

Physically Embodied Conversational Agents as Health and Fitness Companions

Markku Turunen¹, Jaakko Hakulinen¹, Cameron Smith², Daniel Charlton², Li Zhang², Marc Cavazza²

¹Speech-based and Pervasive Interaction Group, Tampere Unit for Computer Human Interaction, University of Tampere, Finland

²School of Computing, University of Teesside, Middlesbrough, United Kingdom

{mturunen, jh}@cs.uta.fi, {c.g.smith, m.o.cavazza, d.charlton, l.zhang}@tees.ac.uk

Abstract

We present a physical multimodal conversational Companion in the area of health and fitness. Conversational spoken dialogues using physical agents provide a potential interface for applications which are aimed at motivating and supporting users. Open source software called jNabServer, which enables spoken and multimodal interaction with Nabaztag/tag wireless rabbits, is presented together with other software architecture solutions applied in the development of the Companion. We also present how the Companion manages interaction with the combination of a dialogue manager and a cognitive model.

Index Terms: spoken dialogue systems, multimodality, software architectures, physical agent interfaces

1. Introduction

Spoken dialogue systems have traditionally focused on task-oriented dialogues, such as making flight bookings or providing public transport timetables. In emerging areas, such as domain-oriented dialogues [1], the interaction with the system, typically modeled as a conversation with a virtual anthropomorphic character, can be the main motivation for the interaction. For example, the EU-funded COMPANIONS-project¹ studies speech-based and multimodal Companions that have a long lasting interaction history with their users [2].

As a part of the project, we are developing a conversational Health and Fitness Companion (H&F), which helps its users in leading more healthy life by daily interactions providing support and guidance. The H&F Companion has different motivations for use compared to traditional task-based spoken dialogue systems. Instead of helping with a single, well defined task, it is a companion, who will provide social support in the everyday activities. The system aims to be a peer rather than an expert system in health related issues.

There are good reasons for using a multimodal spoken dialogue system in such applications. The success of changes in the user's daily habits is mainly a question of motivation. A social and emotional relationship, which can commit a user to the system, is an efficient basis for improving the motivation. Since people build relationships mostly in face-to-face conversations, a physical, multimodal, conversational agent is a potential platform to build such a relationship [3].

1.1. Physical Agent Interfaces

Physical agent interfaces have become increasingly popular in the area of conversational systems. In many cases, they

include rich multimodal inputs and outputs while providing always a physical outlook for the agent. While naturalistic human-like physical robots are under development, especially in Japan, there is room for a variety of different physical interface agents ranging from completely abstract (e.g., simple devices with lights and sound) to highly sophisticated anthropomorphic apparatus. For example, in the study of Marti and Schmandt [4], several toy animals, such as bunnies and squirrels, were used as physical embodied agents for a conversational system. In our previous research [5], we have created an in-door guidance and receptionist application involving a physical interface agent that combines pointing gestures with conversational speech technology.

Some physical agent technology has been commercialized. For example, the wireless NabaztagTM/tag rabbits (<http://www.nabaztag.com/>) have been a huge success and an active user community has emerged around it. They have been used in multimodal interactive research prototypes as well. In this paper we present how we have used the Nabaztag as a multimodal physical interface to create a conversational Health and Fitness Companion. We present the overview of the H&F Companion, the underlying conversational dialogue architecture, and its different components for input and output management including the Nabaztag integration and dialogue and cognitive models. Finally, we present results from the initial evaluation.

2. Health and Fitness Companions

In our definition, the H&F Companion supports healthy eating habits and fitness activities by maintaining records of user's health-related, eating, and exercise information over a long time and providing guidance and feedback on healthy eating habits and various fitness activities on a daily basis. With the current H&F prototype, a user and the Companion make an overall plan for each day, and there can be more detailed conversations during the day. This interaction steers the user to more healthy daily habits.

As an example, the Companion can suggest to the user to walk to work instead of using a car or bus. Planning a day is demonstrated in Example 1. Figure 1 shows the scenario where the dialogue takes place.

C₁: Do you have anything arranged for today?

U₁: I'm playing Squash

C₂: Is that before dinner?

U₂: Yes

C₃: Okay. Are you going to get the train to work?

U₃: Yes

C₄: Right. You could take in a packed lunch to work?

U₄: Yes, that would be fine

¹ <http://www.companions-project.org/>

(Later on in the evening)

C₆: Welcome back! How did your game of Squash go?

U₅: The game was cancelled, so I went to the gym

C₇: Okay. Did you take the train to work?

U₆: Yes, I did

C₈: Right. Did you take in a packed lunch?

U₇: No. I was in a rush so I went to the cafeteria

Example 1: An example dialogue with the H&F Companion, as implemented in the first prototype.



Figure 1: Example from a H&F Scenario video with Nabaztag¹.

There are two other H&F Companion prototypes under development. The mobile Companion follows the user for physical activities, such as jogging, and collects data on the exercises and feeds this back into the main system. The Cooking Companion, placed in a kitchen, helps the user in choosing and cooking food.

Next, we present the overall H&F architecture, and the more detailed dialogue architecture and its components.

3. H&F Architecture

Since the Health & Fitness Companion interaction contains mobile and ubiquitous computing type solutions using novel interface technology, the software architecture requirements are far from trivial. Overall, multimodal pervasive computing applications need different architectural solutions from traditional spoken and multimodal systems, and the need for new theories, models and architectures for speech-based interaction in pervasive computing settings has been identified in the research community [6].

The H&F scenario, as presented in Example 1 and Figure 1, is implemented on top of Jaspis, a generic agent-based architecture designed for adaptive spoken dialogue systems. It has been used in several spoken dialogue systems [7]. In the H&F, this architecture is extended to support interaction with virtual and physical Companions, and the Nabaztag/tag device in particular. Next, we present the principles of the architecture, focusing on the adaptation mechanism and issues relevant for the H&F.

Figure 2 illustrates the H&F system setup. The top-level structure of the system is based on managers, which are connected to the central Interaction Manager using a star topology structure. The Interaction Manager coordinates the

other managers and is responsible for the overall coordination of the interaction. It is similar to certain central components found in other speech architectures, such as the HUB in the Communicator architecture [8], and the Facilitator in the Open Agent Architecture [9]. In addition, the application has an Information Manager that is used by all the other components to store all persistent information. Because of this, all components have access to all information. This is particularly important for dialogue and cognitive model components. Communication between the components is organized according to the client-server paradigm, enabling distribution over a network. Currently, in the H&F we have a set of seven managers in addition to the aforementioned two generic ones. New manager can be added as necessary.

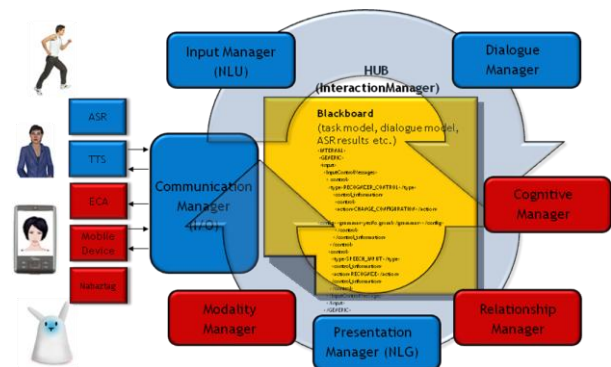


Figure 2: Jaspis based H&F Companion Architecture.

One of the aims in the H&F is to support highly adaptive interaction. System-level adaptation is supported in this architecture by the agents – managers- evaluators –paradigm that is used across all system modules. Tasks are handled by compact and specialized agents. When one of the agents inside a module is going to be selected, each evaluator in the module gives a score for every agent in the module. These scores are then multiplied by the local manager. This gives the final score, a suitability factor, for every agent. This generic system-level adaptation mechanism can be used in flexible ways in different components and systems. Most importantly, it has made it possible to implement the flexible dialogue and cognitive modeling needed in the H&F Companion. Next, we present the key components in more detail.

4. H&F System Components

The most interesting components in the H&F prototype include input and output components, which enable the use of Nabaztag rabbits and other similar physical agents in multimodal conversational spoken dialogue systems. In order to achieve fluent interaction with the users, we present a flexible model for dialogue management and cognitive modeling. It allows a clear separation of dialogue and domain models, still making them interoperate efficiently with each other.

4.1. Input and Output Components

The Communication Manager handles all input and output management. It includes devices and engines that provide interfaces to technology components. Most importantly, in the H&F prototype it includes components to control speech technology components (ASR and TTS) and the Nabaztag agent interface. In addition, the Communication Manager includes agents that take care of low-level input processing,

¹ <http://www.youtube.com/watch?v=KQSiigSEYhU>

such as parsing of the speech recognition results and RFID information from the physical agent.

For speech inputs and outputs, Loquendo™ ASR and TTS components have been integrated into Communication management. ASR grammars are in "Speech Recognition Grammar Specification" (W3C) format and include semantic tags in "Semantic Interpretation for Speech Recognition (SISR) Version 1.0" (W3C) format. Domain specific grammars were derived from a WoZ corpus to rapidly develop baseline for further studies and data collection. The grammars are dynamically selected by the Modality Manager according to the current dialogue state. Grammars can be precompiled for efficiency or compiled at run time when dynamic grammar generation takes place in certain situations. The current versions of recognition grammars have a vocabulary of 1090 words and a total of 436 CFG grammar rules in 39 dynamically selected grammars. In the future, domain specific statistical language models will be studied.

Natural language understanding is using heavily SISR information. These provide a basis for further input processing, where input is parsed against current dialogue state to compile full, logical representations compatible with the planning implemented in the Cognitive Model. In addition, a reduced set of DAMSL dialogue acts [10] is used to mark functional dialogue acts using rule based reasoning.

Natural language generation is implemented using a combination of canned utterances and tree adjoining grammar based generation. The starting point for generation is predicate-form descriptions provided by the dialogue manager. Further details and contextual information are retrieved from the dialogue history, the user model, and potentially other sources. Finally, SSML (Speech Synthesis Markup Language) 1.0 tags are used for controlling the Loquendo™ synthesizer.

4.2. Nabaztag Server

For a physical agent interface, the jNabServer software was created to handle communication with Nabaztag/tags, Wi-Fi enabled robotic rabbits. Nabaztag/tag devices can handle various forms of interaction, from voice to touch (button press), and from RFID 'sniffing' to ear movements. It can respond by moving its ears, by displaying or changing the color of its four LED lights. It can also play sounds which can be music, synthesized speech, and other audio.

By default, Nabaztag/tag communicates with the server of its creator company, Violet. In this case, interaction with Nabaztags is asynchronous due to rather large delays caused by the client-server communication (although this has been turned into a feature of the commercial version, and widely embraced by its users). We created jNabServer to replace the global server so that applications can be developed locally. In the local setup, delays can be as short as milliseconds in best cases, and it is thus compatible with the spoken dialogue interaction of the kind presented in Example 1. Functionality-wise, jNabServer offers full control over the rabbit, including RFID-reading, and makes it possible to use custom programs and technologies to process inputs and outputs, such as the speech recognition and TTS software used in the H&F.

jNabServer includes a very lightweight HTTP-server, and it has a build-in XML application programming interface, so client applications for the jNabServer can be made by using any programming language. For efficient Java-integration, jNabServer offers a plug-in system.

In the H&F Companion, jNabServer was integrated to Jaspis architecture as a set of devices and an engine under the

Communication Manager. This made it possible to use Nabaztag as the embodiment of the H&F Companion to support multi-modal conversational spoken dialogues.

The jNabServer software has been released as open source software¹ to support similar projects. It has been received well by the community and used for several purposes, such as studying the privacy aspects of conversational physical interface agents. Next, we present the interaction level components, the Dialogue Manager and the Cognitive Model, of the H&F Companion.

4.3. Dialogue Management and Cognitive Modeling

Interaction management in H&F is based on close-cooperation of the Dialogue Manager and the Cognitive Model. The Cognitive Model is more than just a simple back-end. It models the domain, i.e., knows what to recommend to the user, what to ask from the user and what kind of feedback to provide on domain level issues. We call this module the Cognitive Model, because it contains what can be considered the higher level cognitive processes of the system. We have separated cognitive modeling from dialogue management. The Dialogue Manager can now focus on interaction level phenomena, such as confirmations, turn taking, and initiative management.

The communication between the Dialogue Manager and the Cognitive Model is based on a dialogue plan. The Cognitive Model provides a plan to dialogue management on how the current task (planning a day, reporting on a day) could proceed. The following example shows how this works in practice:

```
( <plan-item>
  <action>QUERY-PLANNED-ACTIVITY</action>
</plan-item> )
```

C: Good morning. Anything interesting organized for today?

U: I'm going to play football.

```
( <pred>
  <action>PLANNED-ACTIVITY</action>
  <param>ACTIVITY-FOOTBALL</param>
  <param>unknownTime</param>
</pred> )
```

C: Is that football game before dinner?.

U: No, it's after.

```
( <pred>
  <action>PLANNED-ACTIVITY</action>
  <param>ACTIVITY-FOOTBALL</param>
  <param>AFTER-DINNER</param>
</pred> )
```

The Cognitive Model generates and updates a dialogue plan. It is aware of the meaning of the concepts in the plan on a domain specific level and updates the plan according to the information received from the user. The Cognitive Model is implemented in Allegro Common Lisp and it uses Hierarchical Task Networks in the planning process [11]. In the first H&F implementation, the planning domain includes 16 axioms, 111 methods (enhanced with 42 semantic categories and 113 semantic rules), and 49 operators.

Interaction level issues are not directly visible to the Cognitive Model. The Dialogue Manager takes care of conversational strategies. It presents questions to a user based on the dialogue plan, maintains a dialogue history tree and a dialogue stack and communicates facts and user preferences

¹ <http://www.cs.uta.fi/hci/spi/jnabserver/>

to the Cognitive Model. The Dialogue Manager also takes care of error management, supports user initiative topic shifts and takes care of top level interaction management, such as starting and finishing dialogues. Together, the Dialogue Manager and the Cognitive Model have similarities to approaches such as hierarchical task decomposition and dialogue stacks similar to CMU Agenda [12] and RavenClaw [13] systems.

The multi-agent architecture of Jaspis is used heavily on H&F dialogue management; in the current prototype, it consists of 30 different agents, some corresponding to the topics found in the dialogue plan, others related to error handling and other generic interaction tasks. The agents are dynamically selected based on the current user inputs and overall dialogue context, as described in Section 3. Currently this is done with rule-based reasoning. In the future, this will be augmented with machine learning approaches.

5. Conclusions

In this paper, we presented the concept of the Health and Fitness Companion, a dialogue system, which provides new types of conversational interaction. While traditional spoken dialogue systems have been task-based, the Health and Fitness Companions are part of the users' life for a long time, months, or even years. This requires that they are part of life physically, i.e., interactions can take place on mobile setting and in home environment outside of traditional, task-based computing devices. With the physical presence of the interface agent and spoken, conversational dialogue we aim at building social, emotional relationships between the users and the Companion. Such relationships should help us in motivating the users towards healthier lifestyle.

The physical embodiment of the Health and Fitness Companion was enabled by jNabServer. With it, Nabaztag/tag wireless rabbits can be integrated in dialogue systems and other interactive applications. The software has been published as open source software to aid the development of similar applications.

The division of interaction modeling into dialogue management and cognitive modeling was also discussed. In spoken dialogue systems such as the Health and Fitness Companion, interaction modeling becomes complicated, since we must model interaction for a long time, support user modeling, and have a complex domain model, which adapts as the user interacts with the system. The division of dialogue management and the cognitive model has made the development of such complex interaction management more feasible.

We believe that together these developments help us build new kinds of dialogue systems, which can build relationships with their users to support them in their daily lives.

5.1. Initial Evaluation Results and Future Work

An important part of the future work will be the evaluation of the Companions paradigm and of the Health and Fitness Companion. In order to find a baseline for further work and aid further development of the application, initial evaluation experiments were carried out at the University of Teesside. The evaluation involved 20 subjects. Each subject interacted with the Companion in two phases of dialogues similar to Example 1 during a typically 20-minute session.

Using the initial grammars in realistic experimental conditions without any user training or acoustic adaptation, the average Word Error Rate per phase were dialogue 42%

and 44%, the concept error rate was 24%, and the task model completion rate (a correct instantiation of an activity model corresponding to the scenario) varied between 80% and 95%.

The initial results show that even with relatively high WER we can get acceptable task completion rates in this domain, even without confirmation system that we have introduced after the tests. Speaker specific acoustic models and improved grammars should increase WER significantly. In the future evaluation we will focus on subjective evaluation of the system, in particular to find out the user experience of the Companions approach. An important part of this process will be to evaluate the long-term relationship nature of the Companion approach in real usage settings.

6. Acknowledgements

This work is supported by the EU-funded COMPANIONS-project (IST-34434). Nabaztag™ is a trademark of Violet™, who is thanked for authorizing the development of the "jNabServer" software.

7. References

- [1] Dybkjaer, L., Bernsen, N. O., Minker, W. Evaluation and usability of multimodal spoken language dialogue systems, *Speech Communication*, Volume 43, Issues 1-2, June 2004, Pages 33-54.
- [2] Wilks, Y., "Is There Progress on Talking Sensibly to Machines?", *Science*, 9 Nov 2007.
- [3] Bickmore, T. W, Picard, R. W. Establishing and maintaining long-term human-computer relationships *ACM Trans. Computer-Human Interaction* 12, No. 2 (June 2005): 293-327.
- [4] Martí, S. and Schmandt, C. Physical embodiments for mobile communication agents. *Proceedings of the 18th annual ACM symposium on User interface software and technology*: 231 – 240, 2005.
- [5] Kainulainen, A., Turunen, M., Hakulinen, J., Salonen, E.-P., Prusi, P., and Helin, L. A Speech-based and Auditory Ubiquitous Office Environment. *Proceedings of 10th International Conference on Speech and Computer (SPECOM 2005)*: 231-234, 2005.
- [6] McTear, M., *New Directions in Spoken Dialogue Technology for Pervasive Interfaces*. Proc. Workshop on Robust and Adaptive Information Processing for Mobile Speech Interfaces, 2004.
- [7] Turunen, M., Hakulinen, J., Rähä, K.-J., Salonen, E.-P., Kainulainen, A., and Prusi, P. An architecture and applications for speech-based accessibility systems. *IBM Systems Journal*, Vol. 44, No 3: 485-504, 2005.
- [8] Seneff, S., Hurley, E., Lau, R., Pao C., Schmid, P., Zue, V. *Galaxy-II: a Reference Architecture for Conversational System Development*. *Proceedings of ICSLP98* (1998)
- [9] Martin, D. L., Cheyer, A. J., & Moran, D. B. (1999). *The Open Agent Architecture: A frame-work for building distributed software systems*. *Applied Artificial Intelligence: An International Journal*. Volume 13, Number 1-2, January-March 1999 (pp. 91-128).
- [10] Core, M., Allen, J. *Coding Dialogs with the DAMSL Annotation Scheme*, *AAAI Fall Symposium on Communicative Action in Humans and Machines*, Boston, MA, November 1997.
- [11] Cavazza, M., Smith, C., Charlton, D., Zhang, L., Turunen, M. and Hakulinen, J., "A 'Companion' ECA with Planning and Activity Modelling", in *Proceedings of AAMAS08*, 2008.
- [12] Rudnicky, A. and Xu W. An agenda-based dialog management architecture for spoken language systems. *IEEE Automatic Speech Recognition and Understanding Workshop*, 1999, p I-337.
- [13] Bohus, D., and Rudnicky A. (2003) - *RavenClaw: Dialog Management Using Hierarchical Task Decomposition and an Expectation Agenda*, in *Eurospeech-2003*, Geneva,