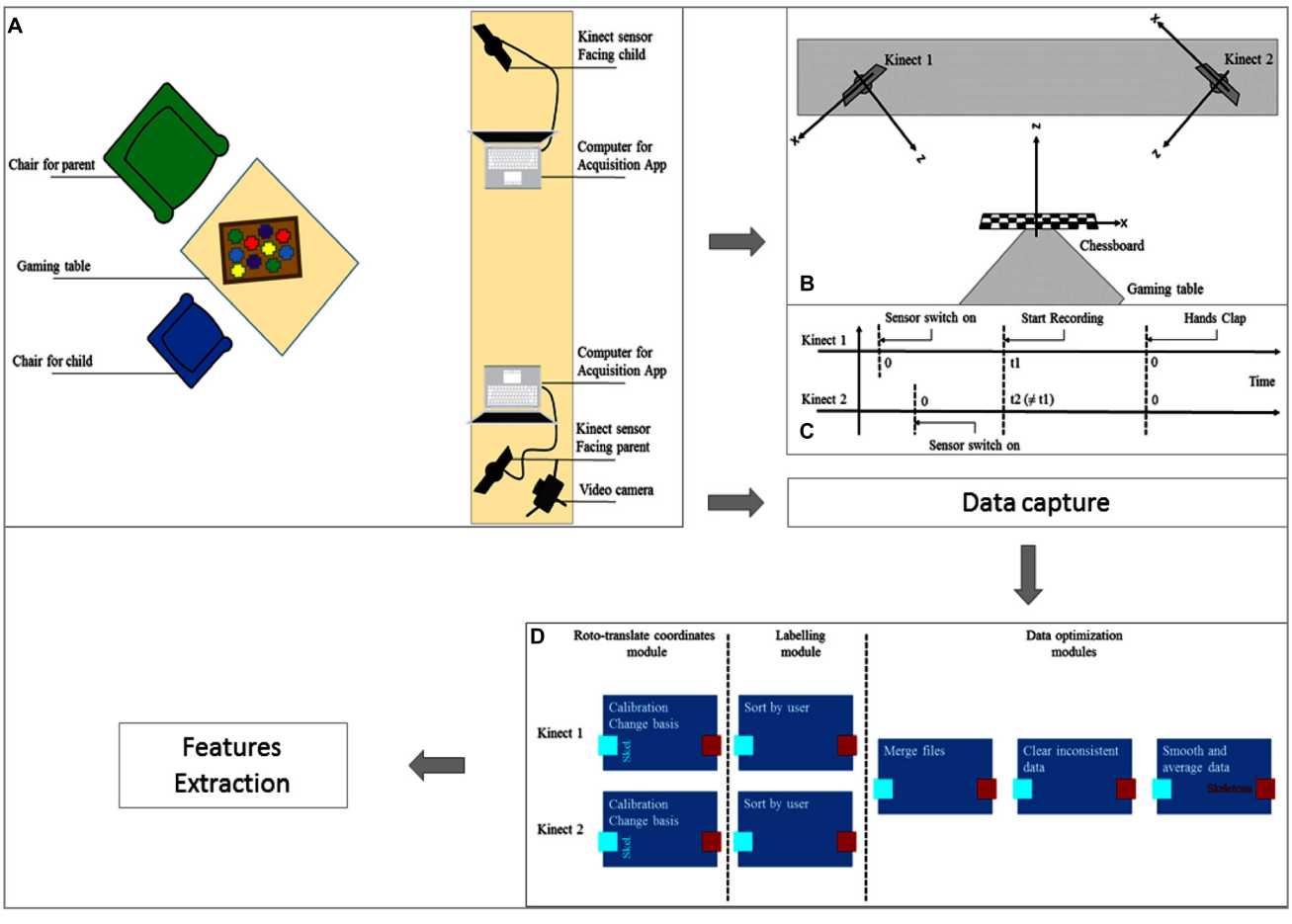


FIGURE 1 | Data recording and extraction. (A) Play room and materials; **(B)** 3D-calibration with a chessboard; **(C)** Time synchronization with a hand clap; **(D)** Skeleton coordinates pre-processing pipeline.

robust and efficient framework and the ability to collect the maximum amount of significant data while remaining easy to use by a non-professional.

As mentioned above, the scene is recorded by two Kinects, low-cost RGB-D sensors designed by Microsoft. These devices, mainly used for gesture recognition, offer the possibility to record many signals from a scene with only one device. The setup incorporates a color camera, depth sensor based on a structured light technique and a microphone array. Coupled with the Microsoft SDK for Kinect, the setup allows the user to directly extract color images, depth images and also 3D coordinates of the skeletons and faces of the participants from a scene in real time. In our case, participants still are too far from the Kinect, so face tracking features are not used. Moreover, as participants are seated, only the upper-body skeleton tracking is activated.

The two Kinects are optimally placed in front of each participant to capture as much information as possible. However, 3D coordinates are obtained in a Kinect centered basis, therefore, trackers record different positions for each sensor. Thus, a spatial calibration of the Kinects is necessary, which is performed by chessboard calibration; a chessboard is placed in the field of view of the Kinects (laid on the gaming table) while the Kinects record the 3D coordinates of significant points of the chessboard (corners of squares). **Figure 1B** shows the calibration step with axis representation. These coordinates will be used later to compute the roto-translation matrix between the two Kinects to transform 3D points tracked into the same spatial basis.

A temporal synchronization is also needed for the Kinects. The internal sensor's clock starts when the device is connected to the computer. As it is impossible to start the two sensors at the exact same time, a temporal synchronization is performed from the microphone outputs. When the Kinects detect a powerful sound for the first time (applause), they record the current timestamp as the beginning of the recording (see **Figure 1C** for a graphical view of the timelines). Then, each Kinect will have the same detection times.

Data captured by the Kinects must be recorded for offline processing. To avoid computer overload during the acquisition (and offer the most efficient recording rate), minimal online processing is performed, and the raw data are saved in a lightweight format. For each sensor, saved data include:

- Color stream in an .avi video file (XVID codec) + timestamp for each image in an .xml file
- Depth stream in an .avi video file (XVID codec) + timestamp for each image in an .xml file
- Audio stream in an audio file (.wav)
- Audio source angle in an .xml file
- Skeleton tracked points (position and orientation) in an .xml file
- 3D calibration data in an .xml file

To facilitate the use of the acquisition application by a non-expert user, a graphical interface has been added. The graphical interface is divided into two windows, one for the visualization of the Kinect stream and the other for parameter management. In the first window, the user can display the Kinect stream, start and stop the recording and also modify the sensor tilt. A message field to display current acquisition status is proposed. In the second window, the user can choose the path to save the recorded data, such as the name of the folder, if tracked skeletons are displayed, or the number of squares on the calibration chessboard. This interface simplifies the use of the acquisition application and allows the verification and correct execution of the recordings.

COMPUTATIONAL ANALYSIS

To extract and analyze the recorded data during the game session, a lightweight framework developed by the IMI2S ISIR group is used (Anzalone et al., 2014b). This framework is a distributed computing software platform that copes with the high level of complexity by simplifying the functional decomposition of the problems through the implementation of highly decoupled, efficient, and portable software. The developers implemented complex solutions using simple, small, and basic operative units that are able to interact between each other. Such basic modules are executed as independent computational units able to solve a particular problem. Inputs and outputs of different modules are then connected to exploit the main, complex problem.

In this study, the IMI2S framework is used to divide records into three segments of data according to the three types of game sessions, preprocess 3D skeleton data and, eventually, to extract behavioral features.

Skeleton preprocessing

As previously described, the use of two RGBD-sensors requires a basis change to obtain 3D coordinates in the same Cartesian space. In addition, to retain a maximum amount of information, data from each sensor must be merged before any treatment. **Figure 1D** presents the pre-processing pipeline for skeleton data from the two displaced sensors. Skeleton data of the parent and child from both sensors are corrected to belong to the same Cartesian space; each skeleton is then labeled, identifying the two users in the scene, the parent and the child. Finally, the data are merged into a unique stream, inconsistent skeletons are suppressed (for example if the tracked skeleton is misplaced) and the data are smoothed through average filtering.

Skeleton processing

After smoothing and cleaning of the skeleton data, several features can be extracted with IMI2S Framework. With 3D coordinates of 10 significant body points for each participant in a unique basis, distances and orientation features can be computed. Many examples of relevant skeleton features will be presented in the section "Results."

Speech processing

We focused on voice activity detection (VAD) estimated through the OpenSmile framework (Eyben et al., 2010). When this feature was combined with the IMI2S framework, we obtained the probability of VAD.

In addition, when the method used by Galatas et al. (2013) was combined with the skeleton localization in space, it was possible to determine an audio source in the 3D space of the

clinical room. Consequently, if a sound was detected, it could be associated with a user, even though distinguishing voices from other sounds (moving a toy, moving chair, etc.) is not currently efficient.

SELECTION OF RELEVANT FEATURES

We deliberately reduce the number of features using a consensus multidisciplinary approach to select the most relevant ones. This was done by going back and forth between engineers and psychologists. First, engineers listed a series of features available from skeleton and audio processing for each partner. Second psychologists discussed with engineers how combining each partner feature could be related to a relevant clinical dimension in terms of communication. We focused on features related to proximity, motor and audio activity, and attention to the task and/or to the partner (see Result). Finally, we determined together higher level features related to synchrony and engagement during the interaction with the aim of selecting a limited number of features for clinical assessment.

RESULTS

The current results focus only on two case reports, one pathological dyad in a severe emotional neglect situation and one control dyad with no interaction difficulty. The pathological dyad is composed of a 25-year-old mother and her 35-month-old boy. The interaction quality is rated as a 45 on the PIRGAS scale (DC 0-3 R). The control dyad is composed of a 29-year-old father and his 19-month-old boy. The interaction quality is rated as a 95 on the PIRGAS scale.

The analyzes were performed for the first phase in the ESPOIR protocol, the free play, where the parent and child are invited to play as they would at home to create as natural of an interaction as possible in the experimental scenario. It should be noted that in these experiments, the psychologist was present in the room with the dyad and stood at the bench between the two computers (see **Figure 1A**). Thus, she was a possible point of attraction during the experiment.

We present successively (1) the clinical assessment; (2) features related to proximity and motor activity; (3) features related to attention to the task and/or to the partner; and (4) participation to the task. Please note that natural interaction does not allow us to extract behavioral features during the entire time of the video session. For instance, data are missing when the child moves from the chair and is off-camera or when he climbs on his parent's knees. A blank or a cross line in figures indicates uncollected data. By convention, results concerning parents are in green, and results concerning children are in blue.

BLIND ASSESSMENT OF THE INTERACTION WITH THE CIB

As expected (**Figure 2**), the control dyad presented significantly higher scores in the CIB positive domains (parental sensitivity, parent limit-setting, child compliance, child engagement, and dyadic reciprocity), and the pathological dyad presented higher scores in the negative domains (Dyadic joint negative state and Child withdrawal). The only domain showing a limited difference was Parent intrusiveness.

PROXIMITY AND ACTIVITY FEATURES

In this paragraph, we present low level features related to physical proximity during the task and motor activity. The main idea is to assess (1) how close partners are to one another and (2) how close partners are to the table where part of the interactions should occur. Several skeleton features have been developed in the IMI2S Framework to extract information concerning the proximity between the parent and child during the game session. Furthermore, these features reveal the general body activity of the participants. **Figure 3** offers a visual representation of (1) the distance between the shoulder center of a participant and the center

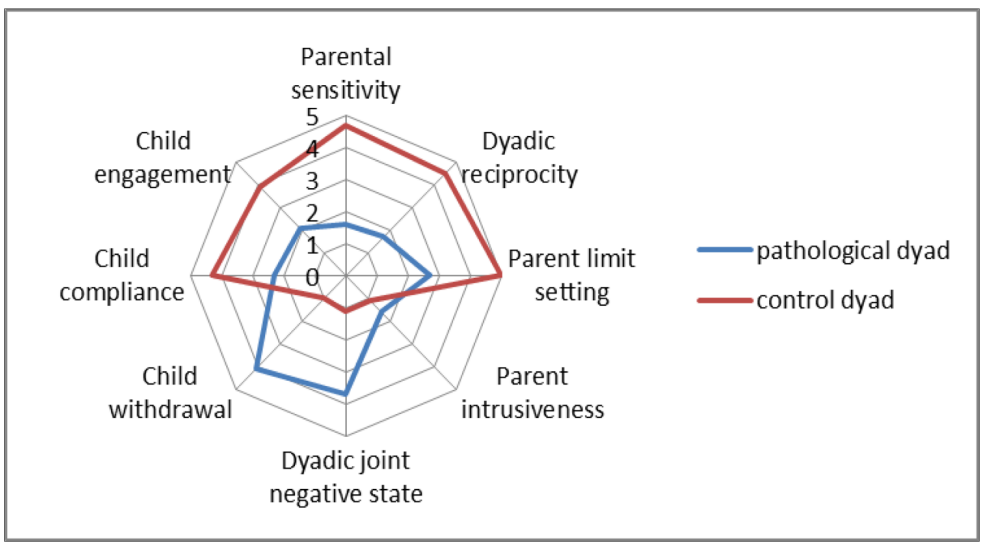


FIGURE 2 | Coding interactive behavior results for the pathological and control dyads.

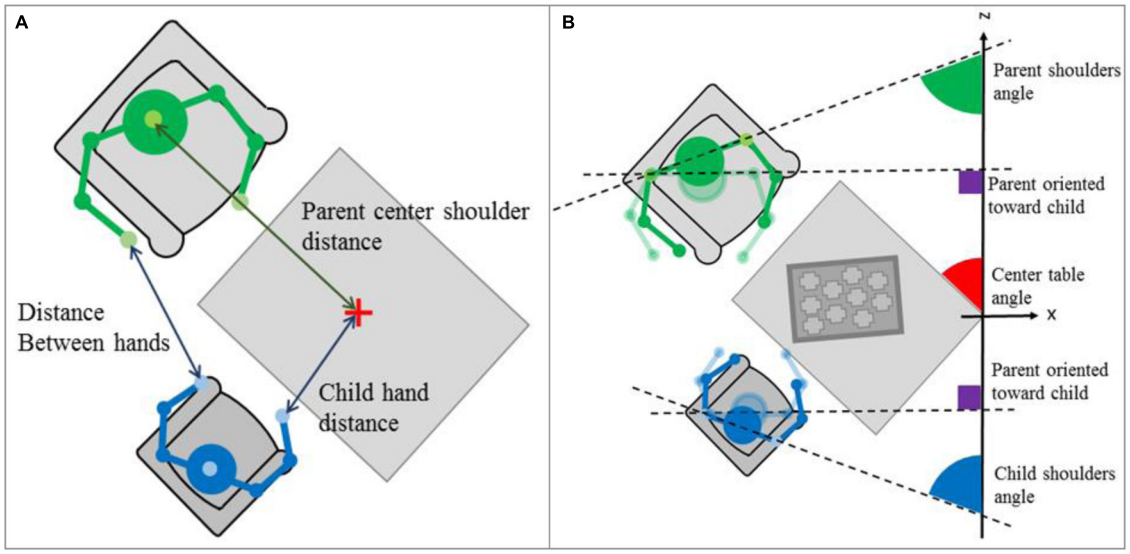


FIGURE 3 | (A) Shoulder center and hand distance to the table center features; **(B)** Shoulder orientation feature – top view representations.

of the gaming table. The shoulder center is the geometrical middle between the left and right shoulders. (2) The distance between each hand of the dyad (parent's left hand-child's right hand and parent's right hand-child's left hand).

The results are presented in **Figures 4A,B**, respectively. The pathological dyad is on the left, whereas the control dyad is on the right. **Table 2** summarizes the conclusions for these features.

ATTENTION TO THE TASK AND THE PARTNER

Here, we present higher-level features related to each partner's attention during the task and whether attention is oriented to the task or to the partner. These features are based on the assumption that if a person's torso faces an area, the person's attention is focused on this area. For example, if the parent's chest is parallel to the table, it indicates that the parent is interested in the action occurring on the table. With the 3D reconstruction from the skeleton features, it was possible to determine the attention of the dyad to the gaming task and the parent's attention to their child and *vice versa* by measuring each partner's shoulder orientation and the relative shoulder orientation during the interaction.

Shoulder orientation results

To determine the torso orientation of a person, the angle between the line formed by the two shoulder points tracked and the line of the z axis has been computed (see **Figure 3B** for a graphical representation). In the current situation, if the person is oriented toward the gaming table, the formed angle will be ∼45° (red line in graphs). Moreover, if the person looks at their partner's spot, the angle will be ∼90° (purple line in graphs). **Figure 4C** displays shoulder orientation for the two dyads.

Relative shoulder orientation results

It is possible to determine the relative orientation between two persons using the same method used for the shoulder orientation. This was defined as the angle between the line formed by the parent's shoulders and the child's shoulders (see **Figure 5** for a graphical representation). Therefore, if parent and child are face to face, the angle will be close to 0° (red line in graphs, **Figure 5C**), while if they are facing the same area, the angle will be oscillate between 45 and 90° (green and purple lines in graphs, **Figure 5B**). The results for this feature are available in **Figure 6**. The interpretations are summarized in **Table 3**.

PARTICIPATION IN THE TASK

In this section, we present higher level features related to synchrony and engagement during the interaction. First, as the shoulder center distance to the table center captures the attention to the task, the hand distance to the table center can express the involvement in the task. Second, by combining distance or audio features with motion energy or speaker localization, we assume that we assessed partner engagement during the interaction.

Hand distance to the table center results

As explained above, the shoulder center distance to the table center captures the attention to the task because the hand distance to the table center can express the involvement in the task. If hands are close to the table, we can assume that the person is playing and therefore involved in the task. Unlike the shoulder centers distance feature, it is not the distance between the centers of the two hands that is studied, but the distance between the closest hand and the center of the gaming table (see **Figure 3A** for a top view representation of the feature). **Figure 7** shows the results for this feature. In the pathological dyad, only the child's hand was close to the table and showed much activity. In contrast, in the control dyad, both the parent's and child's hands were close to the table and showed much activity.

Contribution to global movement

Contribution to the movement determines which partner participates in the global movement, and by studying the distance

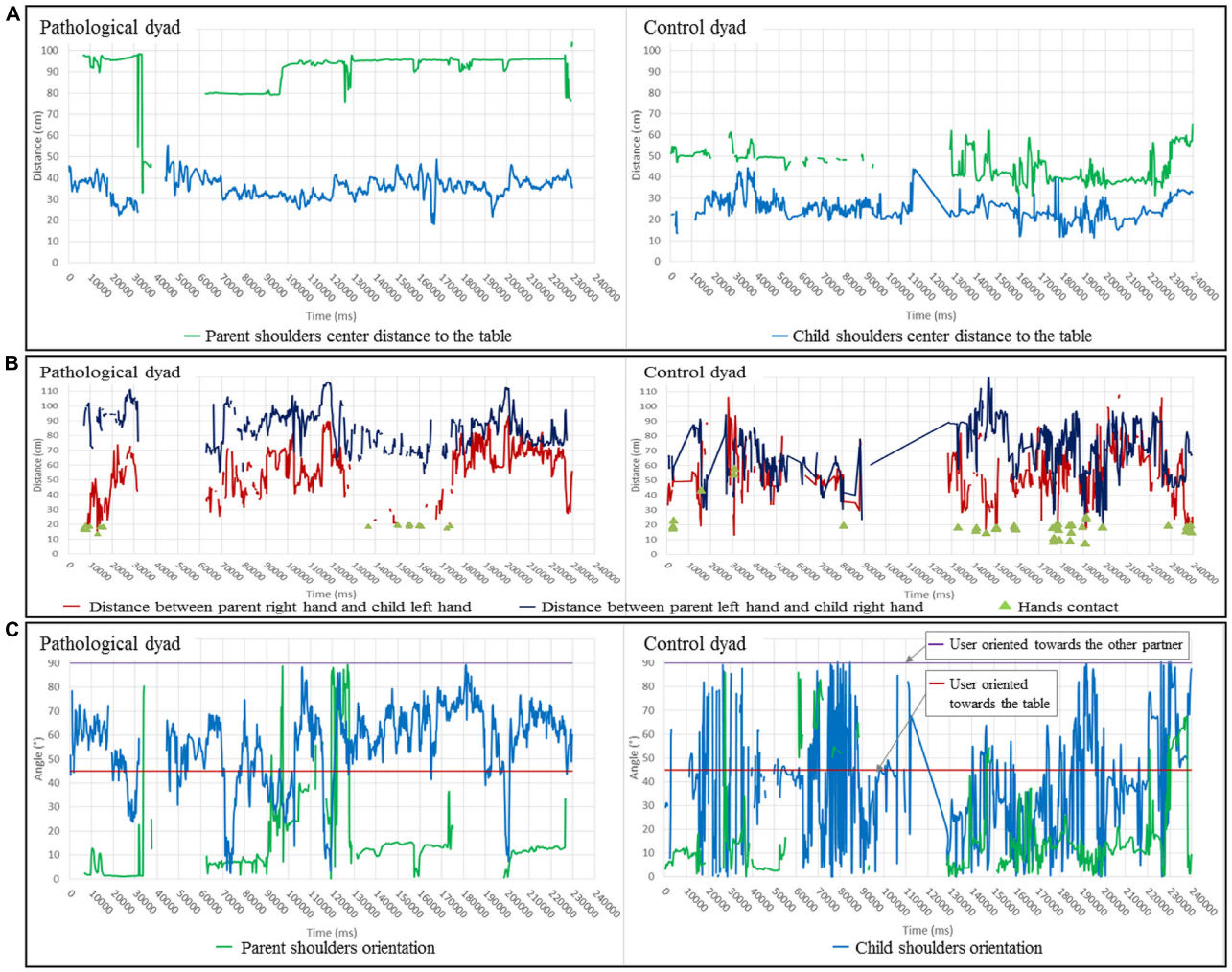


FIGURE 4 | (A) Evolution of the distance between the shoulder center and the table center during the interaction. **(B)** Evolution of the distance between the parent's and child's hands during the interaction. **(C)** Evolution of shoulder orientation during the interaction (Left: pathological dyad; Right: control dyad). A blank or a cross line in figures indicates uncollected data. By convention, results concerning parents are in green, and results concerning children are in blue **(A,C)**.

variations, it is possible to extract the type of movement in which the partner participates (avoidance or approaching). The objective of this feature is to detect when a movement is performed and who initiates it. In other words, if we look only at changes of the distance between the hands of the dyad (**Figure 4B**), we can see that there is some hand activity, but we cannot tell if the variation is due to movement of the parent or the child. To assess who engaged in changes in hand, head or torso distances, we defined a new parameter labeled contribution to the movement. When the distance between two points is tracked, the contribution is defined as the ratio between the velocity amplitude of one point and the sum of the velocity amplitudes of the two points.

This parameter has been computed with the distance between the parent and child heads feature. The results are presented **Figure 8**. At a given time, if the column is completely blue, it means that the current movement is due to the child, and conversely, if it is totally green, the parent is responsible for the movement. Moreover, if the distance (red line) increases, it means that the parent and child move away from each other, and if the distance decreases, they are approaching each other. **Figure 8** shows that in the pathological dyad, the heads were far apart and the child was the leader of the interaction. In contrast, in the control dyad, the heads were close and both the parent and child were the leaders of the changes during interaction, resulting in a motor dialog or movement turn taking. A detailed interpretation of this feature is given in the caption of **Figure 8**.

Sound activity associated with a participant

The sound activity associated with a participant is a feature that parallels the visual modality in the contribution to global movement feature that we described above. In this feature, we combined audio activity with source localization that, in the context of the 3D-reconstruction, determines the speaker.

Table 2 | Proximity and body activity features – results analysis.

	Pathological dyad	Control dyad
Shoulder center distance to the table center	Mother far from the table (average = 80 cm) and is not moving Child near the table (average = 40 cm)	Parent near the table (average = 50 cm)
Distance between hands (please note that partners' asymmetry is also a consequence of the seating position)	Few hand contacts ($N = 20$). Hands are far and distances between parent's left hand-child's right hand (average = 80 cm) and parent's right hand-child's left hand (average = 40 cm) represent partners' asymmetry during the task	Not many hand contacts ($N = 25$). Hands are closer, and more importantly, distances between parent's left hand-child's right hand (average = 60 cm) and parent's right hand-child's left hand (average = 50 cm) break partners' asymmetry
Conclusion	Pathological parent moves less and stays farther from their child than the control parent. Control dyad seems to interact more closely	

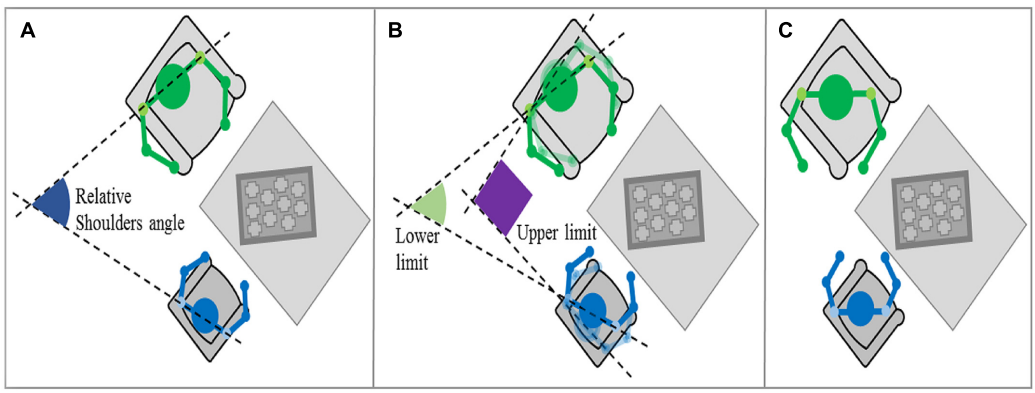


FIGURE 5 | Relative shoulder orientation feature. (A) General case; **(B)** Same point of attention case; **(C)** Face to face case – top view representation.

Figure 9 shows the results and a detailed analysis in the caption. In the pathological dyad, the majority of the sounds were due to the child. In contrast, in the control dyad, both partners contributed to the sound activity, and most importantly, many speech turns occurred, leading to an audio dialog.

DISCUSSION

SUMMARY OF THE RESULTS AND CROSS CORRELATION

We have developed an explorative method to acquire and extract relevant social signals from a naturalistic early parent–infant interaction in a clinical setting. We have extracted various cues from body postures and speech productions of each partner using the IMI2S Framework. Preliminary clinical and computational results for two dyads (one pathological in a situation of severe emotional neglect and one normal control) show that the absence of such interactive social signals indicates behavioral patterns that might be pathologically relevant: the pathological dyad shows dyssynchronic interaction led by the infant whereas the control dyad shows synchronic interaction and a smooth interactive dialog.

The goal oriented aspects (i.e., solving the task) are not affected whereas both the clinical assessment (CIB; **Figure 2**) and the computational feature extraction have revealed clear differences between the pathological and control dyads concerning the body/movement and sound activities of the parent and their involvement in the task and regarding the proximity and joint activity in the dyad. In other words, we can distinguish these two components and provide objective measures for when and how social communication is affected. The pathological parent avoided the activity and the child. This could be interpreted as avoidance of an interaction (Viaux-Savelon et al., 2014), meaning that the parent is less involved in the task and appears to be withdrawn. In contrast, the control dyad was characterized by a clearly distinguishable different dynamic: (1) distances between partners were mediated by movements toward and away from the partner in both the parent and child and (2) the number and regularity of speech turns was high, as in a dialog. These characteristics result in an illustration of synchrony and

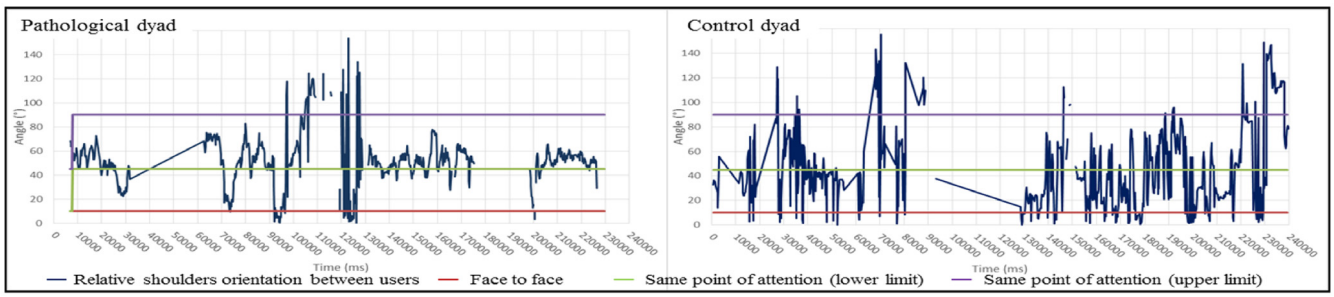


FIGURE 6 | Evolution of relative shoulder orientation during the interaction (Left: pathological dyad; Right: control dyad). In this graph, we report the shoulder orientation according to the relative angle between the two partners' shoulders over time. When the angle is equal to 0°, the partners are facing. When to the angle is 45 to 90°, both shoulders are oriented in the direction of the table that is a point of interest in the given task. In the left graph, the pathological dyad is focused essentially on the task, as partners are facing only three times. In contrast, the control dyad had many face to face positions and showed clear turns between task focusing and other partner focusing. A blank or a cross line in figures indicates uncollected data.

Table 3 | Attention to the task and the partner features – results analysis.

	Pathological dyad	Control dyad
Shoulder orientation	Mother mostly oriented toward table and bench Child focused almost exclusively on the table	Parent moves between table, bench and his child Child moves a lot, focuses on parent, table and bench
Relative shoulder orientation	Dyad focused essentially on the task, just three periods when they are facing	Many face to face interactions Dyad oscillates between task focusing and other partner focusing
Conclusion	Shoulder orientation of the child and parent in the pathological dyad is less mobile than the control dyad. That could be interpreted as a poorer ability to share attention while alternating the focus of attention. The control dyad illustrates a fluid alternation of attention	

engagement switching during harmonious interactions (Delaherche et al., 2012).

The clinical assessment and the computational features do not share the same time scale. By this we mean that the CIB provides a summary of the whole interaction whereas the IMI2S data provides a much a more fine grained scale of the temporal flow. However, we propose the following cross correlation: (i) The "Parental Sensitivity" score of the CIB shows that the parent neglected his child and focused almost entirely on the task in the pathological dyad. CIB "Parental sensitivity" score may be associated with the parent's shoulder distance to the table and the distance between the hands. Indeed, this clinical characteristic

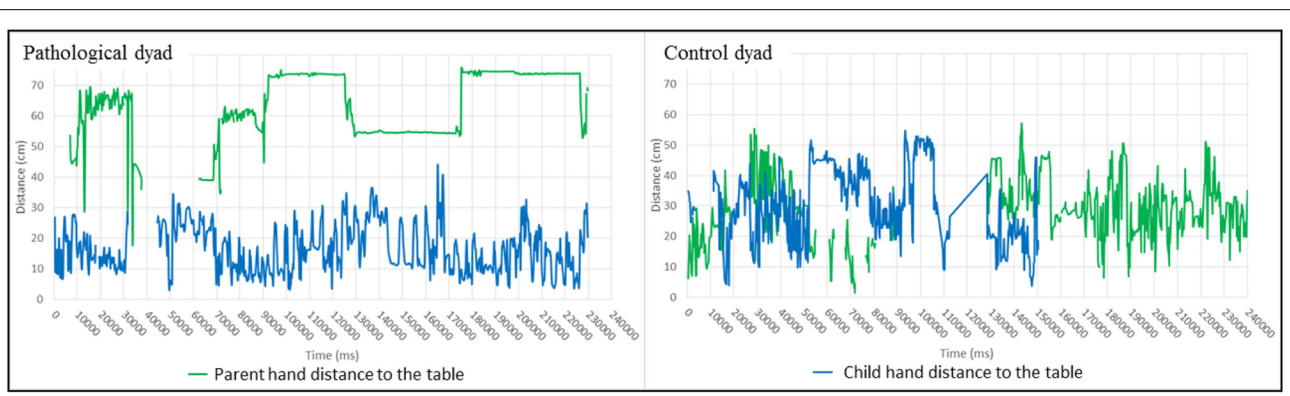


FIGURE 7 | Evolution of the distance between the closest hand and the table center during the interaction (Left: pathological dyad; Right: control dyad). A blank or a cross line in figures indicates uncollected data. By convention, results concerning parents are in green, and results concerning children are in blue.

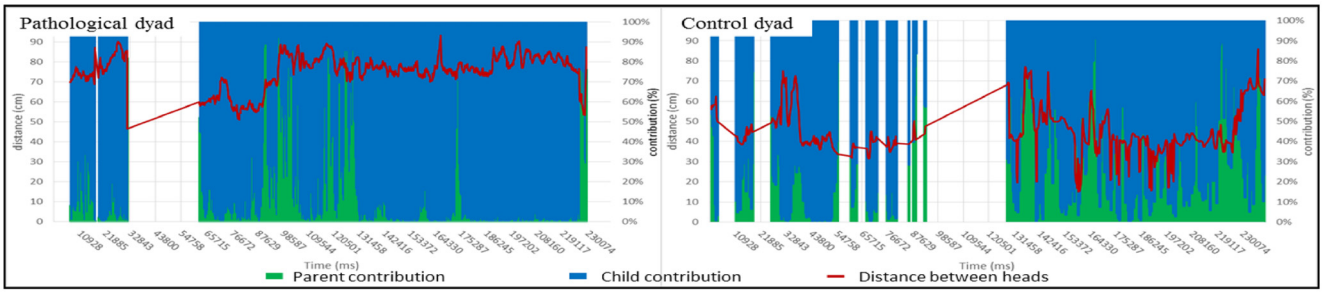


FIGURE 8 | Evolution of the distance between parent and child heads with each partner's contribution to the global hand movement during the interaction (Left: pathological dyad; Right: control dyad). In this graph, we report the distance between the parent's and child's heads with each partner's contribution to the global hand movement during the interaction over time. At the same time, we are able to follow how close or distant partners are and who is moving the most in the previous frames, in other words, who is contributing the most to changing the distance. On the left graph, the pathological dyad showed a large head distance (minimum distance = 50 cm). Movements were initiated mostly by the child, except on two occasions. In contrast, the control dyad showed a smaller head distance (maximum distance = 75 cm). Movement contribution was distributed between the parent and child and the rhythm of the interaction appeared to be a motor dialog with many turns during the course of the interaction. A blank or a cross line in figures indicates uncollected data. By convention, results concerning parents are in green, and results concerning children are in blue.

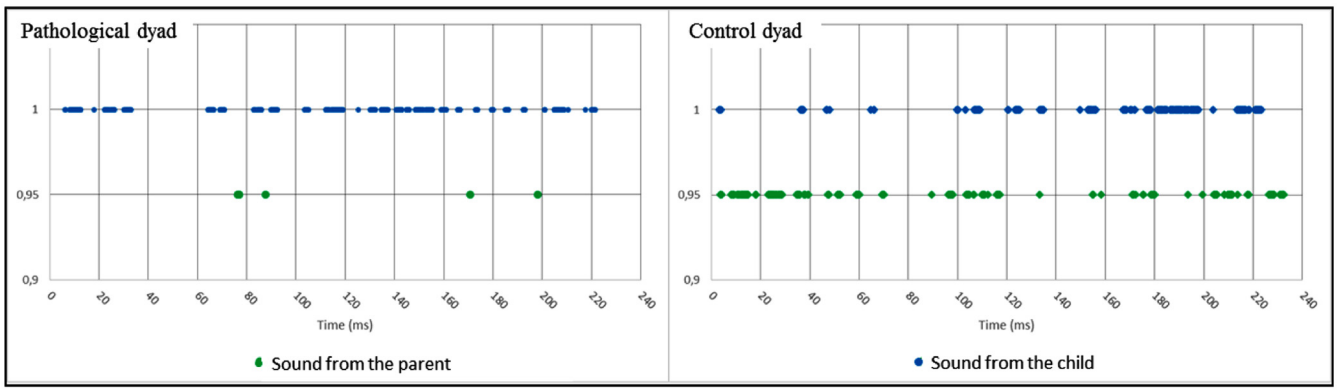


FIGURE 9 | Sound activity by participant during the interaction (Left: pathological dyad; Right: control dyad). In this graph, we combined sound activity and source localization and report sound activity by participant during the interaction over time. In the left graph, the pathological dyad showed a clear disequilibrium. The majority of the sounds were produced by the child. The mother nearly always stayed silent. The dyad only had four speech turns during the entire interaction. In contrast, the control dyad showed no disequilibrium. Sounds were due equally to the child and the parent. Additionally, as in the motor analysis (see Figure), the rhythm resembled a dialog with numerous speech turns. By convention, results concerning parents are in green, and results concerning children are in blue.

could be interpreted as the parent's capacity to remain engaged in the interaction with a proximity adapted to child's movements. (ii) The "Dyadic Reciprocity" score of the CIB clearly distinguishes the two dyads (not much enthusiasm, common involvement, reciprocal affection in the pathological dyad). By definition, a harmonious dyadic reciprocity means smooth and synchronous interaction entailing coordination between partners and intermodality (Feldman, 2007). CIB "Dyadic Reciprocity" may be related to the partners' contributions to movement or speech turns that are equally distributed (**Figures 8** and **9**). (iii) Joint attention (a key item of the CIB "child's engagement" score) can be illustrated by shoulder orientation and relative shoulder orientation (Anzalone et al., 2014a,b). For instance, a parent whose shoulders are oriented toward the same point for a majority of the time (see the pathological dyad in **Figures 4C** and **6**) can reveal a lack of adaptation to the child, preventing the occurrence of joint attention. In contrast, the control dyad showed a large variation of shoulder orientation, which can predict a good adjustment of attention between partners and shared attention (meaning attention of both partners toward a common object) during interactions.

In conclusion, for the current two case reports, computational feature extraction seems to provide the same results as clinical analysis, but allows a finer understanding of interactions by changing the time scale (from a summary of the whole interaction toward a more fine grained scale of the temporal flow) and by providing quantitative features that may be used in large comparison group data or single case longitudinal studies.

LIMITATIONS

Even if the conclusions presented above are promising, the current results are subject to some limitations. First, given the exploratory nature of this study, any generalization of the findings is prevented; only two case-studies are compared, and even if they

are paradigmatic, they cannot be statistically relevant and no statistics was applied. Second, the two cases were not matched for age or gender of the interactive parents but were chosen for their extreme PIRGAS scores. Third, at a group comparison level, it is likely that each pathological dyad would present different patterns of dyssynchrony such as intrusive or under involved styles. In this study, our pathological case was under-involved. Finally, extracted features (skeleton and audio) do not include every facet of the interaction (e.g., motherese). As a consequence, they could not be matched with all the subscores of the CIB.

FUTURE STUDIES

This exploratory study encourages us to pursuing the study of the presented methodology and experimentation in new scenarios. This first work with these two dyads permits us to develop relevant sensor features in a clinical setting and a computational extraction system that can now be tested on a larger population. The next goal will be to accomplish a complete and statistically relevant comparison between the two groups by collecting data from a relevant number of dyads. In our future work, we will be specifically exploring intrusive or under involved parenting because the clinical validity should be tested in these different pathological patterns. We believe that the two features called "evolution of the distance between parent and child heads with each partner's contribution to the global hand movement during the interaction" (**Figure 8**) and "sound activity by participant during the interaction" (**Figure 9**) will be clinically relevant at a group comparison level offering quantitative metrics for under involved parenting. Exploiting low-level signal exchanges allows proposing quantitative metrics without imposing meanings on the signals, which could be not only difficult but also limitative in clinical settings. Various metrics could be investigated ranging from information-based to machine-learning based (Delaherche et al., 2012). Possible metrics could be measuring entropy of individual activities (both infant and caregiver) for individual behavior characterization, mutual information between these activities for inter-personal synchrony characterization. We expect low values of synchrony metrics in pathological dyads whereas it should be higher in harmonious control dyads.

Furthermore, to complete the computational analysis, new features will be implemented in the IMI2S Framework. For example, the video stream recorded with the RGB-D sensor will be analyzed to extract the body activity of each participant or their head orientations (Anzalone et al., 2014a). Additionally, we will include a motherese classifier to better delineate parenting emotional prosody (Cohen et al., 2013). Our future hypothesis would be that these new features will confirm and improve the previous results. In particular, a combination of multimodal features will offer the ability to interpret and understand synchrony and dyssynchrony during early interactions in the context of neglected parenting (Glaser, 2002).

ACKNOWLEDGMENTS

The study was supported by the *Agence Nationale de la Recherche* (ANR-12-SAMA-006), the *Observatoire National de l'Enfance en Danger* and the *Groupement de Recherche en Psychiatrie* (GDR-3557). Sponsors had no involvement in the study design, data analysis, or interpretation of the results.

REFERENCES

Aggarwal, J. K., and Xia, L. (2014). Human activity recognition from 3D data: a review. *Pattern Recognit. Lett.* 48, 70–80. doi: 10.1016/j.patrec.2014.04.011

Altmann, U. (2011). "Investigation of movement synchrony using windowed cross-lagged regression," in *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues, Lecture Notes in Computer Science 6800*, eds A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud, and A. Nijholt (Berlin: Springer), 335–345. doi: 10.1007/978-3-642-25775-9_31

Anzalone, S. M., Tilmont, E., Boucenna, S., Xavier, J., Maharatna, K., Chetouani, M., et al. (2014a). How children with autism spectrum disorder explore the 4-dimension (spatial 3D+time) environment during a joint attention induction task. *Res. Autism Spectr. Disord.* 8, 814–826. doi: 10.1016/j.rasd.2014.03.002

Anzalone, S. M., Avril, M., Salam, H., and Chetouani, M. (2014b). "IMI2S: a lightweight framework for distributed computing," in *Proceedings of the 4th International Conference, Simulation, Modeling, and Programming for Autonomous Robots (SIMPAR)*, Vol. 8810, Bergamo, 267–278. doi: 10.1007/978-3-319-11900-7_23

Batliner, A., Stefan, S., Björn S., Dino, S., Thurid, V., Johannes, W., et al. (2011). Whodunnit – searching for the most important feature types signalling emotion-related user states in speech. *Comput. Speech Lang.* 25, 4–28. doi: 10.1016/j.csl.2009.12.003

Beebe, B., and Lachmann, F. M. (1988). The contribution of mother-infant mutual influence to the origins of self- and object representations. *Psychoanal. Psychol.* 5, 305–337. doi: 10.1037/0736-9735.5.4.305

Berry, T., Koslowski, B., and Main, M. (1974). "The origins of reciprocity: the early mother-infant interaction," in *The Effect of the Infant on Its Caregiver*, Vol. 24, eds M. Lewis and L. A. Rosenblum (Oxford: Wiley-Interscience), 264.

Bilakhia, S., Petridis, S., and Pantic, M. (2013). "Audiovisual detection of behavioural mimicry," in *Proceeding of the Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, Geneva, 123–128.

Bråten, S. (1998). *Intersubjective Communication and Emotion in Early Ontogeny*. Cambridge: Cambridge University Press.

Campbell, N. (2009). *An Audio-Visual Approach to Measuring Discourse Synchrony in Multimodal Conversation Data*. Brighton: ISCA.

Caridakis, G., and Karpouzis, K. (2011). "Full body expressivity analysis in 3D natural interaction: a comparative study, affective interaction in natural environments workshop," in *Proceedings of the ICMI 2011 International Conference on Multimodel Interaction*, Alicante.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., et al. (2006). "The AMI Meeting Corpus: a pre-announcement," in *Machine Learning for Multimodal Interaction, Lecture Notes in Computer Science 3869*, eds S. Renals and S. Bengio (Berlin: Springer).

Chan-Hon-Tong, A., Achard, C., and Lucat, L. (2014). Simultaneous segmentation and classification of human actions in video streams using deeply optimized hough transform. *Pattern Recognit.* 47, 3807–3818. doi: 10.1016/j.patcog.2014.05.010

Cohen, D. (2012). "The developmental being. Modeling a probabilistic approach to child development and psychopathology," in *Brain, Mind and Developmental Psychopathology in Childhood*, eds E. Grralda and J. P. Raynaud (New-York: Jason-Aronson), 3–30.

Cohen, D., Cassel, R. S., Saint-Georges, C., Mahdhaoui, A., Laznik, M.-C., Apicella, F., et al. (2013). Do parentese prosody and fathers' involvement in interacting facilitate social interaction in infants who later develop autism? *PLoS ONE* 8:e61402. doi: 10.1371/journal.pone.0061402

Cohn, J. F. (2010). Advances in behavioral science using automated facial image analysis and synthesis [social sciences]. *IEEE Signal Process. Mag.* 27, 128–133. doi: 10.1109/MSP.2010.938102

Condon, W. S. (1986). "Communication: rhythm and structure," in *Rhythm in Psychological, Linguistic and Musical Processes*, eds J. R. Evans and M. Clynes (Springfield, IL: Charles C Thomas Publisher), 55–78.

Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., and Cohen, D. (2012). Interpersonal Synchrony: a survey of evaluation

methods across disciplines. *IEEE Trans. Affect. Comput.* 3, 349–365. doi: 10.1109/T-AFFC.2012.12

Demouy, J., Plaza, M., Xavier, J., Ringeval, F., Chetouani, M., Périsse, D., et al. (2011). Differential language markers of pathology in autism, pervasive developmental disorder not otherwise specified and specific language impairment. *Res. Autism Spectr. Dis.* 5, 1402–1412. doi: 10.1016/j.rasd.2011.01.026

Dong, W., Lepri, B., Pianesi, F., and Pentland, A. (2013). Modeling functional roles dynamics in small group interactions. *IEEE Trans. Multimed.* 15, 83–95. doi: 10.1109/TMM.2012.2225039

Dumas, G., Nadel, J., Soussignan, R., Martinerie, J., and Garnero, L. (2010). Inter-brain synchronization during social interaction. *PLoS ONE* 5:e12166. doi: 10.1371/journal.pone.0012166

Eyben, F., Wöllmer, M., and Schuller, B. (2010). "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the International Conference on Multimedia, MM '10*, New York, 1459–1462.

Feese, S., Arnrich, B., Troster, G., Meyer, B., and Jonas, K. (2012). "Quantifying behavioral mimicry by automatic detection of nonverbal cues from body motion," in *Proceeding of the Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)*, Amsterdam, 520–525. doi: 10.1109/SocialCom-PASSAT.2012.48

Feldman, R. (2007). Parent–infant synchrony biological foundations and developmental outcomes. *Curr. Dir. Psychol. Sci.* 16, 340–345. doi: 10.1111/j.1467-8721.2007.00532.x

Feldman, R., Greenbaum, C. W., Yirmiya, N., and Mayes, L. C. (1996). Relations between cyclicity and regulation in mother-infant interaction at 3 and 9 months and cognition at 2 years. *J. Appl. Dev. Psychol.* 17, 347–365. doi: 10.1016/S0193-3973(96)90031-3

Fogel, A., Dedo, J. Y., and McEwen, I. (1992). Effect of postural position and reaching on gaze during mother-infant face-to-face interaction. *Infant Behav. Dev.* 15, 231–244. doi: 10.1016/0163-6383(92)80025-P

Galatas, G., Ferdous, S., and Makedon, F. (2013). Multi-modal person localization and emergency detection using the kinect. *Int. J. Adv. Res. Artif. Intell.* 2, 41–46. doi: 10.14569/IJARAI.2013.020106

Gatica-Perez, D. (2009). Automatic nonverbal analysis of social interaction in small groups: a review. *Image Vis. Comput.* 27, 1775–1787. doi: 10.1016/j.imavis.2009.01.004

Glaser, D. (2002). Emotional abuse and neglect (psychological Maltreatment): a conceptual framework. *Child Abuse Negl.* 26, 697–714. doi: 10.1016/S0145-2134(02)00342-3

Harrist, A. W., and Waugh, R. W. (2002). Dyadic synchrony: its structure and function in children's development. *Dev. Rev.* 22, 555–592. doi: 10.1016/S0273-2297(02)00500-2

Hart, S. (2010). *The Impact of Attachment* (Norton Series on Interpersonal Neurobiology). New York: W. W. Norton & Company.

Hogan, P. C. (2011). *The Cambridge Encyclopedia of the Language Sciences*. Cambridge, NY: Cambridge University Press.

Hung, H., and Gatica-Perez, D. (2010). Estimating cohesion in small groups using audio-visual nonverbal behavior. *IEEE Trans. Multimed.* 12, 563–575. doi: 10.1109/TMM.2010.2055233

Keren, M., Feldman, R., and Tyano, S. (2001). Diagnoses and interactive patterns of infants referred to a community-based infant mental health clinic. *J. Am. Acad. Child Adolesc. Psychiatry* 40, 27–35. doi: 10.1097/00004583-200101000-00013

Kuhl, P. K. (2003). Early language acquisition: statistical learning and social learning. *J. Acoust. Soc. Am.* 114, 2445–2445. doi: 10.1121/1.4779344

Lebovici, S. (1985). *Le psychanalyste et l'étude des interactions précoces.* [The psychoanalyst and the study of early interactions.]. *Rev. Fr. Psychanal.* 49, 1307–1329.

Leclère, C., Viaux, S., Avril, M., Achard, C., Chetouani, M., Missonnier, S., et al. (2014). Why synchrony matters during mother-child interactions: a systematic review. *PLoS ONE* 9:e113571. doi: 10.1371/journal.pone.0113571

Marcos-Ramiro, A., Pizarro-Perez, D., Marron-Romera, M., Nguyen, L., and Gatica-Perez, D. (2013). "Body communicative cue extraction for conversational analysis," in *Proceeding of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2013*, Shanghai, 1–8. doi: 10.1109/FG.20120133.6553741

McColl, D., and Nejat, G. (2014). Recognizing emotional body language displayed by a human-like social robot. *Int. J. Soc. Robot.* 6, 261–280. doi: 10.1007/s12369-013-0226-7

Meltzoff, A. N., Kuhl, P. K., Movellan, J., and Sejnowski, T. J. (2009). Foundations for a new science of learning. *Science* 325, 284–288. doi: 10.1126/science.1175626

Paxton, A., and Dale, R. (2014). *Multimodal Networks of Interpersonal Interaction and Conversational Contexts*. Merced: University of California.

Prepin, K., and Pelachaud, C. (2011). "Shared understanding and synchrony emergence synchrony as an indice of the exchange of meaning between dialog partners," in *Proceeding of the International Conference on Agent and Artificial Intelligence (ICAART)*, Vol. 2, Rome, 25–34.

Ramanathan, V., Yao, B., and Fei-Fei, L. (2013). "Social role discovery in human events," in *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013*, Washington, DC, 2475–2482. doi: 10.1109/CVPR.2013.320

Ramseyer, F., and Tschacher, W. (2011). Nonverbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome. *J. Consult. Clin. Psychol.* 79, 284–295. doi: 10.1037/a0023419

Rees, C. (2008). The influence of emotional neglect on development. *Paediatr. Child Health* 18, 527–534. doi: 10.1016/j.paed.2008.09.003

Rehg, J. M., Abowd, G. D., Rozga, A., Romero, M., Clements, M. A., Sclaroff, S., et al. (2013). "Decoding children's social behavior," in *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013*, Portland, OR. 3414–3421. doi: 10.1109/CVPR.2013.438

Ringeval, F., Demouy, J., Szaszak, G., Chetouani, M., Robel, L., Xavier, J., et al. (2011). Automatic intonation recognition for the prosodic assessment of language-impaired children. *IEEE Trans. Audio Speech Lang. Process.* 19, 1328–1342. doi: 10.1109/TASL.2010.2090147

Rolf, M., and Asada, M. (2014). *Visual Attention by Audiovisual Signal-Level Synchrony*. Available at: http://www.er.ams.eng.osaka-u.ac.jp/Paper/2014/Rolf14a.pdf

Saffran, J. R. (2003). Statistical language learning mechanisms and constraints. *Curr. Dir. Psychol. Sci.* 12, 110–114. doi: 10.1111/1467-8721.01243

Saint-Georges, C., Chetouani, M., Cassel, R., Apicella, F., Mahdhaoui, A., Muratori, F., et al. (2013). Motherese in interaction: at the cross-road of emotion and cognition? (A systematic review). *PLoS ONE* 8:e78103. doi: 10.1371/journal.pone.0078103

Salamin, H., Favre, S., and Vinciarelli, A. (2009). Automatic role recognition in multiparty recordings: using social affiliation networks for feature extraction. *IEEE Trans. Multimed.* 11, 1373–1380. doi: 10.1109/TMM.2009.2030740

Sameroff, A. (2009). *The Transactional Model of Development: How Children and Contexts Shape Each Other*. Washington, DC: American Psychological Association.

Sanchez-Cortes, D., Aran, O., Jayagopi, D. B., Mast, M. S., and Gatica-Perez, D. (2013). Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition. *J. Multimodal User Interfaces* 7, 39–53. doi: 10.1007/s12193-012-0101-0

Schuller, B., Vlasenko, B., Eyben, F., Wollmer, M., Stuhlsatz, A., Wendemuth, A., et al. (2010). Cross-corpus acoustic emotion recognition: variances and strategies. *IEEE Trans. Affect. Comput.* 1, 119–131. doi: 10.1109/T-AFFC.2010.8

Stern, D. N. (2009). *The First Relationship: Infant and Mother*. Cambridge: Harvard University Press.

Tronick, E. Z., and Cohn, J. F. (1989). Infant-mother face-to-face interaction: age and gender differences in coordination and the occurrence of miscoordination. *Child Dev.* 60, 85. doi: 10.2307/1131074

Vandenberg, K. A. (2006). *Maternal-Infant Interaction Synchrony Between Very Low Birth Weight Premature Infants and Their Mothers in the Intensive Care Nursery*. Ann Arbor, MI: ProQuest Information & Learning Edition.

Varni, G., Camurri, A., Coletta, P., and Volpe, G. (2009). "Toward a real-time automated measure of empathy and dominance," in *Proceeding of the International Conference on Computational Science and Engineering, 2009. CSE '09*, Vancouver, BC. 4, 843–848. doi: 10.1109/CSE.2009.230

Viaux-Savelon, S., Leclere, C., Aidane, E., Bodeau, N., Camon-Senechal, L., Vatageot, S., et al. (2014). Validation de la version française du Coding Interactive Behavior sur une population d'enfants à la naissance et à 2 mois. *Neuropsychiatr. Enfance Adolesc.* 62, 53–60. doi: 10.1016/j.neurenf.2013.11.010

Vinciarelli, A., Pantic, M., and Bourlard, H. (2009). Social signal processing: survey of an emerging domain. *Image Vis. Comput.* 27, 1743–1759. doi: 10.1016/j.imavis.2008.11.007

Weisman, O., Zagoory-Sharon, O., and Feldman, R. (2012). Oxytocin administration to parent enhances infant physiological and behavioral readiness for social engagement. *Biol. Psychiatry* 72, 982–989. doi: 10.1016/j.biopsych.2012.06.011

Weisman, O., Zagoory-Sharon, O., Schneiderman, I., Gordon, I., and Feldman, R. (2013). Plasma oxytocin distributions in a large cohort of women and men and their gender-specific associations with anxiety. *Psychoneuroendocrinology* 38, 694–701. doi: 10.1016/j.psyneuen.2012.08.011

Yun, K., Honorio, J., Chattopadhyay, D., Berg, T. L., and Samaras, D. (2012). "Two-person interaction detection using body-pose features and multiple instance learning," in *Proceeding of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* Providence, 28–35. doi: 10.1109/CVPRW.2012.6239234

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 16 October 2014; accepted: 24 November 2014; published online: 10 December 2014.

Citation: Avril M, Leclère C, Viaux S, Michelet S, Achard C, Missonnier S, Keren M, Cohen D and Chetouani M (2014) Social signal processing for studying parent–infant interaction. Front. Psychol. ***5****:1437. doi: 10.3389/fpsyg.2014.01437*

This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read, for greatest visibility

COLLABORATIVE PEER-REVIEW
Designed to be rigorous – yet also collaborative, fair and constructive

FAST PUBLICATION
Average 85 days from submission to publication (across all journals)

COPYRIGHT TO AUTHORS
No limit to article distribution and re-use

TRANSPARENT
Editors and reviewers acknowledged by name on published articles

SUPPORT
By our Swiss-based editorial team

IMPACT METRICS
Advanced metrics track your article's impact

GLOBAL SPREAD
5'100'000+ monthly article views and downloads

LOOP RESEARCH NETWORK
Our network increases readership for your article

Frontiers
EPFL Innovation Park, Building I • 1015 Lausanne • Switzerland
Tel +41 21 510 17 00 • Fax +41 21 510 17 01 • info@frontiersin.org
www.frontiersin.org

Find us on