DELIVERABLE 7.4

EXTENDED SYSTEM EVALUATION: UNCERTAIN CONDITIONS

Mary Ellen Foster (HWU), Andre Gaschler (FORTISS), Manuel Giuliani (FORTISS), Amy Isard (UEDIN), Simon Keizer (HWU), Maria Pateraki (FORTH), Ron Petrick (UEDIN), Markos Sigalas (FORTH), Zhuoran Wang (HWU)

Beneficiaries: UEDIN (lead), HWU, FORTISS, FORTH
Workpackage: 7
Description: This deliverable focuses on the modelling and exploitation of uncertainty in the complete JAMES system, with a focus on testing the extended representations of uncertainty developed across all core system components in WP1–WP6, following the experiments reported in Deliverable D7.3. We summarise how uncertainty is represented and used throughout the system, and then report on experiments addressing the addition of uncertainty.

Version: Final
Nature: Report (R)
Dissemination level: Public (PU)
Pages: 17
Date: 2014-07-31
# 1 Summary

The main objective of WP7 is to coordinate all integration activities across the consortium, for implementation on the JAMES robot platform. In particular, WP7 addresses four main tasks: developing a set of interfaces and a runtime environment in which all the technical components of JAMES are able to communicate with each other (Task 7.1), coordinating system development, integration, and testing of the modules produced by the main technical workpackages WP1–6 (Task 7.2), supporting data collection and evaluation studies (Task 7.3), and conducting user evaluations on all versions of the JAMES system (Task 7.4).

The primary task that was active during the final six months of the project was Task 7.4: making use of the system components delivered at Month 36 and described in Deliverable 7.3, the project carried out additional studies centred around the problem of reasoning under uncertainty in the JAMES operating environment. This report brings together details regarding uncertainty processing across all areas of the JAMES system, some of which have been reported elsewhere in other JAMES reports.

# 2 System description

The final architecture of the JAMES bartender system is shown in Figure 1.

![Figure 1: JAMES architecture incorporating Planner/Execution Monitor and Social Skills Executor](image)

Details of all components are given in Deliverable 7.2, Deliverable 7.3, and the respective Month 36 deliverables describing individual core components; no significant changes have been made to the core JAMES components since that time. In the remainder of this section, we summarise how uncertainty is represented in the system’s internal state and how the reasoning components are able to make use of available uncertainty measures.

## 2.1 Uncertainty arising from the vision system

Uncertainty measures for modules of the vision system have been integrated in the last period of the JAMES project, and further used by the state manager components. More specifically, the 3D coordinates of the centroid of the face and hand hypotheses of each agent in the scene, as well as of the object hypotheses are associated...
with a 3x3 covariance matrix as a means of uncertainty. Since tracking is initially performed in 2D, a 2x2 covariance matrix is extracted for each hypothesis, and the 3D coordinates are estimated from corresponding hypotheses centroids in the stereo images (see also D1.1). Since the uncertainty is known in 2D and the 3D coordinates are derived from triangulation, thus using a known function, we formulate the estimation of the 3x3 covariance matrix as a problem of error propagation. Measurements in 2D are considered to be a random vector $x_0$ consisting of image observations in both images with an associated covariance $C_{xx}$ and from those measurements a resulting vector $y_0 = f(x_0)$ in $\mathbb{R}^3$ of 3D measurements is computed. Using a Taylor expansion of $f$ in the vicinity of $x_0$ and following a probabilistic interpretation the uncertainty in 3D by means of the 3x3 covariance matrix $C_{yy}$ is computed according to:

$$C_{yy} = D f(x_0) \cdot C_{xx} \cdot D f(x_0)^T$$

where $D f(x_0)$ is the derivative of $f$ at $x_0$, a 3x4 Jacobian matrix.

### 2.2 Uncertainty arising from linguistic processing

Speech recognition in JAMES is performed using the Microsoft Kinect and the associated Microsoft Speech API for speech recognition in both English and German. The output from the speech recogniser is an ordered list of possible hypotheses, each with a confidence value and an angle of input, which are passed to the natural language interpretation module. Natural language interpretation is then carried out using a bi-directional OpenCCG grammar and an ontology of communicative acts. The result is an $n$-best list of speech hypotheses consisting of a list of communicative action types, associated confidence values for the original speech strings, and the strings that were parsed. The resulting list is sent to the state manager for further processing. More details of this process can be found in Deliverable D2.2.

### 2.3 Representing uncertainty in the state manager

The input provided by the vision and speech processing components to the state manager is noisy and uncertain, so there is an inherent uncertainty about the social state that should be taken into account when choosing actions to perform. However, for simplicity in the previous version of the system, the initial state representation [4] stored only the most likely representation, with no information about the associated confidence. This simplified the initial action-selection task considerably, but also discarded much of the relevant information from the input sensors. We have therefore extended the initial version of the state manager to associate each hypothesis with a confidence score, and to include alternative hypotheses about a customer’s drink order. Initial details of this representation were first presented in Deliverable D3.2 and updated in [2]; an image of the state manager GUI showing a sample state is given in Section 2.3. We summarise the main changes below.

Incorporating multiple hypotheses and confidence scores into the state requires additional processing in the state manager. For vision, we use the values reported by the JAMES computer vision system, which 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. 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; in the current bartender set-up, the ASR speech angle generally determines a single 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 Wang and Lemon [5], which maintains beliefs over user goals based on a small number of domain-independent rules, using basic probability operations. 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.
2.4 Uncertainty reasoning in the social skills executor

The Social Skills Executor has been extended to make use of the updated state representation provided by the state manager. In particular, we have extended the initial representation to include features representing uncertainty information such as the probability of the top hypothesis, the entropy of the state distribution, and the $n$-best list of possible drink orders and their confidence scores. The set of possible system actions has also been extended to include dialogue acts for clarifying drink orders (e.g., “Did you say coke?”, “Did you say blue lemonade or green lemonade?”). The SSE’s decision making process has also been extended with rules for taking such actions: in summary, the criteria for selecting such a clarification action depend on both the confidence of the top drink order hypothesis and the overall entropy of the set of drink possibilities. Initial details of the modified SSE were presented in Deliverable D5.2 and further expanded in [3].

2.5 Uncertainty reasoning in the planner/execution monitor

The high-level planner has also improved its ability to plan using uncertain state conditions reported by the state manager. Since the previous version of the system discarded any state hypothesis other than the top hypothesis, potentially high-likelihood alternatives to the top hypothesis were unavailable to the planner, raising the possibility of less effective (or incorrect) action choices by the planner during plan construction. As a result, we have updated the planner to directly represent the top $n$ state hypotheses from the state manager in the planner’s internal knowledge state representation.

In practical terms, we currently consider uncertain state information on a property by property (feature by feature) basis, by considering the top $n$ mappings for each domain property that account for a significant probability mass in terms of the hypotheses’ associated confidence measures, i.e., \( \{ \langle h_1, c_1 \rangle, \langle h_2, c_2 \rangle, \ldots, \langle h_n, c_n \rangle \} \), such that \( \sum_{i=1}^{n} c_i > \theta \), where \( \theta \) is an empirically defined confidence threshold. At the planning level, such disjunctive state information is represented in the planner’s knowledge state as an “exclusive or” formula of the form \( (h_1|h_2|\ldots|h_n) \), whose disjunctive alternative are ordered from the highest \( h_1 \) to lowest \( h_n \) confidence. Since
the underlying planning system cannot work directly with probabilistic information, the actual confidence measures are discarded. Once such information is available to the planner, it can automatically be used during plan generation without change to the underlying planning algorithm. In practice, such knowledge typically has the effect of introducing additional sensing actions into a plan, to disambiguate between disjunctive alternatives. To aid this process, we also make use of the new domain actions used by the extended SSE, which correspond to information-gathering or clarification acts that the robot can employ to help clarify uncertain beliefs. The details of the planner modifications are given in [1].

3 Evaluations

3.1 Evaluating visual input processing under uncertainty

To obtain estimates on the uncertainty of the torso orientation estimation, with a single confidence value in the range of 0-1, we used data gathered off-line during the quantitative evaluation of the torso orientation in a series of sequences acquired in the lab. The experiments involved different users conducting several poses in various positions relative to the camera positions and orientations. Furthermore, to obtain ground truth information we attached markers on the shoulders of the users.

We stored the quantitative evaluation results in a confusion matrix that encoded the estimated orientation around the y axis for each measured orientation in the range of 0±40 degrees with a step of 5 degrees. The percentages were derived from image sequences of 10,000 image frames and the exact values during online processing are found by interpolation. Moreover, the angular distance from closest person or robot provides a confidence towards which person/or the robot itself the specific user orients his torso.

Similarly, the uncertainty for face orientation estimation is encoded by a confidence score derived by quantitatively analysing the estimated face orientations and intended face orientations in the range of 0±180 degrees. For this analysis a setup was used with points of interest (POIs) placed around the user at equal angular distances of 10 degrees. The user was asked then to look and turn his head at specific POIs and the percentages were derived from image sequences of a total of 7000 image frames. The resulting confusion matrix (Figure 4) encoded the perceived orientation for each intended orientation. As can be seen, the algorithm achieves high success rates for small angles (i.e., the user looks in directions close to the direction of the camera) which are decreased for larger angles (i.e., the user looks away from the camera). The algorithm is able to maintain significant success rates (more than 50%) even for angles up to 120 degrees, where only a small part of the facial patch is visible. Therefore, in the vision module the confidence score for face orientation is extracted only in the range of 0±110 deg.
3.2 Evaluating social skills execution under uncertainty

We carried out a user study testing the above modifications to the SSE, making use of the extended states provided by the SSR. In summary, the results of the study indicated that that the baseline system (without uncertainty) 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 the system avoid serving incorrect drinks. Details of that study are presented in [3].

3.3 Evaluating plan generation under uncertainty

Two evaluations of the planning system were proposed to evaluate the high-level planning component under the various types of uncertainty that could arise from the state manager. The first study was an experiment to explore the scalability (i.e., efficiency) of the plan generation process under conditions of increasing uncertainty and information-gathering clarification (sensing) actions in typical JAMES-style domains. The resulting plans in this study mapped to a common type of contingent plan structure where the number of branches in the plan increased with the number of sensing actions required, decreasing the efficiency of plan generation. The addition of domain-dependent control knowledge could be used to help offset this problem. Also, short interactions that were common to the JAMES domain did not pose a problem. For instance, Figure 5 shows an example of a typical clarification scenario in the JAMES domain. The first evaluation is complete and we are currently preparing a paper on this study.

A second evaluation (a user study) is also planned to verify the quality of plans generated by the planner under uncertain conditions. One observation from the first study described above is that (limited) control knowledge
Figure 5: Plan generation with uncertain information typically gives rise to plans with sensing actions that are used to clarify state properties. The plan on the left is the usual first plan generated by the planner for asking a customer for a drink order and serving the drink. The plan on the right is a modified plan resulting from a replanning operation after the state manager was unable to determine if the customer asked for a blue lemonade or a green lemonade. A clarification action (clarify-lemonade) is added to the revised plan, followed by a pair of conditional branches which act as explicit reasoning points in the plan to resolve the uncertainty.

sometimes appears to be necessary to produce “natural” interactions of the kind humans would use, especially when multiple clarification acts are required. One of the goals of this study is to determine to what extent such plans are required, and whether “less natural” plans have an effect on the perceived or actual interaction quality in typical JAMES scenarios. Due to time constraints, this study is planned to take place shortly after the end of the project.

4 References


A Included papers

The following two papers are included in this deliverable. Note that initial drafts of these papers were included in previous deliverables (D5.2 and D3.2, respectively); here, we include the final versions of both papers.


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.


Abstract: A robot coexisting with humans must not only be able to perform physical tasks, but must also be able to interact with humans in a socially appropriate manner. In this paper, we describe an extension of prior work on planning for task-based social interaction using a robot that must interact with multiple human agents in a simple bartending domain. We describe how the initial state representation developed for this robot has been extended to handle the full range of uncertainty resulting from the input sensors, and outline how the planner will use the resulting uncertainty in the state during plan generation.
Handling uncertain input in multi-user human-robot interaction

Simon Keizer¹, Mary Ellen Foster¹, Andre Gaschler², Manuel Giuliani², Amy Isard³, and Oliver Lemon¹

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.

II. ROBOT BAR TENDER 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.
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 ($\text{scconf} \_\text{thrc}$ and $\text{sconf} \_\text{thru} \text{nc}$ in Table I).

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

$$\text{if} \ (\ \text{conf} \geq \text{conf} \_\text{thrc} \_1 \ ) \ \text{or} \ (\ \text{conf} \geq \text{conf} \_\text{thrc} \_2 \ \text{and} \ \text{entr} < \text{entr} \_\text{thrc}) \ \text{then}$$

select action based on top hypothesis;

(e.g., “Okay, a coke”)

$$\text{else if} \ \text{there is only one drink order hypothesis then}$$

confirm the drink order with the user;

(e.g., “Did you say ‘coke’?”)

$$\text{else}$$

let user choose between top 2 hypotheses;

(e.g., “Did you say ‘green’ or ‘blue’ lemonade?”)

$$\text{end if}$$

<table>
<thead>
<tr>
<th>Description</th>
<th>Threshold</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper confidence threshold</td>
<td>$\text{conf} _\text{thrc} _1$</td>
<td>0.65</td>
</tr>
<tr>
<td>Lower confidence threshold</td>
<td>$\text{conf} _\text{thrc} _2$</td>
<td>0.40</td>
</tr>
<tr>
<td>Entropy threshold</td>
<td>$\text{entr} _\text{thrc}$</td>
<td>0.25</td>
</tr>
<tr>
<td>Parsing confidence threshold (baseline)</td>
<td>$\text{scconf} _\text{thrc}$</td>
<td>0.30</td>
</tr>
<tr>
<td>Parsing confidence threshold (uncertainty)</td>
<td>$\text{sconf} _\text{thrc} _\text{unc}$</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Figure 3 shows an interaction with the uncertainty-aware system from our user study in 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 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.

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
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

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]

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 and sconf 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:

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

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

Fig. 6. Questionnaire for each session.
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: Animorphism (questions 1 and 4), Animacy (question 6), Liking (questions 2, 5, 7, and 8), and Perceived Intelligence (questions 1 and 4). 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 (α > 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 Animorphism 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-enabled 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 $N$ indicates the Z score normalisation function):

\[
\text{Overall} = 4.04 - 3.1 \cdot N(\text{LastDrinkTime}) + 3.04 \cdot N(\text{Duration}) + 0.91 \cdot N(\text{NumDrinks}) - 0.49 \cdot N(\text{Choices}) - 0.36 \cdot N(\text{AskOrder})
\]

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].

**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 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

Planning for Social Interaction with Sensor Uncertainty

Mary Ellen Foster  
School of Mathematical and Computer Sciences  
Heriot-Watt University  
Edinburgh EH14 4AS, Scotland, UK  
M.E.Foster@hw.ac.uk

Ronald P. A. Petrick  
School of Informatics  
University of Edinburgh  
Edinburgh EH8 9AB, Scotland, UK  
rpetrick@inf.ed.ac.uk

Abstract

A robot coexisting with humans must not only be able to perform physical tasks, but must also be able to interact with humans in a socially appropriate manner. In this paper, we describe an extension of prior work on planning for task-based social interaction using a robot that must interact with multiple human agents in a simple bartending domain. We describe how the initial state representation developed for this robot has been extended to handle the full range of uncertainty resulting from the input sensors, and outline how the planner will use the resulting uncertainty in the state during plan generation.

Introduction

A crucial aspect in the design of an interactive system is state management: transforming the noisy, continuous hypotheses produced by the low-level input processing components into a form that can be used as the basis for higher-level action selection by a component such as a planner. Intuitively, states represent a point of intersection between low-level sensor data and the high-level structures used for action selection. Since states are induced from the mapping of sensor observations to property values, the challenge of building an effective state manager rests on defining appropriate mapping functions. A state representation that considers only the highest-confidence inputs is straightforward to maintain and reason with, but discards a great deal of potentially useful information. On the other hand, a representation that takes into account the full set of input possibilities—along with their estimated confidence scores—can be more robust and informative, but requires more sophisticated methods of maintenance and more complex forms of reasoning and planning.

The particular application we consider here is a robot bartender called JAMES (Figure 1), which has the goal of supporting socially appropriate multi-party interaction in a bartending scenario. In particular, the robot’s sensors monitor two primary input modalities: vision and speech. Based on observations about the agents in the bar provided by these sensors, the system maintains a model of the social context, and decides on effective and socially appropriate responses in that context. Key to our approach is the use of a high-level planner for action selection in the robot system, in the place of a traditional interaction manager (Larsson and Traum 2000). Specifically, we use the knowledge-level planner PKS (Petrick and Bacchus 2002; 2004), a choice that is motivated by PKS’s ability to work with incomplete information and sensing actions, since the robot will often have to gather information from its environment (e.g., by asking a customer for a drink order) in addition to performing physical tasks such as handing over drinks.

In this paper, we describe how the initial, deterministic state representation has been extended to incorporate the full data from the robot’s input sensors, and how the planner is using this enhanced representation during plan generation.

State Management with Uncertain Input

The task of the state manager in the robot bartender system is to keep track of information about the agents in the scene: for example, their locations, whether they are currently seeking the bartender’s attention, and their drink orders. The state is derived from the continuous stream of messages produced by the low-level input and output components. In addition to storing low-level sensor information, we also infer additional relations not directly reported by the sensors; for example, we fuse 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 (Foster, Gaschler, and Giuliani 2013).

Since the input provided by the vision and speech processing components is uncertain, there is an inherent uncertainty about the state. However, for simplicity, the state representation used in the initial JAMES system (Petrick and Foster 2013) stored only the highest-probability hypotheses, with no
Table 1: State excerpt, showing both the old discrete representation (highlighted portion) and the new representation

<table>
<thead>
<tr>
<th>Action</th>
<th>Preconditions</th>
<th>Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>seeksAttention(A1)</td>
<td>true</td>
<td>0.75</td>
</tr>
<tr>
<td>seeksAttention(A2)</td>
<td>false</td>
<td>0.45</td>
</tr>
<tr>
<td>lastSpeaker()</td>
<td>A1</td>
<td>1.0</td>
</tr>
<tr>
<td>lastEvent()</td>
<td>userSpeech(A1)</td>
<td>1.0</td>
</tr>
<tr>
<td>drinkOrder(A1)</td>
<td>blue lemonade</td>
<td>0.677</td>
</tr>
<tr>
<td>lastAct(A1)</td>
<td>greet</td>
<td>0.322</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: State excerpt, showing both the old discrete representation (highlighted portion) and the new representation

<table>
<thead>
<tr>
<th>Action</th>
<th>Preconditions</th>
<th>Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>seeksAttention(A1)</td>
<td>true</td>
<td>0.75</td>
</tr>
<tr>
<td>seeksAttention(A2)</td>
<td>false</td>
<td>0.45</td>
</tr>
<tr>
<td>lastSpeaker()</td>
<td>A1</td>
<td>1.0</td>
</tr>
<tr>
<td>lastEvent()</td>
<td>userSpeech(A1)</td>
<td>1.0</td>
</tr>
<tr>
<td>drinkOrder(A1)</td>
<td>blue lemonade</td>
<td>0.677</td>
</tr>
<tr>
<td>lastAct(A1)</td>
<td>greet</td>
<td>0.322</td>
</tr>
</tbody>
</table>

Figure 2: Example PKS action in the bartender domain

Planning under Sensor Uncertainty

To generate plans for the robot, PKS uses a knowledge-level domain model that includes a specification of the physical, sensory, and linguistic (speech) actions available to it. The current domain supports simple interactions with individual agents for ordering drinks from the robot, as well as socially motivated behaviour such as group ordering and multi-party turn-taking. For example, Figure 2 shows the PKS representation for the ask-drink(?a) action ("ask an agent ?a for a drink order"), which is modelled as a sensing action that returns a placeholder (the function drinkOrder) for information that will become known at execution time.

We are currently improving our ability to plan with sensor uncertainty in the states described above. Since PKS does not (currently) work directly with probabilistic representations, we are modelling disjunctive state information like drinkOrder in Table 1 using PKS's ability to use "exclusive or" formula of the form (\(\phi_1 \lor \phi_2 \lor \cdots \lor \phi_n\)) (which is interpreted as "one, and only one, of the \(\phi_i\)s is true").

Acknowledgements

This research has received funding from the European Union's 7th Framework Programme under grant No. 270435.

References


