Towards Optimising Modality Allocation for Multimodal Output Generation in Incremental Dialogue

Nina Dethlefs, Verena Rieser, Helen Hastie and Oliver Lemon[1]

Abstract. Recent work on incremental processing in interactive systems has demonstrated that incremental systems can gain higher responsiveness and naturalness than their non-incremental counterparts and are better perceived by human users. This paper presents a first investigation, based on a proof-of-concept study, into how multimodal information presentation in incremental dialogue systems can contribute towards more efficient and smooth interactions. In particular, we focus on how a combination of verbal and non-verbal output generation can help to reduce the need for self-corrections in a system that has to deal with continuous updates of input hypotheses. We suggest to use Reinforcement Learning to optimise the *multimodal output allocation* of a system, i.e. the idea that for every context, there is a combination of modalities which adequately communicates the communicative goal.

1 Introduction

Traditionally, the smallest unit of processing in interactive systems that triggers a processing module into action has been a complete user utterance. While this facilitates processing and system design, it can lead to inflexible turn-taking and stilted interactions. In contrast, interactive systems with incremental processing align with human-like turn-taking behaviour by defining the *micro-turn* as the smallest unit of processing, which can be seen as the smallest part of an utterance that can be mapped to a dialogue act. This allows them to process input and plan output in parallel and to explore a range of discourse phenomena that occur naturally in human discourse, but that have so far been absent from interactive systems. Among these are backchannel generation, handling of user and system barge-ins, as well as corrections of generated output based on changed user or system knowledge. Several studies have shown that such phenomena can improve the user experience with an interactive system; see e.g. [22, 4] for incremental dialogue management, [18, 8] for turn-taking, [2, 23] for incremental automatic speech recognition, [12, 17, 24, 3] for incremental NLG, and [28] for a study on the impact of real-time feedback on user behaviour. Very recently, incremental processing has also been applied to the information presentation (IP) phase of interactive systems, where it has been combined with machine learning techniques to optimise the timing and order of IP [7] and the timing and occurrence of barge-ins and backchannels [6].

An important advantage resulting from the use of incremental processing is the increased awareness that NLG modules gain of their own generation process: they are able to monitor their own output and, if necessary, e.g. due to updated information coming in from the dialogue manager, modify or self-correct it. Such updates may be necessary in cases where user input hypotheses change during generation (or dialogue processing). As such, incremental NLG has to solve a trade-off between higher system reactiveness versus potentially disturbing self-corrections.

This paper argues that a possible remedy to this problem lies in the combination of different modalities, for example, speech and visual displays on a mobile device. Such multimodality may present a subtle way of communicating the system's current best input hypothesis to the user (and thereby give them a chance to correct it) without mistakenly acting upon it and causing a disruption or delay to the interaction. This hypothesis is based on previous work which has shown that multimodal output generation can increase system robustness to speech recognition errors [10] and decrease user cognitive load [15]. Previous work by [16] has also shown that allowing users to modify their search queries by combing speech and text input can significantly facilitate mobile search in noisy environments.

In this paper, we investigate a model of automatic output generation optimisation that uses *Reinforcement Learning* (RL) to maximise the expected return for the problem of *multimodal allocation* [1], i.e. how to combine output modalities so that they adequately convey a communicative goal in a given context. We present preliminary results from a proof-of-concept study in the domain of restaurant recommendations that compare the *task ease* achieved by our system and a number of hand-crafted baselines in simulated interactions. We discuss the possible advantages and disadvantages of our proposed method with respect to incremental interactive systems in *hands-free, eyes-free* mobile applications.

2 Multimodal Information Presentation

Previous work on multimodal information presentation has investigated rule-based user-tailored content selection [27] and supervised re-ranking techniques [11] for multimodal generation, as well as hierarchical Reinforcement Learning techniques for multimodal dialogue management [20, 5]. However, none of these earlier approaches has considered how multimodal information presentation can be integrated into an incremental model of dialogue processing.

In the following, we extend an earlier model for multimodal IP presented by [19] to incremental multimodal output allocation and show how it can help to avoid frequent self-corrections or output modifications from the system that are the result of dynamically changing input hypotheses. While the benefit of generating fewer self-corrections is not specific to incremental systems, but can be generalised to all interactive systems, we assume here that incremental systems face a particular danger of self-correcting too often due to their increased number of hypothesis updates.

[1] Heriot-Watt University, School of Mathematical and Computer Sciences, Edinburgh, Scotland, email: n.s.dethlefs@hw.ac.uk, v.t.rieser@hw.ac.uk, h.hastie@hw.ac.uk, o.lemon@hw.ac.uk

In Proceedings of the ECAI Workshop on Machine Learning for Interactive Systems: Bridging the Gap Between Language, Motor Control and Vision, Montpellier, France, pages 31-36, 2012.

As a domain of application, we address the information presentation phase in an interactive system for restaurant recommendations, extending previous work by [7], who present an incremental version of the work by [21]. While this previous work has focused on choosing a suitable presentation strategy for verbal presentation, here we focus on choosing the best modality accompanying a list of database hits. We assume that the choice of attributes (i.e. attributes that the user wishes the search to focus on) is determined by matching the types specified in the user input. Attributes include the *cuisine, food quality, location, price range* and *service quality* of a restaurant. The system then performs a database lookup and chooses a multimodal presentation strategy among *verbalOnly* and *combinedModalities*, i.e. visual and verbal output together. Visual output in this context refers to displays, on a screen or mobile device, that inform the user of the system's current best input hypotheses. Figure 1 shows examples of the main types of multimodal presentation strategies. The system does not have the option to present only visual information, since a Wizard-of-Oz study by [19] showed that human wizards never chose this strategy.

3 Optimising Multimodal Output Generation in Incremental Dialogue

3.1 Reinforcement Learning

To optimise the multimodal output generation process within an incremental model of dialogue processing, we define an RL agent as a Markov Decision Process, or MDP, which is characterised as a four-tuple $< S, A, T, R >$, where S is a set of states representing the status of the output generator and all information available to it; A is a set of output generation actions that combine strategies for multimodal IP with handling incremental updates in the system; T is a probabilistic transition function that determines the next state s' from the current state s and the action a according to a conditional probability distribution $P(s'|s, a)$; and R is a reward function that specifies the reward (a numeric value) that an agent receives for taking action a in state s.

Using such an MDP, the output generation process can be seen as a finite sequence of states, actions and rewards $\{s_0, a_0, r_1, s_1, a_1, ..., r_{t-1}, s_t\}$, where t is the time step. Every learning episode falls naturally into a number of time steps at each of which the agent observes the current state of the environment s_t, takes an action a_t and makes a transition to state s_{t+1}. This mechanism also defines the principle for the agent's micro-turn taking behaviour: it checks at each time step whether the state of the environment has changed so that an output action is required, e.g. if new input has come in or old input has been revised. If no particular action is required, e.g. because the user is still speaking, the agent may also decide to do nothing for the moment. Once information has been presented to the user, it is *committed* or *realised*. Here is where the difference between modalities may become most obvious to the user. While verbal output, once communicated to the user, cannot be changed without an explicit self-correction that marks the changed hypothesis, visual output can be changed more straightforwardly through an updated visual display, which may cause less disruption to an interaction.

The ultimate goal of an MDP is to find an optimal policy π^* according to which the agent receives the maximal possible reward for each visited state. We use the Q-Learning algorithm [29] to learn an optimal policy according to

$$\pi^*(s) = \arg\max_{a \in A} Q^*(s, a), \quad (1)$$

States
dataBaseHits {0=none,1=few,2=medium,3=many}
incrementalStatus {0=none,1=holdFloor,2=correct,3=selfCorrect}
modalityStatus {0=none,1=verbalOnly,2=combined}
statusCuisine {0=unfilled,1=low,2=medium,3=high,4=realised}
statusFood {0=unfilled,1=low,2=medium,3=high,4=realised}
statusLocation {0=unfilled,1=low,2=medium,3=high,4=realised}
statusPrice {0=unfilled,1=low,2=medium,3=high,4=realised}
statusService {0=unfilled,1=low,2=medium,3=high,4=realised}
userReaction {0=none,1=select,2=askMore,3=other}
userSilence={0=false,1=true}

Actions

Slot-ordering: presentCuisine, presentFood, presentLocation, presentPrice, presentService,

Incremental: backchannel, correct, selfCorrect, holdFloor, waitMore

Modality: verbalOnly, combinedModalities

Goal State $?, 0, \geq 1, 0 \vee 4, 0 \vee 4, 0 \vee 4, 0 \vee 4, 0 \vee 4, 1, 0 \vee 1$

Figure 2. The state and action space of the learning agent. The goal state is reached when all items (that the user may be interested in) have been presented and the most suitable output modality has been chosen. The goal state is defined with respect to the state variables above, where question marks indicate that the variable's value is irrelevant for reaching the goal state.

where Q^* specifies the expected reward for executing action a in state s and then following policy π^*.

3.2 The State and Action Space

The agent's state space needs to contain all information relevant for choosing an optimal strategy for multimodal output generation and an optimal sequence of incremental actions. Figure 2 shows the state and action space of our learning agent. The states contain information on the incremental, multimodal and attribute presentation status of the system.

The variable 'incrementalStatus' characterises situations in which a particular (incremental) action is triggered. For example, a `holdFloor` is generated when the user has finished speaking, but the system has not yet finished its database lookup. A `correction` is needed when the system has to modify already presented information (because the user changed their preferences) and a `selfCorrection` is needed when previously presented information is modified because the system made a mistake (in recognition or interpretation).

The variables representing the status of the cuisine, food, location, price and service indicate whether the slot is of interest to the user (0 means that the user does not care about it), and what input confidence score is currently associated with its value. Once slots have been presented, they are *realised* and can only be changed through a correction or self-correction.

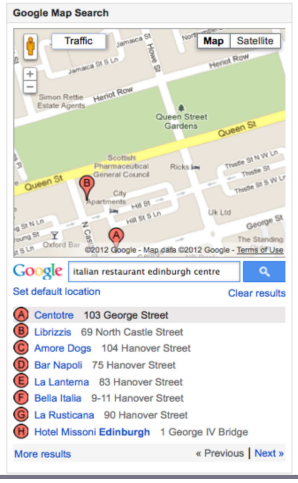
The variable 'userReaction' shows the user's reaction to an IP episode. The user can select a restaurant, provide more information to further constrain the search or do something else. The 'userSilence' variable indicates whether the user is speaking or not. This can be relevant for holding the floor or generating backchannels.

The focus of this paper lies in the optimisation of multimodal output generation for incremental IP settings and is represented by the 'modalityStatus' variable and its accompanying action set of *verbalOnly* and *combinedModalities* (shown in bold-face fonts in Figure 2). The agent will learn to choose the best multimodal output genera-

Verbal Presentation:

I have found the following restaurants that match your query for Italian food in central Edinburgh.
The restaurant Centotre is located in George Street and serves food in the medium price range. The restaurant Librizzis is also in the medium price range. It is located in North Castle Street. The restaurant Amore Dogs is located in Hanover Street. It serves ...

Combined Modalities:



The restaurants Centotre or Librizzis have the best overall match your query. They are located in the city centre of Edinburgh and serve Italian food in the medium price range. The Centotre is located on George Street and the Librizzis is located in North Castle Street.

Figure 1. Examples of the different modalities we are considering for information presentation. The system can choose an exclusively *verbal presentation*, and verbalise all restaurant options it retrieved (left-hand side). Alternatively, the system can choose to *combine verbal and visual* output and present a map of the area along with a list of possible options and a verbalisation of those options that best match the user's query (right-hand side).

tion strategy based on the other available variables, in particular with respect to the (discretised) number of retrieved database hits and the agent's user input confidence scores. We do not consider a *visualOnly* presentation strategy in this paper, since this action was never chosen by human users in the Wizard-of-Oz data that underlies our training environment [20]. In future work, we aim to include such a presentation strategy and investigate its impact within our framework.

The complete state-action space size of this agent is roughly 10 million. The agent reaches its goal state (defined w.r.t. the state variables in Figure 2) when a multimodal output IP strategy has been chosen and all relevant attributes have been presented.

3.3 The Simulated Environment

We train our learning agent in a simulated environment with two components, one for estimating user reactions to multimodal IP strategies and one for simulating dynamically updated input hypotheses within the incremental dialogue setting.

The first component deals with estimating user reactions to a multimodal information presentation strategy which contains the options *verbalOnly* and *combinedModalities*. This simulation component was trained from data (using the simulation described in [20]) and represents user reactions as bi-grams of the form $P(a_{u,t}|IP_{s,t})$, where $a_{u,t}$ is the predicted user reaction at time t to the system's IP strategy $IP_{s,t}$ in state s at time t. We distinguish the user reactions of *select* a restaurant, *addMoreInfo* to the current query to constrain the search and *other*.

While the multimodal IP strategies can be used for incremental and non-incremental output generation, the second part of the simulation deals explicitly with the dynamic environment updates during an interaction. We assume that for each restaurant recommendation, the user has the option of filling any or all of the attributes *cuisine, food quality, location, price range* and *service quality*. The possible values of each attribute and possible confidence scores are shown in Table 1 and denote the same as described in Section 3.2.

At the beginning of a learning episode, we assign each attribute a possible value and confidence score with equal probability. For food and service quality, we assume that the user is never interested in bad food or service. Subsequently, confidence scores can change at each time step. (In future work these transition probabilities will be estimated from a data collection, though the following assumptions are realistic, based on our experience.) We assume that a confidence score of 0 changes to any other value with a likelihood of 0.05. A confidence score of 1 changes with a probability of 0.3, a confidence score of 2 with a probability of 0.1 and a confidence score of 3 with a probability of 0.03. The new states that the agent makes a transition into are uniformly distributed. Once slots have been realised, their value is set to 4. Verbally presented slots cannot be changed then without an explicitly verbalised self-correction. We assume that realised slots change with a probability of 0.1. If they change, we assume that half of the time, the user is the origin of the change (because they changed their mind) and half of the time the system is the origin of the change (because of an ASR or interpretation error). Each time a confidence score is changed, it has a probability of 0.5 to also change its value. The resulting input to the NLG component are data structures of the form *present(cuisine=Indian), confidence=low*.

Attribute	Values	Confidence
Cuisine	Chinese, French, German, Indian, Italian, Japanese, Mexican, Scottish, Spanish, Thai	0, 1, 2, 3, 4
Food	bad, adequate, good, very good	0, 1, 2, 3, 4
Location	7 distinct areas of the city	0, 1, 2, 3, 4
Price	cheap, expensive, good-price-for-value, very expensive	0, 1, 2, 3, 4
Service	bad, adequate, good, very good	0, 1, 2, 3, 4

Table 1. User goal slots for restaurant queries with possible values and confidence scores.

3.4 The Reward Function

The main trade-off that the learning agent needs to optimise is to find the best multimodal information presentation strategy given the number of database hits for the user's query and the confidence scores held for attributes that represent the user's preferences. To learn an action policy for this problem, we use the reward function suggested by [20], which was induced from human data using a multiple linear regression analysis. It aims to optimise *task ease*, which is a combined value of the metrics *The task was easy to solve* and *I had no problems finding the information I wanted*. Human users had originally assigned scores to these metrics in a Wizard-of-Oz study. [2] The reward function is defined as follows.

$$\mathrm{R} = \begin{cases} -20.2 & \times \text{dialogueLength} + \\ 11.8 & \times \text{taskCompletion} + \\ 8.7 & \times \text{multimodalScore} \ . \end{cases} \quad (2)$$

The value for *dialogueLength* here corresponds to the number of dialogue turns until the user has selected a restaurant. The value for *taskCompletion* is a discretised score indicating whether the system has been able to successfully make a restaurant recommendation. It is $+10$ if the user selects a restaurant and -10 otherwise. The value *multimodalScore*, finally, indicates the appropriateness of the chosen presentation strategy estimated from human behaviour in a Wizard-of-Oz study, please see [20] for details. The score is related to the number of database hits presented using each modality through curve fitting. This technique selects the most likely model for the data based on function interpolation. In terms of rewards for a multimodal (or combined) output, it yields a quadratic function that assigns a maximal score to a strategy displaying 14.8 items. This number corresponds to the curve inflection point. For an exclusively verbal presentation, the reward is computed based on a linear function which assigns negative scores to all presented items ≥ 4.

Rewards according to Equation 2 are assigned at the end of an episode, which stretches from the moment that a user specified their initial restaurant preferences to the moment in which they choose a restaurant (or reject all presented choices). In addition, we assign a number of rewards during the course of an episode that are directed at the incremental dialogue setting. The agent receives a reward of 0 whenever the user adds more information to the query, a reward of -10 for generating a (verbal or partially verbal) self-correction, -0.5 for holding the floor and an increasing negative reward for waiting $waiting\_time^2$ (to the power of two), in terms of the number of time steps passed since the last item was presented. This reward is theoretically $-\infty$ so that the agent is penalised stronger the longer it delays to begin the information presentation phase. Using this reward function, the agent was trained for 10 thousand learning episodes.

4 Experimental Results

After training, the agent has learnt the following strategy for multimodal output generation in an incremental dialogue setting. It will choose an exclusively verbal presentation strategy whenever the search has returned few items (up to four) and the confidence in their values is relatively high (or at least medium). For a medium number of items to present (i.e. more than four but less than 30), the agent will choose a combined strategy of verbal and visual output if its confidence in the requested attributes is relatively high. If its confidence is low, it will first only display visual information and delay the verbal presentation as long as possible, waiting for confidence scores to stabilise. The same is true for a large number of items to present. In other words, the agent learns to prefer to include visual information whenever it is not confident (enough) of its current user input hypotheses. In this way, it is able to increase its dialogue efficiency because users are given a chance to restate their preferences when they realise (through a visual display of the system's input hypotheses) that the system is currently working with a wrong input hypothesis. The agent is also able to reduce the number of its own verbal self-corrections (because visual displays can be updated without the need for an explicit correction). Note that due to our incremental setting, the multimodal presentation will typically precede the verbal presentation in order not to interrupt the user while they are still speaking. The system will thus present visual displays representing its current best hypothesis of the user's input and then, once the user has finished speaking, present the retrieved restaurant items verbally.

We designed three baselines to compare our approach with. The first baseline chooses among output modalities randomly, we call this baseline *RandomBase*. This baseline was designed to test whether modality allocation has an impact on task ease, at all. The second baseline was designed to compare our multimodal approach with a system that presents information only verbally. This baseline was used to test whether the visual information that is displayed during processing to inform the user about the system's current hypotheses was indeed helpful to increase task ease and reduce the number of dialogue turns and system self-corrections. We call this baseline *VerbalBase*. Finally, we designed a third baseline which always presents information combing verbal and visual information. We call this baseline *combinedBase*. This baseline tests the added value of incremental modality allocation. Note that all systems, including the baselines, learn to optimise the order of information presentation (as described in [7]) and therefore have a learning curve.

Figure 3 shows the learning curves for the learnt policy and the baselines and compares them according to their average reward (averaged over ten sample runs). The average reward attained by each policy defines their degree of *task ease* as specified in the reward function. As expected, *RandomBase* performs worst and is outperformed by the learnt policy by 44.8% ($p < 0.0001$, according to a t-test). The low performance of this baseline is likely due to its multimodal allocation actions not being sensitive to the number of retrieved database hits nor to the agent's current confidence scores of incoming user input. While the other two baselines also show non-optimal behaviour, their action policies are at least consistent, which in the long run gives them a higher chance of choosing an appropriate modality 'by chance'.

VerbalBase, which presents all information verbally, performs 15.2% worse than the learnt policy ($p < 0.0001$). Again, this baseline fails to take the number of retrieved database hits into account. What is worse, though, is that the policy at times starts presenting results when it is still not confident enough in the user's preferred values. It may thus start to present wrong information to the user and eventually be forced to self-correct, which incurs a high negative reward. While the system has the option to delay the information presentation phase as much as possible by choosing to *waitMore*, the waiting action also incurs an increasing negative reward which eventually forces the agent to start its verbal presentation.

CombinedBase, which always combines multimodal and verbal output, finally performs only 9.9% worse than the learnt policy

[2] Note that even though our setting is not identical to the one used by [20], we assume that the reward function is to an extent transferable to our domain, which is also a slot-filling application with relatively short episodes. In the future, we aim to learn a separate reward function that is specifically tailored towards our incremental setting.

($p < 0.0001$) and is therefore the best performing baseline. The reason is that this baseline is only affected by a non-optimal multimodal allocation, but significantly less by the problem of low confidence in user input hypotheses. The combined modality policy has the option of holding back the verbal presentation until it is confident in its input hypotheses, and is free to modify its visual presentation as much as possible, since a visual display does not need to be self-corrected verbally (and thus does not incur the negative reward associated with a verbal self-correction). [3] The primary source of negative rewards in this setting is therefore the suboptimal multimodal strategy chosen when compared to the human strategies preferred in the Wizard-of-Oz study, based on which we trained our simulation and reward function.

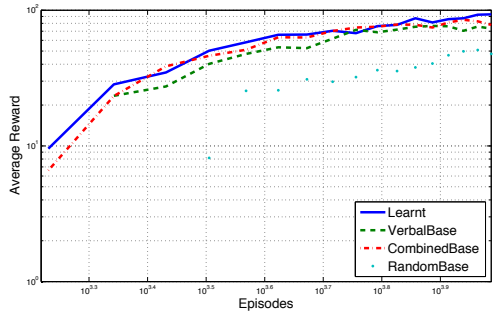


Figure 3. Learning curves indicating the average rewards, i.e. the average degree of task ease, attained by each policy.

5 Discussion

This paper has presented a preliminary investigation into how multimodal output generation can be integrated into incremental dialogue systems that process user input and plan system output in a parallel fashion. Our considerations here have been guided by how the *task ease* of a possible restaurant recommendation application for mobile devices can be optimised, in particular by increasing dialogue efficiency by multimodal display and reducing the number of verbal self-corrections that are caused by dynamically changing user input hypotheses. As is the nature of a proof-of-concept study, results are preliminary and so far based on simulation only. A variety of extensions of this work are possible. Importantly, we have not considered the restrictions that properties of the physical situation, the user or the particular application may pose on the choice of output modality, In in-car applications, for example, if we have indication of a high cognitive load or stress level (e.g. the eyes are fixed on the street) as in [9], the system could delay the presentation until a more suitable situation arrives and, simultaneously, mark the delay by a hesitation signal such as a turn holder. Similarly, we have left the question of user input modalities unaddressed and assumed that users always provide speech input.

The physical location of the user can have an impact on the preferred output modality in several ways. In crowded places, for instance, the system (and the user) may prefer a multimodal display due to the noise conditions that are likely to affect ASR results. Similarly, the system may take the user's current GPS position into account for its database lookup and prefer restaurants that are located close to the user's current location.

In terms of restrictions posed by the user, it is well known that individual users differ with respect to their specific preferences with regard to semantic [25] and lexical-syntactic [26] choices in language production. There is thus reason to expect that individual users will also have preferences for certain output modalities, some preferring verbal presentations, some visual output and combinations of different sorts. As a system 'gets to know' its user better, it may therefore want to increasingly take its particular user's preferences into account when choosing an output modality.

In addition, certain applications may themselves restrict the possible input and output modalities that a system can rely on. Many *hands-free* and *eyes-free* scenarios, such as an in-car mobile device, require the user to use speech only, or buttons that are manufactured into the steering wheel, to specify their search queries, and at the same time, should not be followed by multimodal output of the system that may require the driver to take their eyes off the traffic. On the other hand, previous work has shown that noisy ASR can distract drivers just as much [14], so that finding an appropriate multimodal output combination could amount to a challenging task.

There is also no obvious reason to restrict the user's input modalities to speech only. Instead, previous work has shown that a combination of speech and text input can lead to more efficient interactions when users are allowed to (incrementally) modify their search queries and retrieved results [13]. This can lead to decreased mental demand, perceived effort and level of frustration.

Finally, we have not paid explicit attention to the synchronisation between the different modalities, but have rather assumed that since output modalities are decided at the micro-turn level, they will automatically synchronise at the level of the utterance. While for the present (simulation-based) study, this has not presented a problem, it needs to be determined whether in practice a more principled mechanism for synchronisation is needed. An interesting direction, for example, could be to insert location points of restaurants on a map gradually, as they are presented as speech output in parallel.

6 Conclusion and Future Directions

This paper has presented a proof-of-concept study for optimising multimodal output generation for information presentation for incremental dialogue systems, i.e. systems that perform processing of user input and planning of system output in a parallel fashion. In particular, we have used Reinforcement Learning to optimise the *multimodal allocation* of our system, that is, to find an optimal combination of modalities for every given context. Preliminary results based on a partially data-driven user simulation are promising. They indicate that the agent is able to optimise its modality allocation by choosing an exclusively verbal presentation strategy for few search results and relatively high confidence scores in user input hypotheses. Alternatively, the agent can choose a strategy that combines visual and verbal output for a higher number of search results or situations involving low confidence scores in user input hypotheses. In this way, the resulting dialogues have gained in *task ease*, which was suggested by significantly higher rewards, shorter dialogues and fewer self-corrections which our system produced in comparison to a number of hand-crafted baselines.

In future work, we would like to extend our suggested model and re-train it using a fully data-driven simulated environment and reward function based on a data collection that explicitly addresses incremental discourse phenomena. This would allow us to explicitly

[3] While we did not restrict the number of visual updates in this setting, in practice, such a restriction may be necessary in order not to confuse users.

take the real-time nature of our model into account and not only estimate how input confidence scores change over time, but also how user behaviour changes through the incremental nature of our dialogue framework.

Further possible directions include the use of multiple user input modalities, adaptation to individual users during an interaction using online learning and a comprehensive evaluation of our suggested method using human users in a real-world setting. A further possibility is a data collection in an incremental multimodal setting to learn more about the effects of combining incremental processing and multimodal output generation on human-computer interaction.

ACKNOWLEDGEMENTS

This research has received funding from EC's FP7 programmes: (FP7/2011-14) under grant agreement no. 287615 (PARLANCE); (FP7/2007-13) under grant agreement no. 216594 (CLASSiC); (FP7/2011-14) under grant agreement no. 270019 (SPACEBOOK); (FP7/2011-16) under grant agreement no. 269427 (STAC).

REFERENCES

[1] Elisabeth André and Thomas Rist, 'Presenting through performing: on the use of multiple lifelike characters in knowledge-based presentation systems', *Knowledge Based Systems*, **14**, 3–13, (2001).
[2] Timo Baumann, Okko Buss, and David Schlangen, 'Evaluation and Optimisation of Incremental Processors', *Dialogue and Discourse*, **2(1)**, (2011).
[3] Hendrik Buschmeier, Timo Baumann, Benjamin Dosch, Stefan Kopp, and David Schlangen, 'Combining Incremental Language Generation and Incremental Speech Synthesis for Adaptive Information Presentation', in *Proceedings of 13th Annual SIGdial Meeting on Discourse and Dialogue*, Seoul, South Korea, (2012).
[4] Okko Buss, Timo Baumann, and David Schlangen, 'Collaborating on Utterances with a Spoken Dialogue Systen Using an ISU-based Approach to Incremental Dialogue Management', in *Proceedings of 11th Annual SIGdial Meeting on Discourse and Dialogue*, (2010).
[5] Heriberto Cuayáhuitl and Nina Dethlefs, 'Hierarchical Multiagent Reinforcement Learning for Coordinating Verbal and Nonverbal Actions in Robots', in *Proceedings of the ECAI Workshop on Machine Learning for Interactive Systems: Bridging the Gap Between Language, Motor Control and Vision (MLIS-2012)*, (2012).
[6] Nina Dethlefs, Helen Hastie, Verena Rieser, and Oliver Lemon, 'Optimising Incremental Dialogue Decisions Using Information Density for Interactive Systems', in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Jeju, South Korea, (2012).
[7] Nina Dethlefs, Helen Hastie, Verena Rieser, and Oliver Lemon, 'Optimising Incremental Generation for Spoken Dialogue Systems: Reducing the Need for Fillers', in *Proceedings of the International Conference on Natural Language Generation (INLG)*, Chicago, Illinois, USA, (2012).
[8] David DeVault, Kenji Sagae, and David Traum, 'Can I finish? Learning when to respond to incremental interpretation result in interactive dialogue', in *Proceedings of the 10th Annual SigDial Meeting on Discourse and Dialogue*, Queen Mary University, UK, (2009).
[9] M. Gasic, P. Tsiakoulis, M. Henderson, B. Thomson, K. Yu, E. Tzirkel, and S. Young, 'The effect of cognitive load on a statistical dialogue system', in *Proc. of SIGdial Workshop on Discourse and Dialogue*, (2012).
[10] Alexander Gruenstein, Stephanie Seneff, and Chao Wang, 'Scalable and Portable Web-Based Multimodal Dialogue Interaction with Geographical Databases', in *Proceedings of INTERSPEECH*, (2006).
[11] Hui Guo and Amanda Stent, 'Trainable Adaptable Multimedia Presentation Generation', in *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*, (2005).
[12] Anne Kilger and Wolfgang Finkler, 'Incremental generation for real-time applications', Technical report, DFKI Saarbruecken, Germany, (1995).
[13] Anuj Kumar, Tim Paek, and Bongshin Lee, 'Voice Typing: A New Speech Interaction Model for Dictation on Touchscreen Devices', Austin, Texas, USA, (2012).
[14] Andrew Kun, Tim Paek, and Jeljko Medenica, 'The Effect of Speech Interface Accuracy on Driving Performance', in *Proceedings of INTERSPEECH*, (2007).
[15] Sharon Oviatt, Rachel Coulston, and Rebecca Lunsford, 'When do we interact multimodally? Cognitive load and multimodal communication patterns', in *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*, (2004).
[16] Tim Paek, Bo Thiesson, and Y.C. Ju andBongshin Lee, 'Search Vox: Leveraging Multimodal Refinement and Partial Knowledge for Mobile Voice Search', in *Proceedings of User Interface Software and Technology (UIST)*, (2008).
[17] Matthew Purver and Masayuki Otsuka, 'Incremental Generation by Incremental Parsing', in *Proceedings of the 6th UK Special-Interesting Group for Computational Linguistics (CLUK) Colloquium*, (2003).
[18] Antoine Raux and Maxine Eskenazi, 'A Finite-State Turn-Taking Model for Spoken Dialog Systems', in *Proceedings of the 10th Conference of the North American Chapter of the Association for Computational Linguistics—Human Language Technologies (NAACL-HLT)*, Boulder, Colorado, (2009).
[19] Verena Rieser and Oliver Lemon, 'Learning Effective Multimodal Dialogue Strategies from Wizard-of-Oz Data: Bootstrapping and Evaluation', in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL/HLT)*, (2008).
[20] Verena Rieser and Oliver Lemon, 'Learning and Evaluation of Dialogue Strategies for new Applications: Empirical Methods for Optimization from Small Data Sets', *Computational Linguistics*, **37(1)**, 153–196, (2011).
[21] Verena Rieser, Oliver Lemon, and Xingkun Liu, 'Optimising Information Presentation for Spoken Dialogue Systems', in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, Uppsala, Sweden, (2010).
[22] David Schlangen and Gabriel Skantze, 'A General, Abstract Model of Incremental Dialogue Processing', in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, Athens, Greece, (2009).
[23] Ethan Selfridge, Iker Arizmendi, Peter Heeman, and Jason Williams, 'Stability and Accuracy in Incremental Speech Recognition', in *Proceedings of the 12th Annual SigDial Meeting on Discourse and Dialogue*, Portland, Oregon, (2011).
[24] Gabriel Skantze and Anna Hjalmarsson, 'Towards Incremental Speech Generation in Dialogue Systems', in *Proceedings of the 11th Annual SigDial Meeting on Discourse and Dialogue*, Tokyo, Japan, (2010).
[25] Jette Viethen and Robert Dale, 'The Use of Spatial Relations in Referring Expression Generation', in *Proceedings of the International Conference on Natural Language Generation (INLG)*, (2008).
[26] Marilyn Walker, Amanda Stent, François Mairesse, and Rashmi Prasad, 'Individual and Domain Adaptation in Sentene Planning for Dialogue', *Journal of Artificial Intelligence Research (JAIR)*, **30**, 413–456, (2007).
[27] Marilyn Walker, Steve Whittaker, Amanda Stent, Pretaam Maloor, Johanna Moore, and G Vasireddy, 'Generation and Evaluation of User Tailored Responses in Multimodal Dialogue', *Cognitive Science*, **28(5)**, 811–840, (2004).
[28] Kuansan Wang, 'A Study on Semantic Synchronous Understanding on Speech Interface Design', in *Proceedings of UIST-2033*, Vancouver, Canada, (2003).
[29] Chris Watkins, *Learning from Delayed Rewards*, PhD Thesis, King's College, Cambridge, UK, 1989.