The 23rd IEEE International Symposium on Robot and
Human Interactive Communication
August 25-29, 2014. Edinburgh, Scotland, UK,

Handling uncertain input in multi-user human-robot interaction

Simon Keizer[1], Mary Ellen Foster[1], Andre Gaschler[2], Manuel Giuliani[2], Amy Isard[3], and Oliver Lemon[1]

***Abstract*— In this paper we present results from a user evaluation of a robot bartender system which handles state uncertainty derived from speech input by using belief tracking and generating appropriate clarification questions. We present a combination of state estimation and action selection components in which state uncertainty is tracked and exploited, and compare it to a baseline version that uses standard speech recognition confidence score thresholds instead of belief tracking. The results suggest that users are served fewer incorrect drinks when the uncertainty is retained in the state.**

I. Introduction

Interactive multimodal systems typically consist of components for input processing, state management, action selection and behaviour realisation. In order for such a system to operate robustly in the face of uncertain observations, it is important to explicitly represent the resulting uncertainty in the state and to exploit this in the action selection process. A system that uses only the most likely input hypothesis in maintaining the state is likely to select actions on the basis of incorrect information, and therefore to display undesirable or even unacceptable behaviour. A simple approach for handling uncertain data is to introduce confidence thresholds on the input hypotheses, resulting in system behaviour that can be either too passive (when using high thresholds for accepting an input hypothesis) or too fraught with errors (in the case of lower thresholds). We argue that by taking into account multiple input hypotheses and their confidence scores, the system can make better informed decisions, and—especially when including additional actions aimed at reducing uncertainty—the system will be more robust to uncertain input.

In this paper we present an extended version of the JAMES robot bartender system (see Figure 1), in which the state manager maintains multiple state hypotheses with confidence scores, based on the input hypotheses and their confidence scores provided by the vision and speech processing components. The action selection component is extended with rules that take into account this explicitly represented uncertainty. We have carried out a user study to compare the behaviour of the baseline system that does not handle uncertainty but uses thresholds for accepting input hypotheses or not, to the extended system which does handle uncertainty.

[1]Interaction Lab, Heriot-Watt University, Edinburgh, UK s.keizer@hw.ac.uk
[2]fortiss GmbH, Munich, Germany giuliani@fortiss.org
[3]School of Informatics, University of Edinburgh, Edinburgh, UK amyi@inf.ed.ac.uk

Fig. 1. Two users interacting with the robot bartender

II. Robot bartender system

Figure 2 shows the architecture of our robot bartender system. The Visual Processing component tracks the location and body orientation of multiple customers in the scene, using two calibrated stereo cameras and a Kinect depth sensor. Speech processing consists of speech recognition using the Kinect ASR system and semantic parsing using OpenCCG. The State Manager fuses the audiovisual input stream and maintains a model of the social state; details are presented in Section IV. The Social Skills Executor then selects response actions given social state updates provided by the State Manager, as outlined in Section V. The selected action are then realised via the Output Planner, which sends instructions to the Talking Head Controller (e.g., looking at a particular customer, nodding, and/or speaking) and the Robot Motion Planner. The Robot Motion Planner provides a high-level interface to the physical process of serving of a drink to a customer, along with functions such as idle motions and picking up bottles from arbitrary locations.

III. Speech and Language Processing

For speech recognition, we make use of the Microsoft Kinect for Windows API which produces a series of intermediate hypotheses while recognition is active, and a final n-best list of recognition hypotheses when the end of speech is detected. Each hypothesis has an estimated confidence score, along with an estimate of the sound source angle and the angle confidence. An application-specific speech recognition grammar is used to constrain the recognition process in order to achieve more reliable results and to ensure that the hypotheses can be processed by the parser.

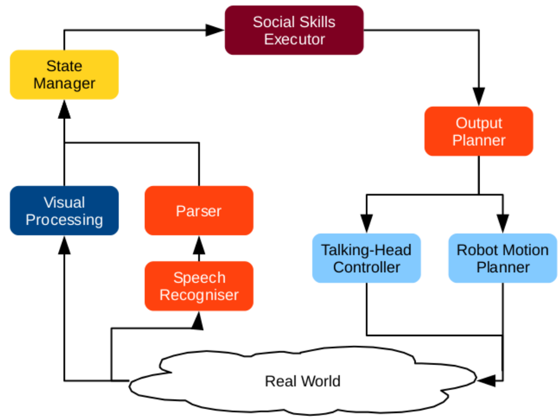

Fig. 2. JAMES system architecture

Once the user speech has been recognised, it must be further processed to extract the underlying meaning. To do this, we parse each hypothesis using a grammar defined in OpenCCG [1], in an attempt to find a full parse. If no full parse is found, we process all substrings of the recognised string, and store the parse of the longest fragment along with its confidence. Finally, after removing any duplicate parses from the list, we convert each parse into a parameterised communicative act, whose types include greeting, thanks, and drink order requests. This list of possible communicative acts is passed to the State Monitoring module along with the original speech recognition string, the fragment string if appropriate, the Kinect confidence score, and the sound source angle and confidence.

IV. State Monitoring with Uncertain Input

In our robot bartender system, the task of the *state manager* is to keep track of the *social state*, which contains information about the customers in the scene: for example, whether they are currently seeking attention from the bartender, and whether they have been served their desired drink. This decision is based on the continuous stream of messages produced by the low-level input and output components. We store all of the low-level information, and also infer additional relations not directly reported by the sensors: for example, we fuse information from vision and speech to determine which user should be assigned a recognised speech hypothesis, and use the vision data to estimate each customer's attention-seeking state [2].

The input provided by the vision and speech processing components is noisy and uncertain: in particular, all signals from the speech recogniser and the vision system include an associated *confidence* value that indicates the estimated reliability of the observation. Also, as noted above, the speech recogniser may in some cases provide multiple alternative hypotheses, each with its own associated confidence value. However, the initial state representation [3] stored only the most likely overall hypothesis, with no information about the associated confidence. This simplified the initial action-selection task considerably, but also discarded potentially valuable sensor information.

We have therefore extended the initial version of the state manager to associate each state hypothesis with a confidence score, and to include alternative hypotheses about a customer's drink order. Incorporating multiple hypotheses and confidence scores into the state requires additional processing in the state manager. The JAMES computer vision system [4] estimates the location, gaze behaviour, and body language of all people in the scene in real time, along with an estimated confidence for each feature; these confidence values are incorporated into the state, and are also used to determine the confidence for derived properties such as attention-seeking. For speech, we use the source angle from the speech recogniser together with the location information from vision to associate the communicative acts with a customer. If the communicative act has to do with ordering a drink, we also update our estimate of the customer's desired drink using the generic belief tracking procedure proposed by [5], which maintains beliefs over user goals based on a small number of domain-independent rules, using basic probability operations: for example, if the customer repeats a request for a Coke, the state-manager confidence for that order will increase, even if the ASR confidence is low for each individual utterance. This allows us to maintain a dynamically-updated list of the possible drink orders made by each customer in the scene, with an associated confidence value for each order. The full details of the updated state manager are given in [6].

V. Action Selection under Uncertainty

The task of the *social skills executor* (SSE) is to decide what action the robot should take next, based on an update of the social state provided by the state manager. In order to exploit the uncertainty information incorporated in the new social state representation, the action selection strategy has been extended to include actions for clarifying the customers' drink orders and rules for when to issue such clarifications.

The decision making process of the SSE consists of two main stages. In the first stage, the SSE decides which of the customers in the scene to focus on in its next action: in particular, it decides whether to engage with a customer seeking attention, whether to politely ask them to wait, or whether to continue its ongoing interaction with them. In case an ongoing interaction is to be continued, the system decides in the second stage which communicative action will be carried out, and whether a drink will be served to the customer. Possible communicative actions include asking the customer for their order (e.g., “What can I get you?”, “What would you like to drink?”), acknowledging an order (e.g., ”Okay, a coke”), serving an order (e.g., “There you go”, “Here is your coke”), addressing social conventions (e.g., greetings, “You're welcome” after a customer thanks the system), and clarifications (e.g., “Did you say coke?”, “Did you say blue lemonade or green lemonade?”).

Since initial tests with the audiovisual input processing system showed that the most important source of uncertainty is in the speech input, the additional clarification actions are focused on reducing this form of uncertainty.

In particular, the second stage of the decision making process of continuing an ongoing interaction with a particular customer has been extended with rules for taking such actions, as shown in Algorithm 1, which uses the empirically-determined thresholds shown in Table I. The criteria for selecting clarifications depend on both the confidence of the top drink order hypothesis and the entropy of the drink order distribution, provided the state manager. The system trusts the top hypothesis if the confidence is either above an upper threshold, or above a lower threshold, combined with a sufficiently low entropy (as a measure of uncertainty about the drink order); a clarification question is generated when these criteria are not met. Note that we also employ a separate minimum confidence threshold on the speech recognition, depending on whether uncertainty processing is enabled (SCONF_THR and SCONF_THR_UNC in Table I).

Algorithm 1 Selecting clarification actions (*conf* refers to the confidence score of the top drink order hypothesis, *entr* refers to the entropy of the drink order belief distribution, and the thresholds used in the experiment are listed in Table I.)

if ($conf \geq$ CONF_THR1) **or**
($conf \geq$ CONF_THR2 **and** $entr <$ ENTR_THR) **then**
  select action based on top hypothesis;
  (e.g., "Okay, a coke")
else if there is only one drink order hypothesis **then**
  confirm the drink order with the user;
  (e.g., "Did you say 'coke'?")
else
  let user choose between top 2 hypotheses;
  (e.g., "Did you say 'green' or 'blue' lemonade?")
end if

TABLE I
THRESHOLDS USED IN SELECTING CLARIFICATIONS

Description	Threshold	Value
Upper confidence threshold	CONF_THR1	0.65
Lower confidence threshold	CONF_THR2	0.40
Entropy threshold	ENTR_THR	0.25
parsing confidence threshold (baseline)	SCONF_THR	0.30
parsing confidence threshold (uncertainty)	SCONF_THR_UNC	0.10

Figure 3 shows an interaction with the uncertainty-aware system from our user studyin which the system successfully clarifies a user order. Figure 4 shows an example interaction with the baseline (uncertainty-unaware) system in which the system misrecognizes the customer's order and serves the wrong drink without confirming the order first.

VI. USER EVALUATION

To assess the impact of the revised action-selection process described above, we carried out a user evaluation. The experiment was similar to that described in [7]: each participant carried out four short drink-ordering transactions with the robot bartender together with a second customer played by a confederate. For comparison, we used the previous version of the action-selection system as a baseline [7]; this version used only the top state hypothesis and did not consider any confidence values. Half of the sessions used the uncertainty-aware system, while the other half used the baseline system; also, to cover a range of scenarios, in half of the sessions, the participant and the confederate each ordered their own drinks, while in the other half, the participant also ordered on behalf of the confederate. We gathered a range of objective measures from the system log files, as well as subjective measures from an online questionnaire.

1) *Customer (A1) enters the scene, seeking attention*
2) **System (looking at A1): "Hello"**
3) *A1 orders*
4) parser: *drink-order(green-lemonade) [0.02]*
5) state: speech input rejected
6) *Customer (A2) enters the scene, not seeking attention*
7) **System (to A1): "What can I get you?"**
8) *A1 answers*
9) parser: *drink-order(blue-lemonade) [0.05]*
10) state: speech input rejected
11) **System (to A1): "What can I get you?"**
12) *A1 answers*
13) parser: *drink-order(lemonade) [0.75]*
14) state: *drink-order(A1)=(blue-lemonade [0.37], green-lemonade [0.37])*
15) **System (to A1): "Did you say 'blue lemonade' or 'green lemonade'?"**
16) *A1 answers*
17) parser: *drink-order(lemonade) [0.53], drink-order(blue-lemonade) [0.53]*
18) state: *drink-order(A1)=(blue-lemonade [0.57], green-lemonade [0.43])*
19) **System (to A1): "Did you say 'blue lemonade' or 'green lemonade'?"**
20) *A1 answers*
21) parser: *drink-order(blue-lemonade) [0.69]*
22) state: *drink-order(A1)=(blue-lemonade [0.66], green-lemonade [0.34])*
23) **System (to A1): "Okay, a blue lemonade"**
24) *System serves blue lemonade to A1*

Fig. 3. Interaction fragment where a user successfully orders blue lemonade

A. Participants

24 participants (21 male), drawn from university departments outside the robotics group involved in developing the bartender, took part in this experiment. The mean participant age was 27.5 (range 21–49), and their mean self-rating of experience with human-robot interaction systems was 3.3 on a scale of 1–7. Seven of the participants had taken part in a previous robot bartender evaluation, while 17 had not. All participants were native or fluent speakers of German.

B. Procedure

Before the experiment, we told participants that their task would be to order a drink from the robot bartender. They were shown the physical form of the robot, but not its interactive behaviour, and were then asked to rate their expectations using a computer-based questionnaire. After they had filled out the questionnaire, we told the participants that they should approach the robot together with another

1) *Two customers enter the scene*
one of them (A1) approaches the bar
2) **System (to A1): "Hello"**
3) *A1 returns greeting*
4) **System (to A1): "What can I get you?"**
5) *A1 tries to order a blue lemonade*
6) parser: *drink-order(coke) [0.332]*
7) state: *drink-order(A1)=(coke [1.0])*
8) **System (to A1): "Okay, a coke"**
9) *system serves A1 a coke*
10) **System (to A1): "Here you go"**
11) *A1 tries to order again but ASR fails*
12) **System (to A1): "Anything else?"**
13) *A1 tries to order again but speaker identification fails*
14) **System (to A1): "Will that be all?"**
15) *A1 responds negatively*
16) **System (to A1): "What can I get you?"**
17) *A1 once again orders a blue lemonade*
18) parser: *drink-order(blue-lemonade) [0.414]*
19) state: *drink-order(A1)=(blue-lemonade [1.0])*
20) **System: "Okay, a blue lemonade"**
21) *system serves A1 a blue lemonade*
22) *A1 thanks the system and leaves*

Fig. 4. Interaction in which the system serves the wrong drink

customer (a confederate). In half of the cases, the confederate approached the bartender with the participant, while in the other half, the confederate remained in the background while the participant ordered on his behalf. Each participant took part in four trials; after each trial, the participant completed another computer-based questionnaire.

C. Independent measures

We manipulated two factors during this study: we varied the use of uncertainty in the system, and also varied whether the confederate ordered for himself or asked the participant to order on his behalf. In a within-subjects design, all participants interacted with the bartender in all four configurations, each in an individually counterbalanced order.

D. Dependent measures

We gathered two classes of dependent measures: objective measures based on the system logs, and subjective measures derived from the pre- and post-experiment questionnaires.

1) Objective measures: The objective measures were based on the dimensions proposed by the PARADISE dialogue evaluation framework [8]. **Task success** was assessed by counting how many drinks were served by the system (maximum 2); **dialogue quality** was measured by counting how many of the user's attempted contributions fell below the speech-recognition confidence threshold, how many times the robot had to ask for a customer's drink order, and—for the system that used uncertainty—how many times it used a clarification; while for **dialogue efficiency**, we computed the time taken to serve the first drink in a trial, the time taken to serve all of the drinks, as well as the total duration of the trial as measured both in seconds and in system turns.

2) Subjective measures: Before the experiment, the participants completed the short subjective questionnaire shown in Figure 5, which is based on the Godspeed questionnaire series [9], a standard user measurement tool for human-robot interaction. On the pre-test, the questions were framed to ask for the users' expectations rather than impressions. After each trial, the participant again completed the Godspeed-based questionnaire, as well as a short questionnaire designed to measure their perceived success and overall impression of the trial (Figure 6). Note that the questions were posed in German; the figures show English translations.

Please rate your impression of the robot:

1.	*Machinelike*	1	2	3	4	5	6	7	*Humanlike*
2.	*Unkind*	1	2	3	4	5	6	7	*Kind*
3.	*Unintelligent*	1	2	3	4	5	6	7	*Intelligent*
4.	*Artificial*	1	2	3	4	5	6	7	*Lifelike*
5.	*Unpleasant*	1	2	3	4	5	6	7	*Pleasant*
6.	*Inert*	1	2	3	4	5	6	7	*Interactive*
7.	*Dislike*	1	2	3	4	5	6	7	*Like*
8.	*Unfriendly*	1	2	3	4	5	6	7	*Friendly*
9.	*Incompetent*	1	2	3	4	5	6	7	*Competent*

Fig. 5. Godspeed questionnaire [9] subset used for evaluation

Q1: What drinks did you order? [2 drinks; coke, green lemonade, or blue lemonade]
Q2: What drinks did you get? [drinks of type coke, green lemonade, or blue lemonade]
Q3: What was your overall impression of this interaction? [1-6 Likert scale]

Fig. 6. Questionnaire for each session.

E. Results

Except where specifically noted below, none of the demographic features of the participants had any significant impact on the results; also, whether the participant ordered for the confederate did not make any significant difference. In this analysis, we therefore concentrate primarily on the effect of varying the action-selection strategy.

The objective results are summarised in Table II, showing the mean results on each measure from the two conditions; the final column shows the significance level from a paired Mann-Whitney test comparing the results from the two versions. Note that the baseline used the same acceptance threshold as in the previous study [7], while the uncertainty-aware version used a lower threshold, as it has a better process for dealing with low-confidence utterances—see Table I, where the thresholds are indicated as SCONF_THR and SCONF_THR respectively. It is therefore not surprising that the baseline version had significantly more user turns discarded due to low ASR confidence. Also, the baseline version never selected choices or confirmations in its output, while—as shown in the table—the uncertainty-aware system generally clarified several times in each trial. This means that the significant difference in system turns is also as expected. The other differences between the systems are more interesting:

TABLE II
OBJECTIVE RESULTS

Measure	Baseline (sd)	Uncertainty (sd)	M-W
Drinks served	1.96 (0.14)	1.72 (0.39)	$p < 0.01$
Low ASR turns	3.2 (1.5)	2.0 (0.84)	$p < 0.001$
Order requests	5.7 (2.6)	5.5 (2.6)	n.s.
Choices	—	2.3 (2.3)	—
Confirmations	—	2.3 (2.0)	—
Time to first drink	49.6 (19.6)	71.3 (58.7)	$p < 0.05$
Time to last drink	94.2 (24.1)	107.7 (61.2)	n.s.
Duration	103.6 (25.3)	122.9 (61.2)	n.s.
System turns	14.1 (3.6)	17.6 (5.0)	$p < 0.05$

TABLE III
SUMMARY OF RESPONSES TO GODSPEED QUESTIONNAIRE

	Pre-test		Post-test	
Category	α	Mean (sd)	α	Mean (sd)
Anthropomorphism	0.77	3.0 (1.1)	0.85	2.6 (1.3)
Animacy	–	3.6 (1.7)	–	3.2 (1.5)
Liking	0.82	5.3 (1.1)	0.91	4.8 (1.2)
Intelligence	0.90	4.5 (1.5)	0.85	3.7 (1.4)

TABLE IV
SUMMARY OF RESULTS TO SESSION QUESTIONNAIRE

Measure	Baseline (sd)	Uncertainty (sd)	M-W
Perceived precision	0.92 (0.26)	0.97 (0.17)	n.s.
Perceived recall	0.90 (0.21)	0.81 (0.33)	n.s.
Overall impression	4.4 (1.0)	3.7 (1.2)	$p < 0.01$

the baseline system served significantly more drinks in a trial (out of a maximum of two), and also served the first drink significantly more quickly. These two results are likely to be related: while the baseline version would immediately act on any recognised drink-order hypotheses (as in Figure 4), the uncertainty-enabled version would make an effort to confirm or clarify any uncertain hypotheses before proceeding (Figure 3); and in some cases, due to input-processing issues, it never achieved sufficient confidence to serve all drinks.

The results on the Godspeed questions are summarised in Table III. We have divided the questions into the high-level Godspeed categories they were drawn from: Anthropomorphism (questions 1 and 4), Animacy (question 6), Liking (questions 2, 5, 7, and 8), and Perceived Intelligence (questions 3 and 9). For each category, on both the pre-test and the post-test, we first computed Cronbach's alpha to test the internal consistency, and then computed the mean response on that category. The experimental manipulation had no significant effect on any of these questions, so Table III simply shows the aggregate responses from the pre-test and from all of the post-tests. As shown, the consistency was generally quite high for all categories ($\alpha > 0.7$), on both pre-test and post-test. The responses on all categories generally decreased from the pre-test to the post-test, with the biggest decrease on the Perceived Intelligence category. This is similar to the score decrease observed on a previous study which also used the Godspeed series as a pre-test [10]. To see whether this pattern was affected by the participants' experience either with HRI systems in general, or with previous versions of the JAMES bartender specifically, we carried out a multiple regression analysis. The only significant effect was that the Anthropomorphism decrease was less for participants with more HRI experience ($R^2 = 0.27, p < 0.01$). In general, this suggests that people's expectations of a robot's interactive capabilities tend to outstrip their actual experience of interacting with it, even when they have previous experience with the same robot.

The results from the additional subjective questionnaire are summarised in Table IV. The top two rows indicate the perceived precision and recall; that is, the proportion of the served drinks that were reported as correct, and how many of the requested drinks were actually be served. Despite the difference in drinks served between the two systems (Table II), there was no significant difference found on these measures; however, note that the precision was somewhat higher for the system with uncertainty enabled, while the recall was higher for the baseline system. Also, the perceived recall was mildly correlated with the number of drinks served ($R^2 = 0.25, p < 0.0001$), while there was no correlation between the number of drinks served and the perceived precision. The bottom row of the table summarises the responses to the final question assessing overall satisfaction with the interaction; and here, the responses for the baseline system were significantly higher than those for the uncertainty-enhanced version.

To test what aspects of the uncertainty-enhanced system affected the users' overall impression of the interaction, we carried out a stepwise multiple linear regression analysis on the subjective results as suggested by the PARADISE procedure [8]. The resulting regression equation is as follows (where $\mathcal{N}$ indicates the Z score normalisation function):

$$\begin{aligned} \text{Overall} \;=\; & 4.04 - 3.1 \cdot \mathcal{N}(\text{LastDrinkTime}) \\ & + 3.04 \cdot \mathcal{N}(\text{Duration}) + 0.91 \cdot \mathcal{N}(\text{NumDrinks}) \\ & - 0.49 \cdot \mathcal{N}(\text{Choices}) - 0.36 \cdot \mathcal{N}(\text{AskOrder}) \end{aligned}$$

In other words, participants' overall subjective scores were higher when the interaction was longer and when more drinks were served, and were lower when the robot took longer to serve all drinks, when it asked more either-or questions, and when it had to repeatedly ask for a drink order. The R^2 value for this equation is 0.235, indicating that it explains about a quarter of the variance in the overall scores. For comparison, the PARADISE analysis on the previous study found that the main contributors to overall satisfaction were the number of drinks served, the system response time, and the number of turns discarded due to low ASR, with a similar R^2 value [7].

F. Discussion

Overall, the results indicate that the baseline system was somewhat faster at serving drinks and also served more of them—however, the responses to the session questionnaire suggest that just because it served a larger number of drinks, that does not mean that it served more correct drinks. Indeed, the additional clarifications made possible by the enhanced state representation can help to avoid serving incorrect drinks. For example, the interaction fragment in Figure 3 demonstrates how the uncertainty-aware system avoids serving the wrong drink by taking into account uncertainty about a customer's order and asking clarification questions. In this same fragment, the baseline system hand would have served the wrong drink with probability 0.5: in line 14, the state contains two order hypotheses, both with confidence 0.37. Since both hypotheses exceed the 0.3 threshold used by the baseline system, it would choose randomly between the two hypotheses, a blue or a green lemonade; whereas in fact, the customer ordered a blue lemonade.

More generally, in cases where the top drink order hypothesis exceeds the 0.3 threshold but is incorrect, the baseline system would fail, whereas the uncertainty-aware system can recover from the misunderstanding. Furthermore, if the confidence of the top drink order hypothesis is in the interval $[0.1, 0.3]$, the baseline system will simply not respond, whereas the uncertainty-aware system will try to clarify the user's order. In practice, however, it turned out that often the baseline system would have served the correct drink right away, whereas the uncertainty-aware system would clarify the order first. This explains why the baseline system served more drinks but sometimes the wrong one, whereas the uncertainty-aware system almost never served the wrong drink, but sometimes did not serve a drink at all, because it failed to accumulate sufficient confidence through clarifications and the user lost patience.

Obviously, the choice of confidence thresholds in selecting response actions plays a vital role. The thresholds used in this study (Table I) were determined empirically but somewhat arbitrarily; it might be that other thresholds would have been more favourable to the uncertainty-aware system. Since tuning such thresholds manually is tedious, in future work we plan to use data-driven methods in which the optimal thresholds are found automatically.

VII. Related work

Recent work in HRI and situated multimodal interaction has seen an increasing interest in handling uncertainty. In particular, the concept of *Value of Information* has been studied as a basis for a system to decide whether to act on the current evidence from multi-sensory data, or to wait for additional information [11]. In [12], an approach to selecting clarification questions is taken, aimed at maximising the reduction of entropy. Ours is a basic approach uses both entropy and top hypothesis confidence scores as criteria for decision making, but does not involve predictions about such measures of uncertainty. However, our aim is to use the data collected in our evaluation to automatically learn an optimal action selection policy as discussed in Section VI-F.

VIII. Conclusions

In this paper we have presented results from a real user evaluation of a robot bartender system which handles uncertainty in speech input by using belief-tracking and generation of clarification questions. This uncertainty-aware system consists of a combination of state estimation and action selection components in which uncertainty due to uncertain input is tracked and exploited. In the evaluation this system was compared to a baseline version that uses standard speech recognition confidence score thresholds instead of belief-tracking and no clarifications.

On the positive side, the results suggest that users are served fewer incorrect drinks when the uncertainty-aware system is used. However, the uncertainty-aware also often unnecessarily clarified the user's order where the baseline system would have served the correct drink right away. Since the deployed confidence thresholds for the decision making process are very hard to tune manually, we plan to use the data collected in this evaluation to automatically optimise an action selection policy that takes into account the uncertainty in the state. Building on previous work on using reinforcement learning for optimising action selection strategies for multi-user human-robot interaction, a learned strategy will have incorporated the optimal thresholds automatically.

Acknowledgment

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 270435, JAMES: Joint Action for Multimodal Embodied Social Systems (james-project.eu).

References

[1] M. White, "Efficient realization of coordinate structures in Combinatory Categorial Grammar," *Research on Language and Computation*, vol. 4, no. 1, pp. 39–75, 2006.

[2] M. E. Foster, A. Gaschler, and M. Giuliani, "How can I help you? Comparing engagement classification strategies for a robot bartender," in *Proceedings of ICMI*, 2013.

[3] R. P. A. Petrick and M. E. Foster, "Planning for social interaction in a robot bartender domain," in *Proceedings of ICAPS*, 2013.

[4] M. Pateraki, M. Sigalas, G. Chliveros, and P. Trahanias, "Visual human-robot communication in social settings," in *Proceedings of ICRA Workshop on Semantics, Identification and Control of Robot-Human-Environment Interaction*, 2013.

[5] Z. Wang and O. Lemon, "A simple and generic belief tracking mechanism for the dialog state tracking challenge: On the believability of observed information," in *Proceedings of SIGDial*, 2013.

[6] M. E. Foster, S. Keizer, and O. Lemon, "Towards action selection under uncertainty for a socially aware robot bartender," in *Proceedings of HRI*, 2014.

[7] S. Keizer, M. E. Foster, O. Lemon, A. Gaschler, and M. Giuliani, "Training and evaluation of an MDP model for social multi-user human-robot interaction," in *Proceedings of SIGdial*, 2013.

[8] M. Walker, C. Kamm, and D. Litman, "Towards developing general models of usability with PARADISE," *Natural Language Engineering*, vol. 6, no. 3&4, pp. 363–377, 2000.

[9] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots," *International Journal of Social Robotics*, vol. 1, no. 1, pp. 71–81, Jan. 2009.

[10] M. Giuliani, R. P. A. Petrick, M. E. Foster, A. Gaschler, A. Isard, M. Pateraki, and M. Sigalas, "Comparing task-based and socially intelligent behaviour in a robot bartender," in *Proceedings of ICMI*, 2013.

[11] S. Rosenthal, D. Bohus, E. Kamar, and E. Horvitz, "Look versus leap: computing value of information with high-dimensional streaming evidence," in *Proceedings of IJCAI*, 2013.

[12] R. Deits, S. Tellex, P. Thaker, D. Simeonov, T. Kollar, and N. Roy, "Clarifying commands with information-theoretic human-robot dialog," *Journal of Human-Robot Interaction*, vol. 2, no. 2, pp. 58–79, 2013.