

A Speech-based and Auditory Ubiquitous Office Environment

Anssi Kainulainen, Markku Turunen, Jaakko Hakulinen, Esa-Pekka Salonen, Perttu Prusi and Leena Helin

Speech-based and Pervasive Interaction group, TAUCHI, Department of Computer Sciences,
University of Tampere, Tampere, Finland

{Anssi.Kainulainen, Markku.Turunen, Jaakko.Hakulinen, Esa-Pekka.Salonen, Perttu.Prusi,
Leena.Helin}@cs.uta.fi

Abstract

This article introduces how speech and non-speech audio can be used in ubiquitous computing office environments. We describe an iterative development of an augmented office environment that helps people in their everyday tasks in office settings. Architectural as well as interaction issues are covered in this paper. We discuss how we have addressed the problems of multimodal data fusion, concurrency and continuity, dynamic content generation and output control, distribution and modularity, which are key elements in building ubiquitous speech based systems.

1. Introduction

Speech technologies have been used successfully in various spoken dialogue applications, such as information systems. Still, the full potential of speech-based interaction has not been utilized. Pervasive computing applications, such as mobile and ubiquitous environments, offer new possibilities and demands for speech-based interaction.

In this paper we present how speech and non-speech audio can be used in ubiquitous computing environments. We present an iterative development of an office environment augmented with services for interactive spoken guidance, unobtrusive group work awareness information, and system-supported speech-based messaging. Our environment has been constructed incrementally using common system architecture. Technology-wise, speech recognition, speech synthesis and speaker recognition are used. In addition, multimodal information sources are used for tasks such as positioning users.

In workplaces it is often important to avoid disturbing people and provide indirect ways of communication to increase awareness of the situation in the office, using unobtrusive methods. This enables direct communication to take place in a meaningful way, when it is appropriate. Speech and non-speech audio provide natural and efficient ways to implement both direct and indirect communication.

In our augmented office environment different ways of communication are used. Subtle indirect methods are used to lessen the information overload yet keeping the information available to trigger more active interaction. An audio messaging application gives colleagues an informal ad hoc method of communicating between people and receiving news and other notifications from the environment. A presence awareness application keeps people informed about each others' activities and presence with minimal cognitive load. These are used together with route guidance and tour guide applications to form the augmented office environment.

In order to enable all of the ways of communication, the system needs to have information about the different applications and the state of the environment. In the following sections we describe the enabling technologies and different applications that our augmented environment includes. We start by introducing the underlying system architecture and the enabling technologies that are used in the environment.

2. Enabling Technologies

The applications and services described in following sections use many technologies embedded to the environment. While previous research [1] has concentrated on separate applications and their interfaces, the applications of our augmented environment can communicate with one another and take advantage of each other's results. Also, much previous research has concentrated on ready-made, complete monolithic infrastructures [2], which are costly, inflexible, and may become quickly outdated. In our opinion, ubiquitous computing environments will be acquired, installed and updated incrementally. Therefore new services and information sources should be easy to add to the existing environment. This means that a flexible platform allowing constant growth for development is needed.

2.1. General system framework

The whole augmented office environment is built on top of the Jaspis system architecture [3]. The applications in the environment are distributed so that certain modules, tasks and services run independently of each other. Functionality of a single service is based on the general agents – managers – evaluators –paradigm [4].

All components inside the system share a common information storage, which enables stateless components, and straightforward information sharing between the applications. In this way, it is more flexible to connect new services and add functionality to the existing applications. This also makes it easier to handle the growing complexity of speech based ubiquitous applications. In addition Jaspis offers standard interfaces to core technologies, such as audio sources, speech and speaker recognizers and so on. Support for abstraction and distribution of dialogue management for different devices is also available [5].

2.2. Environment model

The ubiquitous nature of our assisting office environment requires the system to be aware of the objects and structures of the office space. Depending on the degree of detail this model consists of rooms, corridors and halls, or in more detail doors, windows, furniture and other less stable objects in the envi-

ronment. The model of the environment is based on the description of objects, their attributes and relations between the components.

The model concerning the shape and structure of the office serves all the applications of our environment, therefore it needs to cover structural information for route finding, descriptive vocabulary information for speech-based guidance as well as information concerning the interaction possibilities and input and output devices of each location. A more detailed description of the model is described in [6].

2.3. Positioning and activity information

Positioning of the users is a common problem in mobile and ubiquitous systems. According to our experience, in an office context, reliable positioning of a user has to be done using several techniques and input channels together.

We recognize people's movements using electro-mechanical film (EMFi) sensors placed on the floor. From the sensor's signal we can separate 1) the uneventful baseline from 2) someone walking over the mat, and 3) someone stepping on the mat and staying there. This simple categorization is achieved by lightweight processing. To recognize individual people from the data, we need more complex signal analysis in the manner of previous approaches [7].

The positioning information produced by the sparsely placed EMFi-sensors is not reliable enough as such for tracking individual people's movements. Combined with positioning information gathered from other input devices, such as activity daemons that monitor the usage of mouse, keyboard and application activity and speaker identification results at the interaction points, more precise knowledge of the locations of users is achieved. The data is stored in the shared information storage of the system. This is another example of a need for shared information, cooperating components and iterative growth of the whole infrastructure.

3. Applications of the Augmented Office Environment

Next we present the applications that form the augmented office environment. It consists of three different approaches to help people in everyday office tasks. All of the applications are built using the same system architecture and information sources described in the previous sections. First we introduce speech-based guidance services. Then we present how speech technologies can be used to enhance communication in the environment. Finally, we present how awareness of other people and their activities can be presented in unobtrusive means using environmental non-speech audio.

3.1. Reception and route guidance

Visitors often require help when moving in unfamiliar facilities. Help might be needed in orientating and navigating in large and complex indoor environments like airports [8] and office buildings. In particular, support is needed by special user groups, such as visually-impaired users [9]. In some cases visitors may require additional information concerning artifacts of the environment. This is the case, for example, in museums.

We have developed a speech-based guide called Doorman. It uses speech recognition, speaker verification and speaker identification to recognize the user and his/her

speech. Outputs are handled via puppets that combine pointing gestures and synthesized speech. Figure 1 illustrates the system components and the interaction situation.



Figure 1: Interaction with the Doorman system.

The user interacts with the system for the first time when he/she arrives at the front door of our premises. The system activates the speech recognizer when it detects the presence of the user using an infrared sensor placed outside the premises. The user may also push the doorbell, which causes the system to wake up if the sensors have failed (for example, if there has been a lot of movement recently). A welcome prompt is presented to the user, in which the purpose of the visit or the name of the staff member the visitor is looking for is asked. If entry is successful, visitors are guided to their target locations and/or people. If the purpose of the visit cannot be recognized, or an alleged staff member's identity cannot be verified, a staff member is called to the door to deal with the situation. In the case of a staff member entering the premises, speaker recognition is used to identify the person, after which additional information depending on the user's preferences is given. Example 1 presents two dialogues between the user (U) and the Doorman system (S).

In the case of visitors:

- U: *(Pushes the doorbell)*
 S: *I'm Doorman. Please say the name of a person or a place after the tone.*
 U: *Markku Turunen*
 S: *Welcome. The door is now unlocked.*
 U: *(Enters inside and faces the guidance puppet)*
 S: *The person that you are looking for, Markku Turunen, can be found from room 432. In order to get there turn left. Start near the meeting room. Turn right and head towards the sofa. Markku's office is in front left.*

In the case of staff members:

- S: *(Detects movement and starts recognition)*
 U: *This is Markku Turunen, good morning.*
 S: *(Verifies the speaker correctly.)*
 S: *Welcome Markku. The door is unlocked.*
 U: *(Enters inside and faces the guidance puppet)*
 S: *By the way, you have 12 unread messages in your inbox.*

Example 1: Doorman example dialogues (translated from Finnish).

When the visitor enters inside, the puppet is used to give guidance to the user. The model of the premises is used to generate appropriate route descriptions with required level of

detail [6]. In the current setup the guidance is given in one place, but there may be also other places for route guidance interaction like intersections and lobbies. This way, the guidance is given in small chunks, which are easier to understand and memorize [10]. This brings up the need to distribute the dialogues spatially in the environment. We have implemented a prototype application that locates users moving in our premises and directs the system outputs to the nearest loudspeaker. This way the presented information follows the user.

3.1.1. Additional guidance information

We have applied the guidance system to present objects in a tour guide fashion. It is used for describing our research by explaining the posters placed in our premises. We use positioning and step event recognition to find out user's location, and whether they have stopped in front of an item or just passing by. The description of an item is dynamically constructed and depending on what a person has already seen, the tour guide makes cross references on a variety of topics to other previously seen items.

Currently we are examining more versatile uses for guides in public places such as museums, where they describe artwork. Furthermore, automated guides could bring added value by describing the facts that are not covered in the ordinary tours such as the history of the museum buildings. Interactive advertisements in places like bus stops could be constructed in this fashion. A guide like this could also be used in factories to describe parts of machinery or stages of a process, or offer help in rescue situations. An example of user and context aware tour guide follows.

For sighted users:

*"As you can **see**, the artist uses the same idea of opposites when choosing materials as in the **three previous** sculptures. **Dark** bronze and **light** wool form a strong contrast."*

For visually impaired users:

*"As you can **feel**, the artist uses the same idea of opposites when choosing materials as in the **two following** sculptures. **Coarse** bronze and **soft** wool form a strong contrast."*

Example 2: Two adapted outputs.

In Example 2 the outputs are adapted according to the user's interests. In addition dynamic content creation reacts to accessibility problems as well. This example showed the need for user adaptation and dialogue distribution in time and space. Next we will describe how our augmented office environment supports communication between people.

3.2. Messaging

Messaging services are commonly used in group work for informal and ad hoc communication. Speech-based messaging contains vocal subtleties, emphases and emotions which can not be conveyed in text-based communication such as in [11]. Spoken dialogue technologies can be used to support the communication between the users, and act as a partner in the conversation to bring in added value such as guiding the messages to correct locations. We have implemented a messaging system based on speech and speaker recognition, microphones and loudspeakers placed in corridors, halls and private offices. Microphones in offices are connected to the desktop computers of the group members and they act as triggers to

the dialogue in addition to their primary use. The dialogues in Example 3 demonstrate the interaction between the users and the system.

- U₁: *(opens the microphone connected to his desktop computer) "Message to Anssi"*
 S₁: *(speech and speaker recognition and loudspeaker selection) "Hi Perttu. Do you want to send message to Anssi?"*
 U₁: *"Yes"*
 S₁: *"Please dictate your message. Finish by closing the microphone line"*
 U₁: *"Anssi, you haven't replied to the e-mail I sent you. Are you still at work?" (closes the line)*
 S₂: *(activity check from Anssi's desktop computer and loudspeaker selection) "Anssi, there is a message coming from Perttu." (message playback)*
 S₂: *"Open your microphone to reply?"*
 U₂: *(opens the microphone) "Sorry, I've been reading an article and haven't noticed the e-mail. I'll reply to it shortly." (closes the line)*
 S₂: *"Anssi replied:" (message playback)*

Example 3: Speech-based messaging.

The message system offers a general spoken dialogue interface between the users and the services embedded to the office environment. In addition, group work may benefit from location and context aware messaging applications, in particular if group members are working in different physical locations. Currently we are studying how the system can be more aware of the location and state of its users. With this information, the system can convey the messages using the best possible channels. This example showed the need for supporting multiple dialogue partners and physically distributed dialogues.

3.3. Supporting awareness

In addition to direct messaging we have implemented ways to keep people aware of activities inside their environment in more indirect ways. We aim at bringing information calmly to the periphery of people's attention, in order find less burdening ways to inform people [12].

Presence information is given in a transparent and unobtrusive manner using environmental audio, regarding to the activity data that is collected as presented in section 2.3. The final presentation is a compilation of the general activity of the whole workgroup. We have created soundscapes, in which each person is presented by a sound, for example the singing of a certain bird. Sounds are mixed together dynamically depending on the situation and interleaved in layers in order to create a presentation as natural and calm as possible. Playback is handled by highly directional EMFi-loudspeakers placed in public places and personal offices. The same speakers are shared by other applications described in this paper. In addition to symbolic bird sounds, we also experimented with the actual walking sounds of users [13].

In order to support awareness in more direct ways, we have implemented a speech-based application to present information, such as news bulletins, while people are moving around the premises. These applications show the need for the capability of using different levels of attention, where direct speech and indirect non-speech support each other.

4. Findings

We have presented an assisting office environment that uses speech and non-speech audio for the interaction. Next we will present results and observations gathered during the development of this environment.

We have conducted evaluations on some of the applications. The usability of the Doorman system was evaluated as a part of the development process by performing a Wizard of Oz experiment, in which speech recognition and speaker recognition were simulated [14]. The main findings concerned the way users spoke to the system, and they emphasized the importance of correctly and actively recognizing people's locations, and providing means for place-independent communication.

Based on our experiences, the efficiency of guidance and concurrent events indicate a requirement for a more robust support for concurrent dialogues. The flexibility of control inside the architecture has been a growing need as the amount of components, services and dependencies between modules has multiplied. Some of these findings motivated us to design a new version of the Jaspis system architecture [15].

Early in the development process we discovered that one of the major challenges in ubiquitous environments is the distributed nature of the dialogues. We have developed distributed dialogue management techniques for some of our other spoken dialogue applications that use agents to handle same tasks differently [16]. We have also developed a model that enables physical distribution of dialogues, where high level coordination is done by generating dialogue descriptions using the VoiceXML standard [5]. These descriptions are realized by terminal devices so that low level decisions are done by the devices. Such distribution gives the ability to choose output devices and modalities on the fly. These models and techniques increase general robustness and flexibility of the environment, for example regarding accessibility problems [9].

Finally, regarding to our experience it seems quite likely that ubiquitous computing environments will not appear instantly as single monolithic systems, but slowly over time. Applications, services and systems might have differing aims, interaction styles and technological standards. They will be small increments, sometimes adding, modifying or removing parts of the whole environment. The incremental development of our environment has been made possible by the use of the general system architecture.

5. Conclusions

We have presented how speech-based and auditory interaction can be used in augmented office environments. Speech recognition, speaker recognition and speech synthesis technologies are combined with other technologies and information sources to create applications for guidance, computer-supported communication and providing unobtrusive awareness information. These applications together form the augmented office environment that helps people in their everyday tasks.

6. References

[1] Ishii, H., Wisneski, C., Brave, S., Dahley, A., Gorbet, M., Ullmer, B. & Yarin, P. "ambientROOM: Integrating Am-

- bient Media with Architectural Space. (video)", *Conference Summary of CHI '98*, 1998.
- [2] Kidd, C., Orr, R., Abowd, G., Atkeson, C., Essa, I., MacIntyre, B., Mynatt, E., Starner, T. & Newstetter, W. "The Aware Home: A Living Laboratory for Ubiquitous Computing Research.", *Proc. of International Workshop on Cooperative Buildings*, 191-198, 1999.
- [3] Turunen, M. "Jaspis – A Spoken Dialogue Architecture and its Applications", *PhD dissertation, University of Tampere, Department of Computer Sciences, Report A-2004-2*, 2004.
- [4] Turunen, M. & Hakulinen, J. "Agent-based Adaptive Interaction and Dialogue Management Architecture for Speech Applications.", *Proc. of Fourth International Conference TSD 2001*, 357-364, 2001.
- [5] Salonen, E-P., Turunen, M., Hakulinen, J., Helin, L., Prusi, P. & Kainulainen, A. "Distributed Dialogue Management for Smart Terminal Devices." *To appear in Interspeech 2005*, 2005.
- [6] Prusi, P., Kainulainen, A., Hakulinen, J., Turunen, M., Salonen, E-P. & Helin, L. "Towards Generic Spatial Object Model and Route Guidance Grammar for Speech-based Systems." *To appear in Interspeech 2005*, 2005.
- [7] Headon, R. & Curwen, M. "Recognizing Movements from Ground Reaction Force.", *Proc. of Workshop on Perceptive User Interfaces*, 2001.
- [8] Raubal, M. & Worboys, M. "A Formal Model of the Process of Wayfinding in Built Environments.", in Freksa, C. & Mark, D. (Eds.), *Spatial Information Theory - Cognitive and Computational Foundations of Geographic Information Science, International Conference COSIT '99, Lecture Notes in Computer Science, Vol. 1661*, 381-399, 1999.
- [9] Turunen, M., Hakulinen, J., Rähkä, K-J., Salonen, E-P., Kainulainen, A. & Prusi, P. "An architecture and applications for speech-based accessibility systems.", *IBM Systems Journal, Vol. 44, No. 3, 2005*. (to appear)
- [10] Dale, R., Geldof, S. & Prost, J-P. "CORAL: using natural language generation for navigational assistance.", *Proc. of twenty-sixth Australasian computer science conference on Conference in research and practice in information technology - Volume 16*, 35-44, 2003.
- [11] Schmandt, C. "Everywhere Messaging.", *Lecture Notes in Computer Science, vol. 1707*, 1999.
- [12] Weiser, M. & Brown, J. "The Coming Age of Calm Technology.", *Beyond Calculation: The next fifty years of computing*, 1997.
- [13] Mäkelä, K., Hakulinen, J. & Turunen, M. "The Use of Walking Sounds in Supporting Awareness.", *Proc. of ICAD 2003*, 2003.
- [14] Mäkelä, K., Salonen, E-P., Turunen, M., Hakulinen, J. & Raisamo, R. "Conducting a Wizard of Oz Experiment on a Ubiquitous Computing System Doorman.", *Proc. of International Workshop on Information Presentation and Natural Multimodal Dialogue*, 115 – 119, 2001.
- [15] Turunen, M. & Hakulinen, J. "Jaspis 2 - An Architecture For Supporting Distributed Spoken Dialogues.", *Proc. of Eurospeech 2003*, 1913-1916, 2003.
- [16] Salonen, E-P., Hartikainen, M., Turunen, M., Hakulinen, J. & Funk, J. "Flexible Dialogue Management Using Distributed and Dynamic Dialogue Control.", *Proceedings of ICSLP 2004*, 2004.