Running head: Acoustic User Interfaces for AAL

Title: Acoustic User Interfaces for Ambient Assisted Living Technologies

Stefan Goetze[1], Niko Moritz[1], Jens-E. Appell[1], Markus Meis[2], Christian Bartsch[3] and Jörg Bitzer[1,3]

[1]*Fraunhofer Institute for Digital Media Technology (IDMT), Project group Hearing, Speech and Audio Technology, Marie-Curie-Str. 2, 26129 Oldenburg, Germany,*

[2]*Hearing Centre Oldenburg GmbH, Department Market and Effects Research, Marie-Curie-Str. 2, 26129 Oldenburg, Germany*

[3]*Institute for Hearing Technology and Audiology (IHA), Jade University of Applied Sciences, Ofener Str. 16/19, 26121 Oldenburg, Germany*

Corresponding author:

Stefan Goetze, Email: s.goetze@idmt.fraunhofer.de, Phone: (+49) 441 / 2172-450, FAX: (+49) 441 / 2172-450

Keywords:

Acoustic Interfaces, Speech Recognition, Ambient Assistive Living Technology, User-Centered Design

Acoustic User Interfaces for Ambient Assisted Living Technologies

Stefan Goetze[1], Niko Moritz[1], Jens-E. Appell[1], Markus Meis[2], Christian Bartsch[3] and Jörg Bitzer[1,3]

[1]*Fraunhofer Institute for Digital Media Technology (IDMT), Project group Hearing, Speech and Audio Technology, Oldenburg, Germany,*

[2]*Hearing Centre Oldenburg GmbH, Department Market and Effects Research, Oldenburg, Germany*

[3]*Institute for Hearing Technology and Audiology (IHA), Jade University of Applied Sciences, Oldenburg, Germany*

Abstract

This contribution discusses technologies for acoustic user interaction in ambient assisted living (AAL) scenarios. Acoustic user interfaces allow for a natural and convenient way to interact with technical systems e.g. via sound or speech presentation or via speech input by means of automatic speech recognition (ASR) as well as by detection and classification of acoustic events. Older persons targeted by AAL technologies especially need more easy-to-use methods to interact with inherently complex supporting technology. As an example application for acoustic user interaction a multi-media reminding and calendar system being part of a personal activity and household assistant (PAHA) was designed and evaluated. For this purpose mainly older participants were involved in user studies to continuously evaluate and support the development strictly following a user-centered design process.

The results suggest a wide acceptance of acoustic user interfaces by older users either for controlling inherently complex AAL systems by using robust ASR technologies or as a natural and ambient way of presenting information to the user. However, further research is needed to increase the robustness of ASR systems when using hands-free equipment, i.e. to provide a real ambient way of interaction, and to introduce personalized speech and sound presentation schemes accounting for the individual hearing

capabilities and sound preferences.

1. Introduction

Demographic changes are producing continuous growth of the older population within the near future [1] which raises severe problems in the field of social care [2-4]. Assistive technologies are able to support older persons and, by this, allow for a longer period of independent life and that these persons may stay longer in their own home environments independently [1]. In particular, information and communication technologies (ICT) assisting the older people are of high social relevance [4]. However, since such systems generally are complex to use, the design of simple and intuitive interfaces is of great importance [5, 6]. It has been shown that natural and convenient ways to interact with technical systems are highly desired [7]. One way to realize an intuitive human-machine interaction is using speech. Speech is the most natural form of information exchange between humans and thus, a natural and convenient way to interact with technical systems. Especially for ambient assistive technologies supporting older people inconspicuous control of those systems by spoken commands is favorable [8].

However, the design of successful applications of *intelligent* IT-based assistance requires solutions not only for various purely technical challenges, but also a continuous consideration of user requirements, user acceptance [8], integration into medical and nursing care structures [9] as well as the consideration of economic aspects.

All these issues are addressed at different levels of depth in the German interdisciplinary Lower Saxony Research Network 'Design of Environments for Ageing' (German: ‘Gestaltung altersgerechter Lebenswelten', GAL) [10]. One of the assistive scenarios identified and evaluated in the GAL project is a personal activity and household assistant (PAHA), an ambient reminder system, using various input and output modalities. The development of the PAHA is based on a user-centered design process [11]. By this, the goal is a holistic evaluation of purposes, requirements, feasibility and effectiveness in terms of its acceptance and economic effects in the wider population as well as further improvements and applications.

One important aspect of the PAHA system which will be main topic in this contribution is an acoustic user interaction (AUI) framework for acoustic sound pick-up, processing, enhancement and analysis (e.g. automatic speech recognition) providing functionality for acoustic input and output of assistive systems. The AUI framework is exemplarily used for user interaction with the personal activity and household assistant [12] which can be controlled by spoken commands (amongst other modalities). Furthermore, the AUI framework provides possibilities of acoustic monitoring, i.e. for automatic analysis of the acoustic signals to trigger events e.g. in case of door knocking or an emergency such as an overboiling pot on a kitchen stove or increased coughing of patients. It is therefore used as a generalized front-end for acoustic sound recording, processing and presentation for the GAL-System which is described in [10].

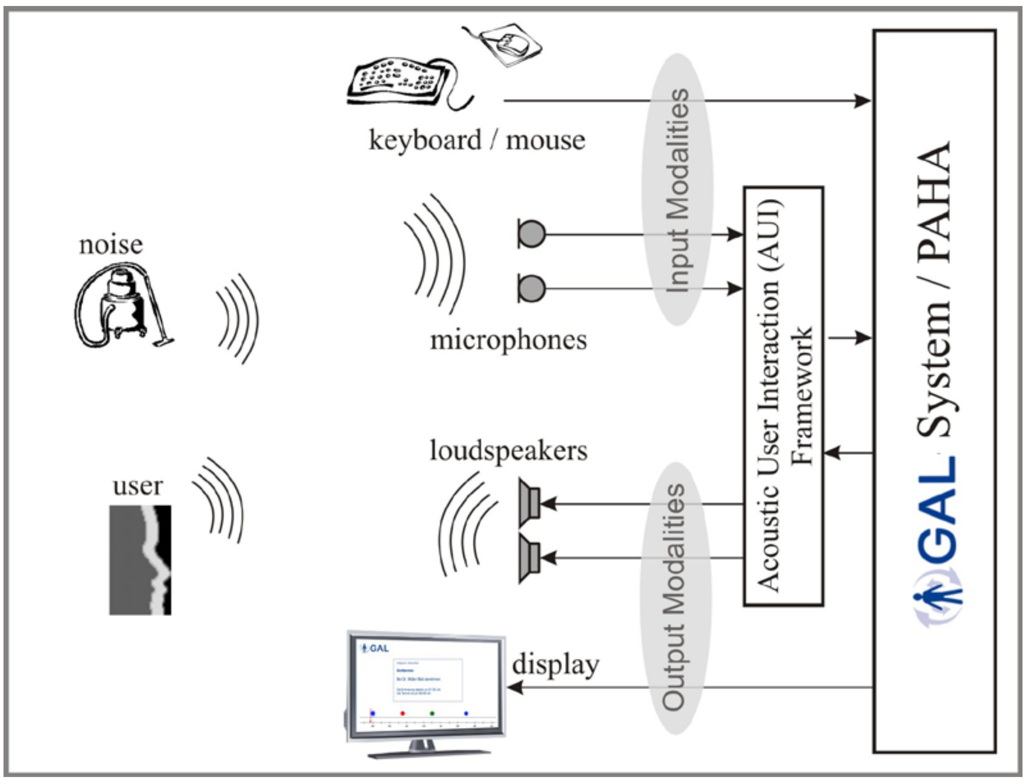

Figure 1: Schematic of the GAL system including the personal activity and household assistant (PAHA), its input and output modalities and the acoustic user interaction (AUI) framework.

Figure 1 shows a general schematic of the PAHA (that is part of the overall GAL system) and the AUI framework. The GAL system comprises several input and output modalities since different studies have shown that assistive systems for older people have to be personalizable and adaptable to the specific preferences and needs of the user [1, 5, 13-15]. Therefore, the PAHA accepts input via conventional mouse or keyboard or, as addressed in this study, via acoustic signal pickup, processing and analysis. Accordingly, the system's output can be presented to the user conventionally on a screen or via light, vibration or acoustic signals. The interested reader is referred to [13] for a detailed information regarding the multimodal user-interfaces of the GAL system.

For the acoustic input several microphones pick up the sound that may contain utterances of the user or other acoustic events that may trigger actions of the GAL-system. A problem that severely degrades the performance of automatic recognition systems is that not only the desired signal parts are picked up by the microphones but also disturbances like ambient noise, e.g. from electronic devices, or signal parts produced by the system itself via its loudspeakers. Thus, an appropriate pre-processing [14-20] of the microphone signals is needed to increase the signal-to-noise ratio (SNR) by removing or attenuating undesired disturbances providing an enhanced signal for the subsequent detection and classification stage. Overall, the AUI framework contains strategies for signal enhancement, acoustic event detection, classification and recognition (cf. Section 5 for a detailed discussion).

In this contribution we will focus on two user studies. The first one determined the user requirements of the PAHA system. These results and the technical constraints lead to a first system of the assistant, which was tested in a second study with a mock-up system. The results showed a large need for a robust speech recognition system. Furthermore, initial results are shown for the ASR system in combination with ambient and hands-free sound pick-up based on the results of the preceding user studies.

The remainder of this paper is organized as follows: The user-requirements, user-studies and the chosen user-centered design process are described in Section 2. The dialog structure of the PAHA is briefly described in Section 3 before Section 4 describes user studies using a mock-up system for the voice-controlled calendar application. The acoustic input and analysis possibilities of the AUI framework that were developed based on these user studies are described and technically evaluated in Section 5. Section 6 concludes the paper.

2. User requirement studies for an ambient reminder system

In research and development projects, prototypes of hardware and software are commonly used for communication between developers and the target group to receive feedback from the prospective users of the product at an early stage of development. A central goal of using such prototypes is to keep the efforts and costs for development at a minimum, since modifications at later development stages are generally more costly and difficult to implement. Therefore, such a user-centered design process [11] was chosen to find solutions based on previously identified user needs for the development of the GAL system and the AUI framework. Here, an inexpensive approach is to offer high-fidelity prototypes in the sense of the "Wizard of Oz" principle [21]: the visual and acoustic level of the prototype can already provide a high level of detail, while the algorithm depth for automatic control and analysis (the so-called mental model) is limited to a minimum. The direct interaction with the system and its feedback of calculation results are invisible for the end-user since an investigator controls the system and forwards feedback to the end-user. This technique is commonly used in the areas of speech recognition and synthesis, which require a high implementation effort in the application logic.

Throughout the development of the system, user studies were made to assess user needs and user feedback with respect to an ambient acoustic user interface for the example application of the appointment calendar [12]. The first study, described in this section, was set up to determine user requirements in a very early state of the development process.

Further studies comprised a user-centered evaluation of the mock-up system for speech recognition to provide clues for further technical implementations (cf. Section 4) and a purely technical evaluation of the running system (cf. Section 5.2).

2.1. Method and sample of the user requirement study

To assess the demands and to define scenarios for input and output of appointments of the planned personal activity and household assistant, 74 test persons were surveyed by means of face-to-face interviews [7]. The age of these test persons ranged from 50-75 years (50% female). They were stratified in groups according to their age (group 1: 50-65 yrs., group 2: 66-75 yrs.; mean age: M=65.8, SD=5.3 yrs.), gender, technology experience (low to moderate vs. high; see [7] for details), and hearing loss (<= 40 dB Better Ear Hearing Loss (BEHL): average of the M4 frequencies 0.5, 1.0, 2.0, and 4.0 kHz vs. > 40 dB BEHL; M=38.8 dB HL, SD=19.7; see [7, 22] for details). They were asked about their current management of appointments and activities and about their future needs for an ambient assistive reminder system.

The statistical analyses for the two user studies were computed with SPSS 18.0. Procedures of significance were established with analyses of variance (ANOVA) for simple comparisons. Complex analyses were computed by means of ANOVAs with repeated measurements (between and within subject designs); adjustments were done with Bonferroni posthoc tests for multiple comparisons. Repeated measurement analyses were corrected with Greenhouse-Geisser. The alpha level was at 5%.

2.2. Analysis of the current state

The analysis of the current state with respect to the management of appointments and activities showed that almost all users used a calendar. Half of the participants took notes on notepads and approximately one third used a notebook. A personal computer (PC) was used by 14%, and 5% of subjects used a smart phone or mobile phone. Analyses of the in-between factor technology experience demonstrated, that the qualities of appointments was not dependent from the technical experience of the users [23]. More than half of the participants used more than one medium for their appointment management in combination.

78% of participants indicated that they did not use any technical systems for monitoring of their homes. Amongst the remaining 22% that used systems for home automation, 42% indicated that they used a burglar alarm system and 58% used household appliances with an acoustic warning function. A condition monitoring system was generally considered as “rather important”. A condition monitoring of possibly open windows, doors and the oven/stove, and also the automatic detection of falls [24-28] was rated as “rather important” to “very important”. Further details of the analysis of the current state are provided in [7, 13].

2.3. Development of the reminder system: input modalities

Concerning the development of an ambient reminder system, the participants were asked for their preferences about input modalities such as keyboard, phone keypad, speech input and handwriting recognition. Possible answers ranged from 1 (“not acceptable at all”) to 5 (“very acceptable”). The results (as shown in Figure 2) indicate that input by speech reached highest acceptance at a score of about 4 (“rather accepted”) followed by keyboard and keypad (“partly accepted” to “rather accepted”). Handwriting recognition was ranked significantly worse compared to all other input modalities (“rather not accepted” to “partly accepted”, cf. Figure 2) tested with repeated measurement ANOVA and with a post-hoc Bonferroni test for multiple comparisons. Repeated measurement analyses were corrected with Greenhouse-Geisser.

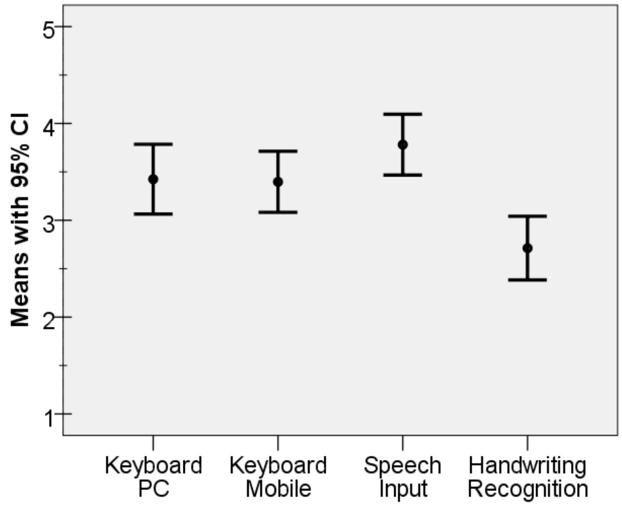

Figure 2: Acceptance of input modalities of the ambient reminder system (means with 95% CI , n=73)

No significant interactions between the input preferences and the variables such as age, gender, technology experience and hearing loss were observed. For statistical details, see [13].

Only 14% of the participants of the study preferred the possibility of entering dates in the entire apartment. The living room (60%), home office and kitchen (80%) were considered to be most useful for entry of appointments by the subjects.

2.4. Development of the reminder system: output modalities

A two stage strategy for presenting reminders to the users was developed for the personal activity and household assistant [13].Reminders are first presented in a non-specific way and in a second stage the concrete date output is given. By this strategy the system is capable to remind e.g. medically relevant appointments in the first stage unobtrusively e.g. by means of light to indicate the category only before a speech output is given in the second stage. For the reminder of the first stage, a non-specific date output, the users were asked to indicate the preferences amongst ambient light, natural sounds, artificial sounds, and vibrations [13]. It was shown that natural (36%) and artificial sounds (29%) were preferred. A similar acceptance could be observed for the ambient light (27%).

Furthermore, the study showed that the degree of hearing impairment has to be considered for the output strategies: people with no or mild hearing impairment in particular preferred sound output, while people with moderate to high degree of hearing loss rather preferred optical modalities. However, in later studies using a mock-up system for the output modalities, it was found, that people with a severe hearing impairment do accept acoustic output modalities if the output level can be individually adjusted to comfortable sound levels [7, 13]. Post-hoc analyses with a younger group of users (50-65 yrs.) showed a preference for sound independent of hearing loss. However, in the older group with medium to high hearing loss, particularly the ambient light was a preferred form of an ambient reminding system. Older users with no or mild hearing loss preferred sounds [7].

Also for the second stage, the concrete date output, a trend towards the preference for acoustic output could be found. In particular, natural language was preferred (27%), followed by the output by a digital picture frame (24%), the display on a (mobile) phone (24%) and a monitor (17%) as shown in Figure 3.

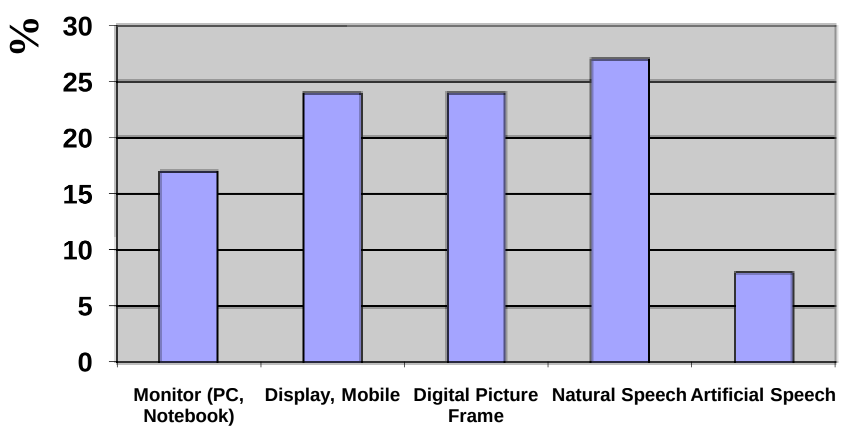

Figure 3: Output modalities of the second reminder for detailed information (Preference in percent, n=70)

Concerning the output of the system, 52% of the users indicated that it would be convenient to have the output throughout the entire apartment. Those who only designated special rooms predominantly named home office, kitchen or living room.

2.5. Discussion of user requirements

Overall, the preceding study on the design of an ambient reminder system provided that acoustic input and output modalities play a key role and are highly accepted by the users for the reminder system. Since these results were obtained based on abstractly formulated scenarios further studies closer to everyday-life situations seemed to be necessary. For this purpose, further user surveys with an executable mock-up system for speech output were carried out which will be addressed in Section 4.

3. Dialog structure of the reminder system

The GAL system is capable of presenting its output to the user either on a TV or computer screen, or by means of light, vibration or acoustic signals [12]. Here, we will focus on the acoustic output of the system since it was found in Section 2 that acoustic input and output is highly demanded by the users. In [12] a detailed discussion of the benefit of multimodal output strategies is given.

Several ideas exist for acoustic reminding systems (cf. [13] and Section 2.4). However, for the input of an appointment without keyboard and mouse, natural speech in conjunction with ASR seems the only possibility to guide the user through all desired data fields. The created dialog structure and an input/output system were used for the mock-up test that will be described in Section 5.1.

The speech output signals of the AUI framework were sentences and single words necessary for the calendar task spoken by a male speaker. The signals were recorded in a sound studio, manually cut, and optimized with regard to their level and were, therefore, of high quality with a high degree of naturalness, even if numbers and words were inserted into the sentences to adopt the acoustic output to the user input. For the final system we also recorded a female voice in order to be able to adjust the system to user preferences.

In the following, the acoustic user-system interaction will be explained for the calendar application dialog whose beginning is exemplarily shown in Figure 4. Acoustic outputs of the system are depicted as right-facing parallelograms (if multiple answers are possible) or diamonds (yes/no-questions) and labeled with loudspeaker symbols. Inputs from the user are depicted as left-facing parallelograms as well as the possible paths of the yes/no questions. ASR pictograms at the left side of Figure 4 indicate speech command input based on one or more microphone signals. The household assistant can be launched with a double hand clapping by the user, which is identified by the implemented acoustic event detection. After starting, the household assistant asks for the desired task, which is “New date.” to start the datebook application. By choosing another predefined task the system would choose a completely different program path, which is symbolized by outwards pointing arrows. After recognizing the user’s command, a request occurs, in which the system asks if all user inputs have been identified correctly. This could be answered by the user through “Yes.” or “No.”. If the answer is “No.” the system jumps back to the previous prompt and otherwise the dialog is continued. At this point, multiple answers are possible too, but every possible answer leads to the same program path, which is symbolized by inwards pointing arrows. For a complete appointment description the system will ask for the category, date, starting and ending time, for a description and if this appointment is repeated. After each question the system will check the ASR results by repeating the input as an acoustical feedback to the user and ask for correctness. The verification of the correct input after each step was chosen, since the frustration level for reentry of information is lower for this strategy compared to a complete check after all information was gathered. The description of the appointment will be saved as audio-files without any ASR technique applied, so the output at reminding time will include the original user utterance instead of the system voice.

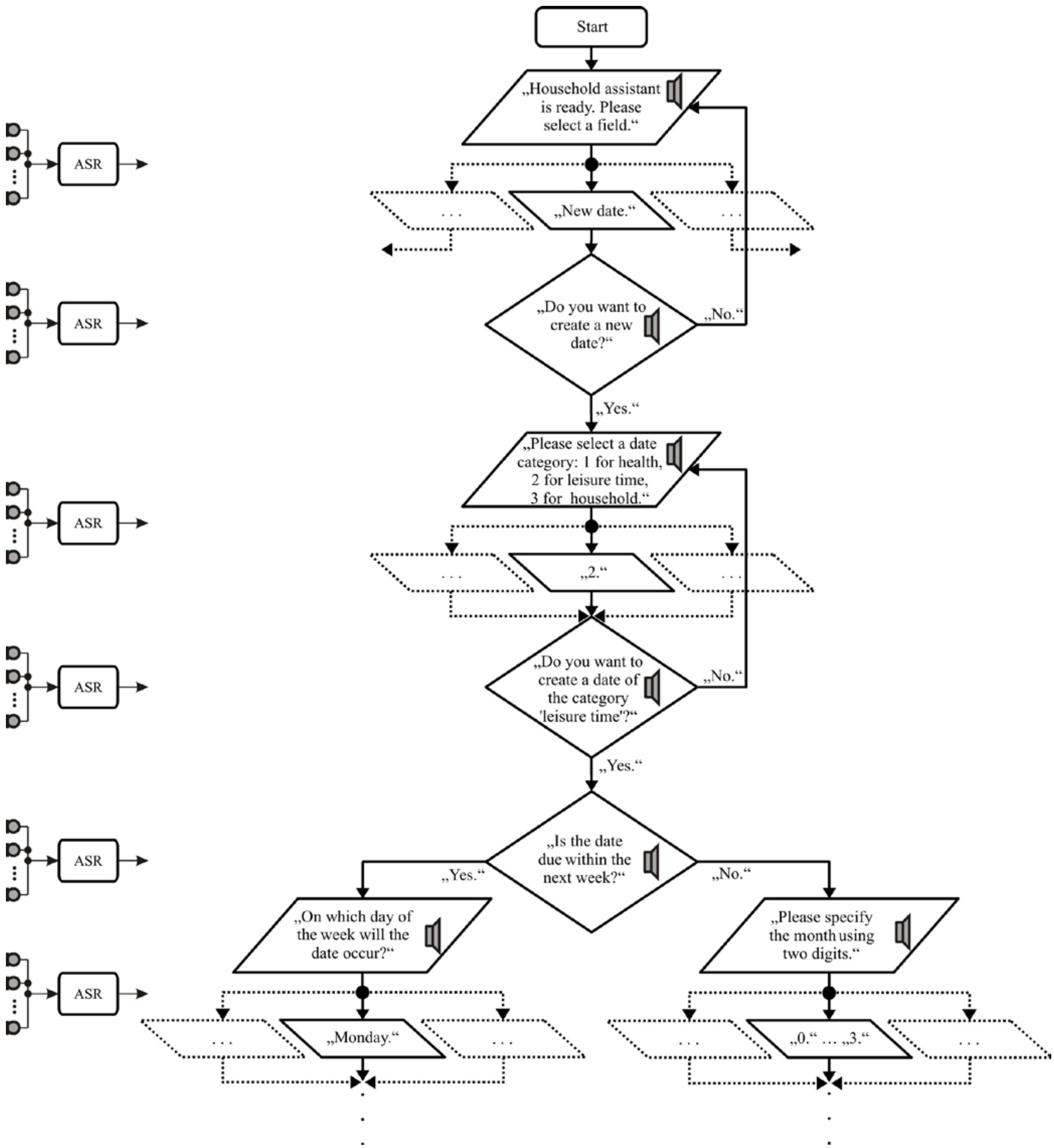

Figure 4: Beginning of the dialog structure of the PAHA input/output dialog.

4. User study with a mock-up system for the voice-controlled calendar application

Since the user requirements described in Section 2 showed a high demand for acoustic input very early in the project the development of the AUI system started immediately. Already at early stages of the development of the AUI, a mock-up system was assembled for testing the user acceptance of the targeted voice-controlled calendar system. The automatic voice output of the system and the speech input by users were controlled by an investigator based on the previously described “Wizard of Oz” principle. This method allowed the user studies to be carried out with well-defined and controlled error rate of the speech recognition system for the calendar application.

The user tests described in this section were divided into three phases with increasing recognition error rate. Erroneous output of the system was feigned intentionally in the third phase to investigate the user’s tolerance regarding errors of the speech recognition system.

4.1. Description of the sample

A total of 12 persons in the age of 63 to 75 yrs. (M=67.5, SD=4.19 yrs.) representing a subsample of the 74 test persons of the study described in Section 2 participated in this study. Again, the subjects for evaluation of the speech input of dates were selected considering age, gender, technology experience and hearing loss to minimize systematic biases (50% for each dichotomous variable).

One group had ‘normal’ hearing abilities (M=16.25 (SD=10.45) dB HL, range from 6.25 to 28.75 dB HL) and the other group was hearing-impaired with a moderate to severe hearing loss of M=57.08 (SD=18.85) dB HL and a range between 40.0 to 82.5 dB HL. The hearing-impaired group was chosen because every second person of age 65+ suffers from hearing deficiencies [29-31]. Thus, every second person targeted in AAL suffers from hearing deficiencies. The hearing-impaired group was subdivided into two groups according to the degree of hearing loss, because it can be assumed that people with a higher degree of hearing loss will have a lower accuracy of pronunciation since the feedback of their own voice is reduced.

The statistical analysis was mainly restricted to descriptive statistics (median), because the aim was to describe the general acceptance with no direct hypotheses. Regarding the recognition rates complex statistical models were tested.

4.2. Experimental procedure

4.2.1. Phase I

The evaluation of the speech input started after each subject received first instructions. According to the design of the study no errors were made by the system in this phase in order to familiarize the subjects got with the acoustic input interface of the household assistant.

4.2.2. Phase II

After a first block of questions regarding phase I errors of the speech input interface were simulated in phase II. To keep the error rate constant for all subjects, two errors were introduced at two different but previously defined positions. Thus, the weekday “Sunday” was incorrectly issued as “Monday” and the input of the time “09:30” was incorrectly issued as “09:20”. The behavior of the system in phase II is closest to the performance of the final system since recognition errors may occur from time to time. Phase II again was followed by a questionnaire to evaluate the properties of the system by the users.

4.2.3. Phase III

The intention of this third test was to determine a limit for tolerance of incorrectly recognized commands. For this purpose two different scenarios were tested with different errors at different positions. The goal was to evaluate whether the position of an error was important for a decision to cancel.

A long and a short test were set up that were carried out by the predefined groups to equal size. For the short version, a repeated error was simulated for the input of the weekday. In this case only the input for the day had to be repeated. In the long version the error occurred for the time input. By this, not only the start time or the end time had to be reentered but both (cf. Section 3). A yellow and a red card were handed out to the subjects that could be used to cancel the test at any time. As in a soccer game the yellow card indicated a warning to the system and the red card immediately ended the trial. The number of tolerated failures was recorded for both cards, in order to determine the maximum number of acceptable errors which occurred during communication with the system.

4.2.4. Results of Phase I

The characteristics of the system as well as the training of the speech recognition system were assessed predominantly positive. User-friendliness, intuitiveness of use, comprehensibility, helpfulness and acceptance were evaluated. These properties were assessed on a scale from 1= ‘not applicable at all’ to 5=‘applies wholeheartedly’.

Very good results could be achieved regarding the user-friendliness of the device. All test persons answered either 4=‘rather applicable‘ (8 persons) or 5=‘applies wholeheartedly‘ (4 persons). The evaluation of the property ‘acceptance’ also showed clear results. Seven persons assessed this category as ‘applies wholeheartedly’, while the rest of the participants chose ‘rather applicable‘. The other three properties were assessed somewhat more heterogeneously. The comprehensibility of the system was judged as negative by one person (‘rather not applicable‘), one person answered ‘rather applicable‘ and ten persons stated that they assessed the system as ‘applicable at all’. The property intuitiveness had a median score of 4 and a range between 2 and 5. The question regarding the helpfulness also showed a positive result: In total ten persons assessed the helpfulness of the speech input system as ‘rather applicable’ (3 persons) or ‘applies wholeheartedly‘ (7 persons).

4.2.5. Results of Phase II

The results only changed gradually if errors were introduced as previously described for phase II. A 2 (within: phase 1 vs. phase 2) by 5 (within: evaluation: user-friendliness, intuitiveness of use, comprehensibility, helpfulness, acceptance) by 2 (age <=65 vs. >= 66 yrs.) by 2 (gender) by 2 (hearing loss: normal hearing vs. moderate to severe hearing loss) analysis with restriction on 2-way interactions, showed no effect of the main within effects ‘evaluation’ and ‘phase’ as well as the interaction ‘evaluation by phase’ (all F-values < 2.2, $p>0.20$. The basic acceptance of the system did not change, indicating that the simulated errors did not have an effect on the user evaluations.

The between main effects age, gender, hearing loss showed no significant main effect regarding the whole model (F values < 1.6, $p > 0.13$). The only significant effect ($F=6.01$, $p=0.04$) was observed with the interaction ‘phase’ by ‘age’ regarding the dimension ‘intuitiveness’: The younger group rated the system from phase I to phase II as more intuitive (phase I: M=3.5, SD=1.4; phase II M=4.0, SD=1.09), whereas the older age group changed the rating of ‘intuitiveness’ from phase I M=4,0 (SD=1,05) to phase II M=3.67 (SD=1.03). The other means regarding the five dimension of the evaluation are summarized in Table I.

Table I: Means and standard deviations of the five evaluation dimensions (phase I and II)

	<=65 yrs. (N=6)		> 65 yrs. (N=6)		Total (N=12)	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
User-friendliness PI	4,50	0,55	4,17	0,41	4,33	0,49
User-friendliness PII	4,83	0,41	4,17	0,41	4,50	0,52
Intuitiveness of use PI	3,50	1,38	4,00	0,63	3,75	1,05
Intuitiveness of use PII	4,00	1,09	3,67	1,03	3,83	1,03
Comprehensibility PI	4,50	1,22	4,83	0,41	4,67	0,89
Comprehensibility PII	4,83	0,41	4,67	0,52	4,75	0,45
Helpfulness PI	4,67	0,52	3,83	1,47	4,25	1,14
Helpfulness PII	5,00	0,00	3,67	1,75	4,33	1,37
Acceptance PI	4,67	0,52	4,50	0,55	4,58	0,52
Acceptance PII	4,67	0,52	3,50	1,64	4,08	1,31
PI= Evaluation phase I, PII=Evaluation phase II, SD=Standard deviation						

For the dimensions ‘user friendliness’, differences between the younger and older age groups were obvious; a ANOVA post-hoc tests revealed a single significant effect in phase II: younger people rated the ASR system more user-friendly in comparison to older people ($F=8.0$, $p=0.02$). Also for the dimension ‘helpfulness’, a tendency from the means could be observed but was not statistically significant. A larger number of subjects is necessary for future research in order to clarify the importance of an improvement, especially for older people.

4.2.6. Results of Phase III

The third phase of this study investigated the limits of acceptable errors (frustration tolerance of the users). Introducing continuous errors as described above lead to interesting results: The amount of acceptable errors ranged from 0 to 17 (median = 6) while entering an appointment before the users quit the test. No age or gender related dependencies were observed regarding the frustration tolerance. Such a high tolerance for most of the test persons was not expected. Nine out of twelve test persons indicated that they would be willing to use the system in their future. However, the willingness to use the system at the time of the study was rather low.

4.2.7. Structured dialog vs. free input of speech

In addition to the previously described studies the question may arise if the suggested dialog described in Section 3 is acceptable for prospective users or if they prefer an alternative input method for the appointments. An alternative could be a free input of speech. Thus, the participants were asked to provide input commands to the system without a structured dialog, which were recorded by microphones. Additionally, they had to mention a preference and to provide reasons for their decision. The free input of appointments was preferred by seven (58%), the structured by five persons. Reasons for and against a free input are summarized in Table II. The advantages of a structured dialog resulted from the structure itself for some participants. By this, it is less likely to forget important details during the input. The detailed and concise output of appointments was also mentioned as an advantage.

Table II: Arguments for a free speech input vs. structured input of appointments.

Arguments for a free input of appointments	**Arguments for a structured input of appointments**
• numerical entry easier as in dialog • input is more familiar (e.g. numbers) • higher flexibility • more familiar and common • more individual • less complicated; appointments can be made in usual style • to structure appointments by yourself is easier than to learn new commands	• similarity to personal organizer • nothing can be forgotten in structured dialog • important topics could be forgotten, unless the system would point out this • easier; you are more concentrated on the information • communication with the technical system is easier than using the free speech input

4.2.8. Assessments of the questions for the training of the speech recognition

A final questionnaire was created to evaluate if the training of a speech recognition system at home is acceptable at all. The variables of the speech training were assessed on a scale ranging from ‘does not apply at all‘ to ‘applies wholeheartedly’ like the three phases of the study described above. Special attention was drawn to the acceptance, duration, operator’s convenience and the intuitive usage.

The training of the speech recognition was assessed to be ‘very good’ in relation to acceptance and operator’s convenience. All twelve subjects rated with ‘rather applicable‘ or ‘applies wholeheartedly‘. Furthermore, the feedback regarding the characteristic intuitive usage was predominantly positive. Ten persons fully agreed, one person answered ‘rather applicable‘ and one ‘rather not applicable‘. Eight persons assessed the training – described in the following, as not too long, one person was rather indecisive and three persons stated that it was too much effort. Furthermore, the test persons were asked if they needed advice for the training or not. Five of twelve persons asked for assistance by the customer services. One person additionally stated that assistance by relatives or friends would also be conceivable.

4.3. Discussion

The characteristics of the systems were mainly assessed as positive. However, the results of the study also indicated that acceptance difficulties in controlled conditions had to be verified further in a demonstrator under more realistic conditions. The results, furthermore, showed that the input of appointments was not intuitive for some users. As shown in Table I the intuitiveness increased only for younger users after using the system for a certain period (phase I to phase II). This benefit was not found in older users. This could be an indication that the dialog control has to be a bit more precise and simplified. The user suggested e.g. to add a command ‘new appointment‘. Thus, some changes in the dialog structure (cf. Section 3) may be necessary. Referring to the questions regarding the utility, it can be pointed out that the system was generally accepted by the participants. However, the current use of such an assistive device was refused by the majority of the interviewed persons. This percentage increased again after using the system. Some tendencies can be observed (e.g. concerning the dimension ‘helpfulness’) from the evaluation in this section. However, to obtain statistically significant results a larger number of subjects is necessary for future evaluations.

Useful information for the further development of the system could be mainly obtained from the free questions. Two test persons suggested giving the system a personal name. The system could then be addressed directly by voice command. Furthermore, great importance was attached on an intelligible pronunciation by the system. This often occurred for hearing-impaired persons, who are dependent on special basic parameters (e.g. intelligible pronunciation) in order to be able to communicate with the system. This aspect clearly has to be addressed for practical applications, because more than 60% of the persons aged 70 yrs. and above show a significant decrease of hearing ability [7]. Further comments of the subjects suggested to rephrase some commands or to make the system for speech input friendlier.

5. The Acoustic User Interaction (AUI) Framework

As found in Section 2, acoustic input and output is desired by the target group of older users. Robust automatic recognition of speech is a desirable input modality, especially if combined with hands-free devices for speech acquisition. The following Section 5.1, therefore, describes the AUI framework including signal acquisition and the ASR system. Two further studies are described in Sections 5.2 and 5.3 based on this ASR system following the chosen user-centered design process. In Section 5.2 results of a study involving older users which is based on the mock-up system already described in Section 4 are presented before results using hands-free equipment are presented in Section 5.3. The results regarding the hands-free ASR system are obtained with younger persons and, thus, are only preliminary in terms of the focus of this paper. However, some general conclusion can already be drawn and involvement of older persons for an ASR system using hands-free equipment is subject to future work. All results in this section are purely technical results in terms of word recognition rates (WRR). Section 5.4 discusses the results concerning the ARS system.

5.1. Acquisition, detection and classification of acoustic input

The input via spoken commands offers great benefits especially for impaired people and in situations in which the hands are not freely available. Furthermore, ASR is hardly evitable if input modalities beneath conventional mouse/keyboard input are desired, such as for ambient technologies. The following section describes the AUI framework that was implemented for the PAHA/GAL system (cf. Figure 1) focusing on robust ASR in hands-free systems.

One great and yet unsolved challenge in ASR which still is an active research field is to ensure a robust transcription of language for various acoustic situations including ambient noise and reverberation. For such adverse acoustic environments, a reliable classification of the spoken utterances is often very difficult. Therefore, the use of hands-free equipment for signal acquisition is still a particular challenge even for the best state-of-the-art ASR systems and signal processing schemes to enhance the signal picked up by the microphone have to be applied [14-20].

A signal processing chain for detection and identification of acoustic events including spoken commands is shown in Figure 5. The signal is picked up by one or more microphones. The preprocessing stage enhances the signal by reducing ambient noise. Afterwards, an acoustic event, which can also be a speech utterance, can be detected and classified.

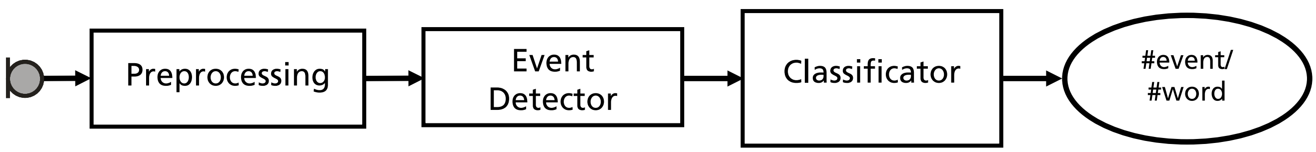

Figure 5: Signal processing chain for detection and identification of acoustic events and speech.

Since the use of multiple microphones allows for more advanced noise reduction schemes [32] several microphones are used in the AUI framework. The position and a picture of the microphone arrays are depicted in Figure 6. They are integrated into the ceiling of an apartment and in two spherical arrays and allow for signal enhancement as well as position estimation of acoustic sources [33] for the purpose of acoustic monitoring.

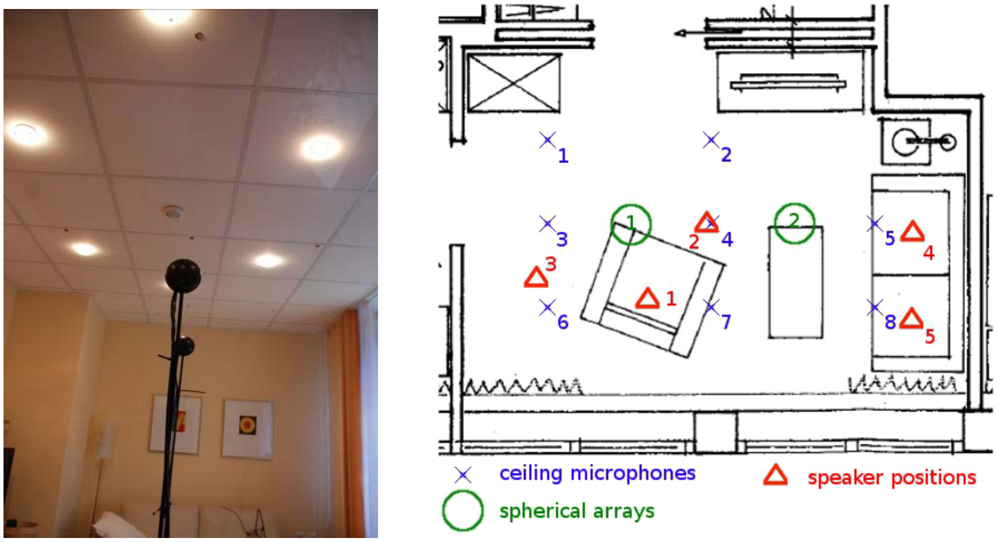

Figure 6: Acoustic signal acquisition of GAL AUI framework. Microphones are placed in the ceiling and in spherical arrays (left panel). Ceiling microphone positions, positions of the spherical arrays and speaker positions are marked in right panel.

The implemented speech recognizer for the AUI/PAHA tests is an isolated word recognition system which is based on a HMM classification framework [34-37] and the most common acoustic features, namely Mel frequency cepstral coefficients (MFCCs) [35-37]. The dialog structure (cf. Section 3) has been defined in advance to keep the vocabulary of the ASR system as small as possible. This was done to minimize the problems for ASR with hands-free equipment to reduce false recognitions. The German words chosen for the dialog are (English translation in brackets)**:**

Table III: The 25 keywords of the dialog system.

Numbers (10)	**Days (9)**	**Commands (6)**
Null (zero)	Montag (Monday)	Termin (appointment)
Eins (one)	Dienstag (Tuesday)	Neu (new)
Zwei (two)	Mittwoch (Wednesday)	Ja (yes)
Drei (three)	Donnerstag (Thursday)	Nein (no)
Vier (four)	Freitag (Friday)	Falsch (false)
Fünf (five)	Samstag (Saturday)	Stopp (stop)
Sechs (six)	Sonntag (Sunday)	
Sieben (seven)	Morgen (tomorrow)	
Acht (eight)	Übermorgen (day after tomorrow)	
Neun (nine)		

A method to reduce the active vocabulary for each recognition step within the dialog structure is to use different grammars for the ASR system (switched grammar technique, SGT). This means that the ASR system can reduce its vocabulary for each recognition step by exploiting the dialog structure described in Section 3. By this, false recognitions can be abated. This technique is a common approach for dialog systems in ASR [35,37].

5.2. ASR performance study under real user conditions using the mock-up system

In Section 4 a user study was described which utilized a mock-up system that simulated the use of an ASR controlled calendar application. For this study an investigator feigned the performance of a speech recognition software. However, an ASR system ran in the background, which was not used for the user study, but whose recognition was independently used to evaluate the performance. This evaluation is described in this section. A close-talk microphone was used for recording the training data and for the speech input by the older subjects. The ASR software was initially trained by six male and six female speakers in the age between 27 and 44 years (average: ~32) which constituted the basic training set (TS1). Thus, the technical system was trained by younger users as it is usual during development of such systems. Besides the basic training set a second training set (TS2) was used which consisted of the basic set and additional training recordings that were taken from 10 of the 12 mentioned older subjects (50% female) during the mock-up study. Purpose of this study is, thus, to evaluate if additional training by older persons is necessary in an ASR system for older persons since their pronunciation may be slightly different than pronunciation of younger persons partly due to slurred speech or hearing disabilities and a different perception of their own voice caused by this.

For the recognition process SGT was used (cf. Section 5.1). For the interpretation of the results it is important to mentioned that incorrect command prompts due to slurred or too slow speech (which had effects on the recognition of the connected digits), were not seen as user errors, but as a failure of the speech recognizer as it would be in a real user scenario.

The speech recognition software was based on linear HMMs with 14 states for each word [34-37]. MFCCs [35-37] were used as acoustic features in this experiment.

An ANOVA with the between-group factors age (age group 1: <=65 yrs. vs. age group 2: 66-75 yrs.), gender (male vs. female), and hearing loss (HL1 vs. HL 2) was computed to show the benefit of the additional training set (TS2) compared to TS1. The factors age and gender were nearly balanced. The results of ten participants are given in Table IV in terms of word recognition rates:

Table IV: Word recognition rates of ASR system for different training methods.

Factor	Groups	Training set 1 Means (SD) in %	Training set 2 Means (SD) in %	p-value (main effects)
Gender	Male (n=5)	94.7 (3.0)	97.8 (1.9)	F=1.3, p=0.307
	Female (n=5)	90.1 (7.3)	93.3 (5.1)	
Age	63-65 yrs. (n=5)	95.7 (1.8)	97.2 (1.9)	F=1.6, p=0.253
	66-75 yrs. (n=5)	90.0 (6.9)	93.9 (5.6)	
Hearing loss	6.3 to 28.8 dB HL (n=4)	93.9 (2.8)	96.6 (1.5)	F<1.0, p=0.567
	40.0 to 82.5 dB HL (n=6)	92.1 (5.6)	95.5 (4.3)	
Total		92.7 (5.6)	95.5 (4.3)	F=17.0, **p=0.006***
p-value=level of significance, F-value=F distribution; n=10				

As shown in Table V a word recognition rate of 94.7% was achieved for male and 90.1% for female users using TS1. The significant difference in WRR between the male and female users is mainly due to the fact that the recognition rate of one female user falls considerably compared to the other results. The reason for the large difference in WRR may be that she used a dentition which influenced the pronunciation and caused unusual noises. This observation and the fact that the ASR performance for older persons (66-75 yrs.) is significantly lower than for the younger user group (63-65 yrs.) leads to the conclusion that it is important to consider the user’s target age when picking the training recordings for an ASR system. Table IV shows that the additional training data of TS2 improved the overall WRR significantly from M=92.7 up to M=95.5%. A reason is that the same persons used for testing were now included in the training set. Thus, the ASR system gets adapted to the users' individual voice characteristics.

Each of the between-group main effects gender, age, and hearing loss was statistically not significant for both of the training sets. The interaction with the repeated measurement factor training set was not significant either (p-values > 0.05). Note, that for the comparisons of the between-group main effects the n of cases was five participants for each group. The data showed an advantage for the male and the younger age group, as well as for the group with the lower degree of hearing loss. In future research a bigger n of cases for each group is necessary to quantify the effects of age, gender, and hearing loss.

5.3. ASR performance for a hands-free equipment with a ceiling microphone

Since the AUI framework is intended to be an ambient system, speech input by means of a close-talk microphone as it was used in the study described in Section 5.2 is not the final goal. Instead, the system is supposed to work with distant, ambient microphones, e.g. in the ceiling as depicted in Figure 6. To test the performance of the speech recognizer for the calendar application together with hands-free equipment, recordings of eight male speakers were collected. Each keyword word recorded 10 times from each speaker synchronously with a ceiling and a headset microphone. During recording the speaker was sitting in the chair shown in Figure 6.

The mean age of the subjects in this study was M=31.8 yrs. (SD=7.1 yrs., range 23-44 yrs.). Although younger subjects were used for this study than for the rest of the paper the obtained results already allow for some general conclusions. A performance study using hands-free equipment with older persons is subject to future work.

In order to determine the ASR performance the so-called cross validation procedure was used: The recordings of seven speakers were used for training and the recording of the remaining speaker was used for testing. All eight possible combinations were tested and the average word recognition rate (WRR) was determined. The architecture of the ASR system used MFCCs [35-37] for the description of speech features and whole word HMMs to model the observations. SGT was not applied in this test to determine the results for a worst case scenario (cf. Section 5.1).

Three scenarios were tested as follows:

a) Recordings of the headset microphone were used for training and the signals for testing were made by ceiling microphone #4 (see Figure 6).

b) Ceiling microphone #4 was used for training as well as for testing.

c) The headset microphone was used for training as well as for testing.

A repeated measurement ANOVA with the three scenarios was computed. The means and standard deviations as well the three comparisons (Bonferroni, adjusted for multiple comparisons) are summarized in Table V. Significant comparisons ($p<0.05$) are indicated with an asterisk.

Table V: Word recognition rates of ASR system for different training methods.

N=8	Scenario a)	Scenario b)	Scenario c)
WRR Means (SD) in %	43.1 (26.5)	97.8 (2.3)	98.9 (2.1)
WRR Range in %	2.4 to 79.6	94.0 to 100.0	94.0 to 100.0
Comparisons	**Scenario a)**	**Scenario b)**	
Scenario b)	* ($p<0.05$)	-.-	
Scenario c)	* ($p<0.05$)	n.s.	
WRR= Word Recognition Rate, SD=standard deviation, n.s.: not significant			

Table V shows the results in terms of word recognition rates WRR. It can be seen that recognition rates are significantly higher in case that recordings for training and testing are produced in the same way (scenario b) and c) compared to the scenario a))

5.4. Discussion

In [37] it was shown that a minimum word recognition rate of 95% has to be achieved to be accepted by users. The reason is that it makes no difference for the user whether the system makes a mistake every twenty words or every hundred words. However, if the recognition rate is only 90% or less, then the application is considered unreliable. The results presented in Section 5.3 for scenario b) show that this level of user acceptance can be reached using the ceiling microphone if in the training recordings the acoustic properties of the later user environment will be considered. Future work will focus on older users for the hands-free case and a robust combination of noise reduction schemes using all available microphones in combination with the ASR system to obtain even higher recognition rates and a higher robustness.

The study described in Section 5.2 involving n=10 older persons showed that training of ARS systems with older subjects is essential. The number of different speakers for training the ASR system was still quite low and, thus, the ASR system could not be considered as fully speaker independent. Although a larger number of participants will be necessary and is subject to future work to obtain statistically more significant results regarding the between factors age, gender and hearing loss it can be seen that age related factors like slurred speech or different pronunciation, partly due to hearing losses, have to be considered for designing ASR systems for older persons.

6. Summary and future work

This contribution discussed the use of acoustic user interfaces focused on automatic speech recognition for ambient assisted living technologies. An acoustic user interface framework was exemplarily discussed for controlling a calendar application that was part of a personal activity and household assistant to support older persons. The development of the system was thoroughly based on a user-centered design approach. We presented different user studies to evaluate the desire and usefulness of acoustical speech input and output for AAL systems.

The system performance was mainly assessed as positive both in terms of technical measures and in terms of user rating. Acoustic input and output of an ambient system was found to be very important and the favorable strategy for interaction with the system. Most participants preferred natural speech as input and output modality for a calendar application independent of age, gender and technology experience.

Future work has to be done to further increase the user friendliness of the system e.g. by enhancing the current dialog structure. A free speech input without structured dialog was favored by nearly half of the users. Technically, the challenging problem of hands-free sound pickup, i.e. recording the sound signals by microphones at positions distant from the sound source (user) has to be tackled further by jointly enhancing pre-processing strategies and the ASR system.

7. References

[1] European Commision Staff. Working Document. Europes Demografic Future: Facts and Figures. Commission of the European Communities; May 2007.

[2] Statistical Federal Office of Germany. Demographic Changes in Germany: Population Development in Germany (German original title: Demografischer Wandel in Deutschland - Heft 1 - Bevölkerungs- und Haushaltsentwicklung im Bund und in den Ländern). Dec. 2007

[3] Statistical Federal Office of Germany. Demographic Changes in Germany: Impacts on Hospital Treatments and People in Need of Care (German original title: Demografischer Wandel in Deutschland - Heft 1 - Auswirkungen auf Krankenhausbehandlungen und Pflegebedürftige im Bund und in den Ländern); 2008.

[4] European Ambient Assisted Living Innovation Alliance. Ambient Assisted Living Roadmap. VDI/VDE-IT AALIANCE Office; 2009.

[5] Alexandersson J, Zimmermann G, Bund J. User Interfaces for AAL: How Can I Satisfy All Users? In: Proc. Ambient Assisted Living - AAL. 2. Deutscher Kongress (German congress). Berlin: Germany; pp 5-9; 2009.

[6] Helal S, Giraldo C, Kaddoura Y, Lee C, El Zabadani H, Mann W. Smart Phone Based Cognitive Assistant. In: Proceeding of The 2nd International Workshop on Ubiquitous Computing and Pervasive Healthcare Applications. Seattle; p 11; 2003.

[7] Meis M, Fleuren T, Meyer EM, Heuten W. User centered design process of the personal activity and household assistant: Methodology and first results. In: Proc. 3. German AAL congress. Berlin: Germany; 2009 Jan.

[8] Meis M, Appell J-E, Hohmann V, v Son N, Frowein H, Öster AM, Hein A. Tele-monitoring and Assistant System for People with Hearing Deficiencies: First Results from a User Requirement Study. Proceedings of European Conference on eHealth (ECEH); pp 163-175; 2007.

[9] Pieper R: The „virtual residential home“ as an option for dementia home care: suggestions for further development of the SOPHIA-project. In: Baehren T (Hrsg.), Ambient Assisted Living, German congress with exhibition, technologies – applications - management; pp 143-149; Berlin, Germany; Feb 30 – Jan 1.; 2008.

[10] Lower Saxony Research Network [Internet]. Design of Environments for Ageing (GAL). [cited 2010 Jul 1] Available from: http://www.altersgerechte-lebenswelten.de/

[11] Sharp H, Rogers Y, Preece J. Interaction Design: Beyond Human-Computer Interaction. 2nd ed. Wiley; 2007.

[12] Frenken T, Brell M, Eichelberg M, Gietzelt M, Martens B, Meyer EM, Spehr J, Wilken O, Hein A. Assistive Technologies for Supporting People with Dementia. AALIANCE conference. Malaga: Spain; 2010 Mar 11-12.

[13] Meyer EM, Heuten W, Meis M, Boll S. Multimodal Presentation of Ambient Reminders for Older Adults In: Proc. 3 German AAL congress. Berlin: Germany; 2010.

[14] Rennies J, Goetze S, Appell J-E. Personalized Acoustic Interfaces for Human-Computer Interaction. In: Ziefle M, Röcker C, editors. Human-Centred Design of E-Health Technologies: Concepts, Methods and Applications. IGI Global. In Press; 2010.
[15] Goetze S, Xiong F, Rennies J, Rohdenburg T, Appell J-E. Hands-Free Telecommunication for Elderly Persons Suffering from Hearing Deficiencies., Proc. 12th IEEE International Conference on E-Health Networking, Application and Services (Healthcom'10); 2010 Jul 1-3; Lyon: France; 2010.
[16] Hänsler E, Schmidt G. Speech and Audio Processing in Adverse Environments. Springer; 2008.
[17] Ephraim Y, Malah D. Speech Enhancement Using a Minimum Mean–Square Error Log-Spectral Amplitude Estimator. IEEE Trans. On Acoustics, Speech and Signal Processing 1985;33(2):443-445.
[18] Goetze S, Mildner V, Kammeyer K-D. A psychoacoustic noise reduction approach for stereo hands-free systems. In: Audio Engineering Society (AES), 120th Convention. Paris: France; 2006.
[19] Goetze S, Kallinger M, Kammeyer K-D. Residual echo power spectral density estimation based on an optimal smoothed misalignment for acoustic echo cancelation. In: International Workshop on Acoustic Echo and Noise Control (IWAENC-2005). Eindhoven: The Netherlands; 2005.
[20] Goetze S, Rennies J, Appell J-E. Intelligent conference systems for natural hands-free communication (German original title: Intelligente Konferenzsysteme für natürliche Freisprechkommunikation). In: Schick A, Meis M, Nocke C, editors. Contributions to psychological acoustics, acoustics in office and buildings. Oldenburg: Germany: Isensee; 2010; ISBN:978-3-89995-710-5.
[21] Lawson SW, Nutter D, Wilson P. Design of Interactive Technology for Ageing-In-Place. In: Stephandidis C, editiors. Universal Access in HCI, Part I, HCII 2007, LNCS 4554. Springer, Heidelberg; 2007. pp 960-967.
[22] Word Health Organization. Grades of Hearing Impairment. Hearing Networks News 2001;1:144-157.
[23] Boll, S., Heuten, W., Meyer, E.M. & Meis, M. (submitted for publication). Development of a Multimodal Reminder System for Older Persons in their Residential Home. Informatics for Health and Social Care, SI Ageing & Technology.
[24] Luštrek M, Kaluža B. Fall Detection and Activity Recognition with Machine Learning. Informatica 2009;33:205-212.
[25] Marschollek M, Nemitz G, Gietzelt M, Wolf KH, Meyer zu Schwabedissen H, Haux R. Predicting in-patient falls in a geriatric clinic: a clinical study combining assessment data and simple sensory gait measurements. Z Gerontol Geriatr 2009;42(4):317-21.
[26] von Renteln-Kruse W, Krause T. Fall events in geriatric hospital in patients. Results of prospective recording over a 3 year period. Zeitschrift für Gerontologie und Geriatrie 2004;37:9-14.
[27] Luštrek M, Kaluža B. Fall Detection and Activity Recognition with Machine Learning. Informatica 2009;33:206-212.
[28] National Ageing and Research Institute. An analysis of research on preventing falls and falls injury in older people: community, residential and hospital settings. Canberra: Australian Government Department of Health and Ageing; 2004.
[29] Davis, A.: Population study of the ability to benefit from amplification and the provision of a hearing aid in 55-74 year old first time hearing aid users. Int J Audiology 42, p.2S39-2S52, 2003
[30] Johansson, M.S.K. and Arlinger, S.D.: Prevalence of hearing impairment in a population in Sweden. Int J Audiology 42, p.18-28, 2003
[31] Uimonen, S., Huttunen K., Jounio-Ervasti, K. and Sorri, M.: Do we know the real need for hearing rehabilitation at the population level? Hearing impairments in the 5- to 75-year old cross-sectional Finnish population. British J Audiology 33(1), p.53-59, 1999
[32] Bitzer, J. & Simmer, K.U. Superdirective microphone arrays. In: M. S. Brandstein & D. Ward (Eds.), Microphone Arrays: Signal Processing Techniques and Applications. Springer, Berlin; pp. 19-38, 2001.
[33] Doblinger, G. (2006). Localization and Tracking of Acoustical Sources. In: Topics in Acoustic Echo and Noise Control. Berlin: Springer, pp. 91-122.

[34] Duda RO, Hart PE, Stork DG. Pattern Classification. Wiley-Interscience. 2nd ed; 2004
[35] Benesty J, Sondh MM, Huang Y. Springer Handbook of Speech Recognition. Springer, New York; 2008.
[36] Rabiner L, Juang B-H. Fundamentals of speech recognition. Prentice-Hall; 1993.
[37] Pfister B, Kaufmann T. Speech processing – Fundamentals and mthods for speech synthesis and speech recognition (German original title: Sprachverarbeitung – Grundlagen und Methoden der Sprachsynthese und Spracherkennung). Springer, Berlin Heidelberg; 2008. p 290.