

Supporting Spoken Assistant Systems with a Graphical User Interface that Signals Incremental Understanding and Prediction State

Casey Kennington and David Schlangen

Boise State University and DSG / CITEC / Bielefeld University
caseykennington@boisestate.edu and david.schlangen@uni-bielefeld.de

Abstract

Arguably, spoken dialogue systems are most often used not in hands/eyes-busy situations, but rather in settings where a graphical display is also available, such as a mobile phone. We explore the use of a graphical output modality for signalling incremental understanding and prediction state of the dialogue system. By visualising the current dialogue state and possible continuations of it as a simple tree, and allowing interaction with that visualisation (e.g., for confirmations or corrections), the system provides both feedback on past user actions and guidance on possible future ones, and it can span the continuum from slot filling to full prediction of user intent (such as GoogleNow). We evaluate our system with real users and report that they found the system intuitive and easy to use, and that incremental and adaptive settings enable users to accomplish more tasks.

1 Introduction

Current virtual personal assistants (PAs) require users to either formulate complex intents in one utterance (e.g., “call Peter Miller on his mobile phone”) or go through tedious sub-dialogues (e.g., “phone call” – *who would you like to call?* – “Peter Miller” – *I have a mobile number and a work number. Which one do you want?*). This is not how one would interact with a human assistant, where the request would be naturally structured into smaller chunks that individually get acknowledged (e.g., “Can you make a connection for me?” – *sure* – “with Peter Miller” – *uh huh* – “on his mobile” – *dialling now*). Current PAs signal ongoing understanding by displaying the state of

the recognised speech (ASR) to the user, but not their semantic interpretation of it. Another type of assistant system forgoes enquiring user intent altogether and infers *likely* intents from context. GoogleNow, for example, might present traffic information to a user picking up their mobile phone at their typical commute time. These systems display their “understanding” state, but do not allow any type of interaction with it apart from dismissing the provided information.

In this work, we explore adding a graphical user interface (GUI) modality that makes it possible to see these interaction styles as extremes on a continuum, and to realise positions between these extremes and present a mixed graphical/voice enabled PA that can provide feedback of understanding to the user *incrementally* as the user’s utterance unfolds—allowing users to make requests in instalments instead of fully thought-out requests. It does this by signalling ongoing understanding in an intuitive tree-like GUI that can be displayed on a mobile device. We evaluate our system by directing users to perform tasks using it under non-incremental (i.e., ASR endpointing) and incremental conditions and then compare the two conditions. We further compare a non-adaptive with an adaptive (i.e., infers likely events) version of our system. We report that the users found the interface intuitive and easy to use, and that users were able to perform tasks more efficiently with incremental as well as adaptive variants of the system.

2 Related Work

This work builds upon several threads of previous research: Chai et al. (2014) addressed misalignments in understanding (i.e., common ground (Clark and Schaefer, 1989)) between robots and humans by informing the human of the internal system state via speech. We take this idea and ap-

ply it to a PA by displaying the internal state of the system to the user via a GUI (explained in Section 3.5), allowing the user to determine if system understanding has taken place—a way of providing feedback and backchannels to the user. Dethlefs et al. (2016) provide a good review of work that show how backchannels facilitate grounding, feedback, and clarifications in human spoken dialogue, and apply an *information density* approach to determine when to backchannel using speech. Because we don't backchannel using speech here, there is no potential overlap between the user and the system; rather, our system can display backchannels and ask clarifications without frustrating the user through inadvertent overlaps.

Though different in many ways, our work is similar in some regards to Larsson et al. (2011), which displays information to the user and allows the user to navigate the display itself (e.g., by saying *up* or *down* in a menu list)—functionality that we intend to apply to our GUI in future work. Our work is also comparable to SDS toolkits such as IrisTK (Skantze and Moubayed, 2012) and OpenDial (Lison, 2015) which enable SDS designers to visualise the internal state of their systems, though not for end user interpretability.

Some of the work here is inspired by the *Microsoft Language Understanding Intelligent Service* (LUIS) project (Williams et al., 2015). While our system by no means achieves the scale that LUIS does, we offer here an additional contribution of an open source LUIS-like system (with the important addition of the graphical interface) that is authorable (using JSON files; we leave authoring using a web interface like that of LUIS to future work), extensible (affordances can be easily added), incremental (in that respect going beyond LUIS), trainable (i.e., can learn from examples, but can still function well without examples), and can learn through interacting (here we apply a user model that learns during interaction).

3 System Description

This section introduces and describes our SDS, which is modularised into four main components: ASR, natural language understanding (NLU), dialogue management (DM), and the graphical user interface (GUI) which, as explained below, is visualised as a right-branching tree. The overall system is represented in Figure 1. For the remainder of this section, each module is explained in

turn. As each module processes input incrementally (i.e., word for word), we first explain our framework for incremental processing.

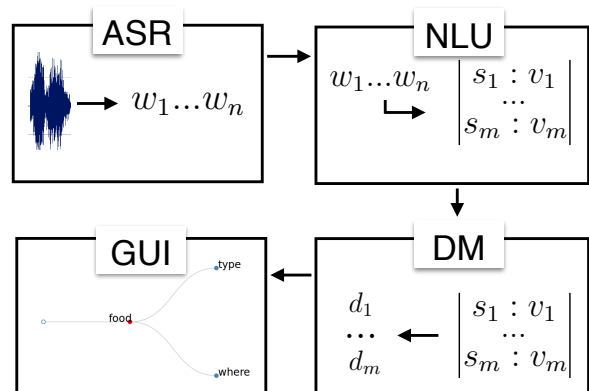


Figure 1: Overview of system made up of ASR which takes in a speech signal and produces transcribed words, NLU, which takes words and produces a slots in a frame, DM which takes slots and produces a decision for each, and the GUI which displays the state of the system.

3.1 Incremental Dialogue

An aspect of our SDS that sets it apart from others is the requirement that it process *incrementally*. One potential concern with incremental processing is regarding informativeness: why act early when waiting might provide additional information, resulting in better-informed decisions? The trade off is *naturalness* as perceived by the user who is interacting with the SDS. Indeed, it has been shown that human users perceive incremental systems as being more natural than traditional, turn-based systems (Aist et al., 2006; Skantze and Schlangen, 2009; Skantze and Hjalmarsson, 1991; Asri et al., 2014), offer a more human-like experience (Edlund et al., 2008) and are more satisfying to interact with than non-incremental systems (Aist et al., 2007). Psycholinguistic research has also shown that humans comprehend utterances as they unfold and do not wait until the end of an utterance to begin the comprehension process (Tanenhaus et al., 1995; Spivey et al., 2002).

The trade-off between informativeness and naturalness can be reconciled when mechanisms are in place that allow earlier decisions to be repaired. Such mechanisms are offered by the incremental unit (IU) framework for SDS (Schlangen and Skantze, 2011), which we apply here. Following Kennington et al. (2014), the IU framework consists of a network of processing *modules*. A typical module takes input, performs some kind of processing on that data, and produces output.

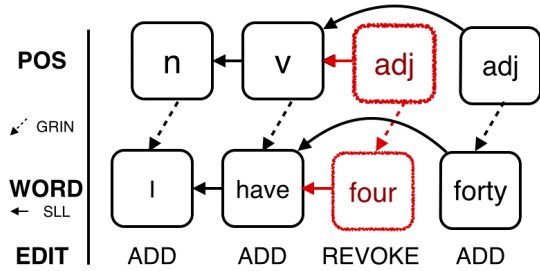


Figure 2: Example of IU network; part-of-speech tags are grounded into words, tags and words have same level links with left IU; *four* is revoked and replaced with *forty*.

The data are packaged as the payload of *incremental units* (IUs) which are passed between modules. The IUs themselves are interconnected via so-called *same level links* (SLL) and *grounded-in links* (GRIN), the former allowing the linking of IUs as a growing sequence, the latter allowing that sequence to convey what IUs directly affect it (see Figure 2 for an example of incremental ASR). Thus IUs can be *added*, but can be later *revoked* and replaced in light of new information. The IU framework can take advantage of up-to-date information, but have the potential to function in such a way that users perceive as more natural.

The modules explained in the remainder of this section are implemented as IU-modules and process incrementally. Each will now be explained.

3.2 Speech Recognition

The module that takes speech input from the user in our SDS is the ASR component. Incremental ASR must transcribe uttered speech into words which must be forthcoming from the ASR as early as possible (i.e., the ASR must not wait for end-pointing to produce output). Each module that follows must also process incrementally, acting in lock-step upon input as it is received. Incremental ASR is not new (Baumann et al., 2009) and many of the current freely-accessible ASR systems can produce output (semi-) incrementally. We opt for Google ASR for its vocabulary coverage of our evaluation language (German). Following, Baumann et al. (2016), we package output from the Google service into IUs which are passed to the NLU module, which we now explain.

3.3 Language Understanding

We approach the task of NLU as a slot-filling task (a very common approach; see Tur et al. (2012)) where an intent is complete when all slots of a *frame* are filled. The main driver of the NLU in

our SDS is the SIUM model of NLU introduced in Kennington et al. (2013). SIUM has been used in several systems which have reported substantial results in various domains, languages, and tasks (Han et al., 2015; Kennington et al., 2015; Kennington and Schlangen, 2017) Though originally a model of reference resolution, it was always intended to be used for general NLU, which we do here. The model is formalised as follows:

$$P(I|U) = \frac{1}{P(U)} P(I) \sum_{r \in R} P(U|R=r) P(R=r|I) \quad (1)$$

That is, $P(I|U)$ is the probability of the intent I (i.e., a frame slot) behind the speaker’s (ongoing) utterance U . This is recovered using the mediating variable R , a set of *properties* which map between aspects of U and aspects of I . We opt for abstract properties here (e.g., the frame for restaurant might be filled by a certain type of cuisine intent such as *italian* which has properties like *pasta*, *mediterranean*, *vegetarian*, etc.). Properties are pre-defined by a system designer and can match words that might be uttered to describe the intent in question. For $P(R|I)$, probability is distributed uniformly over all properties that a given intent is specified to have. (If other information is available, more informative priors could be used as well.) The mapping between properties and aspects of U can be learned from data. During application, R is marginalised over, resulting in a distribution over possible intents.¹ This occurs at each word increment, where the distribution from the previous increment is combined via $P(I)$, keeping track of the distribution over time.

We further apply a simple rule to add in a-priori knowledge: if some $r \in R$ and $w \in U$ are such that $r \doteq w$ (where \doteq is string equality; e.g., an intent has the property of *pasta* and the word *pasta* is uttered), then we set $C(U=w|R=r)=1$. To allow for possible ASR confusions, we also apply $C(U=w|R=r)=1 - ld(w,r)/\max(len(w),len(r))$, where *ld* is the *Levenshtein distance* (but we only apply this if the calculated value is above a threshold of 0.6; i.e., the two strings are mostly similar). For all other w , $C(w|r)=0$. This results in a distribution C , which we normalise and blend with learned distribution to yield $P(U|R)$.

¹In Kennington et al. (2013) the authors apply Bayes’ Rule to allow $P(U|R)$ to produce a distribution over properties, which we adopt here.

We apply an instantiation of SIUM for each slot. The candidate slots which are processed depends on the state of the dialogue; only slots represented by visible nodes are considered, thereby reducing the possible frames that could be predicted. At each word increment, the updated slots (and their corresponding) distributions are given to the DM, which will now be explained.

3.4 Dialogue Manager

The DM plays a crucial role in our SDS: as well as determining *how* to act, the DM is called upon to decide *when* to act, effectively giving the DM the control over timing of actions rather than relying on ASR endpointing—further separating our SDS from other systems. The DM policy is based on a confidence score derived from the NLU (in this case, we used the distribution’s argmax value) using thresholds for the actions (see below), set by hand (i.e., trial and error). At each word and resulting distribution from NLU, the DM needs to choose one of the following:

- `wait` – wait for more information (i.e., for the next word)
- `select` – as the NLU is confident enough, fill the slot can with the argmax from NLU
- `request` – signal a (yes/no) clarification request on the current slot and the proposed filler
- `confirm` – act on the confirmation of the user; in effect, `select` the proposed slot value

Though the thresholds are statically set, we applied OpenDial (Lison, 2015) as an IU-module to perform the task of the DM with the future goal that these values could be adjusted through reinforcement learning (which OpenDial could provide). The DM processes and makes a decision for *each slot*, with the assumption that only one slot out of all that are processed will result in an non-`wait` action (though this is not enforced).

3.5 Graphical User Interface

The goal of the GUI is to intuitively inform the user about the internal state of the ongoing understanding. One motivation for this is that the user can determine if the system understood the user’s intent before providing the user with a response

(e.g., a list of restaurants of a certain type); i.e., if any misunderstanding takes place, it happens before the system commits to an action and is potentially more easily repaired.

The display is a right-branching tree, where the branches directly off the root node display the affordances of the system (i.e., what domains of things it can understand and do something about). When the first tree is displayed, it represents a state of the NLU where none of the slots are filled, as in Figure 3.

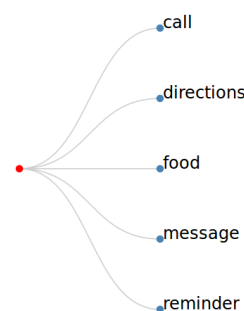


Figure 3: Example tree as branching from the root; each branch represents a system affordance (i.e., making a phone call, reminder, finding a restaurant, leaving a message, and finding a route).

When a user verbally selects a domain to ask about, the tree is adjusted to make that domain the only one displayed and the slots that are required for that domain are shown as branches. The user can then fill those slots (i.e., branches) by uttering the displayed name, or, alternatively, by uttering the item to fill the slot directly. For example, at a minimum, the user could utter the name of the domain then an item for each slot (e.g., *food Thai downtown*) or the speech could be more natural (e.g., *I’m quite hungry, I am looking for some Thai food maybe in the downtown area*). Crucially, the user can also hesitate within and between chunks, as advancement is not triggered by silence thresholding, but rather semantically. When something is uttered that falls into the `request` state of the DM as explained above, the display expands the subtree under question and marks the item with a question mark (see Figure 4). At this point, the user can utter any kind of confirmation. A positive confirmation fills the slot with the item in question. A negative confirmation retracts the question, but leaves the branch expanded. The expanded branches are displayed according to their rank as given by the NLU’s probability distribution. Though a branch in the display can theoretically display an unlimited number of children, we opted to only show 7 children; if a branch had more, the final child displayed as an ellipsis.

A completed branch is collapsed, visually marking its corresponding slot as filled. At any

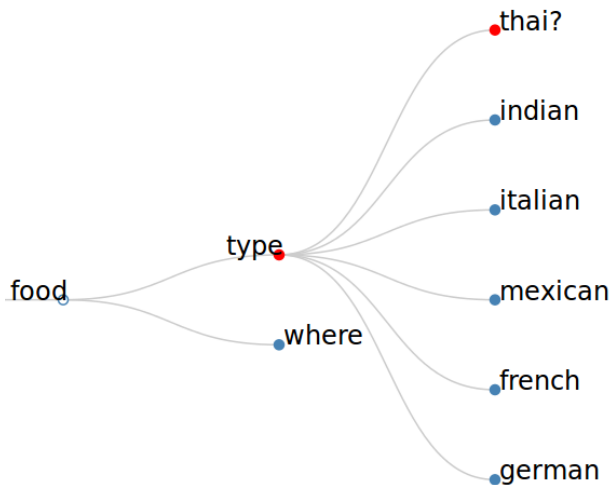


Figure 4: Example tree asking for confirmation on a specific node (in red with a question mark).

time, a user can backtrack by saying *no* (or equivalent) or start the entire interaction over from the beginning with a keyword, e.g., *restart*. To aid the user’s attention, the node under question is marked in red, where completed slots are represented by outlined nodes, and filled nodes represent candidates for the current slot in question (see examples of all three in Figure 4). For cases where the system is in the *wait* state for several words (during which there is no change in the tree), the system signals activity at each word by causing the red node in question to temporarily change to white, then back to red (i.e., appearing as a blinking node to the user). Figure 5 shows a filled frame, represented as tree with one branch for each filled slot.

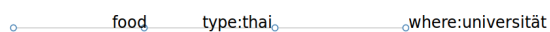


Figure 5: Example tree where all of the slots are filled. (i.e., domain:food, location:university, type:thai)

Such an interface clearly shows the internal state of the SDS and whether or not it has understood the request so far. It is designed to aid the user’s attention to the slot in question, and clearly indicates the affordances that the system has. The interface is currently a read-only display that is purely speech-driven, but it could be augmented with additional functionalities, such as tapping a node for expansion or typing input that the system might not yet display. It is currently implemented as a web-based interface (using the JavaScript D3 library), allowing it to be usable as a web application on any machine or mobile device.

Adaptive Branching The GUI as explained affords an additional straight-forward extension: in order to move our system towards adaptivity on the above-mentioned continuum, the GUI can be used to signal what the system thinks the user might say next. This is done by expanding a branch and displaying a confirmation on that branch, signalling that the system predicts that the user will choose that particular branch. Alternatively, if the system is confident that a user will fill a slot with a particular value, that particular slot can be filled without confirmation. This is displayed as a collapsed tree branch. A system that perfectly predicts a user’s intent would fill an entire tree (i.e., all slots) only requiring the user to confirm once. A more careful system would confirm at each step (such an interaction would only require the user to utter confirmations and nothing else). We applied this adaptive variant of the tree in one of our experiments explained below.

4 Experiments

In this section, we describe two experiments where we evaluated our system. It is our primary goal to show that our GUI is useful and signals understanding to the user. We also wish to show that incremental presentation of such a GUI is more effective than an endpointed system. We further want to show that an adaptive system is more effective than a non-adaptive system (though both would process incrementally). In order to best evaluate our system, we recruited participants to interact with our system in varied settings to compare endpointed (i.e., non-incremental) and non-adaptive as well as adaptive versions. We describe how the data were collected from the participants, then explain each experiment and give results.

4.1 Task & Procedure

The participants were seated at a desk and given written instructions indicating that they were to use the system to perform as many tasks as possible in the allotted time. Figure 6 shows some example tasks as they would be displayed (one at a time) to the user. A screen, tablet, and keyboard were on the desk in front of the user (see Figure 7).² The user was instructed to convey the task presented on the screen to the system such

²We used a Samsung 8.4 Pro tablet turned to its side to show a larger width for the tree to grow to the right. The tablet only showed the GUI; the SDS ran on a separate computer.

that the GUI on the tablet would have a completed tree (e.g., as in Figure 5). When the participant was satisfied that the system understood her intent, she was to press space bar on the keyboard which triggered a new task to be displayed on the screen and reset the tree to its start state on the tablet (as in Figure 3).

The possible task domains were *call*, which had a single slot for *name* to be filled (i.e., one out of the 22 most common German given names); *message* which had a slot for *name* and a slot for the *message* (which, when invoked, would simply fill in directly from the

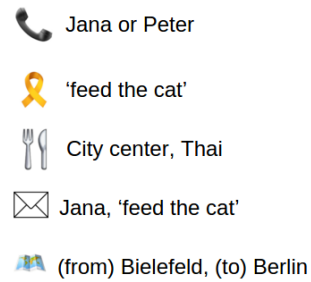


Figure 6: Examples of tasks, as presented to each participant. Each icon represents a specific task domain (i.e., call, reminder, find a restaurant, leave a message, or directions).

ASR until 1 second of silence was detected); *eat* which had slots for *type* (in this case, 6 possible types) and *location* (in this case, 6 locations based around the city of Bielefeld); *route* which had slots for *source* city and the *destination* city (which shared the same list of the top 100 most populous German cities); and *reminder* which had a slot for *message*.

For each task, the domain was first randomly chosen from the 5 possible domains, and then each slot value to be filled was randomly chosen (the *message* slot for the *name* and *message* domains was randomly selected from a list of 6 possible “messages”, each with 2-3 words; e.g., *feed the cat*, *visit grandma*, etc.). The system kept track of which tasks were already presented to the participant. At any time after the first task, the system could choose a task that was previously presented and present it again to the participant (with a 50% chance) so the user would often see tasks that she had seen before (with the assumption that humans who use PAs often do perform similar, if not the same, tasks more than once).

The participant was told that she would interact with the system in three different phases, each for 4 minutes, and to accomplish as many tasks as possible in that time allotment. The participant was not told what the different phases were. The experiments described in Sections 4.2 and

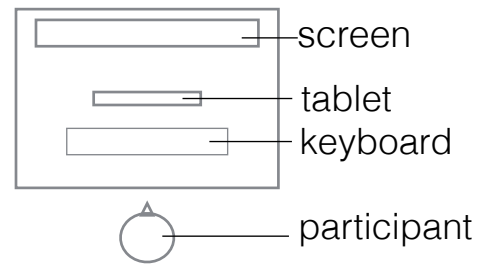


Figure 7: Bird's eye view of the experiment: the participant sat at a table with a screen, tablet, and keyboard in front of them.

4.3 respectively describe and report a comparison first between the Phase 1 and 2 (denoted as the *endpointed* and *incremental* variants of the system) in order to establish whether or not the incremental variant produced better results than the endpointed variant. We also report a comparison between Phase 2 and 3 (*incremental* and *incremental-adaptive* phases). Phase 1 and Phase 3 are not directly comparable to each other as Phase 3 is really a variant of Phase 2. Because of this, we fixed the order of the phase presentation for all participants. Each of these phases are described below. Before the participant began Phase 1, they were able to try it out for up to 4 minutes (in Phase 1 settings) and ask for help from the experimenter, allowing them to get used to the Phase 1 interface before the actual experiment began. After this trial phase, the experiment began with Phase 1.

Phase 1: Non-incremental In this phase, the system did not appear to work incrementally; i.e., the system displayed tree updates after ASR endpointing (of 1.2 seconds—a reasonable amount of time to expect a response from a commercial spoken PA). The system displayed the ongoing ASR on the tablet as it was recognised (as is often done in commercial PAs). At the end of Phase 1, a pop up window notified the user that the phase was complete. They then moved onto Phase 2.

Phase 2: Incremental In this phase, the system displayed the tree information incrementally without endpointing. The ASR was no longer displayed; only the tree provided feedback in understanding, as explained in Section 3.5.

After Phase 2, a 10-question questionnaire was displayed on the screen for the participant to fill out comparing Phase 1 and Phase 2. For each question, they had the choice of *Phase 1*, *Phase*

2, *Both*, and *Neither*. (See Appendix for full list of questions.) After completing the questionnaire, they moved onto Phase 3.

Phase 3: Incremental-adaptive In this phase, the incremental system was again presented to the participant with an added user model that “learned” about the user. If the user saw a task more than once, the user model would predict that, if the user chose that task domain again (e.g., *route*) then the system would automatically ask a clarification using the previously filled values (except for the *message* slot, which the user always had to fill). If the user saw a task more than 3 times, the system skipped asking for clarifications and filled in the domain slots completely, requiring the user only to press the space bar to confirm it was the correct one (i.e., to complete the task). An example progression might be as follows: a participant is presented with the task *route from Bielefeld to Berlin*, then the user would attempt to get the system to fill in the tree (i.e., slots) with those values. After some interaction in other domains, the user sees the same task again, and now after indicating the intent type *route*, the user must only say “yes” for each slot to confirm the system’s prediction. Later, if the task is presented a third time, when entering that domain (i.e., *route*), the two slots would already be filled. If later a different route task was presented, e.g., *route from Bielefeld to Hamburg*, the system would already have the two slots filled, but the user could backtrack by saying “no, to Hamburg” which would trigger the system to fill the appropriate slot with the corrected value. Later interactions within the *route* domain would ask for a clarification on the *destination* slot since it has had several possible values given by the participant, but continue to fill the *from* slot with *Bielefeld*.

After Phase 3, the participants were presented with another questionnaire on the screen to fill out with the same questions (plus two additional questions), this time comparing Phase 2 and Phase 3. For each item, they had the choice of *Phase 2*, *Phase 3*, *Both*, and *Neither*. At the end of the three phases and questionnaires, the participants were given a final questionnaire to fill out by hand on their general impressions of the systems.

We recruited 14 participants for the evaluation. We used the Mint tools data collection framework (Kousidis et al., 2012) to log the interactions. Due to some technical issues, one of the participants

did not log interactions. We collected data from 13 participants, post-Phase 2 questionnaires from 12 participants, post-Phase 3 questionnaires from all 14 participants, and general questionnaires from all 14 participants. In the experiments that follow, we report objective and subjective measures to determine the settings that produced superior results.

Metrics We report the subjective results of the participant questionnaires. We only report those items that were statistically significant (see Appendix for a full list of the questions). We further report objective measures for each system variant: total number of completed tasks, fully correct frames, average frame f-score, and average time elapsed (averages are taken over all participants for each variant; we only used the 10 participants who fully interacted with all three phases). Discussion is left to the end of this section.

4.2 Experiment 1: Endpointed vs. Incremental

In this section we report the results of the evaluation between the *endpointed* (i.e., non-incremental; Phase 1) variant vs the incremental (Phase 2) variant of our system.

Subjective Results We applied a multinomial test of significance to the results, treating all four possible answers as equally likely (with Bonferroni correction of 10). The item *The interface was useful and easy to understand* with the answer of *Both* was significant ($\chi^2(4, N = 12) = 9.0, p < .005$), as was *The assistant was easy and intuitive to use* also with the answer *Both* ($\chi^2(4, N = 12) = 9.0, p < .005$). The item *I always understood what the system wanted from me* was also answered *Both* significantly more times than other answers ($\chi^2(4, N = 14) = 9.0, p < .005$), similarly for *It was sometimes unclear to me if the assistant understood me* with the answer of *Both* ($\chi^2(4, N = 12) = 10.0, p < .005$). These responses tell us that though the participants did not report preference for either system variant, they reported a general positive impression of the GUI (in both variants). This is a nice result; the GUI could be used in either system with benefit to the users.

Objective Results The *endpointed* (Phase 1) and *incremental* (Phase 2) columns in Table 1 show the results of the objective evaluation. Though the average time per task and fscore for the endpointed variant are better than those of the

	endpointed	incr.	adaptive
tasks	105	122	124
frames	46	46	59
fscore	0.81	0.74	0.80
time	19.1	19.6	19.5

Table 1: Objective measures for Experiments 1 & 2: count of completed tasks, number of fully correct frames, average fscore (over all participants), and average elapsed time per task (over all participants).

incremental variant, the total number of tasks for the incremental variant was higher.

Manual inspection of logs indicate that participants took advantage of the system’s flexibility of understanding instalments (i.e., filling frames incrementally). This is evidenced in that participants often uttered words understood by the system as being negative (e.g., *nein/No*), either as a result of an explicit confirmation request by the system (e.g., *Thai?*) or after a slot was incorrectly filled (something very easily determined through the GUI). This is a desired outcome of using our system; participants were able to repair local areas of misunderstanding as they took place instead of needing to correct an entire intent (i.e., frame). However, we cannot fully empirically measure these tendencies given our data.

4.3 Experiment 2: Incremental vs. Incremental-Adaptive

In this section we report results for the evaluation between the *incremental* (Phase 2) and *incremental-adaptive* (henceforth just *adaptive*; Phase 3) systems.

Subjective Results We applied the same significance test as Experiment 1 (with Bonferroni correction of 12). The item *The interface was useful and easy to understand* was answered with *Both* significantly ($\chi^2(4, N = 14) = 10.0, p < .0042$). The item *I had the feeling that the assistant attempted to learn about me* was answered with *Neither* ($\chi^2(4, N = 14) = 8.0, p < .0042$), though *Phase 3* was also marked (6 times). All other items were not significant. Here again we see that there is a general positive impression of the GUI under all conditions. If anyone noticed that a system variant was attempting to learn a user model at all, they noticed that it was in Phase 3, as expected.

Objective Results The *incremental* (Phase 2) and *adaptive* (Phase 3) columns in Table 1 show

the results for the objective evaluation for this experiment. There is a clear difference between the two variants, with the adaptive showing more completed tasks, more fully correct frames, and a higher average fscore (all three likely due to the fact that frames were potentially pre-filled).

4.4 Discussion

While the responses don’t express any preference for a particular system variant, the overall impression of the GUI was positive. The objective measures show that there are gains to be made when the system signals understanding at a more fine-grained interval than at the utterance level, due to the higher number of completed tasks and locally-made repairs. There are further gains to be made when the system applies simple user modelling (i.e., adaptivity) by attempting to predict what the user might want to do in a chosen domain, decreasing the possibility of user error and allowing the system to accurately and quickly complete more tasks. Participants also didn’t just get used to the system over time, as the average time per episode was fairly similar in all three phases.

The open-ended questionnaire sheds additional light. Most of the suggestions for improvement related to ASR misrecognition and speed (i.e., not about the system itself). Two participants suggested an ability to add “free input” or select alternatives from the tree. Two participants suggested that the system be more responsive (i.e., in *wait* states), and give more feedback (i.e., backchannels) more often. For those participants that expressed preference to the non-incremental system (Phase 1), none of them had used a speech-based PA before, whereas those that expressed preference to the incremental versions (Phases 2 and 3) use them regularly. We conjecture that people without SDS experience equate understanding with ASR, whereas those that are more familiar with PAs know that perfect ASR doesn’t translate to perfect understanding—hence the need for a GUI. A potential remedy would be to display ASR with the tree, signalling understanding despite ASR errors.

5 Conclusion & Future Work

Given the results and analysis, we conclude that an intuitive presentation that signals a system’s ongoing understanding benefits end users who perform simple tasks which might be performed by a PA. The GUI that we provided, using a right-branching

tree, worked well; indeed, the participants who used it found it intuitive and easy to understand. There are gains to be made when the system signals understanding at finer-grained levels than just at the end of a pre-formulated utterance. There are further gains to be made when a PA attempts to learn (even a rudimentary) user model to predict what the user might want to do next. The adaptivity moves our system from one extreme of the continuum—simple slot filling—closer towards the extreme that is fully predictive, with the additional benefit of being able to easily correct mistakes in the predictions.

For future work, we intend to provide simple authoring tools for the system to make building simple PAs using our GUI easy. We want to improve the NLU and scale to larger domains.³ We also plan on implementing this as a standalone application that could be run on a mobile device, which could actually perform the tasks. It would further be beneficial to compare the GUI with a system that responds with speech (i.e., without a GUI). Lastly, we will investigate using touch as an additional input modality to select between possible alternatives that are offered by the system.

Acknowledgements Thanks to the anonymous reviewers who provided useful comments and suggestions. Thanks also to Julian Hough for helping with experiments. We acknowledge support by the Cluster of Excellence “Cognitive Interaction Technology” (CITEC; EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG), and the BMBF Kogni-Home project.

Appendix

The following questions were asked on both questionnaires following Phase 2 and Phase 3 (comparing the two most latest used system versions; as translated into English):

- The interface was useful and easy to understand.
- The assistant was easy and intuitive to use.
- The assistant understood what I wanted to say.
- I always understood what the system wanted from me.
- The assistant made many mistakes.
- The assistant did not respond while I spoke.

³Kennington and Schlangen (2017) showed that our chosen NLU approach can scale fairly well, but the GUI has some limits when applied to larger domains with thousands of items. We leave improved scaling to future work.

- It was sometimes unclear to me if the assistant understood me.
- The assistant responded while I spoke.
- The assistant sometimes did things that I did not expect.
- When the assistant made mistakes, it was easy for me to correct them.

In addition to the above 10 questions, the following were also asked on the questionnaire following Phase 3:

- I had the feeling that the assistant attempted to learn about me.
- I had the feeling that the assistant made incorrect guesses.

The following questions were used on the general questionnaire:

- I regularly use personal assistants such as Siri, Cortana, Google now or Amazon Echo: Yes/No
- I have never used a speech-based personal assistant: Yes/No
- What was your general impression of our personal assistants?
- Would you use one of these assistants on a smart phone or tablet if it were available? If yes, which one?
- Do you have suggestions that you think would help us improve our assistants?
- If you have used other speech-based interfaces before, do you prefer this interface?

References

- Gregory Aist, James Allen, Ellen Campana, Lucian Galescu, Carlos Gallo, Scott Stoness, Mary Swift, and Michael Tanenhaus. 2006. Software architectures for incremental understanding of human speech. In *Proceedings of CSLP*, pages 1922—1925.
- Gregory Aist, James Allen, Ellen Campana, Carlos Gomez Gallo, Scott Stoness, and Mary Swift. 2007. Incremental understanding in human-computer dialogue and experimental evidence for advantages over nonincremental methods. In *Pragmatics*, volume 1, pages 149–154, Trento, Italy.
- Layla El Asri, Romain Laroche, Olivier Pietquin, and Hatim Khouzaimi. 2014. NASTIA: Negotiating Appointment Setting Interface. In *Proceedings of LREC*, pages 266–271.
- Timo Baumann, Michaela Atterer, and David Schlangen. 2009. Assessing and Improving the Performance of Speech Recognition for Incremental Systems. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 380–388, Boulder, USA, jun.

- Timo Baumann, Casey Kennington, Julian Hough, and David Schlangen. 2016. Recognising Conversational Speech: What an Incremental ASR Should Do for a Dialogue System and How to Get There. In *Proceedings of the International Workshop Series on Spoken Dialogue Systems Technology (IWSDS) 2016*.
- Joyce Y Chai, Lanbo She, Rui Fang, Spencer Ottarson, Cody Littley, Changsong Liu, and Kenneth Hanson. 2014. Collaborative effort towards common ground in situated human-robot dialogue. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 33–40, Bielefeld, Germany.
- Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13(2):259–294.
- Nina Dethlefs, Helen Hastie, Heriberto Cuayáhuitl, Yanchao Yu, Verena Rieser, and Oliver Lemon. 2016. Information density and overlap in spoken dialogue. *Computer Speech and Language*, 37:82–97.
- Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. 2008. Towards human-like spoken dialogue systems. *Speech Communication*, 50(8-9):630–645.
- Ting Han, Casey Kennington, and David Schlangen. 2015. Building and Applying Perceptually-Grounded Representations of Multimodal Scene Descriptions. In *Proceedings of SEMDial*, Gothenburg, Sweden.
- Casey Kennington and David Schlangen. 2017. A Simple Generative Model of Incremental Reference Resolution in Situated Dialogue. *Computer Speech & Language*.
- Casey Kennington, Spyros Kousidis, and David Schlangen. 2013. Interpreting Situated Dialogue Utterances: an Update Model that Uses Speech, Gaze, and Gesture Information. In *SIGdial 2013*, number August, pages 173–182.
- Casey Kennington, Spyros Kousidis, and David Schlangen. 2014. InproTKs: A Toolkit for Incremental Situated Processing. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 84–88, Philadelphia, PA, U.S.A. Association for Computational Linguistics.
- Casey Kennington, Ryu Iida, Takenobu Tokunaga, and David Schlangen. 2015. Incrementally Tracking Reference in Human / Human Dialogue Using Linguistic and Extra-Linguistic Information. In *HLT-NAACL 2015 - Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings of the Main Conference*, pages 272–282, Denver, U.S.A. Association for Computational Linguistics.
- Spyridon Kousidis, Thies Pfeiffer, Zofia Malisz, Petra Wagner, and David Schlangen. 2012. Evaluating a minimally invasive laboratory architecture for recording multimodal conversational data. In *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog, Interspeech Satellite Workshop*, pages 39–42.
- Staffan Larsson, Alexander Berman, and Jessica Villing. 2011. Multimodal Menu-based Dialogue with Speech Cursor in DICO II +. In *Computational Linguistics*, number June, pages 92–96.
- Pierre Lison. 2015. A hybrid approach to dialogue management based on probabilistic rules. *Computer Speech and Language*, 34(1):232–255.
- David Schlangen and Gabriel Skantze. 2011. A General, Abstract Model of Incremental Dialogue Processing. In *Dialogue & Discourse*, volume 2, pages 83–111.
- Gabriel Skantze and Anna Hjalmarsson. 1991. Towards Incremental Speech Production in Dialogue Systems. In *Word Journal Of The International Linguistic Association*, pages 1–8, Tokyo, Japan, sep.
- Gabriel Skantze and Samer Al Moubayed. 2012. IrisTK : a Statechart-based Toolkit for Multi-party Face-to-face Interaction. In *ICMI*.
- Gabriel Skantze and David Schlangen. 2009. Incremental dialogue processing in a micro-domain. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics on EACL 09*, (April):745–753.
- Michael J. Spivey, Michael K. Tanenhaus, Kathleen M. Eberhard, and Julie C. Sedivy. 2002. Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*, 45(4):447–481.
- Michael Tanenhaus, Michael Spivey-Knowlton, Kathleen Eberhard, and Julie Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science (New York, N.Y.)*, 268(5217):1632–1634.
- Gokhan Tur, Li Deng, Dilek Hakkani-Tür, and Xiaodong He. 2012. Towards deeper understanding: Deep convex networks for semantic utterance classification. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 5045–5048. IEEE.
- Jason D Williams, Eslam Kamal, Mokhtar Ashour, Hani Amr, Jessica Miller, and Geoff Zweig. 2015. Fast and easy language understanding for dialog systems with Microsoft Language Understanding Intelligent Service (LUIS). In *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue. 2015.*, pages 159–161. ACL – Association for Computational Linguistics, sep.