An Intelligent House Control Using Speech Recognition with Integrated Localization

Jozef Ivanecký, Stephan Mehlhase, Margot Mieskes

European Media Laboratory
Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany
{jozef.ivanecky, stephan.mehlhase, margot.mieskes}@eml.org

Abstract. In the past few years great technological progress have been made in mobile devices performance and capability as well as in mini computers world. Mobile phones with wireless interface and advanced operating systems are becoming today's standard. Recent years showed, that they are going to be accepted also by elderly people and they are becoming an inseparable belonging of their every day's life. Since so called intelligent houses are moving from "special" to "common" world, together with today's mobile devices they open the door for many innovations, among others also speech recognition in the house.
In this paper, we describe our efforts towards introducing a simple speech recognition system for voice control of intelligent houses. Introduction to intelligent house environment is given with focus on elderly and disabled people. We also analyse the use of automatic speech recognition in such environments. The main focus is on the design and implementation of the simple speech recognition system and the integration of localization service in such environment.

1 Introduction

In this paper we present a user interface for controlling home equipment such as lights, blinds or heating via speech. In the research area of Ambient Assisted Living (AAL) the question of how to provide the user with an easy to accept and easy to use interface/device is still going on. Some suggest the TV as a device that is readily availbale and accepted by people [8]. But it has the draw-back that it is not mobile and it does not allow for a speech interface, which has emerged as a preferred input method for assistant systems [10]. [8] stated that assistant systems had the following requirements:

- light weighted
- simple and intuitive to use
- adaptable to physical and psychological changes
- offers various input methods like speech and touchscreens
- reliable

Therefore, we propose a speech interface for controlling home devices that runs on mobile phones. The mobile phone addresses several of the previously mentioned requirements in that it is light weighted, simple and intuitive to use and

newer mobile phones, especially so-called smartphones also offer touchscreens. Furthermore, mobile phones are very common and elderly people also use it for emergency calls. In the past it has been noted that a common platform should be used or that proposed solutions should be made available on a variety of platforms [1]. As our solution is available for several different mobile phones it accomplishes the latter of these goals.

Our user interface runs on the mobile phone as an additional application that allows the user to interact with their home devices. The microphone is only activated as the respective button is pushed, which addresses another issue raised in AAL applications: privacy [2]. In environments where microphones are set to always-listening modes this is a major issue, as the microphones are constantly recording. This is avoided by giving the user the control over the microphone.

In a preliminary study we have found out that it is rather cumbersome for users to always specify the room where they want something to happen (e.g. "turn on the lights in the living room") if they are already sitting in the specified room. Therefore, a localisation method is added to the application.

This paper is structured as follows: We will first introduce the environment that our application is based on (see Section 2). Next, we will give some details on the topic of Speech Recognition in general (see Section 3) and specifically in the home environment. The next section deals with the problem of localisation in the house and our proposed solution (Section 4). Finally, we will show the setup for the technical realization and some experiments we performed both with the user interface as well as with the localization (see Section 5). A summary that contains our conclusions and suggestions for future work finalizes this paper (Section 6).

2 House Environment

So called intelligent or automated houses today are equipped by default with a central control system. Such a system is able to control and monitor many devices, like lights, shutters, doors, the heating and others. They are usually based on KNX/EIB or similar technology. The control of such a system is usually done with switches similar to those in "normal" houses[1]. Beside, there is also a graphical user interface which allows the same functionality as standard switches but also opens the door to more advanced control and monitoring features.

Such a graphical user interface (GUI) is mostly integrated in to the wall at some fixed place—for example right beside the entrance of a house. It can also be accessible with a personal computer or via some kind of tablet PC, which allows usage from almost anywhere. However, the tablet PC is not being carried all the time with the user and can still be relatively heavy for a disabled or elderly users. If they do not have a simplified user interface, they can not be considered as user friendly for elderly people even despite individual adaptation to the user. It is not possible to use them also for other purposes, such as localization, or in

[1] Recently wireless EIB/KNX switches have become available that allow upgrading existing houses with this technology without need to replace all the cabling.

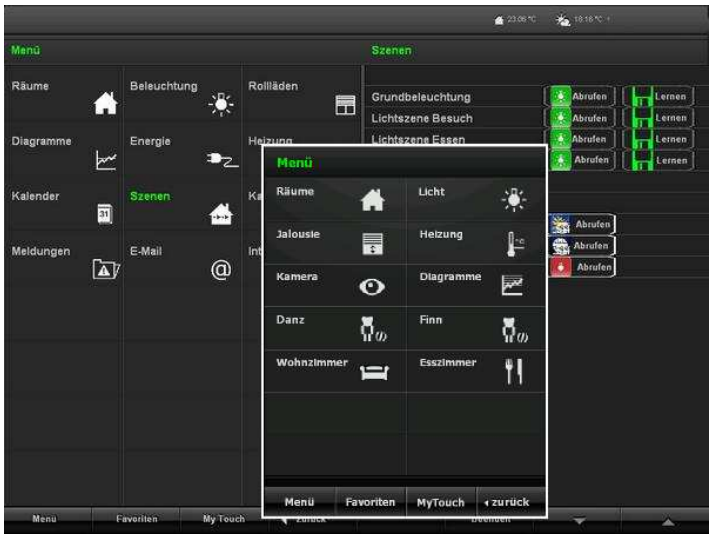

Fig. 1. Two different GUI examples for EIB/KNX system in a house.

emergencies because when they are really needed, they are very likely not with the user (see also the list of requirements presented in Section 1). On Figure 1 are two examples of such GUIs showing controls and monitoring for lights, shutters and the heating.

3 Speech Recognition in House

Speech recognition is able to minimize most of the problems mentioned in previous section. It can be used in a house with elderly or disabled people, to control different devices [5, 10]. The presented solutions are usually suffering by one of the following problems:

- They are reliable, but they are very complex and expensive. Reliable systems require a large microphone array, speaker localization, echo cancellation, noise filtering and several other advanced techniques. Such complex systems are still more in a research state than reality today.
- They are cheap, but completely unreliable. A solution with one microphone in the room is usually possible to use only in a quiet room in a certain position with a strong voice. Any other louder acoustic activities in the same room make the system completely unusable.

3.1 Overall design

To design and implement a simple, reliable, inexpensive, and especially today, widely usable speech recognition system for an intelligent house, we rely upon a

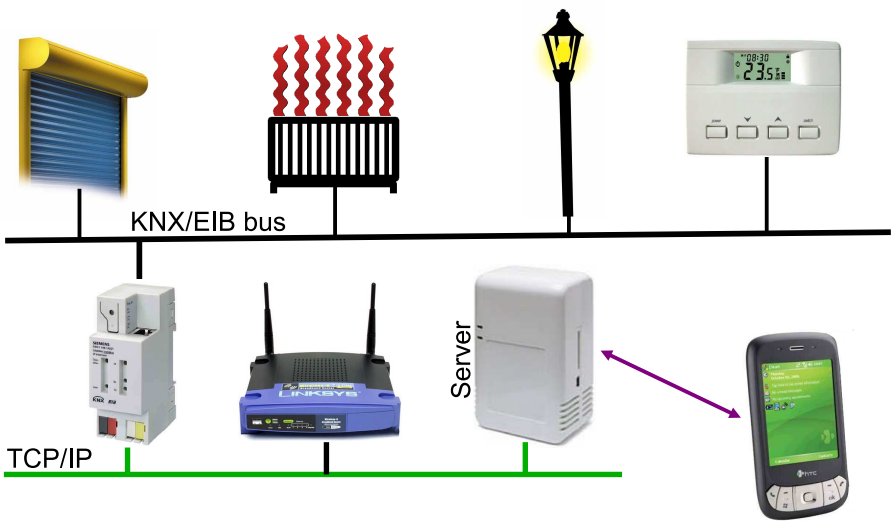

Fig. 2. Voice enabled house control architecture.

few technological achievements. For one, the significantly increased performance of mini PCs. Pocket-sized computers can today easily outperform normal desktop PCs from some years ago [1]. Most of the middle-class and state-of-the-art mobile devices today are equipped also with wireless network support. Besides that, a change in user behaviour has occurred: The mobile phones are already accepted and being used through all age groups [1].

Mobiles phones today are powerful enough to run system for automatic speech recognition. Unfortunately, the huge variety of mobile devices prevents to design a low cost speech recognition software running on all available mobile phones reliably. Indeed it is much easier to design a simple application which is just recording the speech signal. Applying a client–server approach, the recorded speech signal from the mobile device can be send to a recognition server for processing. As already pointed out such a "server" can today also be a cheap device which has a similar size as the mobile phone itself.

Such hardware equipment allows to make very quickly any intelligent house voice enabled. The entire architecture is shown in Figure 2. The user says the voice command in to a mobile phone. The mobile phone records the command and send it to the "server" using the available wireless network. The server will process the speech signal. After the recognition, the result is interpreted to generate the proper command for the house and also sent back to the mobile phone for visual feedback. The final command is sent to the KNX/EIB network via an interface. The entire system is working in real time and the action derived from the speech command takes place immediately.

3.2 Speech Recognition

Speech recognition under the technical conditions described above and controlling the utilities in an intelligent house have two important and positive features which results in high reliability of entire system:

1. The recorded speech signal has a very good quality. The mobile phone is acting as a close-talk microphone. In general, mobile phones have very good audio input hardware in contrast to many other hand-held devices where audio input is designed only as an optional feature.
2. The set of the commands for the house control is relatively small. The number of controlled utilities in average house is usually around 50. For this reason the speech recognition system can be grammar based and still very robust[2].

The grammar based recognition system obviously requires designing a grammar. Since each house is different, each house needs also an individual grammar. Fortunately, the group of the devices (lights, shutters, heating, ...) as well as group of available commands (switch on/off, up/down, dim, ...) is relatively small. Therefore we were able to design a fixed grammar, where during the adaptation for a particular house it is "just" necessary to add the existing devices with their real names (Peter's room, garden light, ...). The additional subsystem responsible for the interpretation of the voice commands is individualized in the same way.

All the changes necessary for one specific house can be done on the "server". The mobile phone is running a universal speech recording software and can be used in any house where such a server-based recognizer is installed.

The exemplified grammar in Figure 3 accepts for example the following commands:

- *"Könntest du bitte die Beleuchtung im Garten einschalten?"* (Would you please turn on the light in the garden?)
- *"Das Licht im Garten an"* (Light in the garden on)

Different mobile phones are recording the audio signal in different qualities. Although for a human ear in all cases the audio quality is very good and the differences are barely noticeable, it matters for the speech recognition process. Because the standard API for each device was used, it is not clear, if this a hardware or software limitation of the particular mobile phone. So far we tested our application on Android, Blackberry and iPhone.

4 Localization in House

As already pointed out in introduction, localization of the user is an important factor in order to create intuitive user interfaces. From the user's point of view

[2] One of the requirements for running on "small/slow" devices in real time. Language models based system require usually huge amount of memory and more powerful CPU as well.

```
[$prefix] $loc_garten $actionSchalten
   {out.device=rules.loc_garten.device; out.action=rules.actionSchalten.action;}
 | ([bitte] [$prefixMach]|[$prefixMach] [bitte]) $loc_garten $actionSimple
   {out.device=rules.loc_garten.device; out.action=rules.actionSimple.action;}
 | [$prefixMach] $loc_garten $actionSimple [bitte]
   {out.device=rules.loc_garten.device; out.action=rules.actionSimple.action;}
 ;
$loc_garten = ($lampe (im | in dem) Garten | [die] Gartenlampe)
   {out.device="L_Garten";}
 ;
$lampe = [das] Licht | [die] Beleuchtung
 ;
$prefix = (Würden Sie|Könnten Sie|Würdest du|Könntest du) [bitte]
 ;
$prefixMach = Mache|Mach | Machen Sie | drehe|dreh | schalte|schalt
 ;
$actionSimple = (an|ein) {out.action="ON";}
 | aus {out.action="OFF";}
 ;
$actionSchalten = (einschalten|anmachen|anschalten) {out.action="ON";}
 | (ausschalten|ausmachen) {out.action="OFF";}
 ;
```

Fig. 3. Example of a simple grammar for switching a garden light.

it is not very comfortable to always use the name of the room for addressing the devices in his current location, e. g. if the user is in the living room, he does not want to say: "*Lights to 50 per cent in the living room!*", but rather "*Lights to 50 per cent!*". To implement such functionality it is necessary to have a localization system available in the house.

The topic of indoor localization is still a research topic as pointed out by [6]. Various methods have been examined as for example GSM, FM or ultrasound [12, 9, 4]. [3] states that for localization wireless solutions should be preferred, as they are easy to install and also existing buildings can be equipped with this. Following this reasoning, we suggest to use Wireless LAN (WLAN) routers for this purpose. These are readily available and if the household should not be equipped with a WLAN router they are available at very little costs. They also do not suffer from problems stated by [3] concerning energy supply, as the access point sits somewhere in the flat/house and is directly plugged to the electricity network. Additionally, [14] showed that WLAN can also be used in environments with several floors with little additional effort. Therefore, WLAN based positioning also supports one main goal of AAL research, namely to allow elderly to stay in their homes for as long and as independent as possible. The problem of localization using WLAN has been well researched [13, 7, 14] in the past years, and [13] gives an overview about the topic.

The very little costs of the WLAN-based approach is even further lowered by the fact that WLAN has become ubiquitous within the last decade. Therefore, many users already have some WLAN hardware at home (e. g. internet routers).

As reported in [13], localization by one single access point does not yield a satisfactory recognition of the room. As mentioned, WLAN devices have become ubiquitous, so in most households you can discover several different WLAN access points from neighbouring flats. For the purpose of localization it might be sufficient if some neighbours have WLAN devices, which, we assume, is the case in most urban areas today. Even if there were not enough neighbouring networks available, WLAN equipment is still cheap and easy to setup for localization. As the application recording the spoken command of the user runs on his mobile device, it is obvious to do the localization with the device itself.

The localization using WLAN relies on several measurements. During these scans the available access points and their signal strength are recorded. These so called *fingerprints* are then fed into different learning algorithms for training (*calibration*). As recording these fingerprints might be a tedious and too complex process for the user, the application needs to record these fingerprints in an intuitive and easy manner.

The signal strength recorded in the fingerprints can be influenced by various factors: People that are "in the way", or furniture that has been moved around. If neighbouring WLAN networks are sensed during calibration, also factors out of control of the user, like new devices connected by neighbours or access points or furniture being physically moved to different places in the neighbour's apartment, may decrease recognition performance. These changes in the WLAN infrastructure can clearly not be anticipated in advance by the user. These problems call for a *recalibration* from time to time. So the (re)calibration process must also be integrated in an intuitive way into the application.

4.1 Technical Realization

The speech recognition application is running on the mobile device and sends the recorded audio together with the deduced room to the server. Therefore, the application initiates a network scan to create a fingerprint of the current position. This activity starts as soon as the user starts recording his command. While the audio recording is running, the scan terminates. The algorithm to deduce the room from the generated fingerprint runs and the data is made available as metadata during the request to the recognition server. The server will then use this information to disambiguate the spoken command.

In case the network scan takes longer than an average spoken command, a periodically initiated scan in the background could be realized in order to decrease the delay for the user. The network is scanned every x seconds, and the most recent results will be used. However, this approach is in general not recommended as it drains on the battery power of the device.

The different learning algorithms (e. g. k nearest neighbours, Bayesian inference, artificial neural networks, support vector machines) have different requirements on computing power and might therefore not be realizable on a mobile

Out of grammar utterances	14.93 %
Sentence accuracy	55.56 %
Action accuracy	91.23 %

Table 1. Out of grammar utterances, sentence and action accuracy for evaluation period one month and four different users.

device. In such case the raw fingerprints could be sent to the server during the calibration as well as during the recognition phases. The server, which is doing the speech recognition anyway, will then recognize the room the user is in. This could also be a solution for the aforementioned delay problems.

5 Experiments

In the following part we are presenting preliminary results from our experiments, which were used to confirm the described approach is feasible.

5.1 Speech recognition

To test and evaluate the implemented solution we installed the entire system into real houses. After adaptation to the house environment, as described in Section 3, the system was passed to the householder for real usage. The users were not informed about the available commands. They were asked to talk to the system as they wish.

After one month we downloaded all speech commands, which were saved with the householder's consent, and transcribed them. In the Evaluation, we did not focus on the speech recognition accuracy, but on the action accuracy. For example if a user said: "*Die Beleuchtung in der Küche einschalten*" and the system recognized: "*Licht in der Küche einschalten*", then from a recognition point of view it is incorrect, but from an action accuracy point of view it is correct, as the same action would be triggered. We also analysed out of the grammar sentences to improve the grammar to be able to cover bigger variety of utterances. In Table 1 are results for out of the grammar utterances, sentence accuracy and action accuracy for an evaluation period of 1 month with 4 different users depicted.

The result for out of the grammar (OOG) utterances is high, but is caused by the fact, that users did not get any initial instructions. A closer look at the OOG utterances distribution in time, we can clearly observe, that most of them appear shortly after system installation. For more detailed results a longer evaluation period is needed. For sentence and action accuracy, out of the grammar utterances were removed from evaluation pool.

On the first look, 55.56 % sentence accuracy may seem very small, but it resulted in a 91.23 % action accuracy. We analysed the recognition errors and

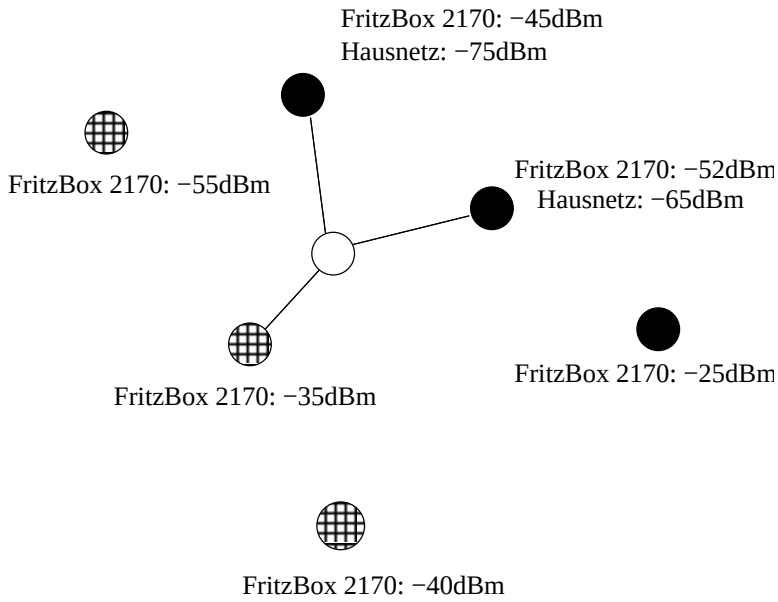

Fig. 4. Example for the kNN algorithm.

most of the errors in prepositions like "im" or "in" or incorrectly recognized articles. Such errors are not influencing the action accuracy rate and are mostly not noticed by the user. It is also important to note, that almost 30 % of utterances were spoken by a non native speaker. Recognition errors that resulted in faulty actions usually lead to the user to re-try.

Besides measuring accuracy, we asked the householder about their personal satisfaction with a free-form questionnaire. In all cases the reported satisfaction can be summarized as very high.

5.2 Localization

To evaluate the feasibility of our localization approach we used a simple Android-based application which scans the available networks. If asked, it stores the most recent scan as a fingerprint associated with the room the user selected. It then uses a simple k-nearest neighbours (kNN) approach to match a recorded fingerprint to a room. The idea behind kNN is to find those k fingerprints recorded during the calibration phase that have the closest distance to the current fingerprint. The distance is measured as a euclidean distance on the signal strength vectors (as described in [13]). If found, the room is returned as a result which was assigned to the majority of the k votes. For example, in the situation depicted in Figure 4, the reference point (white) would have been measured with -50 dbm for the access point named "FritzBox 2170" and -35 dbm for the access point named "Hausnetz". From all the recorded fingerprints available the $k = 3$ nearest fingerprints would be searched (the ones connected to the reference point).

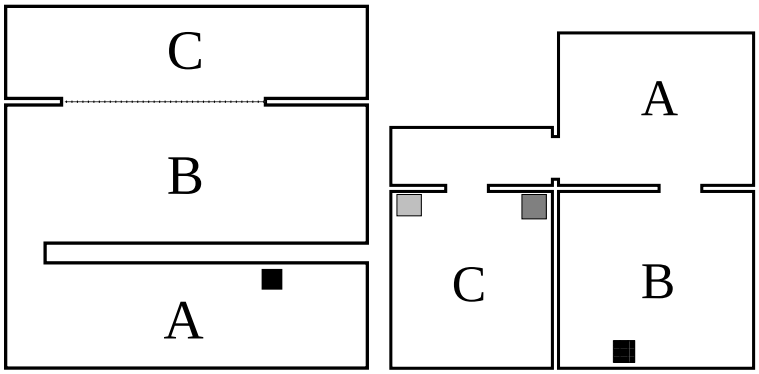

Fig. 5. Schematic floor plans of the experimental houses

In the example the black fingerprints would be assigned to one room and the patterned for another. The reference point would be assigned to the black room as there are more black than patterned neighbors close to it. For more details on kNN see [11, 13].

The results of the experiments prove that it is not problem from a performance point of view to run the room recognition on a mobile device. The chosen algorithm is not very complex and easy to apply. The performance is fine on a recently bought phone. Literature has already mentioned that one Wireless network access point is not enough to make proper localizations [13]. We observed the same behaviour under lab conditions: With rooms close by each other the error rate is quite high.

We conducted two experiments in real-world apartments, where several WLAN access points by neighbours were available. The first apartment (Figure 5(a)) consisted of tow rooms, between which there is a quite thick wall. They are connected by a few stairs, but without a door. The second room consisted of two areas which were separated by a wooden double wing door which was open. The household contains one router in room A. From the neighbours there were detected up to 6 different WLAN networks in the rooms depending on where the mobile device was located. In each room the kNN algorithm was calibrated using 4-7 fingerprints per room. As expected, the results were much better than the laboratory test with only one WLAN access point. We anticipated that the confusion between B and C would be higher than between A and B. In the other house, the situation was a little bit different. We had 3 rooms as depicted in Figure 5(b). In this household there are three routers, one in room B, one in a room below room C (lighter grey) and one in a room above room C (darker grey). In addition, there were more than 10 WLAN networks available through neighbours. We expect this to be the usual case in urban areas. There is no door between room A and B, but doors between A and C and the hallway. The algorithm was calibrated by 10-15 fingerprints per room. The observed accuracy

was even better than in the first case and made it possible to clearly distinct the individual rooms.

In total the recognition of the rooms was acceptable and, as only a very simplistic implementation was used, there is much room for improvement.

6 Summary

In this paper we have described various aspects of the simple and reliable speech recognition system for the voice control of house utilities. We showed, that such system has not only acceptable action accuracy and can be easily used without any special training, but because of integratation in to the mobile phone it has potential to be used for elderly or disabled people.

By adding WLAN-based localization the ease of use for the speech control could be further improved. As the localization is only activated when the communication with the house control is activated, privacy is at the control of the user. Experiments in a real testing environment demonstrated the feasibility of our approach, but also unveiled the need for future broader scale evaluation of the localization to find out whether the localization algorithm has to be improved. In one test environment it worked very well in the current setup. Therefore, it is necessary to find out whether more sophisticated methods (e. g. using artificial neural networks) for localization, which would also be more compute intensive pay off.

References

1. van Bronswijk, J.E., Kearns, W.D., Normie, L.R.: Ict infrastructures in the aging society. International Journal of the fundamental aspects of technology to serve the ageing society – Gerontechnology 6(3) (2006)
2. Caine, K.E., Fisk, A.D., Rogers, W.A.: Benefits and privacy concerns of a home equippend with a visual sensing system: A perspective from older adults. In: Proceedings of the Human Factors and Ergnomics Society 50th Annual Meeting (2006)
3. Flöck, M., Litz, L.: Aktivitätsüberwachung in bestandswohnung mit einfach nachrüstbarer basisausstattung. In: Proceedings of the 3rd German Ambient Assisted Living Conference, *Berlin, Germany 26th – 27th January 2010* (2010)
4. Hightower, J., Borriello, G.: Particle filters for location estimation in ubiquitous computing: A case study. In: Proc. of the 6th International Conference on Ubiquitous Computing (2004)
5. Hummes, F., Qi, J., Fingscheidt, T.: Untersuchung verschiedener parameter für die sprecher-lokalisierung mittels akustischer sensoren in wohnräumen. In: Proceedings of the 3rd German Ambient Assisted Living Conference, *Berlin, Germany 26th – 27th January 2010* (2010)
6. Kranz, M., Fischer, C., Schmidt, A.: Comparative study of dect and wlan signals for indoor localization. In: Proceedings of the 8th IEEE International Conference on Pervasive Computing and Communications *Mannheim, Germany 29th March – 2nd April 2010*. IEEE (2010)

7. Ladd, A., Bekris, K., Rudys, A., Marceau, G., Kavraki, L., Wallach, D.: Robotics-based location sensing using wireless ethernet. In: Proceedings of the 8th ACM International Conference on Mobile Computingand Networking (MOBICOM) *Atlanta, USA 23rd – 28th September 2002*. ACM (2002)
8. Lienert, K., Spittel, S., Stiller, C., Roß, F., Ament, C., Lutherdt, S., Witte, H.: Seniorenbefragung zum assistenzsystem weitblick – ergebnisse einer bedarfsanalyse. In: Proceedings of the 3rd German Ambient Assisted Living Conference, *Berlin, Germany 26th – 27th January 2010* (2010)
9. Matic, A., Papliatseyeu, A., Osmani, V., Mayora-Ibarra, O.: Tuning to your position: Fm radio based indoor localization with spontaneous recalibration. In: Proceedings of the 8th IEEE International Conference on Pervasive Computing and Communications *Mannheim, Germany 29th March – 2nd April 2010*. IEEE (2010)
10. Meis, M., Fleuren, T., Meyer, E., Heuten, W.: Nutzerzentrierte konzeptentwicklung des persönlichen aktivitäts- und haushaltsassistenten: Methodologie und erste ergebnisse. In: Proceedings of the 3rd German Ambient Assisted Living Conference, *Berlin, Germany 26th – 27th January 2010* (2010)
11. Russell, S., Norvig, P.: Artificial Intelligence A Modern Approach. Pearson Education, 2nd edn. (2003)
12. Varshavsky, A., de Lara, E., Hightower, J., LaMarca, A., Otason, V.: Gsm indoor localization. Pervasive and Mobile Computing 6(6) (2007)
13. V.Patmanathan: Area Localization using WLAN. Master's thesis, KTH Stockholm (2006)
14. Wang, H.Y., Zheng, V.W., Zhao, J., Yang, Q.: Indoor localization in multi-floor environments with reduced effort. In: Proceedings of the 8th IEEE International Conference on Pervasive Computing and Communications *Mannheim, Germany 29th March – 2nd April 2010*. IEEE (2010)