Collaborative User Responses in Multiparty Interaction with a Couples Counselor Robot

Dina Utami
College of Computer and Information Science
Northeastern University
Boston, USA
dinau@ccs.neu.edu

Timothy Bickmore
College of Computer and Information Science
Northeastern University
Boston, USA
bickmore@ccs.neu.edu

***Abstract*—Intimate relationships are integral parts of human societies, yet many relationships are in distress. Couples counseling has been shown to be effective in preventing and alleviating relationship distress, yet many couples do not seek professional help, due to cost, logistic, and discomfort in disclosing private problems. In this paper, we describe our efforts towards the development a fully automated couples counselor robot, and focus specifically on the problem of identifying and processing "collaborative responses", in which a human couple co-construct a response to a query from the robot. We present an analysis of collaborative responses obtained from a pilot study, then develop a data-driven model to detect end of collaborative responses for regulating turn taking during a counseling session. Our model uses a combination of multimodal features, and achieves an offline weighted F-score of 0.81. Finally, we present findings from a quasi-experimental study with a robot facilitating a counseling session to promote intimacy with romantic couples. Our findings suggest that the session improves couples intimacy and positive affect. An online evaluation of the end-of-collaborative-response model demonstrates an F-score of 0.72.**

Keywords— Human-Robot Interaction, User Studies, Multiparty Interaction, Collaborative Responses, Turn Taking

I. Introduction

As robots become increasingly ubiquitous in our social world, there is increasing need for them to interact with multiple people simultaneously. Romantic couples counseling represents an ideal research testbed for robot multiparty interaction research, since it requires the recognition and production of many complex and subtle linguistic, interactional, and multimodal behaviors, while tightly constraining the number of users and the spatial configuration of the users relative to the robot. Automated couples counseling is also an important research endeavor in its own right: relationship distress and divorce in Western societies are very common, yet most people who suffer from relationship distress do not seek help due to a variety of barriers that automated counselors may be able to overcome [1] [2].

Collaborative responses, in which one person finishes another's sentence or corrects or elaborates the other's response, is common when dyads who share significant knowledge—such as romantic couples—are asked a question about their shared experience. The recognition and understanding of such collaborative responses have been understudied in the multiparty interaction, but are important to address in the development of systems that interact with users who know each other well. The processing of these responses is challenging at both the linguistic and interactional levels. Techniques from incremental speech processing may provide a promising approach to understanding these co-constructed responses. However, before linguistic analysis can even be addressed, a prerequisite interactional problem must be solved, namely the identification of when a collaborative responses has ended. While several researchers have investigated multimodal approaches to identifying end-of-turn for single users, identifying end-of-turn for collaborative responses from multiple users has not been investigated to date.

In this paper we report the ongoing development and evaluation of a couples counselor robot, and focus on the development and evaluation of a data-driven approach to identifying the end of collaborative responses by a romantic couple to questions by the robot during a counseling session. We first review relevant prior work, then describe our research platform for couples counseling. We then present a pilot study used to gather data for model building, followed by a description of our model-building approach and offline evaluation. Finally, we present the results of a validation study in which the end-of-collaborative-turn model is used in real time during couples counseling, with its outputs compared to a human judge. We also report the impact of the robot-driven couples counseling interaction on couples' intimacy.

II. Related Work

Romantic couples interact with each other and third parties in variety of complex ways, including collaborative responses to queries. We briefly review communication and sociolinguistic studies of interactive behavior between couples and on collaborative responses in conversation, as well as research on multiparty interaction with robots.

A. Communicative Behaviors in Interpersonal Relationship

There are many verbal and nonverbal behaviors exhibited by intimate couples and close friends that are associated with various dimensions of interpersonal relationships. Perhaps the most studied dimension of interpersonal communication is immediacy (described as intimacy, warmth, and closeness), which is indicated by various signals such as close proximity, touch, forward-leaning, eye contact and gaze, and use of verbal backchannels [3]. Goffman studied ways in which people conduct their relations in public and described several behaviors that may indicate that pairs of persons are "with" each other or "in a relationship". For example, he suggests that hand-holding is a "tie-sign" which "*contains evidence about their relationship*" and that *"these tie-signs not only inform that the relation is anchored, but provide some information about its name, its terms, and its stage"* [4]. Mandelbaum suggests that the appearance of "withs" can be also observed during conversation. She conducted a conversation analysis of couples sharing stories about events in which they participated together and, through this, produce the appearance of being "with" each other. In the stories examined, the couples shared the beginning of a telling with one participant first using an indirect bid to start a collaborative story (a "remote" move), which the other then exhibits an uptake on (a "forward" move), followed by a "ratification" by the first partner. Once this has occurred, the shared story can be told. Mandelbaum also observed frequent instances of collaborative corrections, requests for verification, and complementary telling [5].

Outside of intimate relationships, differences in communicative behaviors have also been observed in conversation between strangers, acquaintances, and friends. Planalp and Benson demonstrated that observers are able to discriminate audiotaped conversations between friends and acquaintances based on the number of references to mutual knowledge and continuity, the number of interruptions, and the distribution of floor time [6]. Tickle-Degnen and Rosenthal propose a model of rapport that deepens over time, and consists of three components, each with associated nonverbal conversational behaviors: mutual attentiveness, or perceived interest by the other; positivity, or mutual friendliness and caring; and coordination in interaction. The relative importance of these components is predicted to vary throughout the course of a relationship, with coordination increasing and positivity decreasing [7]. Taking this theory as inspiration, Cassell et al. found that when strangers are giving directions, they used more explicit acknowledgment and used more nonverbal behavior related to coordination (e.g. head nods and mutual gaze) than when friends are giving directions [8]. Schulman and Bickmore demonstrated that conversational behaviors change over time as a function of interaction history (e.g. frequency, pattern, purpose) as the interactants' interpersonal relationship (e.g. trust, intimacy, working alliance) change. In a longitudinal study of weekly face-to-face conversations between a certified exercise trainer and clients, they found that participants had faster articulation rates on discourse markers, used fewer posture shifts, gaze away more during speech, and smile and frown less in later conversations compared to initial interactions [9]. Finally, several autonomous systems have been developed to quantify the level of rapport based on multimodal behaviors of the speakers and listener [10, 11].

B. Collaborative Production in Conversation

In multiparty interactions, commonly, conversational participants change roles from speakers, to addressees, and listeners. However, when multiple speakers share common ground, they can also collaboratively produce a response during a single speaking turn. Several terms have been used to describe this phenomena including "collaboratively built sentences", "sentences-in-progress", "joint production", "co-constructions", and "conversational duet" [12]. The types of collaborative production we have observed in our couples counseling research are closest to the notion of "conversational duet" introduced by Falk in 1980. A "conversational duet" is a multiparty conversation in which *"two or more persons may participate as though they were one, by talking to an audience in tandem for both (or sometimes one) of them about the same thing, with the same communicative goal"*. A duet may occur under the following set of conditions: 1) the speakers have mutual knowledge and are equally competent to talk about the topic at hand; 2) the speakers share similar communicative goals; 3) the speakers are addressing a mutual third party (not each other); and, 4) the speakers intend that each of their contributions counts on both of their behalf. According to Falk, a dueter's speaking turn consists of multiple sub-turns. One reason why Falk considers duetters contributions as sub-turns is that they are not treated as interruptions, and duetters are often treated as if they were one speaker by their audience [13].

Joint production commonly occurs in a conversation between people who are deeply empathic with each other, such as people who are *"engaged in long-standing relationships or close working conditions, are in frequent proximity on a regular basis, or otherwise, establish rapport"* [14]. Several explanations are offered to answer the question of why joint production occurs. Sacks sees the main function of collaborative production as social: *"The fact that there is a job that any person could clearly do by themselves, provides a resource for members for permitting them to show each other that whatever it is they're doing together, they're just doing together to do together"* [15]. Ferrara classified four types of joint productions based on their communicative function: utterance extensions, helpful utterance completions, predictable utterance completions, and invited utterance completions. Additionally, in therapeutic discourse, joint production is often used to build rapport and show empathy [14].

C. Turn-taking with Conversational Robots and Agents

Conversational turn-taking is a complex multi-modal process in which myriad cues, including gaze, speech, hand gesture, and prosody, are used to coordinate the behavior of speakers [16-19] so that they do not speak at the same time. One of the general hypotheses from this work is that the overall strength of a turn-taking signal is a function of the number of cues involved [20]. Several computational models have been developed to enable humans to engage in conversational turn-taking with dialogue systems, virtual agents and robots, using automated recognition and production of these cues. Some researchers have explored specific cues associated with particular aspects of turn-taking, including identification of the end of a user's turn [21], or when an agent should take or give the turn [22-25].

Within the dialog system community, several researchers have developed computational models of turn-taking that enable

users to engage virtual agents in real-time dyadic face-to-face conversation [26-29]. For example, Raux, et al. introduced a non-deterministic finite state models of turn taking that uses a cost matrix and decision theoretic principles to select turn-taking actions [26]. Selfridge, at al. presented a model that treats turn-taking as a negotiation process, learns the importance of turn-taking behaviors through reinforcement learning, and supports mixed-initiative interactions [29].

D. Turn-taking in Multiparty Interaction

Conversational human-robot turn-taking is particularly challenging in a multiparty setting. The presence of more than one interlocutor means that the robot must be able to identify which interlocutor takes the turn when a turn is yielded, and to differentiate when a turn is yielded to the robot or to another interlocutor [30]. One signal that has been shown to be useful in detecting whether or not a robot was addressed by a human speaker is gaze [31]. However, because gaze detection is often difficult in conversational settings, many developers use head pose as a surrogate for gaze direction and have shown it to be a reliable approach [32]. Other features that have been used to detect when a turn is yielded to a robot includes lexical features and prosody (speech rhythm and vocal effort) [33].

Other than addressee detection, another challenging problem in turn-taking is optimizing the timing of system responses. The simplest approach to detect end-of-turn is to use a fixed silence threshold, however, when the threshold is set too high, response latency increases. Conversely, setting the threshold too low will increase interruption by the robot. To overcome this problem, several models based on data in human-human interaction [34, 35] or human-machine interaction [21, 36, 37] have been proposed to guide dialogue systems make turn-taking decision. For example, Raux et al developed an algorithm to dynamically set the endpointing threshold to detect the end of user utterance in a dyadic interaction. The model uses several features from discourse, semantics, prosody, and timing to detect whether a silence indicates the end of turn. The proposed method reduces latency by up to 24% over a fixed threshold baseline [21]. In multiparty interaction setting, Skantze et al. developed a data-driven model for robot participating in a multiparty card sorting game with multiple users to make turn-taking decisions. Using multimodal features, including head pose, speech, and card movements, they were able to detect the timing for robot responses with a weighted F-score of 0.88 [38].

III. Research Platform

We have developed a common testbed for our research studies on robot-driven couples counseling. The robotic counselor is a humanoid head developed by Furhat robotics (Fig.1). It has an animated face, back-projected on a translucent mask, mounted on a two degree-of-freedom mechanical neck that allows it to direct its attention using eye gaze and head pose [39].The robot's speech is generated using the Cereproc speech synthesizer and lip movement is synchronized using viseme callbacks from the text-to-speech engine. Eyebrow raises (for emphasis) and gaze toward/away from users (for turn taking) are generated using BEAT [40]. Dialogue is modeled in hierarchical transition networks using a state chart-based XML formalism. Our long-term objective is to develop a fully-automated system. However, currently the rest of the robot's behavior (e.g. orienting head pose towards speaker, natural language understanding, and updating system state after detecting end of turn) is driven by a research assistant in a Wizard of Oz framework. Participants are seated in fixed chairs in front of the robot and wear headset microphones. The session is video recorded from different angles using four cameras, and a Microsoft Kinect is used for recording the couple.

Screening for a history of domestic violence and an IRB-approved safety protocol minimizes the risk of threats, abuse, or violence during sessions, and well-defined procedures should they occur.

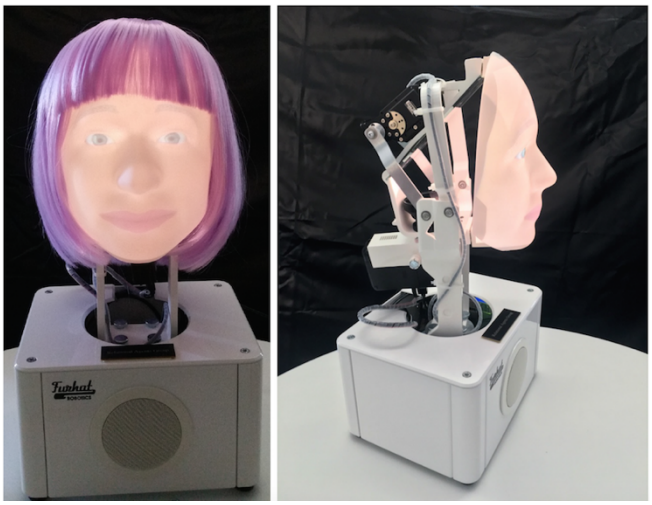

Fig. 1. Robotic Counselor

IV. Data Collection Study

To develop a data-driven model for identifying the end of collaborative responses in couples counseling, we analyzed and annotated data from a pilot study using our research platform, in which we had 16 couples interact with the robot. We reported user's acceptance and satisfaction of the system from self-reported questionnaire in [41]. However, data regarding participant's interactional behavior has not been published. In this interaction, the robot introduced two relationship communication skills: active listening (a method to understand a speaker's message accurately and to communicate that understanding to the speaker) and effective speaking (a method to communicate feelings and needs in a clear, non-judgmental manner). For each of these skills, the robot began by describing the skills and then asked the couple to role-play these skills with the robot and with each other. In the pilot study, all sessions were videotaped, with one session dropped due to technical error, which left us with 15 sessions to analyze.

A. Types of Collaborative Responses

We analyzed videotapes of our robotic counselor interviewing 15 couples about the history of their relationships to understand the types of collaboration occurring during these responses. All couples (15) answered all four of the robot's open-ended rapport-building questions, yielding 60 responses to analyze. Of the 60 responses, 57.5% were produced jointly. In general, we observed two kinds of joint responses: invited collaboration (33.3%) and spontaneous collaboration (66.6%) (Table I).

TABLE I. TYPES OF COLLABORATIVE RESPONSE

(a) Invited collaboration		
Robot	:	And, when was that?
C1,P2	:	Um it's well I wanna say [gaze towards partner] a year and three quarters [raised pitch].
C1,P1	:	Yeah. Yeah, February twenty fif teen. Twenty yeah February twenty fifteen. Something like that
Robot	:	Can you tell me what you did on your first date?
C3, P1	:	What was our first date?
C3, P2	:	It's a good question.
C3	:	We um I think we knew each other for a long time before we went on what you might call a first date. Um. We have many shared interest.
(b) Forwarding		
Robot	:	So, how did you two meet?
C1, P2	:	Well, it's kind of a long storyline. [gaze towards partner] Do you wanna have a go with that one?
C1, P1	:	It's so complicated but basically we were friends online before we had a chance to encounter at a coffee shop and then we ended up talking online and then scheduling to meet officially and then it went from there.
(c) Sentence Completion & Extension		
Robot	:	So, how did you two meet?
C9, P1	:	I was hitchhiking um in Brookline
C9, P2	:	Brookline
C9, P1	:	and Steve picked me up
C9, P2	:	on Harvard street in front of what used to be the um they have frivolous Friendly's
C9, P1	:	Friendly's
C9, P2	:	and I picked her up. I thought she'd be easy. I was wrong.
(d) Corral Response		
Robot	:	So, the two of you have been together for a while then?
C15, P1	:	Yes!
C15, P2	:	That's it.

In an invited collaboration, one person elicits a response from the other. One strategy used by a first speaker is to make a question or a statement (complete or incomplete) followed by a gaze change towards the partner (second speaker) [Table 1a]. Frequently, the statements end with a word stretch (syllable elongation) and a slight change in pitch. Essentially, these utterances are questions masquerading as statements [14]. Not all invitations for collaboration are followed by uptakes. When the elicitation fails, the first speaker directs his/her gaze back to the robot and completes his/her own utterance, if it is not complete. Another type of collaboration that we observed is what Madelbaum described as "forwarding" [5]. In this class of collaborative response, one partner answers the robot's question and then directs it to his/her partner [Table 1b]

In a spontaneous collaboration, the second person contributes without an explicit elicitation or request from the first. One common type of spontaneous collaboration that we observed is sentence completion. This happens when a speaker has trouble recalling a fact and generates a long pause mid-sentence (word search). The partner helps complete the sentence by filling in the blank (Table 1c). On several occasions, even when a speaker utterance is complete, a second speaker was observed to join in by extending the first speaker's utterance to add more information [Table 1c]. Another type of spontaneous joint response we observed is a "corral response", where both partners answers the robot's questions almost simultaneously [Table 1d].

V. DATA DRIVEN MODEL FOR DETECTING END OF COLLABORATIVE RESPONSES

A. Data Annotation

Our end goal is to build a supervised machine learning model to detect when couples have finished their collaborative speaking turn and the robot can take the floor. The findings from our corpus analysis suggest that to do so, the system essentially needs to perform three tasks: detect the end of the speakers' utterance, detecting whether the utterance is addressed to the robot or not, and predicting if a second speaker will barge-in.

Our approach is similar to previous studies on turn taking mentioned above where Inter-Pausal Units (IPUs) with a duration of 200ms or higher are treated as possible end of turns. Audio recordings from the 120 responses in our pilot study were automatically segmented and manually corrected. Each speech segment between IPUs were manually transcribed and separated into different layers, each for one interactant. Speech prosody was automatically extracted using Praat [42].

To provide ground truth for building a supervised model, a set of 226 decision points from the 15 dialogues were manually annotated for end of collaborative turn by two annotators. The annotator made a binary decision: "Yes" and "No". The first annotation was based on the Wizard's turn taking decision executed in real-time during a counseling session. Decision points that are closest to the wizard's action are labeled as "Yes" while the rest are labeled as "No". The second annotation was performed by two research assistants who watched the interaction video up to each decision point and decided whether it was an end of turn or not. The inter-rater reliability between the two annotators was 0.84, indicating good agreement. Additionally, we coded participants' gaze behavior from the video as being directed at the robot vs. away from the robot. All annotations were performed using ANVIL [43].

TABLE II. DATA SUMMARY

Number of participants	15 couples
Number of question-response pair	60
Average duration	1 minutes 7 seconds
Number of decisions	226

B. Baseline

The dataset is imbalanced with 73.4 % labeled as "No". The majority-class baseline yields a weighted F-score of 0.62.

C. Features

We computed a total of 16 features in five categories: voice activity, gaze, syntax, context, and prosody.

1) Voice Activity Features: Since there are two possible possible speakers at any time, one basic feature is voice activity from either interactant. Note that in the case of speech overlap, we include points where one speaker has stop speaking but the other has not. In our data, all (58) decision points where anyone is still speaking is labeled as "No", suggesting that the system should not take a turn at that point.

2) Gaze Features: Several studies have demonstrated that gaze is an essential cue for turn taking. Speakers use gaze for turn taking, turn yielding and selecting the next speaker. Thus, we expect gaze to be an important feature for our model. Our gaze features includes last speaker's gaze, interlocutor's gaze, and joint head pose. We found a significant difference in participants' gaze target depending on their conversational roles ($x^2(2)=7.03$, $p<0.01$). The joint head pose feature describes whether any or all participants are looking at the robot, and has values of "neither", "one-interactant", and "both-interactants". Additionally, gaze has also been correlated with intimate behavior, and past research indicates there may be a subtle gender difference [3]. In our corpus, we found that female participants gaze at their partner more than male participants, while they were speaking or listening ($x^2(2)=104.8$, $p<0.01$). Thus, we also includes female participant gaze and male participant gaze in our gaze feature set. We expect that the female participant gaze will have less discriminating power than the male participant gaze.

3) Syntax Features: To achieve joint production, interlocutors must pay close attention to the ongoing syntactic structured of the current utterance under construction for them to be able to interject their own words [14]. In our corpus, all sentence completions occur in mid sentence, thus, sentence incompleteness may indicate a possible location for interlocutors to barge-in. Syntax incompleteness may also suggest that a speaker has not finished his/her thought, thus, a pause may indicate hesitation rather than end of turn. Part-of-speech (POS) has been shown to be a useful feature in detecting end of turn [34, 36]. Our syntax features include last word POS, second last word POS, and last two word POS. These features are automatically extracted using the Stanford POS tagger [44].

4) Context Features: Our context features include the robot's previous dialog act, the number of spoken words in current IPUs, the number of spoken words since robot's last utterance, and the number of pauses since the robot's last utterance. The length of couples' responses differ depending on the robot's dialogue act. For example, story elicitation (e.g., *"How did you two meet?"*) generated responses with an average length of 19.8 words, while a declarative question (e.g., *"So, the two of you have been together for a while then?"*) generated responses with an average length of 3.3 words. There are also many more pauses in user responses to story elicitation than to other types of query. The more pauses and the longer the utterances the more likely a pause indicates an end of turn. *"For dueters, whenever one partner speaks, the other can and often does speak next."* [13]. Therefore, we included another feature that indicates whether both partners have produced an utterance since the robot's last utterance.

5) Prosodic Features: As prosodic features, we use mean pitch and pitch slope of the final 200ms voiced region extracted using Praat [42]. Pitch was sampled at 100ms to estimate F_0. Since we are interested in tracking prosody change (not the absolute value), the values were then log transformed and z-normalized for each participant.

D. Offline Evaluation

We evaluated the contribution of each feature category individually, as well as in combination, in detecting end of collaborative turn. We evaluated Bayes Net, Support Vector Machine (SVM), and Random Forest algorithms in the Weka toolkit [45]. All numeric features were standardized for scaling. All evaluations were performed using 10-fold cross-validation.

Results: Using gaze features alone can improve the weighted F-score to 0.77 (Table III). Ranking of attribute importance, based on impurity decrease and the number of nodes in the random forest tree using that attribute, for gaze features from highest to lowest was: joint gaze (0.13), male gaze (0.11), last speaker gaze (0.1), interlocutor gaze (0.08), and female gaze (0.07). As we predicted, the female gaze features ranked lower than male gaze features. Another feature that results in higher F-score when used individually was the context features, with an F-score up to 0.70.

TABLE III. MODEL EVALUATION USING F-SCORE METRIC

Features	Bayes Net	SVM	Random Forest
Voice Activity	0.62	0.62	0.62
Gaze	0.77	0.77	0.78
Syntax	0.62	0.62	0.62
Context	0.69	0.70	0.70
Prosody	0.62	0.62	0.62
Context + Syntax	0.70	0.65	0.65
Context + Gaze	0.78	0.74	0.70
Gaze + Prosody	0.78	0.78	0.75
Prosody + Gaze + Context	0.78	0.78	0.72
All features	0.81	0.81	0.74

The syntax features did not improve performance when used in isolation. One possible explanation is that we have many examples in our corpus in which speakers' utterances were complete but were addressed to their partner or addressed to the robot but were extended by their partner. Therefore, while syntax completeness may indicate the end of an individual turn (sub-turn), it may not be the end of a collaborative turn (compound turn). However, used in combination with other features, they can increase the overall F-score.

Using all features combined, the SVM and the Bayes Net classifier provided the best performance, with a weighted F-score of 0.81.

VI. ONLINE VALIDATION STUDY

To validate our end-of-collaborative-response model in a real-time setting, and to further our exploration of robot-driven couples counseling, we conducted another quasi-experimental

study using our research platform. In this study, the model was evaluated during the rapport-building phase of the session, using the same four elicitation questions by the robot used in the previous study. In addition, the counseling part of the interaction was revised with intimacy building exercises led by a robot. The study was approved by the University IRB.

A. Procedure

Participants were recruited from an online job-posting website and our university portal, and were required to be at least 18 years old, able to speak and read English, and have been romantically involved with their current partner for at least two years.

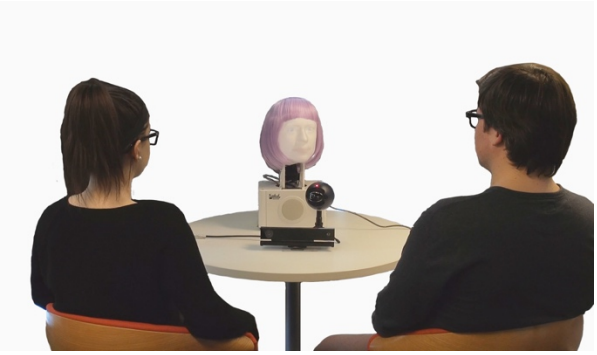

Fig. 2. A Couple Interacting with the Robot Counselor

After couples were consented, they were brought to separate rooms to fill out a baseline questionnaire, which includes the Conflict Tactic Questionnaire (CTS) to screen out couples with any history of domestic violence [46].

Eligible couples proceeded to have a 30-minute counseling session with the robotic counselor. Couples are seated in chairs in front of the robot that they can interact with the robot and each other (Fig. 2). Following the rapport-building questions, two positive relationship techniques are introduced: a gratitude exercise and "caring days". During the gratitude exercise, couples were asked to recall and share three recent positive behaviors of their partner. Caring Days is techniques used in Behavioral Couples Therapy (BCT) that has been shown to be effective [47]. Each person is asked to think of a request for a behavior that their partner can perform to show that they care. The behavior must be positive, specific, manageable, and not related to a topic of recent conflict. For each exercise, the robot first explains the rationale for the exercise, asks the couple to practice the exercise, and then provides feedback. The session ends with a summary and contract (verbal commitment from each partner) to perform the behavior outside of the session.

Following the session, couples filled out self-report outcome measures. A recorded semi-structured interview was then conducted to ask them about their experience.

B. Measures

In addition to socio-demographic measures, the following measures were collected at baseline (T0) and immediately after the session (T1).

- **Positive and Negative Affect Scale (PANAS)**, collected at T0 and T1 to assess changes in emotional state during the session. The scale ranges from 1 (not at all) to 5 (extremely) [48].
- **Inclusion of Other in the Self (IOS)**, collected at T0 and T1, to assess changes in couples' perceived interpersonal closeness during the session. It is a seven-point graphic (visual analogue) scale [49].
- **Closeness and Intimacy**, collected at T0 and T1, to assess changes in current feelings of intimacy using a semantic differential scale. Ten adjective-pairs are listed in opposition (e.g. estranged-intimate, distant-close) and rated on a 7-point scale [50].
- **Attitudes Toward Robotic Facilitator**, collected at T1 (Table IV). The scale ranges from 1 (not at all) to 7 (very much).
- **Enjoyment of the interaction**, collected at T1, to assess the enjoyment of the interaction, with a four item composite scale, adapted from [51]. The scale ranges from 1 (not at all) to 7 (a great deal).
- **Perceived partner's responsiveness**, collected at T1, to assess partner's responsiveness during interaction with a four item composite scale ("My partner seemed to really listen to me.", "My partner seemed interested in what I am thinking and feeling.","My partner was on 'the same wavelength with me'.", "My partner was responsive to my questions/answers.",on a scale of 1:Not true at all to 7:Very true).

C. Results

We recruited 24 volunteers (12 couples) for the study, of which 2 couples were screened out because of a history of physical abuse, which left us with 10 couples completing the study. The average age of the 20 participants was 29 years old (range 19 to 61). All but one couple were heterosexuals and most (9) had never participated in a couples counseling session. Of the 10 couples, 6 were seriously dating (do not date other people), 2 were in cohabiting relationships, and 2 were married and were living together. The length of the couples' relationships ranged from 2 to 24 years.

1) Positive and Negative Affect: A paired Wilcoxon Rank Sum test showed a significant decrease of participants' negative affect (PRE:M=1.42, SD=0.31 vs. POST:M=1.21, SD=0.25; W=118, $p<0.01$) and a near significant increase on participants' positive affect (PRE:M=3.91, SD=0.88 vs. POST:M=4.15, SD=0.75; W=26, $p=0.057$) during the session.

2) Inclusion of Others in Self: We did not find significant differences before (M=5.95, SD=1.1) and after session (M=6.1, SD=1.21) on self-reported interpersonal connectedness.

3) Closeness and Intimacy: We found a significant increase in self-reported intimacy during the session with the robot (PRE:M=6.13, SD=1.11; POST:M=6.45, SD=1.25), paired Wilcoxon signed rank W=X, $p<0.01$.

4) Attitudes towards Robot: Attitudes towards the robot were generally positive across participants. Participants were satisfied with the robotic facilitator and the counseling experience. They trusted and liked the robot, expressed a desire to continue working with the robot, and thought that the robot was effective (Table IV). The internal consistency of the questionnaire items was acceptable (Cronbach's $\alpha=0.82$)

5) Enjoyment of Interaction: Couples generally enjoyed the interaction with the robot and with each other (M=6.15, SD=0.92).

6) Partner's Responsiveness: Participants rated their partner's responsiveness high (M=6.76, SD=0.36).

TABLE IV. FACILITATOR RATINGS

Ratings of Robotic Facilitator (Anchors 1: Not at all; 7: Very much)	**Mean (SD)**
How satisfied are you with the facilitator?	5.9 (1.29)
How effective was the facilitator at leading the session?	5.85 (1.18)
How helpful was the facilitator in getting you involved in the interaction?	5.75 (1.02)
How much do you like the facilitator?	5.5 (1.7)
How much do you trust the facilitator?	5.5 (1.76)
How interesting was the facilitator?	6.05 (1.15)
How much would you like to continue working with the facilitator?	5.5 (2.09)
How satisfied are you with the interaction experience?	5.9 (1.17)

D. Intimate Behaviors

Participants actively engaged in the positive relationship exercises guided by the robot. All participants complied with the robot's request to share gratitude with each other and came up with a Caring Day request to practice at home. During these exercises, and throughout the counseling period, we observed several examples of intimate behaviors, including touching, comforting, handholding, caressing, back rubbing, hugging, and kissing (Fig 3). We also observed several verbal intimate behaviors, such as caring statements (e.g., "I love you"), intimate self-disclosure, and compliments (e.g., "She's such a kind and understanding person"). This indicates that participants were comfortable with the robot and the experimental setup.

Post-session interviews also indicated that participants had a positive impression about the robot and the overall experience. Some of the words used to describe the interaction were "*interactive*"[C1, P2], "*natural*"[C9, P1], and "*fun*"[C10, P2]. They especially liked that the robot was tracking who is speaking and made eye contact with them: "*the tracking and eye contact and nodding made it seem more real. It's giving the impression that we were being heard*" [C10, P1]. Participants felt that interacting with the robot is "*like interacting with a real person* […] *her responses, eyes. I feel like I'm talking with someone*" [P14,P2]. However participants felt that the interaction was too structured, and that the robot's responses were very generic: "*Most of the conversation was very generic, but the answers were very much like, so, when you said this thing you get this thing...*" [C2, P2] and "*It's like talking to Siri*" [C2, P1]

When asked if they would prefer a human counselor, participants give mixed responses. On the one hand they felt that the robot is non-judgemental and unbiased: "*[the] robot is unbiased towards either gender so it's better than a human*" [C1, P2] and "*I'm not sure I would have been comfortable with a human.*" [C10, P1]. On the other hand, human counselors are perceived to be "*more credible*" [C9, P1] and better at "*reading your expression*"[C9, P1].

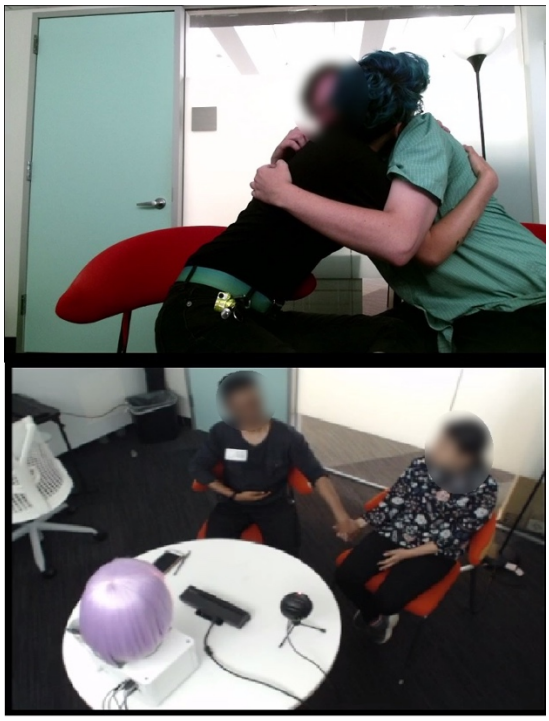

Fig. 3. Intimate Behaviors

Regarding the technical aspects of the system, participants said that they were a little uncomfortable with the headset but were okay with it: "*the headset was small*" [C14, P2]. Participants did notice the camera but only in the beginning before the session begun: "*she* [robot] *takes all of our attention*" [C1, P2]. Participants also noticed several instances where the robot's responses are delayed (when the wizard was typing a response): "*sometimes it's a little slowish […] but perhaps it's a processing thing*" [C14, P2]. And, they wished for more grounding signals when the robot is listening, "*sometimes when we talk we're like - is she hearing us?*" [C2, P1]

In addition to being more interactive, participants thought that interacting with the robot was a better alternative than reading a self-help material and practicing by themselves because they felt that having a neutral third party facilitating the session helped structure the interaction: "*having someone who is new to the situation who is just playing the neutral field I think helps*" [C9, P1]

Participants felt that the advice given by the robot was "*helpful*"[C2, P1], "*credible*" [C14, P2], "*insightful*"[C1, P2], and helped them learn to express affection more explicitly: "*there are many thing which he didn't say to me before and today in front of Julia he said it*" [C10, P1]. While, for some couples, the skills introduced in the session were not new, they felt that it was a good reminder: "*sometimes other things overshadow the things that happen in the beginning when you are in the honeymoon stage and I think it's really important to go back*" [C9, P1] They also feel that the homework given during the session was "*simple, easy to remember, and measurable*" [C10, P2].

VII. Online Model Evaluation

We implemented the model described in Section 5 in our robotic couples counseling system to provide real-time decisions regarding whether each 200ms silence following rapport-building questions by the robot should be classified as end-of-collaborative-turn or not. The final decision for end of turn is made by the wizard and the wizard is blinded to the model's decision. To evaluate the model's performance, we compare the model's end of turn decisions with those made by the wizard in real time during each session. The comparison was made offline using system log.

The online system was implemented using an open source dialogue system framework IrisTK, which consists of a set of modules that send and listen to events [52]. The Kinect was used to track users' head location and rotation. We calculated angular distance between the direction of the user's head pose and the robot's head (Fig. 4). The system classifies users' gaze target as "at robot" when the angular distance is less or equal to 25 degrees, otherwise as "away from robot".

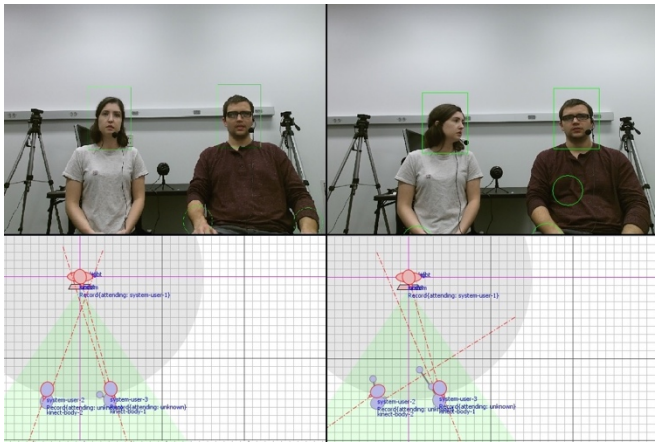

Fig. 4. Kinect Tracking Users' Head Location and Rotation

User speech from the headset is recognized using Google's cloud-based speech recognizer. To obtain users' fundamental speech prosody, we obtained a speech sample prior to the session and analyze it using Praat [42]. Prosodic features are obtained using an approach similar to that described in the Section 5.C. Syntax features are extracted automatically by processing recognized speech through the Stanford POS tagger [44]. The model makes a end-of-turn decision whenever a silence with a maximum of 200ms period is detected using an energy-based Voice Activity Detector (VAD).

We tested the model in half (5) of the counseling sessions, with the wizard still controlling the system. During the sessions, the system made 65 decisions. For each decision point, we coded it with a value of "True", if the system's prediction agrees with the wizard's, and "False" otherwise. The F-score for the online model was 0.72, lower the offline evaluation f-score, but higher than the majority baseline f-score.

VIII. Discussion

We describe an ongoing research effort to develop a robotic counselor for romantic couples and results from a quasi-experimental study that demonstrates that the robot is effective at increasing intimacy and positive affect among couples who interact with it. All participants engaged in the exercises led by the robot, and overall they were satisfied with the experience. We observed several examples of intimate behavior during the session, which suggests that participants were comfortable sharing and expressing intimacy in front of the robot. In addition, several participants showed intense emotional reactions (e.g., crying during the session) as a response to what their partner said during the exercise.

We also report on the development and evaluation of a data-driven model to identify the end of collaborative responses to questions from the robot counselor, with a resulting weighted average F-score of 0.81 for offline evaluation using hand-coded features, and 0.72 for a real-time model using features derived from noisy sensors in use during counseling sessions. Collaborative responses represent a novel research problem in multiparty interaction, and we describe several types of collaborative responses made by couples in our studies. We found that gaze direction is very useful for collaborative turn-taking decisions, and that male participants' gaze is more useful than the female participants' gaze. Regarding the design of turn-taking systems in counseling, the findings from our interviews suggest that participants tolerate system delays due to processing. Thus, a conservative model that minimizes interruption by the robot is preferred.

Our study has many limitations, beyond the small convenience sample used. Our quasi-experimental study was a step towards developing a fully-automated system and set out to assess the acceptability of our robotic counselor advising couples in different aspects of intimate relationships, while incrementally automating parts of the system. A true efficacy study must ultimately be performed in a randomized, controlled trial. Finally, we recognize that we have only implemented a tiny fraction of what human counselors do, and referring to our robot as a true couples "counselor" at this point is perhaps a stretch.

IX. Future Work

The study presented in this paper is one step towards the development of a fully-automated robotic couples counselor. We plan to continue incrementally automating the system to support basic communication processes, such as grounding, as well as specific couples counseling techniques, such as reflection and feedback. We are particularly interested in the challenges in developing a model for collaborative turn-taking [53] and automated assessment of exercise fidelity with noisy sensor data, including imperfect speech recognition and missing data. We would also like to extend our counseling to multiple sessions and include more couple counseling techniques.

The general trend of declining relationship satisfaction is worrisome, both in and of itself, but also because relationship distress has been associated with many health problems, including depression, anxiety and alcohol abuse. While couples counseling has been shown to alleviate these problems, access to professional help is often limited. Robotics couples counselors could provide help to many couples, especially asymptomatic couples, who are looking for ways to maintain their relationship satisfaction.

Acknowledgment

The authors thank Lou Kruger, Teresa O'Leary, Elise Mason, Prem Shah, and Jee-Eun Heo for their assistance in conducting the study.

References

[1] W. Halford and D. Snyder, "Introduction to Special Series on Universal Processes and Common Factors in Couple Therapy and Relationship Education," *Behavior Therapy,* vol. 43, pp. 1-12, 2012.

[2] M. I. Morrill, C. Eubanks‐Fleming, A. G. Harp, J. W. Sollenberger, E. V. Darling, and J. V. Cördova, "The Marriage Checkup: Increasing access to marital health care," Family Process, vol. 50, pp. 471-485, 2011.

[3] P. A. Andersen, "Nonverbal immediacy in interpersonal communication," Multichannel integrations of nonverbal behavior, pp. 1-36, 1985.

[4] E. Goffman, Relations in public: Transaction Publishers, 2009.

[5] J. Mandelbaum, "Couples sharing stories," Communication Quarterly, vol. 35, pp. 144-170, 1987.

[6] S. Planalp, "Friends' and acquaintances' conversations II: Coded differences," Journal of Social and Personal Relationships, vol. 10, pp. 339-354, 1993.

[7] L. Tickle-Degnen and R. Rosenthal, "The Nature of Rapport and Its Nonverbal Correlates," Psychological Inquiry, vol. 1, pp. 285-293, 1990.

[8] J. Cassell, A. J. Gill, and P. A. Tepper, "Coordination in conversation and rapport," in Proceedings of the workshop on Embodied Language Processing, 2007, pp. 41-50.

[9] D. Schulman and T. Bickmore, "Changes in verbal and nonverbal conversational behavior in long-term interaction," presented at the Proceedings of the 14th ACM international conference on Multimodal interaction, Santa Monica, California, USA, 2012.

[10] Y. Matsuyama, A. Bhardwaj, R. Zhao, O. Romeo, S. Akoju, and J. Cassell, "Socially-aware animated intelligent personal assistant agent," in Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2016, pp. 224-227.

[11] R. Zhao, T. Sinha, A. W. Black, and J. Cassell, "Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior," in International conference on intelligent virtual agents, 2016, pp. 218-233.

[12] B. Szczepek, "Functional aspects of collaborative productions in English conversation," InLiSt No 17, 2000.

[13] J. Falk, "The conversational duet," in Annual Meeting of the Berkeley Linguistics Society, 1980, pp. 507-514.

[14] K. Ferrara, "The interactive achievement of a sentence: Joint productions in therapeutic discourse," Discourse Processes, vol. 15, pp. 207-228, 1992.

[15] H. Sacks, "Lectures on conversation, volumes I and II. Edited by G. Jefferson with Introduction by EA Schegloff," ed: Oxford: Blackwell, 1992.

[16] S. Duncan, "On the structure of speaker-auditor interaction during speaking turns," Language in Society, vol. 3, pp. 161-180, 1974.

[17] C. Goodwin, "Achieving Mutual Orientation at Turn Beginning," in Conversational Organization: Interaction between Speakers and Hearers, ed New York: Academic Press, 1981, pp. 55-89.

[18] A. Kendon, "Some functions of gaze direction in social interaction," Acta Psychologica, vol. 26, pp. 22-63, 1967.

[19] H. Sacks, E. A. Schegloff, and G. Jefferson, "A Simplest Systematics for the Organization of Turn-Taking for Conversation," Language, vol. 50, pp. 696-735, 1974.

[20] S. Duncan, "Some signals and rules for taking speaking turns in conversations," Journal of Personality and Social Psychology, vol. 23, pp. 283-292, 1972.

[21] A. Raux and M. Eskenazi, "Optimizing endpointing thresholds using dialogue features in a spoken dialogue system," in Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue, 2008, pp. 1-10.

[22] N. Ward, O. Fuentes, and A. Vega, "Dialog prediction for a general model of turn-taking," presented at the Interspeech, 2010.

[23] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den, "An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs," Language and Speech, vol. 41, pp. 295-321, 1998.

[24] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," Computer Speech and Language, vol. 25, pp. 601-634, 2011.

[25] A. Hjalmarsson, "The additive effect of turn-taking cues in human and synthetic voice," Speech Communication, vol. 53, pp. 23-35, 2011.

[26] A. Raux and M. Eskenazi, "A finite-state turn-taking model for spoken dialog systems," in Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2009, pp. 629-637.

[27] A. Raux and M. Eskenazi, "Optimizing the turn- taking behavior of task-oriented spoken dialog systems," ACM Transactions on Speech and Language Processing, vol. 9, pp. 1-23, 2012.

[28] K. R. Thórisson, "Natural turn-taking needs no manual: Computational theory and model, from perception to action," in Multimodality in language and speech systems, ed: Springer, 2002, pp. 173-207.

[29] E. O. Selfridge and P. A. Heeman, "Importance-Driven Turn-Bidding for spoken dialogue systems," in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010, pp. 177-185.

[30] D. Traum, "Issues in Multiparty Dialogues," in Advances in Agent Communication. vol. 2922, F. Dignum, Ed., ed: Springer Berlin Heidelberg, 2004, pp. 201-211.

[31] R. Vertegaal, "The GAZE groupware system: mediating joint attention in multiparty communication and collaboration," presented at the Proceedings of the SIGCHI conference on Human Factors in Computing Systems, Pittsburgh, Pennsylvania, USA, 1999.

[32] M. Katzenmaier, R. Stiefelhagen, and T. Schultz, "Identifying the addressee in human-human-robot interactions based on head pose and speech," presented at the Proceedings of the 6th international conference on Multimodal interfaces, State College, PA, USA, 2004.

[33] E. Shriberg, A. Stolcke, and S. Ravuri, "Addressee detection for dialog systems using temporal and spectral dimensions of speaking style," presented at the INTERSPEECH-2013, 2013.

[34] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den, "An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs," Language and speech, vol. 41, pp. 295-321, 1998.

[35] L.-P. Morency, I. de Kok, and J. Gratch, "Predicting listener backchannels: A probabilistic multimodal approach," in International Workshop on Intelligent Virtual Agents, 2008, pp. 176-190.

[36] R. Meena, G. Skantze, and J. Gustafson, "Data-driven models for timing feedback responses in a Map Task dialogue system," Computer Speech & Language, vol. 28, pp. 903-922, 2014.

[37] D. Bohus and E. Horvitz, "Decisions about turns in multiparty conversation: from perception to action," presented at the Proceedings of the 13th international conference on multimodal interfaces, Alicante, Spain, 2011.

[38] M. Johansson and G. Skantze, "Opportunities and obligations to take turns in collaborative multi-party human-robot interaction," in Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2015, pp. 305-314.

[39] S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström, "Furhat: A Back-Projected Human-Like Robot Head for Multiparty Human-Machine Interaction," in Cognitive Behavioural Systems, A. Esposito, A. M. Esposito, A. Vinciarelli, R. Hoffmann, and V. C. Müller, Eds., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 114-130.

[40] J. Cassell, H. H. Vilhjálmsson, and T. Bickmore, "BEAT: the Behavior Expression Animation Toolkit," in Life-Like Characters: Tools, Affective Functions, and Applications, H. Prendinger and M. Ishizuka, Eds., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 163-185.

[41] D. Utami, T. W. Bickmore, and L. J. Kruger, "A robotic couples counselor for promoting positive communication," in Robot and Human Interactive Communication (RO-MAN), 2017 26th IEEE International Symposium on, 2017, pp. 248-255.

[42] P. Boersma, "Praat, a system for doing phonetics by computer," Glot international, vol. 5, 2002.

[43] M. Kipp, "Anvil-a generic annotation tool for multimodal dialogue," in Seventh European Conference on Speech Communication and Technology, 2001.

[44] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit,"

in Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, 2014, pp. 55-60.

[45] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," ACM SIGKDD explorations newsletter, vol. 11, pp. 10-18, 2009.

[46] M. A. STRAUS, S. L. HAMBY, S. BONEY-McCOY, and D. B. SUGARMAN, "The Revised Conflict Tactics Scales (CTS2)," Journal of Family Issues, vol. 17, pp. 283-316, 1996.

[47] C. W. LeCroy, P. Carrol, H. Nelson-Becker, and P. Sturlaugson, "An experimental evaluation of the caring days technique for marital enrichment," Family Relations, pp. 15-18, 1989.

[48] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: The PANAS scales," Journal of Personality and Social Psychology, vol. 54, pp. 1063-1070, 1988.

[49] A. Aron, E. N. Aron, and D. Smollan, "Inclusion of Other in the Self Scale and the structure of interpersonal closeness," Journal of Personality and Social Psychology, vol. 63, pp. 596-612, 1992.

[50] N. Alea and S. Bluck, "I'll keep you in mind: The intimacy function of autobiographical memory," Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, vol. 21, pp. 1091-1111, 2007.

[51] R. L. Dunn and A. I. Schwebel, "Meta-analytic review of marital therapy outcome research," Journal of Family Psychology, vol. 9, pp. 58-68, 2015-02-03 1995.

[52] G. Skantze and S. A. Moubayed, "IrisTK: a statechart-based toolkit for multi-party face-to-face interaction," presented at the Proceedings of the 14th ACM international conference on Multimodal interaction, Santa Monica, California, USA, 2012.

[53] R. Reiter-Palmon, T. Sinha, J. Gevers, J.-M. Odobez, and G. Volpe, "Theories and models of teams and groups," Small Group Research, vol. 48, pp. 544-567, 2017.