



WIRELESS WORLD

RESEARCH FORUM

Advanced User-Interface Technologies for Mobile Applications

A. Steinhage, Infineon Technologies, Munich, Germany

Abstract—This paper presents the current state of our research on novel user-interfaces (UI) for embedded applications. Our approach is based on the conviction that the growing complexity of embedded devices calls for new forms of man-machine interaction. Guided by the role model of human-to-human dialogs, we aim at an anthropomorphic assistive interface which is based on natural communication channels such as speech, gesture and mimics.

Index Terms—User-interface, natural communication, virtual personal assistant.

INTRODUCTION

Investigating the way in which customers interact with current high-tech devices two interesting observations can be made: First, the number of features and functions increases permanently. This is obvious in the domain of mobile communication where cell phones of the current generation feature voice-, text- and video-communication, web-surfing, personal information management, recording and playback of music, photos, video and FM radio reception. However, also the feature list of home entertainment devices (e.g. TV set top boxes), household appliances (e.g. computerized washing machines) and even cars (e.g. navigation systems, on-board entertainment etc) grows from generation to generation.

The other observation concerns the diversity of customer groups: a decade ago, mobile communication, for instance, was a domain for a small group of technologically experienced people only. Today, a cell phone is common equipment for almost everyone including the large group of technological laymen.

Taking these two developments into account, it is obvious that the average user becomes more and more overwhelmed with

the operation of the devices' functions [1]. In the mobile communication industry, this effect is known as the Usability Gap [2]. This phenomenon is expected to have a growing negative influence on the sales figures: already now the average customer uses less than fifty percent of the features of his/her cell phone. Accordingly, the manufacturers will have growing problems to convince the user of the necessity of new functions if they are too complicated to be operated. This problem is not unique to mobile communication: the manufacturers of feature-loaded DVD-recorders, digital photo-cameras, washing machines and modern cars stuffed with electronics suffer from the same effect.

An obvious way out of this dilemma is the improvement of the user interface (UI) and the integration of assistive functions which are easy to operate and adapt autonomously to the user's preferences and level of expertise. The underlying complexity of the functions should be hidden from the user and the devices should be operated through a standardized interface using robust natural communication channels. The ideal assistive interface would be able to run an interactive dialog with the user based on speech, gesture and mimics.

However, the realization of such an advanced user interface requires large efforts in research and development: first of all, the potential features and functions of future devices need to be identified. This can only be done by joining forums like the WWRF, where all the players in the specific domain are present.

The next step is the requirement analysis of assistive user interfaces. Here, the focus must be put on the robustness, as most applications in mobile communication and



WIRELESS WORLD

RESEARCH FORUM

automotive deal with high noise levels and varying visual conditions. This robustness can be achieved by exploiting the inherent redundancy when integrating many different communication channels.

Therefore, not only the algorithms for each channel but also their integration contribute to the computational complexity of the interface. However, the algorithms for vision and speech processing share many common functions such as filter operations or other parallel instructions. Hence, it is necessary to identify these common functional elements to be able to exploit the task- and data-parallelism of the full system. This is done by modelling, profiling and benchmarking the applications on the PC platform.

Current general-purpose processors for desktop PCs are powerful enough to run such an assistive interface. However, their price and power consumption forbid their usage in the embedded devices we focus on. Therefore, we use the detailed application-profiles and -benchmarks as guide during the design phase of an innovative embedded application processor. This processor will run not only the assistive UI but also provide functionality for the various multimedia applications predicted for future mobile phones.

Finally, it is clear that an assistive user interface must be equipped with a certain amount of artificial intelligence to lead a sensible dialog with the human user. Therefore, some effort must be spent on the development of a system which equips the assistive interface with anthropomorphic capabilities such as reactivity, adaptivity and context awareness.

Within the paper on hand we will focus on one of the steps of the aforementioned research, i.e. the natural communication channels. For the example of face tracking we show that robust image processing can be done even with low quality sensors. Further, we show how an animated humanoid 3D head can be controlled by that face tracker to result in a virtual "person" equipped with gaze following. This animated model builds the basis for a *Virtual Personal Assistant* (VPA): an intelligent

anthropomorphic companion in future mobile phones. Then, we will give some results on our research on audiovisual speech recognition and we will present a simple approach to the formalization of the behavioral organization for such a VPA. Finally, we will give first results of the requirement analysis for the described algorithms.

Natural Communication Channels

Classical user-interfaces such as keypad and display of a mobile phone are sufficient as long as the number of functions to control remains of the same order of magnitude as the number of control elements. However, shrinking devices and growing number of features enforce keys and switches each carrying multiple functions. Therefore, the operation becomes more and more cumbersome and the need for new robust, standardized and user-friendly means of interaction arises. A role-model for such an interaction is a human dialog: based on the integration of multiple partly redundant communication channels such as speech, gesture, mimics and gaze, a high level of robustness is achieved. The usage of these channels is standardized among humans of the same cultural group. No adaptation or initial configuration is needed. The dialog can directly begin without "reading a manual of operation".

Given these advantages, the implementation of natural communication channels in man-machine interfaces appears to be required to solve the usability gap problem. Many IT companies conduct research in this direction and some products are already on the market: speaking navigation systems, voice dialing, speech controlled car radios and even video games that detect the player's hand movements. What is still missing, however, is the integration of many channels, into one interface. Only when speech, gesture and mimics are perceived simultaneously, a high level of robustness can be achieved. A car radio which is switched on depending on the spoken keyword "radio", for instance, might



WIRELESS WORLD

RESEARCH FORUM

unintentionally become activated during a conversation about radios. Including a channel which detects the driver's gaze or pointing, for instance solves this problem: only when the driver looks or points towards the radio and simultaneously utters the keyword, the device is switched on.

In the following, we will describe our approach to this type of speech- and image processing.

Image and speech processing

We have developed a robust pointing- and face tracking algorithm based on the fusion of color-, movement- and shape-cues (for details see [3]). For the standard situation of a single user observed by a camera (e.g. the driver in a car or the user of a mobile phone), the center of a skin-colored image region which has the oval shape of a human head is defined as the face position. This position is tracked over time by means of a low-pass filter. This algorithm turns out to be robust enough to follow the user's face in a wide range of illumination conditions.

We implemented the algorithm on a PocketPC® as "Virtual Cameraman" (see Fig. 1).

A patent is pending for the application as hands-free videophone: while the user is walking through the room, his/her face is



Fig.1. "Virtual Cameraman" on a PocketPC®: From a camera video the region containing the user's face is extracted and tracked over time. If only this portion is transmitted, a video-conferencing partner has the impression the camera is actively following the head. tracked from the camera image of the mobile

phone standing on the desk. Only the video's interesting portion containing the face needs to be transmitted to the other mobile dialog partner. This saves bandwidth and takes the small display sizes of mobile phones into account.

Using a similar image processing algorithm, which additionally takes into account the speed of moving skin colored objects in the video, the pointing direction of a user can be detected. We have implemented this mechanism in a demonstration we call "Smart Cockpit": a user can manipulate switches and controls on a virtual car cockpit displayed on a large screen (see Fig. 2).

As third input channel, we have implemented an audiovisual speech recognition system which fuses two complementary channels: an ordinary



Fig.2. "Smart Cockpit": Using an image processing algorithm based on skin-color and movement speed, a user's pointing gestures are recognized robustly. In this example, interaction elements on a car dashboard can be controlled.

Hidden-Markov-Model (HMM) – based speech recognizer and a lip-reading algorithm [4]. This algorithm works by extracting a feature vector based on the motion of specific regions around the user's mouth (see Fig. 3).

This feature vector is combined with the ordinary audio-based feature vector and fed into a single Hidden Markov Model (HMM) based speech recognizer.



WIRELESS WORLD RESEARCH FORUM



Fig.3: Motion-based tracking of lip regions.

By integrating the input from vision and audition in training and test on a simple keyword recognition task, we could achieve a remarkable improvement of the recognition rate in noisy environment [5] (see Fig. 4).

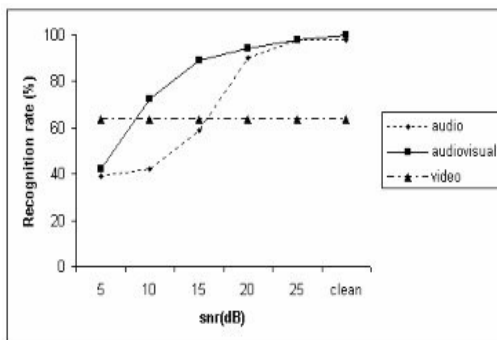


Fig.4: Audiovisual recognition rate for different noise levels compared with audio only and vision only

Animated head model

The final Virtual Personal Assistant will be an animated 3D model of a human head. To this end, we have integrated the aforementioned natural input channels into a commercially available 3D model from the company DigiMask® (see Fig. 5). By means of an ActiveX™-player we can control the position, the angular orientation and the facial expression of the model in real time. Furthermore, we can generate lip-synchronous speech output using the Microsoft™ text-to-speech-engine.



Fig.5: Animated anthropomorphic 3D model of a human head.

Finally, we have equipped the pose control of the head with the aforementioned face tracking algorithm. By means of this, the head can track the user's face with its gaze [6].

As described, we have implemented the following capabilities of the VPA so far:

- audiovisual speech recognition
- lip-synchronous speech output
- gaze following
- recognition of pointing
- mimics and head pose control

The next steps will be the recognition of the user's mimics and gaze.

Behavioral organization

As described in the introduction, the VPA can only fulfill its assistive function if its general behavior is appropriate for the current situation. To this end, we separate the overall behavior of the VPA into a set of so-called *elementary behaviors* [6]. Examples for these elementary behaviors are actions like "speaking", "listening", "following the user's gaze", "starting applications on the mobile device" etc. We assign an activity variable $a_i \in [0..1]$ to each of these elementary behaviors which controls its current state of activation (1=fully active, 0=not active). The current behavioral state of the VPA can then be described by the activity pattern of all a_i . Mathematically spoken, the a_i span a space (the behavioral space) in which the current behavioral state of the VPA is represented by a point. The overall behavior of the VPA in time (i.e. the sequence of



WIRELESS WORLD

RESEARCH FORUM

behaviors), corresponds then to a smooth trajectory in the behavioral space (see Fig. 6).

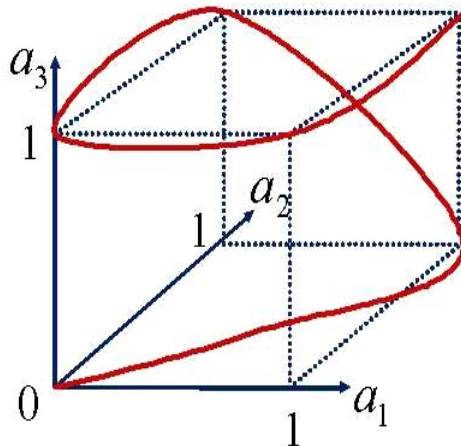


Fig.6: Visualization of the behavioral space spanned by the activity of each elementary behavior.

In the figure, an example for three elementary behaviours is depicted. This formalism captures both, elementary behaviours which can be partly activated (e.g. VPA is smiling) and behaviours which are binary (e.g. invoking an email client or not). In the latter case, we will define a threshold for a_i , above which the corresponding elementary behaviour will be activated.

Further, we define so-called sensor contexts $s_i \in [0...1]$ which encode the current state of the VPA's sensors on the level of the elementary behaviours. An example is the recognition of a specific spoken command (e.g. "start email client!") which is indicated by the speech recognition module setting the corresponding $s_j=1$. The vector of all s_j represents the current state of the sensor system with respect to the elementary behaviours. In this view, $s_j=1$ indicates that all pure sensor-based requirements for activating elementary behaviour a_i are fulfilled.

Some elementary behaviours will have a constant $s_j=1$. For instance, the speech engine of the VPA may become active independent of the current sensor situation. However, the activation of a specific

elementary behaviour may depend on the previous or simultaneous activity-state of another elementary behaviour.

Finally, each elementary behavior and each sensor context is equipped with a parameter p_i which can be used to receive and send additional information. For instance, the behaviour "track the user's face position" can be equipped with the current x/y position of that face in the video image. For other elementary behaviours, p_i may remain unused or may hold a default value only.

This formalism allows to generate complex behavioral sequences and the interdependencies between sets of elementary behaviors in a mathematical way. In this mathematical architecture, which is known as the "Dynamic Approach to Behavioral Organization" [7], the variables a_i , s_i and p_i become state variables of coupled dynamical systems. The nonlinear nature of these dynamics allows for a simulation of intelligent decision making as phase changes of the underlying dynamics. Whereas the general formalism is well known, extensive research has to be conducted to find the concrete parameterization of the VPA's behavioral repertoire.

The Virtual Personal Assistant

The final implementation of the VPA, which will look as in Fig. 7, will provide several assistive functions to the user:

- Interactive manual: The VPA offers spoken help about the operation of the mobile phone and answers corresponding questions by the user.

- Text to speech: The assistant reads out emails, appointments or web content to the user. This function is ideal for hands-free operation.

- Audiovisual answering machine: The VPA autonomously decides which calls to let through to the user and on which it will react as audiovisual answering machine. This decision is based on the current context (e.g. let no calls through when the user is currently in a meeting indicated by the list of appointments stored in the phone) and on previous experience (e.g. the user has never



WIRELESS WORLD

RESEARCH FORUM

picked up calls from a certain caller).

- Advanced Videophone: Videotelephony and -conference system in which instead of the full video streams showing the participants, only low bandwidth mpeg4 Facial Animation-Parameters (FAPs) are transmitted. The participants are rendered as realistic 3D head model on the screen and the mimics and head pose is morphed during the conversation according to the FAPs. This application is developed by our partner Fraunhofer Heinrich Hertz Institute (HHI) in Berlin. An integration with the VPA's 3D head model is possible so that the VPA can take part in a video conference and serve as the personal secretary of one human participant.

- Keyword driven web search: The assistant searches the internet for keywords spoken by the user. The result is read back at the user for hands-free information acquisition.



Fig.7: The VPA searches the internet for operation instructions of a ticket machine and reads back the result to the user.

Although the speech channel plays a major role in these examples, also mimics and gaze input is evaluated: a frowning face of the user can be detected and interpreted as dissatisfaction with the latest action of the VPA. This information can serve as training

signal for the VPA's learning process. The gaze of the user provides helpful information when deciding between text- or speech output: when the user looks at the display, the speech output may be switched off and be replaced by text output only. A fully fledged implementation of such a VPA can be regarded as an artificial personal secretary.

System Requirements and Functional Characterization

We have analyzed the described algorithms with respect to their computational complexity and functional contents. Expressed in Million Instructions per Second (MIPS) we arrive at the following rough figures:

- Text to Speech ~ 150
- 3D Animated Head ~ 800
- Keyword Spotting ~ 200
- Lip-Reading ~ 200
- Face Tracking ~ 200
- Behavioral Organization ~ 100

And we have some estimates for the functions, which are not yet implemented:

- Mimics Recognition ~ 300
- Scene/Object Recognition ~ 1500

We estimate the computational complexity of the fully fledged VPA to be around 10.000 MIPS.

It is obvious that under the special requirements for mobile devices, only novel embedded SoC architectures can provide the performance required for such a complex embedded application like the VPA.

Such a novel SoC architecture will have to exploit the inherent task- and data parallelism of the VPA's multiple speech- and vision algorithms. However, different tasks may share common subtasks and use a common set of library functions. We have extracted a list of these library functions shared by several different subtasks. This library is currently ported to a format suitable for bit- and cycle accurate performance tests on a virtual prototype of an SoC architecture



WIRELESS WORLD

RESEARCH FORUM

appropriate to run a fully fledged VPA. We expect a first prototype of this new application processor in silicon by the end of 2005.

Summary

The growing number of features and functions and the increasing diversity of user groups call for new robust and user friendly embedded man-machine interfaces. Integrating multiple communication channels and assistive agents that can lead a natural dialog with the human user can increase the market impact of future embedded devices. Similar to a dialog between humans, the inherent redundancy of many communication channels enhances the robustness of the overall system.

Given the particular requirements of embedded mass market devices with respect to power consumption, price and spatial dimensions and the computational complexity of the described assistive interface, novel embedded SoC processor architectures are required. Profiling and benchmarking user-interface algorithms for two fields of application, mobile communication and driver assistance, we support the design of such advanced SoC architectures for embedded devices.

Extensive research is required particularly to generate and organize the behavior of virtual assistants.

ACKNOWLEDGMENT

We are grateful to W. Hemmert, A. Techmer, M. Sanchez, D. Alonso and J.P. de la Cruz for their contribution to this work.

Parts of the described research activities are funded by the German Ministry for Research and Education (BMBF) within the project "VisionIC".

REFERENCES

- [1] K. Crisler et al., "Considering the User in the Wireless World", *IEEE Communications Magazine*, vol. 42 no. 9, pp. 56–62.
- [2] "Guidelines for Generic User Interface Elements for Mobile Terminals and Services", European Telecommunications Standards Institute (ETSI), Human Factors, User-Interfaces, Number DEG 202 132, DEG/HF-00041, 2004, Available: <http://www.etsi.org>
- [3] A. Steinhage, "Tracking Human Hand Movements by Fusing Early Visual Cues", *Proc. of the 4th Workshop on Dyn. Perception*, Bochum, Germany Nov. 2002, Aka-publishing
- [4] J.P. de la Cruz, "Motion based lipreading" (Master's thesis), Dep. of electronic eng., Univ. of Granada, Spain, June 2002
- [5] M. Sanchez, J.P. de la Cruz, "AudioVisual Speech Recognition Using Motion Based Lipreading", *Proc. of the 8th Int. Conference on Spoken Language Processing ICSLP (Interspeech)*, Sunjin Printing Co. (ISSN: 1225-441x), Korea, October 2004
- [6] D. Alonso, M. Sanchez, "An Approach to a Multimodal Man-Machine Communication System", *Proc. of the 9-th International Conference Speech and Computer (SPECOM'2004)*, St. Petersburg, Russia, 22. Sep. 2004
- [7] A. Steinhage, T. Bergener. „Learning by doing: A dynamic architecture for generating adaptive behavioral sequences". In *Proceedings of the Second International ICSC Symposium on Neural Computation*, NC'2000, pp. 813–820, 2000