Multimodal Agent Interfaces and System Architectures for Health and Fitness Companions

Markku Turunen[1], Jaakko Hakulinen[1], Olov Ståhl[2], Björn Gambäck[2], Preben Hansen[3], Mari C. Rodríguez Gancedo[3], Raúl Santos de la Cámara[3], Cameron Smith[4], Daniel Charlton[4] and Marc Cavazza[4]

[1]Tampere Unit for Computer Human Interaction, University of Tampere, Finland, [2]SICS, Swedish Institute for Computer Science AB, [3]Telefonica I+D, Spain, [4]School of Computing, University of Teesside, Middlesbrough, United Kingdom

Abstract. Multimodal conversational spoken dialogues using physical and virtual agents provide a potential interface to motivate and support users in the domain of health and fitness. In this paper we present how such multimodal conversational Companions can be implemented to support their owners in various pervasive and mobile settings. In particular, we focus on different forms of multimodality and system architectures for such 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 modelled as a conversation with a virtual anthropomorphic character, can be the main motivation for the interaction. The aim of the EC-funded COMPANIONS-project is to create speech-based and multimodal Companions that have a long lasting interaction history with their users [2].

As a part of the COMPANIONS-project, we are developing a conversational Health and Fitness (H&F) Companion, which helps its users for a healthy lifestyle by providing daily 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. Next, we present example dialogues from interaction with the H&F Companion.

Dialogue 1: During breakfast

C_1: Do you have anything arranged for today?
U_1: I'm playing Squash
C_2: Is that before dinner?
U_2: Yes
C_3: Okay. Are you going to get the train to work?
U_3: Yes
C_4: Right. You could take in a packed lunch to work?
U_4: Yes, that would be fine

Dialogue 2: Later on in the evening

C_5: Welcome back! How did your game of Squash go?
U_5: The game was cancelled, so I went to the gym
C_6: Okay. Did you take the train to work?
U_6: Yes, I did
C_7: Right. Did you take in a packed lunch?
U_7: No. I was in a rush so I went to the cafeteria

Dialogue 3: A physical exercise with the Companion

C_8: Hello, what kind of exercise would you like to do?
U_8: I would like to do some running
C_9: Ok, running it is. Select start when you are ready.

Dialogue 4: After 12 minutes

C_{10}: You have been running for 12 minutes. The distance is 2.5 kilometres; the current pace is 5 minutes per kilometre.
U_9: Okay, please play some relaxing music

Dialogue 5: Cooking with the Companion

C_{11}: Good, let's start preparing the ingredients. You should remove the stalk and seeds from the Red Pepper then chop it up.
U_{10}: And I should chop up the Onion now too?
C_{12}: Yes, that's right, peel and chop the onion. You could peel and crush the garlic too afterwards.

As seen in examples, the H&F Companion interacts with its owner using a spoken conversational dialogue. The interaction is implemented using several multimodal embodied agent interfaces. In the following, we first describe the different types of interaction the users can have with the H&F Companion. Second, software architectures used to implement each of these interfaces are detailed. Finally, conclusions are drawn.

2 COMPANION INTERFACES

There are good reasons for using a multimodal spoken dialogue system to implement H&F Companions. 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 improv-

ing the motivation. Since people build relationships mostly in face-to-face conversations, a multimodal conversational embodied agent is a potential platform to build such a relationship [3]. The H&F Companion interacts with users in three main forms; as Home Companion, which provides general support for healthy lifestyle, as Cooking Companion, which focuses on food, and as Mobile Companion, which support users during physical activities, such as jogging.

2.1 Home Companion Interface

The Home Companion interface provides advice on user's daily activities in order to encourage a healthier lifestyle. Such a system needs to support flexible dialogue, while supporting sufficient knowledge in relevant domains (daily activities, food and nutrition, basic exercise physiology) to build detailed activity models. The system gathers information from the user and makes suggestions on daily basis. The H&F Companion resides in the home and communicates with the user in two main dialogue phases; a planning phase where the system talks about the coming day with the user, and a reporting phase where the user's actual activities are assessed with reference to what was agreed on earlier.

The Home Companion interface, illustrated in Figure 1, and demonstrated in Dialogues 1 and 2, uses a Nabaztag/tag WLAN rabbit [4] as a physical agent interface. Nabaztag provides audio output and push-to-talk input, moves its ears and operate four coloured lights to signal its status.

Figure 1. Home Companion interface.

2.2 Cooking Companion Interface

One of the aims of the complete H&F Companion is to provide tips for a healthier lifestyle, and one of the key aspects is a correct nutrition. The Cooking Companion is a virtual embodiment of a dietary advisor, which will: 1) Present a set of possible recipes to the user, based on availability of ingredients or home delivery food. 2) Inform her/him of the appropriateness of each choice, based on available information such as the user's likes and dislikes, authorised medical and nutritional databases, hers/his current physical condition, special dietary requirements (e.g., allergies) and the physical exercise scheduled by the rest of the H&F Companion. 3) Help in the preparation of the selected recipe using multimedia and/or dialogue.

As an example of the dialogue flow, among the possibilities presented in 1) a user might want to eat a seafood paella. The system could then inspect all of the aforementioned parameters and conclude advising the user in stage 2) to prepare vegetable paella, based on his current exercise schedule and high blood fat levels (see Dialogue 5 for an example). In stage 3), the system could instruct the user on how to prepare the dish using videos and speech.

Figure 2. Cooking Companion interface.

In Figure 2, the current interface for the Cooking Companion prototype is shown. The system augments plain spoken dialogue with the help of an ECA, powered by a third party engine [5] and a finger-operated touchscreen interface. The interaction is to be multimodal, intertwining speech utterances and finger pointing to the UI elements. The ECA, modelled as a photorealistic human, includes some advanced gesturing capabilities that have been demonstrated to make the avatar both more engaging and understandable [6].

2.3 Mobile Companion Interface

The Mobile Companion interface, as seen in Figure 3, runs on Windows Mobile devices and can be used during outdoor exercise activities such as walking, jogging or cycling. The Mobile Companion can download the plan of the day the user has agreed on with the Home Companion. The Mobile Companion will then suggest an exercise, based on the user's current location, time of day and the plans made earlier, or, if there are no suitable exercises, asks the user to define one, as in Dialogue 3. Once an exercise has been agreed upon, the Companion asks the user to start the exercise and will then track the progress (distances travelled, time, pace and calories burned) using a built-in GPS receiver. While exercising, the users can ask the Companion to play music or to give reports on how the user is doing, as shown in Dialogue 4. After the exercise, the Companion will summarize the result, and upload it to the Home Companion so it can be referred to later on.

As can be seen in Figure 3, the Mobile Companion's graphical user interface consists of a single screen showing an image of the Nabaztag rabbit, along with a speech bubble. The

graphics include Nabaztag graphics and the same synthesized voice as in the home system to help users associate the two interfaces. All spoken messages are also shown as text in the speech bubble. The user can provide input via voice, by pressing hardware buttons on the mobile device, and in some situations, by tapping on a list of selections on the touch screen. Screen-based input is used, for example, when ASR errors occur, to perform error correction.

Figure 3. Mobile Companion interface.

3 COMPANION ARCHITECTURES

The different H&F Companions and their multimodal embodied interfaces, as presented in the previous section, take place in different physical environments, use different hardware, and require different software architectures. Thus, the overall H&F Companion platform consists of multiple system architectures specialized in various aspects of the interaction. Still, they use common resources and exchange information in order to achieve fluent interaction. This is illustrated in Figure 4.

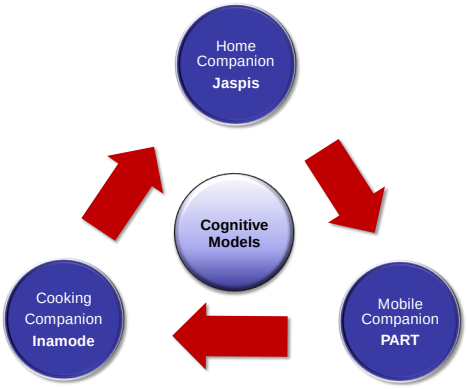

Figure 4. H&F Companion architectures and components.

As illustrated in Figure 4, the overall H&F Companion platform consists of four main components. First, there are three different system architectures suitable for (i) conversational interaction with physical embodied agents used in the Home Companion, (ii) spoken and tactile interaction with virtual embodied agents used in the Cooking Companion, (iii) mobile context-aware interaction used in the Mobile Companion. Second, all these architectures use the same shared cognitive models. Next, we present the cognitive model and the architectures in more detail.

3.3 Cognitive Modelling

In order to make the interaction with the H&F Companion coherent, we need to provide a shared cognitive model for the Companion. Currently, the focus is on an activity model, which decomposes the day into a series of activities for an office worker during a typical working day. These activities cover transportation to/from work, post-work leisure activities and meals (both in terms of the food consumed and how this food is obtained). We use a Hierarchical Task Network (HTN) planner to generate the activity model in the form of an AND/OR graph.

The planning phase involves the system operating through a global interaction cycle, integrating dialogue with planning to construct the user's activity model. This process is illustrated in Figure 5. The first step consists in generating an initial activity model for the user, based on default knowledge, along with any previously captured information on user preferences. Throughout this interaction cycle, each time the Planner generates a candidate activity model, it generates a corresponding dialogue plan, which is then used as a basis for the dialogue to enquire about user preferences and make suggestions in relation to the planned activities.

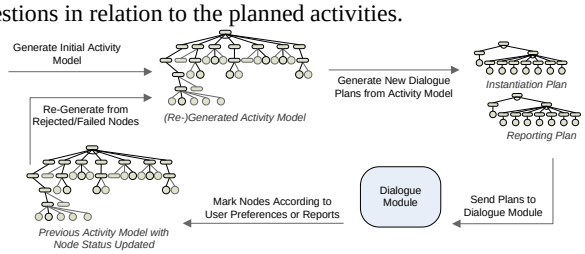

Figure 5. Activity model cycle.

User responses are used to update the activity model, with the user utterance mapped to the predicates used within the planning domain and semantic categories associated with the domain's methods. If a user response validates part of the existing activity model, that part of the model is marked as 'planned' and will not appear in the dialogue plan. If the user response is incompatible with the current activity model, either through explicit rejection or stating of a conflicting preference, it remains 'unplanned'.

The cycle continues with the activity model being regenerated. Those tasks that have been accepted, that is, marked as 'planned', are preserved while those 'unplanned' parts are replanned making use of the latest preferences provided by the user. An updated dialogue plan is then generated and the dialogue with the user continues until the user has agreed a fully 'planned' activity model.

The reporting phase is accomplished in a similar manner with a dialogue plan asking the user questions about what they did and the user utterances being used to produce an activity model with completed activities.

The planning domain includes 16 axioms, 111 methods (enhanced with 42 semantic categories and 113 semantic rules), and 49 operators. The Planner has been implemented in Allegro Common Lisp under Windows XP and communicates with the dialogue architectures using TCP sockets through which it sends XML structures corresponding to the format used by the dialogue controller. The cognitive modelling of the H&F Companion is presented in detail in [7;8].

3.1 The Home Companion

The H&F Companion is implemented on top of Jaspis, a generic agent-based architecture designed for adaptive spoken dialogue systems [9]. In COMPANIONS, the Jaspis architecture has been extended to support interaction with virtual and physical Companions, and the Nabaztag/tag device in particular [10]. The top-level structure of the system is based on managers, which are connected to the central Interaction Manager using a star topology structure. In addition, the application has an Information Manager that is used by all the other components to store and share information. Because of this, all components have access to all information. Communication between the components is organized according to the client-server paradigm, enabling distribution over a network. As illustrated in Figure 6, the H&F contains seven managers in addition to the aforementioned two generic ones.

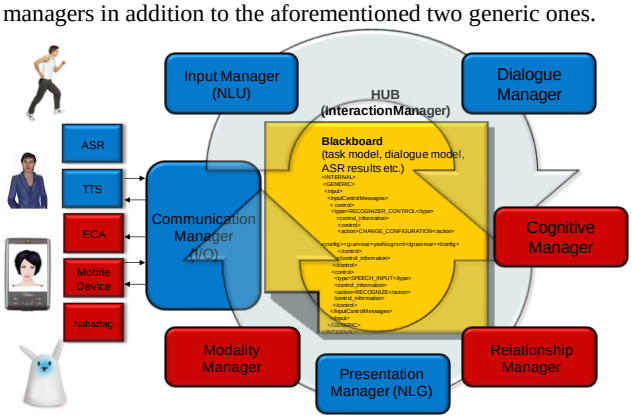

Figure 6. The Home Companion Architecture.

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 located in modules and coordinated by managers. When one of the agents inside a module, such as dialogue manager, 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, which gives the final score, a suitability factor, for every agent.

As an example, the multi-agent architecture of Jaspis is used heavily on dialogue management; in the current prototype, there are 30 different dialogue 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. Currently this is done with rule-based reasoning by three different evaluators. In the future, this will be augmented with machine learning approaches.

For speech input and output, Loquendo™ ASR and TTS components have been integrated into the Communication Manager. 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. During the development, data collected in user studies has been used to improve the grammars. This is further discussed in the last section with initial results from the first user studies. The grammars are dynamically selected by the Input 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 grammar size is about 1400 words and a total of about 900 grammar rules. The vocabulary coverage is balanced across the four relevant domains part of the activity model: transportation, physical activity, leisure and food. In the future, domain specific statistical language models will be studied.

Natural language understanding is using heavily SISR information. It 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 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.

For a physical agent interface, the jNabServer software was created to handle communication with Nabaztag/tag. By default, Nabaztag/tag communicates with the server of its creator company, Violet. 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.

Interaction management in H&F is based on close-cooperation of the Dialogue Manager and 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 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 of dialogues. The overall dialogue architecture is illustrated in Figure 7, and presented in detail in [11].

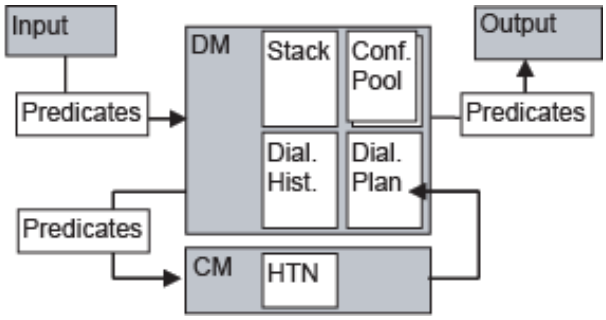

Figure 7. The Home Companion Dialogue Architecture.

3.2 The Cooking Companion

The Cooking Companion is built on Telefónica I+D's digital-home oriented framework, Inamode, which supplies both a high-level communication layer and modules aiding in the dialogue management and other functions.

The communications follow a loose hierarchy in which the architecture does not comply or require the developer to adhere to a complex and rigid set of rules, but instead a simple, chat-room like communication scheme is used using streamlined XML/JSON messaging over standard TCP sockets. Language and OS independent, this scheme allows us to quickly prototype interconnecting heterogeneous software modules, both own and third-party produced, by attaching a small communictaion stub (the Coupler). These Couplers are connected to repeating hubs (Concentrators) that echo the message to all of the other connected parties. A third element, Relayers, act as a interconnecting medium between each Concentrator's 'chatroom'.

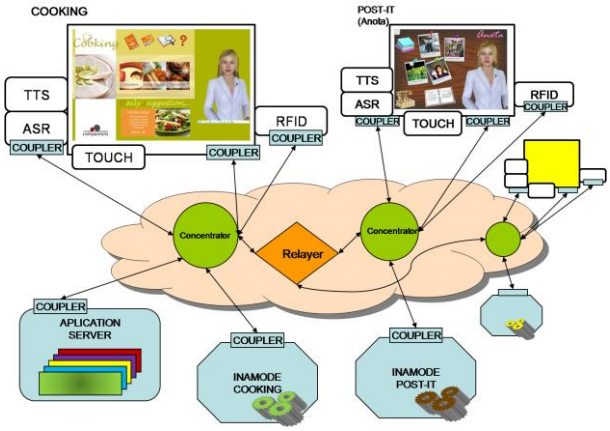

Figure 8. Cooking Companion Architecture.

The application modules built for Inamode (such as 'Cooking' or 'Post-it' in Figure 8), follow a consistent architecture and provide the multimodal interaction management. This is an augmented version of a traditional dialogue manager, in which input and outputs are not only speech but multimodal - including tactile interaction and even some inference on the user's emotions in the input side and generation of an appropriate speech utterance and body/facial expressions for the ECA in the output. This system is more in-depth described in [12].

3.3 The Mobile Companion

The Mobile H&F Companion, as illustrated in Figure 9, runs on a Fujitsu Siemens Pocket LOOX T830 device, which has a 416 Mhz XScale processor, 128 MB RAM and a built-in GPS receiver. The device runs Windows Mobile 5. The Companion consists of two components running on the mobile device, a Java midlet that handles the main application logic and user interface, and a speech server that performs ASR and TTS. The components communicate via a TCP socket. The Java midlet sends "commands" to the speech server, requesting, for instance, ASR input, and the server sends back the result or possibly some error message. The Java midlet is built using the PART library [13], and uses the Hecl scripting language [14] for dialogue management. The speech server uses Loquendo™ ASR and TTS libraries for embedded devices, and SRGS 1.0 grammars in XML format. Grammars can be loaded dynamically on request by the Java midlet, but all grammars must be pre-compiled and installed on the mobile device beforehand. In the current version, the Mobile Companion makes use of seven different grammars that are switched dynamically based on the dialogue context. The total vocabulary size is about 100 words.

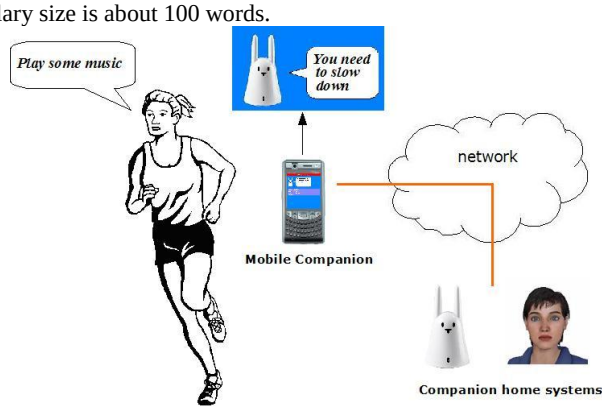

Figure 9. Mobile Companion Architecture.

While the Mobile Companion itself is running as a stand-alone system, the communication with the Home Companion requires the Mobile Companion to have access to the Internet, for instance via WLAN or 3G/GPRS. The Mobile Companion maintains a persistent data store, in which information can be saved in-between sessions. Currently, this store is used to save various information about the user (name, age, gender, weight, etc.), as well as the exercise results. Saving the exercise result allows the Companion to compare the progress of an exercise with previous exercises of the same kind. For instance, if the Companion knows that the user is currently cycling from home to work, it can provide feedback on how the user is doing compared to previous sessions. This allows for status messages like "You are currently 1 minute and 23 seconds behind yesterday's time". The Mobile Companion is described in detail in [15].

4 DISCUSSION AND CONCLUSIONS

In this paper we have presented an overview of conversational multimodal Companions in the area of health and fitness. We have presented usage scenarios, multimodal interfaces, and underlying system architectures for Companions. We believe that our approach, loose integration and interoperability of different multimodal interfaces and system platforms based on shared cognitive models, will be useful to study the Companions paradigm and similar conversational ambient intelligence systems, since any single platform or interface paradigm alone is not able to meet all requirements in this complex domain.

In addition to technical issues, we are studying how people interact with their Companions. With the use of synthetic, physical, and mobile embodied agent interfaces we try to have a better understanding of the practical benefits and problems of these different approaches. Here, we study how people

accept the different agent interfaces, and how they are used when interacting with the Companion over a long time. This long-term relationship is crucial for the paradigm.

In addition for generating new knowledge on Human – Companion interaction, we are producing concrete software components to implement Companions. Some parts of this work have been released for public use. The jNabServer software [16] has been released as open source software 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. Furthermore, Jaspis [17] and PART [13] architectures are open source software. Together, they can be used to construct similar distributed multimodal applications with virtual, physical, and mobile conversational agents.

4.1 Initial Evaluation

In order to gain understanding on the Companions paradigm, we will conduct user studies for the prototypes both in laboratory and field settings. In the first user studies we have focused on out-of-box functionality to set a baseline for further studies. This evaluation, conducted at the University of Teesside, involved 20 subjects who interacted with the Home Companion in both planning and reporting phases. They were briefly introduced to the concept of the Companion approach and the scenario and provided with a set of slides illustrating (via images of activities and food types) what was known to the system. To avoid bias they were not shown examples of possible utterances nor allowed to witness experiments with other subjects.

Our results are summarised in Table I. The fact that no user training or speaker adaptation was carried out, along with the realistic experimental conditions, explains the level of the Word Error Rate (WER). The Concept Error Rate (CER) is lower indicating some resilience to misrecognition of portions of the user utterance. The Task Completion was high, with, on average, 80% of the Activity Model being correctly instantiated in planning dialogues and 95% being correctly instantiated in reporting dialogues. This is better than the corresponding per utterance CER as the system was able to eliminate some errors over the course of the entire dialogue. Also worth noting is that reporting dialogues tended to involve simpler user utterances, such as basic confirmations, than those in the planning phase which reflects in the smaller average utterance length and higher task model completion rate.

Table I: Evaluation Results

Dialogue	WER	CER	Task Model Completion	Utterance Length
Planning	42%	24%	80%	4.2s
Reporting	44%	24%	95%	3.3s

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 evaluation currently under progress, in addition to dialogue metrics, we also collect subjective evaluations, in particular to find out the initial user experience of the Companions approach. An important part of the evaluation process in the future will be to evaluate the long-term relationship nature of the Companion approach in real usage settings.

ACKNOWLEDGEMENTS

This work was funded by the Companions project (www.companions-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-034434.

REFERENCES

[1] Dybkjaer, L., Bernsen, N. O., Minker, W., Evaluation and usabi ity of multimodal spoken language dialogue systems, Speech Communication, 43, 1-2, June 2004, pp. 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] Nabaztag/tag, Violet. http://www.nabaztag.com.

[5] Haptek Player, Haptek, inc. http://www.haptek.com.

[6] B. López Mencía, A. Hernández Trapote, D. Díaz Pardo de Vera, D. Torre Toledano, L. Hernández Gómez, and E. López Gonzalo, "A Good Gesture: Exploring nonverbal communication for robust SLDSs," IV Jornadas en Tecnología del Habla, Zaragoza, Spain, 2006.

[7] Cavazza, M., Smith, C., Charlton, D., Zhang, L., Turunen, M., Hakulinen, J., A 'Companion' ECA with Planning and Activity Modelling, Proc. of 7th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2008), Padgham, Parkes, Müller and Parsons (eds.), May, 12-16., 2008, Estoril, Portugal, pp. 1281-1284, 2008.

[8] Smith, C., Cavazza, M., Charlton, D., Zhang, L., Turunen, M., Hakulinen, J., Integrating Planning and Dialogue in a Lifestyle Agent, Proc. of the 8th Int. Conf. on Intelligent Virtual Agents (IVA 2008), LNAI 5208, pp. 146–153, (2008).

[9] Turunen, M., Hakulinen, J., Räihä, 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, 2005, pp. 485-504.

[10] Turunen, M., Hakulinen, J., Smith, C., Charlton, D., Li, Z., Cavazza, M. Physically Embodied Conversational Agents as Health and Fitness Companions. In Proceedings of Interspeech 2008 (to appear).

[11] Hakulinen, J., Turunen, M., Smith, C., Cavazza, M., Charlton, D. A Model for Flexible Interoperability between Dialogue Management and Domain Reasoning for Conversational Spoken Dialogue Systems. 4th International Workshop on Human-Computer Conversation, 2008. (to appear).

[12] Hernández, A., López, B., Pardo, D., Santos, R., Hernández, L., Relaño, J. & Rodríguez, M.C. (2008), "Modular definition of mult modal ECA communication acts to improve dialogue robustness and depth of intention", In The First Functional Markup Language Workshop Workshop at AAMAS 2008 (FML 2008).

[13] PART - Pervasive Applications RunTime, http://part.sourceforge.net/.

[14] The Hecl Programming Language, http://www.hecl.org/.

[15] Ståhl, O., Gambäck, B., Hansen, P., Turunen, M., and Hakulinen, J. A Mobile Fitness Companion. 4th International Workshop on Human-Computer Conversation, 2008 (to appear).

[16] jNabServer – java-based Nabzatag/tag Server. http://www.cs.uta.fi/hci/spi/jnabserver/.

[17] Jaspis - Framework for Adaptive Multimodal Applications. http://www.cs.uta.fi/hci/spi/Jaspis/.