Development of a Location-Aware Speech Control and Audio Feedback System

Lasse Kaila
Tampere University of Technology
Department of Electronics
Tampere, Finland
lasse.kaila@tut.fi

Jouni Hyvönen, Markus Ritala, Ville Mäkinen,
Jukka Vanhala
Tampere University of Technology
Department of Electronics
Tampere, Finland
{jouni.hyvonen, markus.ritala, ville.makinen,
jukka.vanhala}@tut.fi

Abstract—**Speech is a natural form of interaction for people, and smart environments offer good possibilities for experimenting with new kinds of interaction with users and devices. Speech control can offer more intuitive and easy forms of interaction with the environment compared to traditional graphical or physical user interfaces. This paper describes the development, implementation and current ongoing work on a location-aware speech control and audio feedback user interface.**

Keywords: speech control, audio feedback, context-awareness, smart environments

I. Introduction

A smart environment contains devices, sensors, actuators and user interfaces, all connected to each other through networks [1]. A main advantage that smart homes are able to offer is the ability to transfer the user interface (UI) from a device to another location, for example controlling the washing machine from a mobile phone would be possible [2]. UIs can also be grouped together (e.g. to a webpage) or completely transformed to another form (e.g. opening a door by double-tapping on the floor). Graphical and other physical UIs each have their advantages, but there are situations where a more intuitive, natural form of interaction with the smart environment would be useful. SCARS (Speech Control and Audio Response System) is an experimental speech control system that enables human-computer interaction by accepting spoken commands from users and returning spoken feedback from the computer. It is also aware of the user's location, bringing contextual information into the interaction scenario.

II. Speech Control User Interfaces

Speech is a natural form of interaction for people, and using speech as a form of interaction allows users to focus their attention elsewhere as they do not have to concentrate on reading text on a screen, pressing buttons etc [3]. As the amount of complexity in graphical user interfaces and the amount of available functions in a smart environment increase a more intuitive way of interfacing with the home system is welcome. This is especially important when considering mobile devices and smart clothing where UIs are usually restricted in size; with small screens and tiny buttons usability is often reduced drastically [4]. In an ideal situation a speech control interface would allow hands-free operation, something that would be very useful for example when coming home from the grocery store carrying large shopping bags or when physical controls are out of reach.

Speech control in smart environments has been a popular research subject for years, and tools for interpreting human speech and generating artificial speech steadily keep improving. One experiment in the Aware Home in Georgia, USA, the Family Intercom [5] researched inter-family interaction with a location-aware speech control system. The Family Intercom was used to communicate between family members either inside the home or to another remote family member. The goal was to integrate voice communication capabilities into everyday objects in a home. The Intelligent Room [3] is another project that utilised speech recognition for interacting with the computer in a conference room, making it possible to give commands to the computer or ask for certain information. The Technical Research Centre of Finland (VTT) has also conducted research in analysing audio content and environmental sounds [6]. In an experiment they attempted to recognise seven different contexts (e.g. car, elevator, rock music, speech) by their auditory footprints. The MPEG-7 standard [7] was used to categorise and describe the audio content captured in each situation and researchers were able to achieve a context detection accuracy of 80 %.

III. Research Motivation

The Department of Electronics has been researching smart environments since 1999 and the institute has a laboratory, the Smart Home (shown in Fig. 1), dedicated to smart home research [1]. The Smart Home is a testing space designed to resemble a normal apartment with a bedroom, kitchen, living room, bathroom, sauna etc. Hidden beneath structures, removable ceiling and wall tiles etc. are many different kinds of sensors and actuators. These can be monitored and controlled using multiple user interfaces, connected together by a home network and a central computer. User interfaces in the smart home include graphical UIs, a mobile phone UI, wall-mounted touch-sensitive panels and traditional light switches and buttons. With the availability of sufficient processing

978-1-4244-3304-9/09/$25.00 ©2009 IEEE

power and decent speech recognition software a speech control and audio feedback system was set up. SCARS [4] was designed in order to test how a location-aware speech control interface would work in a smart environment. It was not designed to become the sole UI in the Smart Home but instead it would complement other existing UIs. By combining wireless microphones, ceiling-mounted speakers and a positioning system it would be possible for users to give voice commands and receive auditory feedback. Voice commands would be dependent on the user's location, and auditory feedback would also be directed to a speaker near the user.

Figure 1. The Smart Home laboratory.

IV. Design Issues

For speech recognition systems it is naturally vital to capture the speech from users with the best quality possible. Speech recognition systems can easily become unusable if the level of ambient noise becomes too high or if there are other factors that interfere with the voice of the speaker. A fundamental design decision is whether microphones are used for constant monitoring of speech or if they are activated by the user in some way (e.g. by pressing a button in the same way as when using a walkie-talkie). The former way was used in the Intelligent Room [3], where users were required to say a pre-determined phrase (in this case it was "Computer") that initiated the speech recognition sequence. For example users wanting to control lights would say "Computer, lights on". However, in a larger environment, such as an apartment or a house constant monitoring of speech becomes more complicated as the amount of microphones and audio processing required becomes significantly larger. Also, in the case of multiple users in a space it becomes difficult to identify who is merely talking and who is giving commands. For context-aware applications it would be beneficial to know the identity of the user, and an identifiable portable microphone unit could offer this possibility. Requiring users to press a button in order to give a speech command to the system is not very practical from the users' point of view but on the other hand it allows the system (and the user) to know exactly when a command is sent to it, greatly improving recognition accuracy. These are few reasons why the handheld unit-type solution with a pushbutton was considered for the Smart Home.

Context awareness is a key component in a smart environment [8]. For example the user's location can reveal a lot of his/her intentions and when data has been collected for a longer period of time it can reveal certain trends and patterns. This is why location information was integrated in the SCARS system, with different kinds of technologies being evaluated in order to get a rough estimate of the user's location. In addition to SCARS experiments with another kind of context-aware control device have been made in the Smart Home [9]. It is a wireless home remote control that can be used to control all devices in the home using a joystick and menus. The remote also uses the SCARS infrared positioning infrastructure.

V. System Description

SCARS hardware, shown in Fig. 2, consists of five parts: the receiver unit, wireless microphones, a positioning system, ceiling-mounted speakers and a software module. All hardware has been custom built at the university, being specially built to suit the needs of the Smart Home and to integrate into the other parts of the infrastructure.

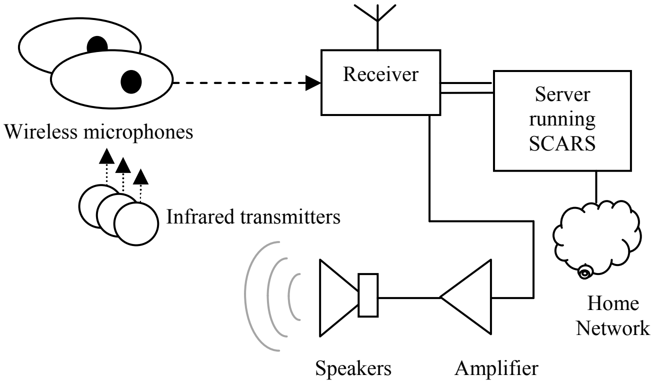

Figure 2. SCARS hardware.

The handheld unit, shown in Fig. 3, consists of a main board, rechargeable battery, FM radio transmitter module, pushbutton, microphone and an infrared receiver. Battery life is sufficient for one day's use, and when the unit is not in use it can be placed into a recharging station. The button in the handheld unit is used for activating audio transmission; when it is pressed a signal is sent to the receiver unit and an audio channel is opened. Now the user is free to talk into the microphone, and the audio signal is transmitted to the receiver and further to the computer for processing. There are a total of three handheld units in the Smart Home, and each has a unique 16-bit identifier code which allows multiple simultaneous active users (however only one audio channel is available, so users cannot talk into the microphone at the same time). The identifier code also makes it possible to personalise units or to grant them different levels of access.

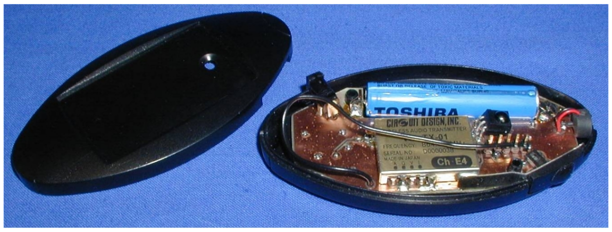

Figure 3. SCARS Handheld Unit.

The positioning system consists of multiple ceiling-mounted infrared diodes that transmit an ID code every few seconds. The code is room-specific, and also hallways and the front door have their own ID code. The ID code is received by an infrared receiver on the handheld unit and transmitted FSK-modulated over the FM radio link to the receiver unit, which decodes the ID and transmits a location to the server. This way there is no need to implement a separate data transmission channel or radio.

The receiver unit is located in a closet in the bedroom in the Smart Home. It consists of a FM receiver unit with an antenna, audio relays and connectors to the computer (RS-232 and audio in/out). The receiver is responsible for receiving analog audio from the handheld unit and relaying it to the server. It also uses the location code received from the handheld unit to direct audio feedback (received from the server) to an appropriate speaker by controlling relays and audio outputs. The receiver unit can also accept commands from the server and direct audio to a desired location, for example when someone is at the front door or in the kitchen.

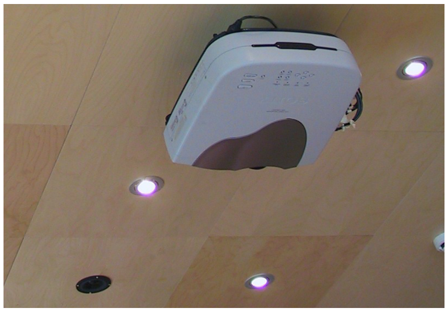

Figure 4. Ceiling showing lights and speakers. Infrared diodes are placed between ceiling tiles.

The audio hardware consists of speakers and amplifiers. Speakers, one for each location ID code, are mounted in the ceiling and connected to a multi-channel amplifier. There is a total amount of seven speakers and they are driven by two four-channel audio amplifiers. Thus it is possible for the system to direct audio to any (or multiple) speaker in the Smart Home, depending on the situation.

The server, a miniature-sized PC, runs the server software for the Smart Home, controlling devices and user interfaces. It also runs speech recognition software, Philips FreeSpeech Viva and a SCARS controlling module. FreeSpeech Viva was chosen because at the time it was one of very few speech recognition software that was able to recognise Finnish words. A set of pre-selected commands (approx. 20) were recorded to the software together corresponding actions (key presses) that would be captured by the SCARS module. When audio is received from the microphones it is first processed by the speech recognition software, compared to pre-recorded commands and if a match is made it sends a key press to the SCARS module. The SCARS module then reads the code, communicates the appropriate task to the Smart Home server software (e.g. turn on lights in the living room) and directs an audio confirmation to a speaker near the user. Audio feedback is created with a text-to-speech synthesizer program (Mikropuhe), which allows SCARS to read aloud any strings in text format. This feature can also be used to generate notifications to the user (e.g. reminders, alarms), personalised greetings or output of information coming from another UI or device. One application using this feature is the fingerprint scanner on the front door of the Smart Home; when the user is recognised the fingerprint scanner software sends a text string (containing the user's name and a random greeting) to SCARS, which in turn opens the front door. The greeting message can be for example "Welcome John, nice day today" and it will be played from the speaker above the front door.

VI. Using SCARS

SCARS is used by grabbing the microphone, pressing and holding the button on the handheld unit and waiting for a beep from the speaker. Now the system is ready to accept a command. The user speaks a command into the microphone, e.g. "turn on lights" and releases the button. The command is processed by the computer, lights turned on in the room where the user is located, and a confirmation message played from the speaker "living room lights turned on". Commands that are recognised by the software include direct commands, such as turning on/off specific groups of lights, controlling window blinds, opening the front door etc. Enquiries are also possible, for example it is possible to ask what the temperature outside is, is the sauna ready etc. Some commands are location sensitive, so commanding "turn on lights" would turn on lights in the location where the microphone has been detected, whereas "turn on lights in the kitchen" would always control the kitchen lights. If the server is set to play music the audio output will automatically follow the user, changing output speaker to the one that is closest to the user.

VII. Preliminary Findings

SCARS has been implemented in the Smart Home and the system was tested by groups of students attending the "Modern User Interface Electronics "– course [10]. In general test subjects were positively surprised of how easy it was to learn to use SCARS, and voice commands were also successfully recognised even among users of different ages and gender. Users found the response time to be too long, according to them there is a noticeable delay (from ~200 ms up to a second) from pressing the button on the handheld unit until the confirmation beep is heard from the speakers. This can result in users being confused about when they can talk into the microphone and when it is OK to release the button. Portability of the microphone was decent but the test subjects also agreed that it would not be practical to carry a microphone around with them in everyday life. Possibly the largest issue was the dependability of the positioning system; sometimes it took a long time for the system to figure out where the microphone is located, and near the border of two rooms it can give false readings. Because the system relies on infrared transmission it is also possible that something blocks the light beam, preventing the system from getting a location signal. If the location has been incorrectly sensed audio feedback will be heard from another room and location-aware commands will also perform actions in the wrong room. Practical problems arose also in conjunction with controlling A/V equipment in the living room. If the sound level of a movie playing was too loud there it became rather difficult to lower the volume or give any other commands with the microphone. Commands related to A/V equipment have since been removed from SCARS as they are more likely to be issued from the universal remote on an internet tablet anyway. Another workaround would be to have the system automatically mute the amplifier when the button on the handheld unit is pressed.

VIII. Further Development

According to these test results and from personal experiences there seem to be good reasons for developing the system further. First, the vocabulary of the speech recognition software has to be greatly increased by creating new commands and several aliases for existing ones. Then it would not matter as much if the spoken word did not exactly match the recorded data. Second, the handheld unit should be improved and possibly integrated into a wearable badge or a possibility to use the microphone in a mobile phone added. There are also other benefits of integrating wearable computing with smart environments [11] and since our research group is also involved in smart clothing this presents more interesting opportunities. An integrated microphone would relieve users from carrying around extra pieces of equipment, hands free operation could be enabled by eliminating the button and introducing other ways of initiating dialogue with the computer. Last, the part that requires improvement the most is the positioning system. There are numerous alternatives for replacing infrared beacons and receivers, fortunately the accuracy requirement is not terribly big for this application (as it is enough to have a rough estimate on in which room the user is located). However there are other applications that would largely benefit from an accurate, dependable infrastructure for locating users, and thus more accurate technologies should be taken into consideration. Currently tests are conducted in the Smart Home using capacitive positioning with floor tiles as receivers/transmitters. This system would detect users walking around with an accuracy of about 30 cm and require little extra hardware. However with several users roaming around in the home it becomes difficult to identify them and knowing which one of them is talking.

Another way of obtaining rough location estimates would be to use Received Signal Strength Indication (RSSI) – information from a radio transceiver that is able to obtain this information. RSSI is an estimate on how strong the received radio signal is, and it can be used for determining how far away the transmitting device is (e.g. a server or network node). By using multiple reference signals it is possible to triangulate the position of the mobile node. Bluetooth would be a good candidate to replace the FM radio (although it would consume more power) and it would also support RSSI. However it would seem that RSSI is by nature very unreliable due to the effect of obstacles, structures and various other radio signals, and recent studies have shown that positioning based on RSSI is too unreliable to be of any use in this application [12].

Replacement and upgrade of other devices and wired networks is also underway in the Smart Home, and the development for SCARS fits well inside this time frame. The plan is to make the Smart Home context-aware by using various sensors, contextual information gathered by users' routines, actions and habits and a adaptive, learning home controller software.

IX. Conclusion

This paper presents the practical implementation of a location-aware speech control and feedback system for smart environments. The goal was to create an alternative way of controlling devices in a home environment and add more mobility and freedom compared to existing user interfaces. In addition, location information is used in conjunction with speech commands, making the system aware of the user's location. A prototype system has been built and set up in the Smart Home laboratory and tested, being in use every day. Findings from using the system have shown that the concept is viable and that there are situations where speech control and audio feedback is useful when interacting with the smart home environment. Further development, however, is needed to make positioning more accurate and the functionality of the system more reliable and valuable.

References

[1] L. Kaila, J. Mikkonen, A-M. Vainio, J. Vanhala, "Open architecture for practical implementation of smart homes", in: proceedings of Telecommunications, Networks and Systems (TNS 2007), Lisbon, Portugal, July 3-8, 2007.

[2] J. Plomp, "Generic UI for interaction with future home environments" In: proceedings of the Dreaming for the future conference, Helsinki, Finland, 2001.

[3] M. H. Coen, "A prototype intelligent environment", in: Proceedings of the first international workshop on cooperative buildings, integrating information, organization, and architecture (Cobuild '98), pp. 41-52, Lecture Notes In Computer Science, Vol. 1370, 1998.

[4] J. Hyvönen, "Speech control system for intelligent environment", M.Sc. thesis, Tampere University of Technology, 2003.

[5] K. Nagel, C.D. Kidd, T. O'Connell, A. Dey, G.D. Abowd, "The Family Intercom: Developing a Context-Aware Audio Communication System", Lecture Notes in Computer Science, pp. 176-183, Springer Verlag, 2001.

[6] S-M. Mäkelä, "Environmental sound analysis", Interactive home seminar, Oulu, Finland, 2001.

[7] International organisation for standardisation, MPEG-7 Overview, available at: http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm.

[8] A. K. Dey, "Understanding and using context", Personal and Ubiquitous Computing Journal, Volume 5 (1), 2001, pp. 4-7.

[9] M. Ritala, T. Tieranta, J. Vanhala, "Context aware user interface system for smart home control", in: proceedings of the home oriented informatics and telematics 2003 (HOIT2003), The Networked Home and the Home of the Future, California, USA, April, 6-8, 2003.

[10] Modern user interface electronics course, usability studies conducted by groups of students, Tampere University of Technology, 2003, unpublished.

[11] L. Kaila, "Expanding smart clothing with smart environments", in: Proceedings of the Doctoral Colloquium session, IEEE International Symposium on Wearable Computers (ISWC 2005), pp. 5-7, Osaka, Japan, October 18-21, 2005.

[12] K. Benkic, M. Malajner, P. Planinsic, Z. Cucej, "Using RSSI value for distance estimation in wireless sensor networks based on ZigBee", in: Proceedings of the 15th Internatnional Conference on Systems, Signals and Image Processing (IWSSIP 2008), pp. 303-306, 25-28 June 2008.