Multimodal Media Center Interface based on Speech, Gestures and Haptic Feedback

Markku Turunen[1], Jaakko Hakulinen[1], Juho Hella[1], Juha-Pekka Rajaniemi[1], Aleksi Melto[1], Erno Mäkinen[1], Jussi Rantala[1], Tomi Heimonen[1], Tuuli Laivo[1], Hannu Soronen[2], Mervi Hansen[2], Pellervo Valkama[1], Toni Miettinen[1], and Roope Raisamo[1]

[1] University of Tampere, Tampere Unit for Computer-Human Interaction, Finland
{firstname.surname@cs.uta.fi}

[2] Tampere University of Technology, The Unit of Human-Centered Technology , Finland
{firstname.surname@tut.fi}

Abstract. We present a multimodal media center interface based on speech input, gestures, and haptic feedback (hapticons). In addition, the application includes a zoomable context + focus GUI in tight combination with speech output. The resulting interface is designed for and evaluated with different user groups, including visually and physically impaired users. Finally, we present the key results from its user evaluation and public pilot studies.

Keywords: Speech, haptics, gestures, multimodal interaction, media center.

1 Introduction

Multimodal interaction may bring many benefits for interaction with digital home appliances and their digital content. The so-called media centers, which combine the commonly used digital media (television broadcasts, music, photographs etc.) to a single device/application, form one potential area. There are plenty of such systems, both hardware solutions and software applications for personal computers and game consoles. Typically, they are used with a remote controller, which, in many cases, makes the interaction clumsy or even too complex for some users.

In this paper we present how speech input, gestures, and haptic feedback can provide an efficient interface for media centers. Furthermore, we present how they can be made accessible for different user groups. For blind users, speech output and haptic feedback provides full access to the information, while the zoomable GUI makes it accessible for visually-impaired (partly seeing) users. Finally, speech input combined with voice activity and blow detection makes the interface usable for physically impaired people who cannot use their hands at all.

2 Media Center Application

As a part of the project TÄPLÄ (Ambient Intelligence based on Sound, Speech and Multisensor Interaction) we are developing multimodal interaction methods for home environments. After a large consumer survey [3], we have focused on developing a multimodal user interface for a media center. This application area is rapidly becoming popular in homes, and provides opportunities and challenges for user multimodal interaction [1,2]. Our media center provides users full control over digital television content, including an advanced electronic program guide (EPG).

Our solution is based on a low-cost PC (Athlon X2 3800) running the media center server software, a mobile phone (Nokia N95) as an interface device, a wireless access point to connect these together, and a high-definition 40" television (Figure 1). The main application is written in C# and runs under Windows XP. In it includes speech recognition and speech synthesis. The mobile device has embedded gesture recognizer, speech recognizer, haptic feedback controller, speech synthesizer, and a GUI. All mobile technology components are native Symbian applications while the GUI and main logic is written in MIDP 2.0.

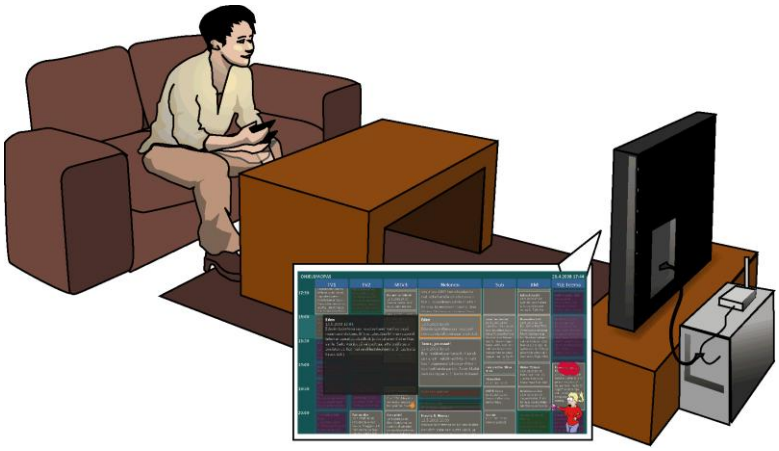

Fig. 1. Media center setup with the EPG user interface.

3 Multimodal User Interface

The system has two graphical displays; the large television screen and the mobile phone display. Both devices have also speech output capability, and the mobile device provides haptic feedback. As an alternative for the mobile phone, we have implemented a wheelchair mounted wireless microphone solution for physically impaired users. It is, however, out of the scope of this paper.

The main GUI on the television screen consists of sections for different application functionality. Here, we focus on the EPG interface. It consists of a matrix, where

columns represent channels, and rows represent time slots. Cells are individual television programs. In order to utilize the high-definition resolution for different users, we apply several Focus + Context techniques to the interface. First, the columns and rows near the center of the display are slightly enlarged. Second, a strong enlargement is applied for the active program, while transparency is used to make the content under the enlarged area visible. Third, the GUI is freely zoomable from weekly overviews to extreme close-ups. Finally, overlaid animated icons give guidance and feedback for gesturing and speech input. The mobile phone GUI applies similar techniques, originally developed in different domain.

Users have full control over the media center via its speech user interface (SUI). It includes navigation in the application (e.g., "*Go to program guide*") and EPG (*"Show Monday afternoon"*), and watching media ("*Show Documentary channel*"). In EPG, it is possible to record multiple programs with a single utterance (*"Record all the Tom the Tractor shows this week*"), and to highlight programs based on their genre ("*Show me all the children programs*"). These commands were found highly appropriate among the test users in the public pilot studies and user evaluations. The SUI is implemented with context-free grammars, the vocabulary being about 900 words. There are two recognizers: an embedded recognizer running on the mobile phone, and a server-based recognizer. The choice of the recognizer is a balance between speed, vocabulary size and accuracy. Two variations of the 'push-to-talk' approach were implemented. In the first approach, the user presses a key on the mobile device to speak. The second approach combines gesture recognition with voice activity detection: the voice activity detector is activated when a user raises the phone to an upright position in front of his or her mouth.

The mobile phone keypad and gestures can be used for navigation and selections either independently or combined. Navigational gestures are made by moving the mobile phone in specific patterns. The accelerometer-based gesture recognizer supports tilting of the phone for moving the selection up and down on the screen, swinging left and right for left and right movements, swinging forward for selection, upwards for cancel, and shake to view help. Alternatively, the different orientations of the mobile phone can alter how the keypad works. In the vertical/down orientation keys are used to move selection in the EPG, in the vertical/middle orientation keys move the EPG display area, and in the horizontal orientation keys perform zooming. A combination of rule-based methods and Hidden Markov Model (HMM) based statistical model is used for gesture recognition, similar to the approach presented in [5]. The swing and shake gestures are recognized using HMMs, while the other gestures are recognized with rule-based methods.

Haptic feedback is given using the vibration component of the mobile phone. We defined a markup language to generate different kinds of haptic patterns, hapticons. Patterns are series of pulses with differences in direction of the intensity, motor, and delays between the pulses. With a set of ten control parameters (e.g., start/end intensity), it is possible to create rather sophisticated rhythmic patterns. We defined in total nine different hapticons. Recognition results, user interface actions, and the state of the application are mapped into these patterns. For example, the phone vibrates with different patterns when it recognizes a gesture, and when it receives a speech recognition result.

For visually-impaired users (who are heavy users of digital media, including digital television), the system includes a tight integration between the Focus + Context GUI and synthesized speech output. In a nutshell, the content of the active area is read out loud by the speech synthesizer when the item is activated. The spoken content is not the same as the text on display, since speech and text have different strengths and weaknesses. For example, speech outputs should use full sentences to keep the message easily comprehensible, and to allow efficient browsing, the most important information must be spoken first. The interface is based on our earlier work with mobile multimodal interfaces.

3 Pilot Studies, User Evaluations

We arranged a long-term public pilot study between June 2008 and March 2009. A mock living room with the media center was built inside the local media museum Rupriikki (Tampere, Finland). Museum visitors could freely use the system, and provide feedback using a web questionnaire. Several user studies were also conducted with museum visitors. In order to study the expectations and user experiences of different modalities, we also arranged a controlled user experiment with 26 participants in a laboratory. We used a subjective evaluation method called SUXES. The results (which are described in detail in [4]) show that the perceived quality of the speech input surpassed the upper limit of user expectations, indicating that users were very positively surprised by the performance of the speech interface. Furthermore, the low amount of out-of-vocabulary sentences shows that the restricted speech interface can be used efficiently in this domain. Users were also positive towards the application in general. In the future, we will evaluate the application in long-term pilot studies with physically and visually-impaired users. Finally, we will release the system for general public to collect feedback and usage statistics from a large amount of real users.

References

1. Ibrahim, A., and Johansson, P. 2003. Multimodal Dialogue Systems: A Case Study for Interactive TV. Carbonell, Noelle; Stephanidis, Constantine (Eds.) Universal Access. Theoretical Perspectives, Practice, and Experience, 7th ERCIM International Workshop on User Interfaces for All, Revised Papers. Springer, LNCS, Vol. 2615. 209-218.
2. Wittenburg, K., Lanning, T., Schwenke, D., Shubin, H., and Vetro, A. 2006. The prospects for unrestricted speech input for TV content search. In Proceedings of the Working Conference on Advanced Visual interfaces (AVI '06). ACM, New York, NY, 352-359.
3. Soronen, H., Turunen, M., and Hakulinen, J. Voice Commands in Home Environment - a Consumer Survey. In Proceedings of Interspeech 2008: 2078-2081, 2008.
4. Turunen, M., Melto, A., Hella, J., Heimonen, T., Hakulinen, K., Mäkinen, E., Laivo, T., and Soronen, H. User Expectations and User Experience with Different Modalities in a Mobile Phone Controlled Home Entertainment System. In Proceedings of MobileHCI 2009.
5. Schlömer, T., Poppinga, B., Henze, N., and Boll, S. 2008. Gesture Recognition with a Wii Controller. In Proc. TEI 2008, ACM Press, 11-14.