To appear in Proc. of the IJCAI-05 Workshop on
Knowledge and Reasoning in Practical Dialogue Systems

Assessing the Impact of Adaptive Generation in the COMIC Multimodal Dialogue System

Mary Ellen Foster and **Michael White**
Institute for Communicating and Collaborative Systems
School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW United Kingdom
{M.E.Foster,Michael.White}@ed.ac.uk

Abstract

We describe how information from the dialogue history and the user model is incorporated into the output-planning process of the COMIC multimodal dialogue system, and present the results of experiments analysing the impact of both of these factors on the generated descriptions. The results of the experiments confirm and extend previous results by showing that both forms of adaptation make a perceptible difference in the generated speech. They also point to ways in which the system output can be improved to take further advantage of these information sources; in particular, they indicate the importance of concessions to negative preferences to the perceptibility of user-model tailoring.

1 Introduction

In this paper, we describe how information from the dialogue history and the user model is used to adapt the output-planning process of the COMIC multimodal dialogue system, and present the results of experiments analysing the impact of both of these factors on the generated descriptions.

COMIC[1] (COnversational Multimodal Interaction with Computers) is an EU IST 5th Framework project combining fundamental research on human-human interaction with advanced technology development for multimodal conversational systems. The multimodal dialogue system built as part of the project adds a dialogue interface to a CAD-like application used in sales situations to help clients redesign their bathrooms. The input to the system includes speech, handwriting, and pen gestures; the output combines synthesised speech, an animated talking head, deictic gestures at on-screen objects, and direct control of the underlying application.

There are four main phases in the full COMIC system. First, the user specifies the blueprint of their own bathroom, using a combination of speech input, pen-gesture recognition and handwriting recognition. Next, the user chooses a layout for the sanitary ware in the room. After that, the system guides the user in browsing through a range of tiling options for the bathroom. Finally, the user is given a three-dimensional virtual tour of the finished bathroom.

[1] http://www.hcrc.ed.ac.uk/comic/

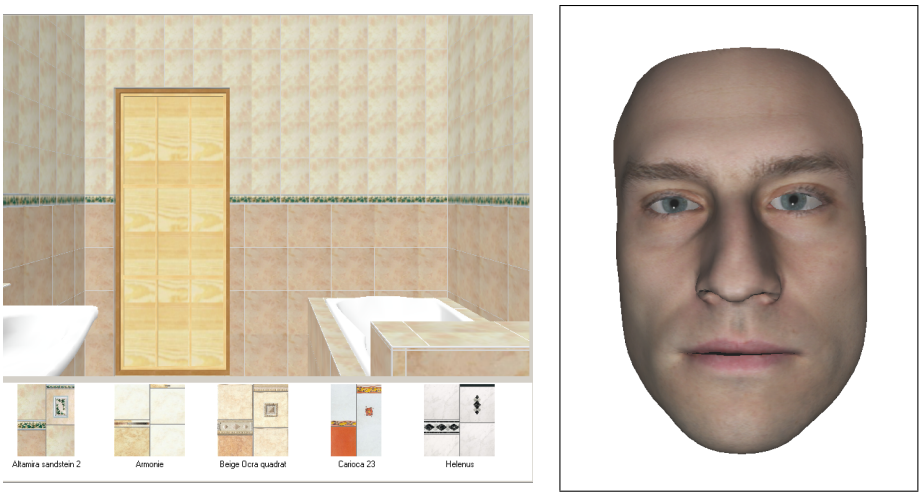
We will concentrate on the third phase of the interaction, where the system helps the user to explore the space of tiling options. The metaphor for this part of the system is one of *guided browsing*, and the goal is for the user to develop a better understanding of the range of possibilities and to have a clearer idea of the features that they like and dislike. The next step is for the user to work with a human sales consultant to fine-tune the exact tiling solution to use in their new room.

The rest of this paper is arranged as follows. In Section 2, we describe how adaptive output was generated and evaluated in several previous systems, and compare them with the COMIC system. In Section 3, we then give details of how adaptive output is created in COMIC using information from the dialogue history and the user model. Then, in Section 4, we describe experiments designed to assess the individual impact on the generated output of each of these knowledge sources. Finally, in Section 5, we summarise the results of the evaluations, and make recommendations for adaptive generation in future systems of this type.

2 Evaluating Adaptive Output

The approach we took to adaptive output in COMIC was inspired by the approaches taken in the GEA [Carenini, 2000], M-PIRO [Isard *et al.*, 2003], and MATCH [Walker *et al.*, 2002] systems. In this section, we review the types of adaptation employed in each of these systems, and describe how the impact of the adaptation was assessed. At the end of the section, we compare and contrast the types of adaptation in COMIC with those in the earlier systems, and give an overview of the evaluation reported in this paper.

The GEA system was designed to generate evaluative arguments tailored to user preferences in the domain of real estate, in the context of a graphical data-exploration environment. When generating a description of an option, the user preferences influenced the features that were included, the ordering of those features, and the use of scalar adjectives and adverbs in the text. A task-based evaluation of the generated arguments [Carenini and Moore, 2001] took the following form. While the subjects were using the data-exploration system, a new house was added to the options. A subject was informed about this new house in one of three ways: in some cases, they were just notified that there was a new house, with no description. In other cases, they were given a textual description of the new option; the description was either tailored



"*[Look at screen]* THIS DESIGN *[circling gesture]* uses tiles from SPHINX TILES's HELENUS series. As you can see, there are GEOMETRIC SHAPES on the DECORATIVE tiles. It is in the MODERN style."

Figure 1: COMIC interface and sample output

to their preferences, or not tailored at all. The effectiveness of the three presentation strategies was assessed by examining the user's behaviour after the presentation of the new option. The authors found that tailored descriptions were significantly more effective than both non-tailored descriptions and no descriptions, while non-tailored descriptions were not any better than no descriptions.

The M-PIRO system generates dynamic descriptions of museum artifacts, with the goal of instructing the user. In the evaluation described by [Karasimos and Isard, 2004], subjects read a series of textual descriptions of artifacts that were generated either with or without sophisticated text structuring. There were two forms of text structuring, which were either both enabled or both disabled together: combining semantically-related propositions into complex sentences (aggregation), and using common attributes to make links between the current object and any previously-seen objects (comparison). In this study, the subjects performed significantly better on a factual recall task when the descriptions they had read included aggregation and comparison; they also gave higher subjective scores to the generated texts that used these structuring techniques.

The MATCH system generates multimodal descriptions and comparisons of restaurants in New York City, tailored to the user's preferences. The user preferences control the restaurants that are recommended to the user, as well as the features of the selected restaurants that are included in the output. [Walker *et al.*, 2004] evaluated the user tailoring by directly asking subjects to judge the quality of the generated output. Subjects were presented with a series of responses to user requests that were tailored either to the subject's own preferences or to the preferences of some other user; the responses were presented in text first and then in speech. They found that subjects significantly preferred descriptions tailored to their own user model, with similar results for both speech and text.

In COMIC, output is adapted both to the user preferences (as in GEA and MATCH) and to the dialogue history (like M-PIRO). The primary output format in COMIC is synthesised speech, like MATCH and unlike GEA and M-PIRO. The guided-browsing task—in which the system helps the user explore and understand a space with which they are initially unfamiliar—is most similar to the task of M-PIRO; in GEA and MATCH, the role of the system is more one of helping the user to find a relevant option in a space with which they are already familiar. For this paper, we chose to assess directly whether the adaptations were perceptible to users, as in MATCH; we leave for future work questions of whether the adaptation has an impact on on task performance and user satisfaction.

3 Output Planning in COMIC

Figure 1 shows the interface for the guided-browsing phase of the COMIC system, along with a typical example of the output generated in this phase (where small capitals indicate pitch accents in the speech). The output combines the following modalities:

- Synthesised speech, generated using the OpenCCG surface realiser [White, 2004; 2005a; 2005b] and synthesised using a custom Festival 2 voice [Clark *et al.*, 2004] with support for APML prosodic markup [de Carolis *et al.*, 2004].
- Facial expressions and gaze shifts of the talking head.
- Deictic gestures at objects on the bathroom-application screen, using a simulated mouse pointer.

All of the information about tile designs in the COMIC system is stored in an ontology, which is linked to the dialogue history and the user model. The dialogue manager makes use

of these resources to choose the next set of designs to describe to the user, while the presentation planner uses them to select and structure the content of the descriptions that are generated.

3.1 Information and Knowledge Sources

Ontology Information about the available tile designs is stored in an ontology represented in DAML+OIL.[2] The ontology contains catalogue information including the manufacturer and series name, style, colours, and decoration, as well as any canned descriptive text associated with each design.

Dialogue History For each design in the ontology, the dialogue history keeps track of whether it has been mentioned in the dialogue. It also records the properties of each design that have been described to the user; note that a property may be described directly (e.g., *this design is classic*) or indirectly (e.g., *here are some classic designs* when pointing to several designs). The dialogue history also stores the identity of the last design that was described.

User Model As in [Carenini, 2000; Walker *et al.*, 2002; Moore *et al.*, 2004], user preferences are represented in COMIC using a model based on multi-attribute decision theory. Multi-attribute decision models are based on the notion that, if anything is valued, it is valued for multiple reasons. The value of a particular design for a user is computed as the weighted sum of its value on a number of primitive features; to create a specific user's model, we must therefore set the weights of the attributes and the evaluation function for each individual attribute.

The tile-design user model in COMIC is made up of four features: style, colour, designer, and decoration. These features are represented in a one-level tree. The evaluation function for each of these features assigns a score between 0 and 1 to every possible value of that attribute. A score of 0.5 represents a neutral evaluation; scores above or below that value indicate that the user respectively likes or dislikes that value, with the distance from 0.5 corresponding to the strength of that preference. For the colour and decoration attributes, which may have multiple values on a design, the score is computed by combining the individual attribute values.

The user-model manager supports two types of queries. It can produce an overall evaluation of a set of designs, to help the dialogue manager to choose options that are relevant to the user; it can also produce a detailed evaluation of a single design with scores on each individual attribute, to help the presentation planner create descriptions focussing on the options that are most important for that user. The detailed evaluation of a design is created by retrieving its features from the ontology and computing the evaluation of each feature; the overall evaluation is the weighted sum of the individual evaluations.

There are two ways that a user model can be defined in COMIC. One possibility is that a complete model is created and stored offline, before the dialogue begins, and is then selected and loaded at the start of the interaction. The other possibility is that the dialogue begins with a neutral user model, which then gets updated during the course of the interaction to take into account emerging preferences. In the current system, this is implemented by increasing the score for attribute values that the user specifically requests. For example, if the user asks for designs with blue tiles, the score for the colour blue is increased in the model for the remainder of the dialogue. There is no support in the current COMIC system for dynamically decreasing scores during the interaction, or for explicitly discussing the user model as part of the dialogue.

3.2 Planning Adaptive Output

When browsing through tile designs, the user may request designs with a particular feature—for example, *Show me designs with blue tiles*. To choose the next set of tile designs to present in response to such a user request, the dialogue manager selects all designs from the ontology that meet the user's criterion, and then uses the dialogue-history and user-model information to rank these designs, favouring those that have not been seen before and those with higher overall scores in the user model.

The dialogue manager then sends a message to the presentation planner that it should describe a particular design to the user. This message contains only a high-level request to describe a particular design, optionally including any features that must be included in the description for dialogue-flow reasons; it is up to the presentation planner to flesh out such a specification into a full description.

The presentation planner selects and structures content from the ontology to meet the dialogue-manager specification, and then creates a logical form for each sentence in the turn. The logical forms are then sent to the OpenCCG surface realiser, which produces prosodically-annotated text that is sent to the Festival speech synthesiser. Festival creates the waveform for the synthetic speech and returns the timing information for the words and phonemes in the text. This timing information from Festival is then used to set the schedule for output in the other modalities (talking-head behaviours, gestures at objects on the screen). When a complete schedule has been prepared for all of the output channels, the presentation planner starts the output. Full details of this process are given in [Foster and White, 2004; Foster, 2005].

The presentation planner has three main decisions to make when creating the content of a design description: choosing the features of the design to include, structuring the selected content appropriately, and choosing the eventual surface form of the resulting text. Information from the dialogue history and user model influences all of these decisions, as follows.

The maximum length of a description is normally three facts about the design being described. The presentation planner always includes any features specifically requested by the dialogue manager; it then chooses the rest of the content (up to the maximum length) based on information from the user model and dialogue history, as follows. First, it includes all features that have not been previously mentioned and that have a non-neutral evaluation in the user model, in decreas-

[2] `http://www.w3.org/TR/daml+oil-reference`

```
<!--
"Once again it is modern, but here the tiles
are from the Carioca collection by Aparici."
-->
<messages>
  <msg type="same-different">
    <slot name="same">
      <msg type="prop-has-val"
           same-as-last="true" prop="has_style">
        <slot name="object" value="Tileset4"/>
        <slot name="value" value="modern"/>
      </msg>
    </slot>
    <slot name="different">
      <msg type="prop-has-val"
           same-as-last="false"
           prop="has_designer-series">
        <slot name="object" value="Tileset4"/>
        <slot name="designer" value="Aparici"/>
        <slot name="series" value="Carioca"/>
      </msg>
    </slot>
  </msg>
</messages>
```

Figure 2: Combined text-planning messages

ing priority of user-model score, breaking ties arbitrarily. If there is still space in the description, it then includes any other features that have not previously been mentioned, again in an arbitrary order. Finally, the planner may choose to include a feature with a positive evaluation even if it has previously been described, if it is needed to offset an otherwise entirely negative description. If the result of this process is that no features are selected (because all of the features of the design have already been described), the planner chooses the most highly-ranked previously-mentioned feature to include.

Once the content of a description has been selected, user-model and dialogue-history information is also used to help create an overall structure for the description. When ordering the messages, for example, features with a positive evaluation are generally put earlier in the description. The contextual information is also used to combine adjacent messages into complex sentences where possible. We can combine a feature that is common between the current design and the previous one with one that differs between the two; Figure 2 shows the text planner's internal XML representation of two messages that have been combined in this way. A feature with a high user-model evaluation can be combined with one with a lower evaluation in a similar way.

Finally, the user model and dialogue history also impact the surface form of the description. For example, if a description is to include a fact that we have already told the user, we signal this repetition with words such as *as I said before* or *as I mentioned earlier*. If two facts have been combined to make a comparison with the previous design as in Figure 2, we use a structure such as *Once again X, but here Y*. If we are mentioning a property that we know the user does not like, we add words such as *though* or *although* to the sentence.

4 Experiments

We conducted two studies to assess the effectiveness of the adaptive generation described above, looking separately at the impact on the generated output of each of the two knowledge sources; we did not address the dialogue manager's selection of designs to describe. Like [Walker *et al.*, 2004], we used an "overhearer" paradigm in which the subjects watched and listened to recorded interactions between the system and a user and judged the quality of those interactions.

As pointed out by [Whittaker and Walker, 2005], an overhearer evaluation provides several advantages for evaluating multimodal dialogue systems. It allows judgements to be gathered during the course of a dialogue rather than at the end, and allows multiple alternative dialogue strategies to be compared in the same dialogue context. It also avoids any possible problems with speech recognition and language understanding, and allows the evaluation to be run on computers not powerful enough to support the full system. However, it does have the disadvantage that it measures only perception, rather than behaviour change, task performance, or user satisfaction; in Section 5, we discuss how these factors could be assessed.

The interactions were all synthesised in advance, and were presented to the subjects using a modified version of the COMIC system that was able to play back scripted output. User input was provided by playing recordings of a user making requests to the system. This allowed us to ensure that every subject saw and heard exactly the same version of each system turn. The output modules in these experiments were the speech synthesiser and an output-only emulator of the bathroom-design application able to display tile designs and animate pointer gestures; the talking head and full bathroom application were not used.

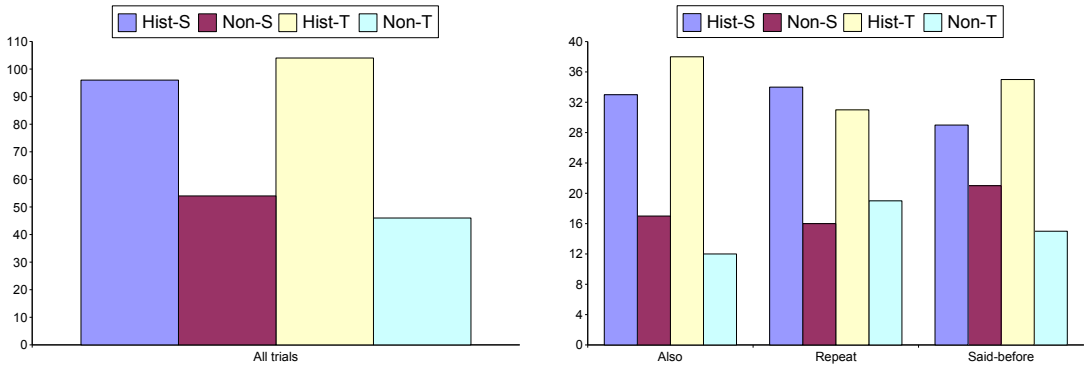
In both experiments, the basic structure was as follows. Subjects saw and heard two possible versions of a short interaction between a user and the COMIC system. Subjects were then asked to choose which version was better, based on seeing and hearing the presentation. They were then shown the transcripts of both versions of the interaction (in a format similar to Figure 3), and were asked again to choose between the versions. Even though written text is not an output modality of COMIC, we chose to present the transcripts in order to assess whether there were aspects of the adaptation that would have had an effect, but that were missed due to intelligibility issues with the synthesised speech.

Note that our experiment design differs somewhat from that used by [Walker *et al.*, 2004]: the MATCH subjects judged all of the outputs first using text, and then judged them all again in speech, and answered Likert-scale questions to assess the quality of each output individually. We presented the speech first because speech is the primary modality for the COMIC system; the text presentation was shown afterwards to allow the modalities to be compared. We chose to do a forced-choice study instead of using Likert scales because, for a transient medium like speech, direct judgements are difficult to make consistently, especially when the differences are subtle (cf. [Baker *et al.*, 2004; Rocha, 2004]).

(1) a. **User:** Tell me about this design *[click on Lollipop]*
b. **COMIC:** Here is a family design. Its tiles are from the Lollipop collection by Agrob Buchtal. The colours are blue and green.
c. **User:** What about this one *[click on Asterix]*
d. **COMIC:** Here we have another design in the family style. Its tiles are from the Asterix collection by Villeroy and Boch. It has comic motifs on the decorative tiles.

(2) a. **User:** Tell me about this design *[click on Lollipop]*
b. **COMIC:** This design is in the family style. The tiles are from the Lollipop collection by Agrob Buchtal. The tiles have a blue and green colour scheme.
c. **User:** What about this one *[click on Asterix]*
d. **COMIC:** This design is in the family style. Its tiles are from the Asterix collection by Villeroy and Boch. As you can see, there are comic motifs on the decorative tiles.

Figure 3: **Also**-class dialogue with dialogue history enabled (1) and disabled (2)



(a) Choices by presentation modality

(b) Choices by dialogue class

Figure 4: Graphs of dialogue-history results

The two experiments were run consecutively in the same session. Subjects were recruited through an email to the Informatics departmental student mailing list, and were compensated for their participation in the experiment. There were 25 subjects (20 male, 5 female) for the dialogue-history evaluation; due to technical difficulties, only 23 of those subjects were able to complete the user-model study as well. The full study took approximately half an hour to complete.

4.1 Dialogue-History Study

Materials and Presentation We created a set of six short dialogues between COMIC and the user, using a neutral user model. Each dialogue had two system turns. The tile designs used in each dialogue were selected arbitrarily, except in those cases where it was necessary that the designs selected have some factor in common. We generated two versions of each dialogue: one in which the second system turn made use of the context provided by the first turn, and one in which it did not.

The set consists of two dialogues in each of the following three classes, representing the three different ways in which the dialogue history can affect the generated output:

Also COMIC makes links between features in the current description and those in the preceding description.

Repeat Whenever possible, COMIC avoids repeating information that it has already said about a design.

Said-before When COMIC does repeat previously-mentioned information—for example, because there is nothing new to say about a design—it signals the repetition.

Figure 3 shows both versions of one of the dialogues that was generated in class **Also**: (1) makes use of the dialogue history, while (2) does not. The primary dialogue-history-based difference between the two versions is highlighted; all of the other surface differences occur because they were independently generated by the full COMIC presentation planner, which incorporates variation into its planning process.

All 25 subjects in this experiment were shown the same six dialogues, each in an individually randomised order; the order of versions within each dialogue was counterbalanced randomly so that they saw the dialogue-history version first in three trials, and second in the other three. Subjects were instructed to pay attention to how the system responded to the user's requests and how it kept track of what had already been said in the conversation. After they had seen both versions

(3) *(target)* Here is a family design. As you can see, the tiles have a blue and green colour scheme. It has floral motifs and artwork on the decorative tiles.

(4) *(other)* Here is a family design. Its tiles are from the Lollipop collection by Agrob Buchtal. Although the tiles have a blue colour scheme, it does also feature green.

(5) *(neutral)* This design is in the family style. It uses tiles from Agrob Buchtal's Lollipop series. There are artwork and floral motifs on the decorative tiles.

Figure 5: Output generated for three different user models

of each complete dialogue, subjects were asked the following question: *Which conversation had a more natural sequence of turns?* They chose initially based on the speech presentation, and then chose again after reading the transcripts.

Results and Discussion In general, the subjects chose the versions generated with the dialogue history enabled more often than those with it disabled, in both text and speech, as shown in Figure 4(a). Both differences are significant at the $p < 0.001$ level in a binomial test. Figure 4(b) divides the counts by dialogue class. To analyse statistical significance in this post-hoc analysis of the divided counts, we modified the required p value using a Bonferroni correction for multiple comparisons; the required significance value to achieve $p < 0.05$ overall is $p < 0.017$ on each individual test. All of the differences were significant at this level except for **Said-before** in speech and **Repeat** in text ($p \simeq 0.16$ and 0.06, respectively).

There was little overall difference between the choices made across the modalities; that they were not equal is probably due to intelligibility issues with the speech. For example, on the **Also**-class dialogue shown in Figure 3, preferences were essentially at chance when subjects chose on the basis of the speech (12–13; $p \simeq 0.65$). However, when subjects were able to read the text, there was a trend[3] for the dialogue-history version (18–7; $p \simeq 0.02$). As highlighted in the figure, the primary difference between the two versions is the single use of the word *another*; this was evidently difficult to pick up based on the speech, but was noticed often when subjects were able to read the transcripts. Adaptations with more obvious surface impacts—e.g., saying *as I said before* to signal a repetition, or describing multiple common properties of two designs—were perceived at a similar rate in text and speech.

4.2 User-Model Study

Materials and Presentation For this part of the experiment, four random user models were generated. Each model was generated by selecting three feature values that the user liked—which were given an evaluation of 0.8—and five values that the user disliked—which were given an evaluation of 0.2. All other values were given the default evaluation of 0.5. Note that the feature weights were equal in all four user models. One of the random models is shown in Figure 6.

For each model, an individual dialogue with the system was then created. Each dialogue started with the system's default selection of designs, and consisted of eight user requests and system responses. The user requests were selected to be plausible for the target user model, and the dialogue-manager also made its choices based on that target model. Four additional versions were then generated of each system output in each dialogue: one version based on the preferences of each of the other three random models, as well as a version based on a neutral user model (all evaluations 0.5). Figure 5 shows three versions of a system turn: (3) is generated based on the user model in Figure 6, (4) reflects the preferences of one of the other user models, while (5) is based on a neutral model.

Feature	Likes	Dislikes
Colour	blue, beige	pink
Style		modern, classic
Decoration	floral motifs	geometric shapes
Designer		Porcelaingres

Figure 6: Sample user model

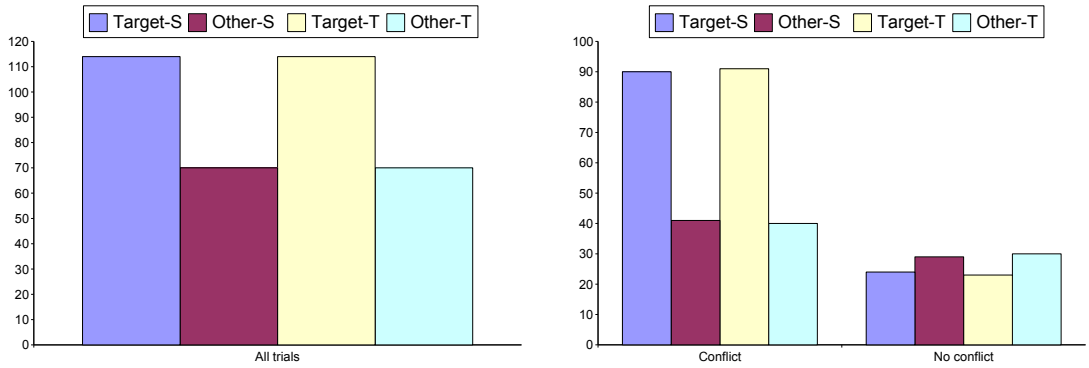
Subjects were assigned to one of the four target user models in rotation. The target model was shown on screen throughout the study, in a window similar to Figure 6. Subjects were asked to read through the user preferences before beginning, and to keep them in mind when making their choices. Two versions of each system turn in the dialogue were played—the version generated for the target model, and one of the other versions. The target model version was compared with versions for each of the other models twice, in an individually randomly-chosen order; the order within the trials was counterbalanced so that the target model version was seen first in four trials and second in the other four.

After seeing and hearing both versions of each system turn, a subject was asked the following question: *Which COMIC output was more appropriate for this user?* As in the dialogue-history study, subjects chose first on the basis of the speech presentation, and then again after reading the text.

Results and Discussion As shown in Figure 7(a), the overall results were similar to those in the dialogue-history evaluation: subjects generally chose the presentation generated for the target model over the one generated for the other model, using both speech and text. These preferences were both significant at $p < 0.001$.

In the MATCH evaluation [Walker *et al.*, 2004], the trials were subdivided based on the distance between the target model and the other model. However, the distance measure used there was based on the difference between the feature

[3]The necessary Bonferroni adjusted value for $p < 0.05$ overall significance is $p < 0.0083$ on each of the six individual instances.



(a) Choices by presentation modality

(b) Choices by trial type

Figure 7: Graphs of user-model results

weights, and is therefore not applicable to the current study in which the weights were the same in all the models. Instead, we divided the trials based on conflicting use of explicit concessions to negative preferences in the two descriptions. For example, the highlighted sentence in (4) has the concession *although the tiles have a blue colour scheme*, as if the user disliked the colour blue; in contrast, (3) has no such concession. This resulted in the following two categories of trials:

Conflict There was at least one conflicting concession across the two versions.

No conflict There were no concessions at all, or the concessions were the same in both versions.

Note that, due to the way the user models were selected, the majority of trials fell into the **Conflict** class. Figure 7(b) shows the counts for each of these classes. There is a very significant preference for the target description ($p < 0.00001$) in the **Conflict** class, but there is no significant preference either way in the **No conflict** class; the results are similar in both modalities.

The positive preferences do have an effect on the content that is selected for a description and the order in which it is presented: features that the user likes will always be included in the description, and are generally placed nearer the beginning (e.g., compare the content selection in descriptions (3) and (5) in Figure 5). However, this effect was not noticed by the subjects in this study, perhaps because they were basing their judgements on a random user model instead of their own preferences.

5 Conclusions and Future Work

These studies together demonstrate that the types of adaptation employed in the COMIC presentation planner do have a perceptible positive effect on the system output. Our results confirm and extend the results of previous studies, and point to ways in which the interaction could be improved to take better advantage of the information stored in the dialogue history and user model.

The results for the dialogue-history tailoring show that it does have a perceptible effect in all cases. This both confirms and refines the results of the M-PIRO study, which presented the output in text and did not examine the effects of aggregation and comparison independently. We also found that with spoken output, dialogue-history adaptations that affect only the surface form may occasionally be missed if the synthetic speech is difficult to understand. However, when the text can be read, such adaptations are perceptible, so with improved speech synthesis they should also have a perceptible effect.

When the output is tailored according to a user model, subjects generally chose the target versions, using both speech and text; this is in line with the results of the GEA and MATCH evaluations. The results of this study demonstrate that the effect of the user-tailored generation in COMIC is only noticeable when there are conflicting concessions to negative preferences in the two versions. Such concessions are also included in the output of GEA, but not in that of MATCH. In general, our subjects could not tell the difference between a description intended to be positive and one intended to be neutral; this may be partly due to the fact that the subjects in our study were basing their judgements on a randomly-generated user model, rather than on their own preferences as in GEA and MATCH. Note that the positive preferences are also used by the dialogue manager to select designs that are likely to be relevant to the user; the effect of this selection was not studied here, but the results of the MATCH evaluation suggest that it would also be perceptible.

While these studies demonstrate that users are able to notice the adaptive generation in COMIC, we have not yet shown that the adaptation has any effect on interactions with the full COMIC system. In future work, we hope to examine this question, using measures such as subjective user-satisfaction scores, objective measures of dialogue quality, and scores on a factual recall task, as in [White *et al.*, 2005]. COMIC could support such an evaluation of the dialogue-history adaptations in its current form; however, a full evaluation of the user-model adaptations would require some modifications to the COMIC system.

To get a true picture of the impact of the user tailoring, the system would have to respond based on the user's actual preferences. In GEA and MATCH, the user model is obtained before the interaction begins, by filling in a form. Note that the domains of these systems—real estate and restaurants, respectively—are ones where users can be expected to have a clear idea of their likes and dislikes before using the system. In contrast, with COMIC, it is not realistic to gather a full user model before the interaction begins, as the purpose of guided browsing is to help users become familiar with the range of available options. Thus, the most natural way to gather preferences would be through the course of the interaction. In the current system, the model is optionally updated by increasing the score for explicitly-requested features; however, there is no way to query or modify the user model that is updated in this way, and the system does not learn any negative preferences.

The system could take more initiative in inferring a user's preferences based on their browsing behaviour, possibly using the work of [Carberry *et al.*, 1999] as a basis. For example, if the user has rejected a series of proposed tile designs that all have green tiles, the system could hypothesise that the user does not like the colour green, and update the model accordingly. Instead of directly updating the user model based on these hypothesis, with no feedback, the system could also ask a confirmation question such as *It appears that you don't like green much.* This would give the user the opportunity to confirm or reject the system's beliefs about their preferences, and should help avoid getting "painted into a corner" prematurely. In the future, we hope to explore extended user-modelling and interaction capabilities in this way, and to assess the effect of the adaptively-generated output on user interactions with the system.

Acknowledgements

This work was supported by the COMIC project (IST-2001-32311). We thank Johanna Moore, Jon Oberlander, John Lee, Andrea Setzer, Roberta Catizone, and the other members of COMIC for helpful discussions, and the anonymous reviewers for their useful comments on the first draft of this paper.

References

[Baker *et al.*, 2004] R. Baker, R.A.J. Clark, and M. White. Synthesizing contextually appropriate intonation in limited domains. In *Proceedings of 5th ISCA workshop on speech synthesis*, 2004.

[Carberry *et al.*, 1999] S. Carberry, J. Chu-Carroll, and S. Elzer. Constructing and utilizing a model of user preferences in collaborative consultation dialogues. *Computational Intelligence*, 15:185–217, August 1999.

[Carenini and Moore, 2001] G. Carenini and J. Moore. An empirical study of the influence of user tailoring on evaluative argument effectiveness. In *Proceedings of IJCAI 2001*, 2001.

[Carenini, 2000] G. Carenini. *Generating and Evaluating Evaluative Arguments*. PhD thesis, Intelligent Systems Program, University of Pittsburgh, 2000.

[de Carolis *et al.*, 2004] B. de Carolis, C. Pelachaud, I. Poggi, and M. Steedman. APML, a mark-up language for believable behaviour generation. In H. Prendinger, editor, *Life-like Characters, Tools, Affective Functions and Applications*, pages 65–85. Springer, 2004.

[Clark *et al.*, 2004] R.A.J. Clark, K. Richmond, and S. King. Festival 2 – build your own general purpose unit selection speech synthesiser. In *Proceedings of 5th ISCA workshop on speech synthesis*, 2004.

[Foster and White, 2004] M.E. Foster and M. White. Techniques for text planning with XSLT. In *Proceedings of NLPXML 2004*, 2004.

[Foster, 2005] M.E. Foster. Interleaved planning and output in the COMIC fission module. In *Proceedings of ACL 2005 Workshop on Software*, 2005.

[Isard *et al.*, 2003] A. Isard, J. Oberlander, I. Androtsopoulos, and C. Matheson. Speaking the users' languages. *IEEE Intelligent Systems*, 18(1):40–45, 2003.

[Karasimos and Isard, 2004] A. Karasimos and A. Isard. Multi-lingual evaluation of a natural language generation system. In *Proceedings of LREC 2004*, 2004.

[Moore *et al.*, 2004] J. Moore, M.E. Foster, O. Lemon, and M. White. Generating tailored, comparative descriptions in spoken dialogue. In *Proceedings of FLAIRS 2004*, 2004.

[Rocha, 2004] N.F. Rocha. Evaluating automatic assignment of prosodic structure for speech synthesis. Master's thesis, Department of Theoretical and Applied Linguistics, University of Edinburgh, 2004.

[Walker *et al.*, 2002] M.A. Walker, S. Whittaker, A. Stent, P. Maloor, J.D. Moore, M. Johnston, and G. Vasireddy. Speech-plans: Generating evaluative responses in spoken dialogue. In *Proceedings of INLG 2002*, 2002.

[Walker *et al.*, 2004] M.A. Walker, S.J. Whittaker, A. Stent, P. Maloor, J. Moore, M. Johnston, and G. Vasireddy. Generation and evaluation of user tailored responses in multimodal dialogue. *Cognitive Science*, 28(5):811–840, September–October 2004.

[White *et al.*, 2005] M. White, M.E. Foster, J. Oberlander, and A. Brown. Using facial feedback to enhance turn-taking in a multimodal dialogue system. In *Proceedings of HCI International 2005*, 2005. To appear.

[White, 2004] M. White. Reining in CCG chart realization. In *Proceedings of INLG 2004*, 2004.

[White, 2005a] M. White. Designing an extensible API for integrating language modeling and realization. In *Proceedings of ACL 2005 Workshop on Software*, 2005.

[White, 2005b] M. White. Efficient realization of coordinate structures in Combinatory Categorial Grammar. *Research on Language and Computation*, 2005. To appear.

[Whittaker and Walker, 2005] S. Whittaker and M. Walker. Evaluating dialogue strategies. In W. Minker, D. Bühler, and L. Dybkjær, editors, *Spoken Multimodal Human-Computer Dialogue in Mobile Environments*. Kluwer Academic Publishers, 2005.