Voice Controlled **Assis**tive Care and Communication Services for the Home

D4.3 – Field Trials Evaluation Report

Project Number:	AAL-2010-3-106
Coordinator	Manfred Tscheligi AIT Austrian Institute of Technology GmbH, AT
Category:	Deliverable (*public*)
Reference:	D4.3; *v1.0; 28-02-2015*
Status:	Final
Responsible:	AIT
Participants:	AIT, APHP, EURAG, I&S, IT
Related to:	WP4: T4.3

Co-funded by the AAL Joint Programme

Abstract

This deliverable describes the results of the field trial of the vAssist system prototypes. The aim of the vAssist project is to offer a voice-controlled home care and tele-medicine system for older adults in their home environment. The integrated services will be available on already existing devices at home, such as TV, PC, Smartphone or tablet.

The evaluation will emphasize on the following points: A) System and User Evaluation: This part evaluates the system against the use cases defined in WP 2. The vAssist system is evaluated and assessed in real world environments showing relevance, impact and importance (including weaknesses, strengths and improvements) as well as QoE factors. B) Business Model Evaluation: Further, the acceptance of the developed service delivery models is evaluated.

End-users of the defined target groups are invited to evaluate the system. Results of the field trials will contribute to the subsequent development and modification processes of the vAssist system.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

1 Introduction

The D4.3 Field Trials Evaluation Report presents the results of the field trial of the vAssist system prototypes within real life scenarios based on the data collection of the user requirement analysis, the first and the second lab trials and the evaluation and assessment plan for the lab and field trials. The overall goal of the user evaluations in the field is to ensure that issues related to Quality of Experience and Quality of Service factors are detected, to investigate the influence of the vAssist system on the users' life, e.g. their quality of life, and to evaluate the business model.

This deliverable describes the selected sample of test participants and an analysis of the evaluations together with recommendations for the improvement of the system prototypes.

1.1 Background

The following section summarizes the background of the document and indicates related deliverables.

1.1.1 User requirements

D2.1 (User requirements), chapter 8 summarizes the results of the user requirements analysis. vAssist has to deal with primary users (seniors) and secondary users (formal and informal caregivers). Devices to be used to interact with vAssist are mobile devices (at home and out of home) and static devices (at home). The expected services are: making phone calls, writing emails, SMS, MMS to relatives, friends, etc., managing the contact information and sharing information such as photos etc. with family and friends. Well-being services are also required by the users, including emergency functionalities (calls, tele-alarm), recording and reporting of medical data, drug intake reminder and drug diary, electronic pill jars and the information exchange between user and health professionals.

Although voice based interaction is preferred, a text or graphical user interface should be in place to allow for monitoring and/or controlling the interaction process.

1.1.2 Service provider requirements

Document D2.2 (Service provider and business requirements) describes the requirements of service providers. The key requirements of service providers are the request for a maximum of flexibility that should be offered by the architecture to enable the integration of existing services with minimal effort and to enhance or replace vAssist modules during the lifecycle of the vAssist platform in order to respond to business needs as fast as possible.

1.1.3 System definition

Document D2.3 (System definition) describes the technical aspects and logical structure of the vAssist platform and its individual system components designed on the basis of the results of the user requirements analysis. It includes the specification of the service interaction processes and their integration into scenarios. The interaction of the individual components and their interfaces are described in detail in order to allow each partner to start the development of components or to go into a more detailed level of specification as required by the components, according to the partner's development process model.

1.1.4 Scenario definition

Document D2.4 (Scenarios definition) illustrates the business and service scenarios based on user requirement data from the target groups. Business scenarios refer to purchase, setup and the maintenance of the vAssist system. Service scenarios encompass contact management, audio call, E-Mail, SMS, MMS, internet/information search, calendar/reminder, video call, well-being diary, cognitive games, fall detection and actimetrics. The selected scenarios outline the functionality of the vAssist system and the interaction with the future end-users based on the information collected from end-users during the requirement phase. Moreover, the scenarios are an important supportive instrument for the technical definition of the system architecture that is done in parallel in D2.3 (System Definition).

1.1.5 Lab Trial Evaluations

D4.2 (Lab Trial Evaluation Report) describes the results of the first and the second lab trial of the vAssist system prototypes. The focus of the evaluation process is on different tasks and services, which are available in the first and the second system prototype. End-users of the defined target groups were invited to evaluate the system. Results of both iterations of lab trials contributed to the subsequent development and modification processes of the vAssist system.

1.1.6 Evaluation and assessment plan

The D4.1 (Evaluation and assessment plan for the lab and field trials) presents the roadmap for all evaluation processes of the vAssist prototypes within real life scenarios based on the data collection of the user requirements analysis. The deliverable defines the general procedural methods for the implementation of the lab (T4.2) and field trials (T4.3) and specifies the goals, strategies, tasks and subtasks of the whole evaluation process.

1.2 Scope of this deliverable

D4.3 (Field Trial Evaluation Report) presents the results of the field evaluation of the vAssist system by end-users. Chapter 2 describes the implementation and procedure. Results of the quantitative measurements and qualitative analysis reported separately for each trial site as well as implications for the overall system design are presented in chapter 3. Chapter 4 provides a closing summary and draws a conclusion taking all field evaluations into account.

2 Evaluation procedure and study setup

The methodology and test setup of the field evaluation are described in D4.1 (Evaluation and Assessment Plan for the Lab and Field trials). The vAssist system was evaluated and assessed in real world environments showing relevance, impact and importance of defined scenarios (including weaknesses, strengths and improvements) as well as QoE factors. Further, impact on the users and the acceptance of the developed service delivery models was evaluated.

The evaluation took place in Austria, France and Italy. Table 1 gives an overview of the time plan.

Table 1: Overview time plan

Month:		**November**		**December**					**January**					**February**		
Calendar week:		46	47	48	49	50	51	52	1	2	3	4	5	6	7	8
Trial Site																
Austria	Group 1:	■	■	■	■	■	■	■	■							
	Group 2:					■	■	■	■	■	■	■	■			
	Group 3:													■	■	■
France						■	■	■	■	■	■	■	■	■	■	■
Italy				■	■	■	■	■	■	■	■	■	■	■	■	■

At all trial sites, the field study started with a kick-off workshop where the participants were trained in the use of vAssist and the applied data gathering methods. User factors, such as socio-demographic data, experience with technology, basic information regarding language behavior etc., were collected at the end of the workshop. During the field phase the users were asked to fill in questionnaires, diaries applying the Day Reconstruction Method [5] and Critical Incidents Technique [2] and to conduct weekly telephone interviews. Furthermore, a problem-centered interview took place in Austria and France at the end of the field trial. Hence, a mixed-methods approach was applied to cross validate and enhance the gained insights and their credibility and get a more detailed picture of the perception, drawbacks and strengths of the system.

In Austria participants received a financial allowance after the participation in the study.

2.1 Summary of the Methodology

To evaluate the vAssist system selected aspects of the taxonomy by Wechsung et al. [11] were operationalized. Fig. 1 shows an overview of the taxonomy and indicates the respective evaluation foci.

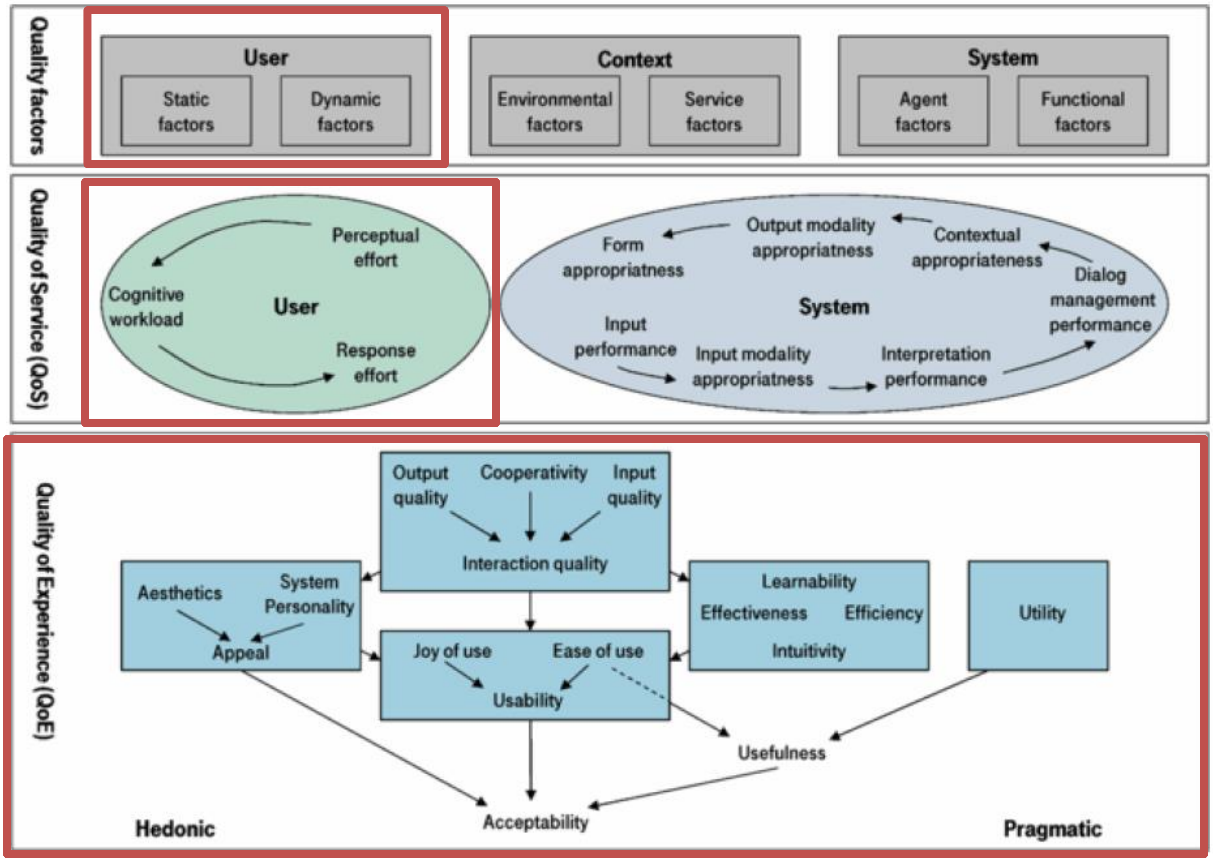


Fig. 1: Taxonomy of QoS and QoE aspects of multimodal human–machine interaction [11]

In order to gain detailed and balanced insights, quantitative and qualitative methods were applied. For the quantitative measurement, the following questionnaires were filled in by the participants.

- SAM [1]
- TAM3 [10]
- UEQ [6]
- SASSI [4]
- NASA-TLX [3]
- WHOQOL-BREF [9]

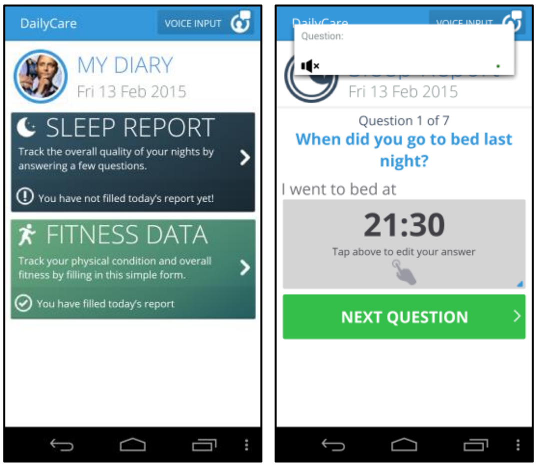
The Critical Incidents Technique [2] the Day Reconstruction Method (DRM) [5] and telephone interviews offered self-reported data which was collected during the field trial. A problem-centered Interview (PCI) [8] was conducted at the end of the field study. Details regarding the operationalization and methods can be found in D4.1 Evaluation and assessment plan for the lab and field trials.

2.2 System prototype and system usage during the field phase

The following section sums up the system usage during the field trial. A detailed description can be found in D4.1 (Evaluation and Assessment Plan for the Lab and Field trials).

In Austria and France the PillBox application, the DailyCare application and the communication services were tested. In addition, France included the Ambulatory Terminal. The DailyCom was evaluated in Italy.

DailyCare: The sleep report had to be filled in daily. In contrast, the fitness data report was not required if no specific physical activity was performed. The participants could choose whether they want to fill in the data via speech or touch interaction.



Track the overall quality of your nights by answering a few questions.

You have not filled today's report yet!

Track your physical condition and overall fitness by filling in this simple form.

You have filled today's report

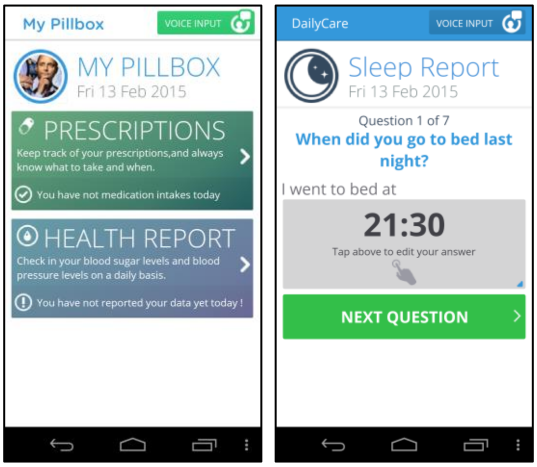
When did you go to bed last night?

I went to bed at

Tap above to edit your answer

Fig. 2: Daily Care application

PillBox: Due to ethical and security issues, the medication reminder was not tested during the field trials. The input of the blood pressure data and the blood sugar level was entered once a week based on defined fictitious data to respect the participants' privacy. As with the DailyCare application, the participants could choose between speech and touch interaction.



Keep track of your prescriptions,and always know what to take and when.

You have not medication intakes today

Check in your blood sugar levels and blood pressure levels on a daily basis.

You have not reported your data yet today !

When did you go to bed last night?

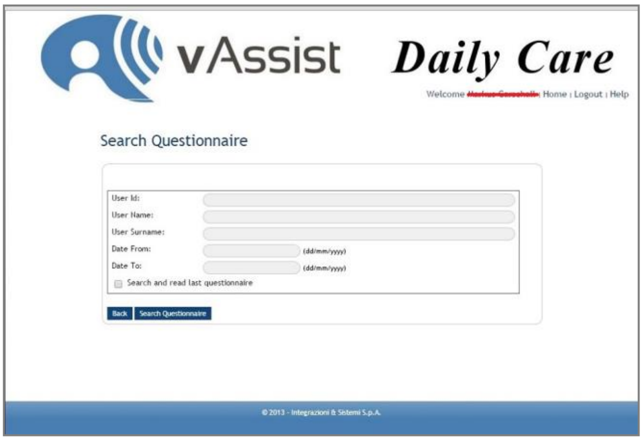
I went to bed at

Tap above to edit your answer

Fig. 3: PillBox application

DailyCom: For the evaluation of the DailyCom, the primary users filled in questionnaires via speech, providing the necessary data to monitor the state of health and the comfort in their home environment from a minimum of once per week to a maximum of once per day. The DailyCom also provides a messaging service which could be used as often as wished. The DailyCom application offers speech interaction but no touch interaction.

The web application of the DailyCom was used by the secondary users. With the template management function, secondary users created, edited and deleted the templates for the questionnaires provided via the DailyCom app. With the questionnaire management function, secondary users reviewed, searched, and saved the questionnaires filled in by primary users.


vAssist
Daily Care
Welcome [redacted] | Home | Logout | Help
Back
Search Questionnaire
© 2013 - Integrazioni & Sistemi S.p.A.

Search Questionnaire

User Id:
User Name:
User Surname:
Date From: (dd/mm/yyyy)
Date To: (dd/mm/yyyy)
□ Search and read last questionnaire

Fig. 4: DailyCom web application

During the field trial, the secondary users constantly monitored the patients and analyzed the results of the questionnaires. If some abnormal or unclear responses were recognized, the secondary users contacted the primary users and scheduled a visit for further investigation if necessary.

Ambulatory Terminal: Due to ethical and security issues, fall detection service of the Ambulatory Terminal (AT) was not tested during the field trail. The main task regarding the AT was wearing the device as often as possible during the trial to simulate an ideal usage of this kind of device. In addition, concerned participants executed a pulse rate measure once per month. This task required wearing the pulse rating sensor during two hours at least (plugged ear clip) without stopping daily activities. Further, each participant activated an emergency button twice during trial period, without specific date or time indication.

3 Results

In this section the results of the field trial evaluation of the vAssist system, according to the setup described in D4.1 (Evaluation and Assessment Plan for the Lab and Field trials) are reported.

3.1 User factors

The following table sums up the most relevant user factors and describes the sample.

Table 2: Description of sample and pre-interview results

Data	Austria	France	Italy
Participants	19	9	10
Age (range)	66 - 84 years	61 - 87 years	63 - 70 years
Gender	8 ♀ 9 ♂ (2 Persons decided not to furnish particulars)	7 ♀ 2 ♂	5♀ 5♂
Retiree	19/19	9/9	9/10
Experience with touchscreen devices	14/19	7/9	6/10
Physical constrains	5/19	0/10	5/10
Frequency doctor visits	11 less than once a month 4 at least once a month 5 never	6 less than once a month 2 at least once a month 1 never	3 once a week 1 once a month 7 by appointment only
Perceived stress level	1.8 (4-point Likert scale)	1.2 (4-point Likert scale)	“Low”
Reduction of doctor visits requested	10/19	0/10	4/10

3.1.1 User factors in detail

A detailed user factors description per trial site explains how the test groups of the field study is characterized.

Austria

A total of 19 people participated in the field study. Eight persons described themselves as female, nine as male and two persons refused to answer the question. The participants are between 66 and 84 years old. The average age is 71.25 years (SD= 5.04). All participants are retired and they worked in different areas. Almost half of the group worked in technical professions and as teachers; the rest of the group was employed in different fields of work such as office management, trainer, manager, etc.

All but two participants (one acquired the German language already 45 years ago) are native German speakers and most of them speak a rather weak dialect. Only two participants indicated to speak a rather strong dialect.

All participants are right-handed reflecting the fact that within this generation, the left hand is considered to be the so called “bad hand” and people had to learn to use the right hand.

Regarding the ownership of new technologies, ten participants own a feature phone and already twelve of 19 participants own a smartphone indication that some participants own both devices. 15 own a PC or laptop. The tablet is only available for four persons at home. Eight of 19 participants have their own navigation device. This number seems to be a rather high due to the fact that not all of them own a car.

Concerning the usage of a mobile phone, all participants of the field study are able to initiate calls and write SMS. Some (7/19) also play games on the phone, listen to music (10/19) or download music (9/19). All participants are able to adjust the settings according to their needs.

Initiating calls with a smartphone via contact list can be done by only 14 of 19 participants, 13/19 can manage it via calling list. Writing SMS seems not to be as easy as with a mobile phone as a lower number of participants is able to execute that task (13/19). Further, fewer participants are able to take photos (10/19), write an E-Mail (11/19) and use apps (11/19). Searching and installing new apps can be done by eight of 19 participants. About half of the participants are in a position to use the following functions: photos, videos, alarms, reminders, notes, weather function, navigation, music, internet and settings.

Almost all participants are able to boot a computer, to play and to create documents, save and print them, surf the internet, download programs, edit and print photos, etc. The stated computer usages skills are probably due to the fact that half of the participants were technicians and teachers who were accustomed to the usage of these devices in their work.

Participants who own a tablet are well versed with the main functions.

About half of the participants have already used speech commands. One person wrote a SMS. Writing E-Mails and searching via Google was already done by four people. Eight participants already initiated a call via speech command.

The majority of participants (14/19) stated that they do not have any difficulties in handling touch screens. Only five persons indicate having problems, four of them, because they have the so called "fat finger" problem. One participant mentioned a problem after a hand surgery exerting pressure for touch interaction. Another participant suffers from a tremor.

Concerning the health status of the participants, about half of person (11/19) sees a medical doctor less than once a month, four of them see the MD at least once a month, five never visit a MD. Nevertheless, all participants state that there are not too many physical appointments. But if possible, half of them would like to reduce the amount of physical appointments. On average, the physical appointments cause just low stress for the participants; only two of them perceive it as a huge burden.

Only one participant neither needs glasses nor lenses, seven persons are farsighted, eight nearsighted and there are diseases such as astigmatism and cataracts which are a typical symptom of old age. On average, the participants assess their visual faculty as rather good. Two participants are wearing a hearing aid. Basically, they rated their hearing ability as rather good.

Eight of the participants monitor their well-being. Some of them are documenting their blood pressure (nine persons) and one person monitors the physical activity. The sleeping behavior is not an issue to be observed; only one person does it. The blood sugar is measured by three participants (suffering from diabetes).

Regarding medical equipment, 13 persons are in possession of a blood pressure gauge and five of them own a blood glucose meter. Cardiovascular diseases are the most common ones in Austria, so the presence of a blood pressure gauge at home is often represented in Austrian households, while blood glucose meters are only in the hands of people with diabetes and not used for prophylaxis.

<u>France</u>

A total of 9 people participated in the field study – seven described themselves as female and two as male. The participants are between 61 and 87 years. The average age is 70.44 years (SD= 8.52). All participants are already retired. Before, they were working in different areas, such as technology engineering, nurses, management consulting, teachers or secretary.

All but one participant are native French speakers and most of them speak a rather weak dialect. The remaining participant indicated to speak a rather strong dialect (natively Luxembourgish speaker, but she has learned French as second language since her childhood).

Five participants are right-handed, two others ones are left-handed, and the last one described himself as able to use both hands.

With regard to the ownership of new technologies, five participants own a feature phone; already four of nine participants have got a smartphone. All of them own a PC or laptop, while the tablet is only available for four persons at home. Two of nine persons have their own navigation device. Only four participants are without smartphones.

All participants of the field study are able to initiate calls and write SMS with a mobile phone. Most of them (5/9) also play games on the phone, but just a minority listens to music or downloads music

(2/9). Seven participants are in a position to adjust the settings according to their needs on the phone but two participants are not able to execute this task.

Initiating calls with the smartphone via contact list can be done by all participants (5/9 participants own a smartphone), but one of them can’t manage to call somebody via the calling list. Writing SMS seems to be as easy as with the normal mobile phone (5/5), but sharing pictures via MMS seems more difficult since only two participants are able to execute this task. All participants are able to take photos (5/5), write E-Mails (5/5), but only two participants use apps. Searching and installing new apps can be done by three participants. All participants could easily set an alarm clock and surf the internet but other functions, such as setting reminder, making a video, creating notes, using weather forecast, navigation app or changing settings, are used by only half of smartphones’ owners. One participant is able to edit pictures and to listen and load music on his device.

Almost all participants are able to boot the computer, to play games and to create a Word document, save and print them, or to surf the internet. Only five participants are able to use more advanced functionalities, such as downloading programs, editing and printing photos or adjusting settings; three participants are able to copy videos but no participant is able to edit this kind of media. Most of participants have learnt using a computer at work and they are still using it since they are retired.

Participants who own a tablet are well versed with the main functions (write and send SMS and E-Mails, using, searching and installing new apps, taking and editing pictures) but other functions are less used.

Five participants are already experienced with speech commands. One person had already written E-Mails. Searching via Google in the internet was already done by three people, while two participants have already initiated a call via speech input.

All participants state that they do not have any difficulties with touch screens. Even if some of them indicated suffering of arthritis, coping strategies seem to be unconscious. For example, one participant described sensitivity issues at her fingertips during a meeting but she did not report a physical restriction in the questionnaire.

Regarding the health status of the participants, the majority (6/9) sees a medical doctor less than once a month, two of them see the MD at least once a month and one participant never visit a MD. They all confirm that there are not too many physical appointments. Consistently, no one would like to reduce the physical appointments. On average, physical appointments are not described as too stressful (1.2), only one of participants perceive it as a moderate burden

Only one participant neither needs glasses nor lenses. Five persons are farsighted, four participants are nearsighted and there are diseases such as astigmatism, AMD or cataracts for four participants. On average, the participants assess their visual faculty as rather good (3.62). Three participants are wearing a hearing aid. Basically, they rated their hearing ability as slightly less good (3) than their visual acuity.

Monitoring of well-being is interesting for eight of the nine participants. Two persons have to monitor their blood pressure and half of them document their physical activity (walking, gym, stretching, and sport). Additionally, two participants take care of their nutrition and three included social/cultural activities as a field of well-being monitoring. The sleeping behavior is not an issue to be observed. Only one person records these data. The blood sugar is measured by five participants but only as control measures (once per year in specialized laboratory). The blood pressure is also inspected frequently by additional five persons with personal blood pressure gauge or by the physicians in charge.

Concerning medical equipment, six participants are in possession of a blood pressure gauge but none of them owns a blood glucose meter. Cardiovascular diseases are the most common ones in France, so the presence of a blood pressure gauge at home is often represented in French households while blood glucose meters are only in the hands of people with diabetes (none of participants).

Italy

In Italy, ten persons participated in the field trial, five female and five male persons aged between 63 and 70 years. All but one participants are already retired. Before, they worked in different areas: teaching, medical, administrative and commercial activities. Most of the participants do not speak a strong dialect and all but one are right-handed.

Five participants indicated that they had experience with touchscreen devices before the field trial. Regarding the usage of a mobile phone, all participants do not have any problems initiating calls and writing SMS. Some (2/10) also play games on the phone but no one listens or downloads music. Five participants are in a position to adjust the settings according to their needs on the phone. Five participants indicate that they know how to surf the internet.

Concerning the health status of the participants, seven users usually see a medical doctor by appointment only, three users once a week and one user once a month. None of the participants indicate to have any particular problems of hearing and of sight. Further, all but one participants rate their Quality of Live as good.

3.2 Quantitative Results

The following section shows the quantitative evaluation of the vAssist system and results regarding the Quality of the Service and the Quality of Experience. Further, the Quality of Life metrics are presented.

3.2.1 Quality of Service metrics

To evaluate the Quality of the Service we focus on the user interaction performance aspects. In line with Wechsung et al. [11] the NASA Task Load Index (NASA-TLX) [2] was applied to quantify the effort required from the user interacting with the system.

Fig. 5 shows the mean ratings of the task load for the vAssist system with speech interaction on a scale starting with 0 (low effort) up to 100 (high effort) and the confidence intervals (p=0.95) for Austria and France.

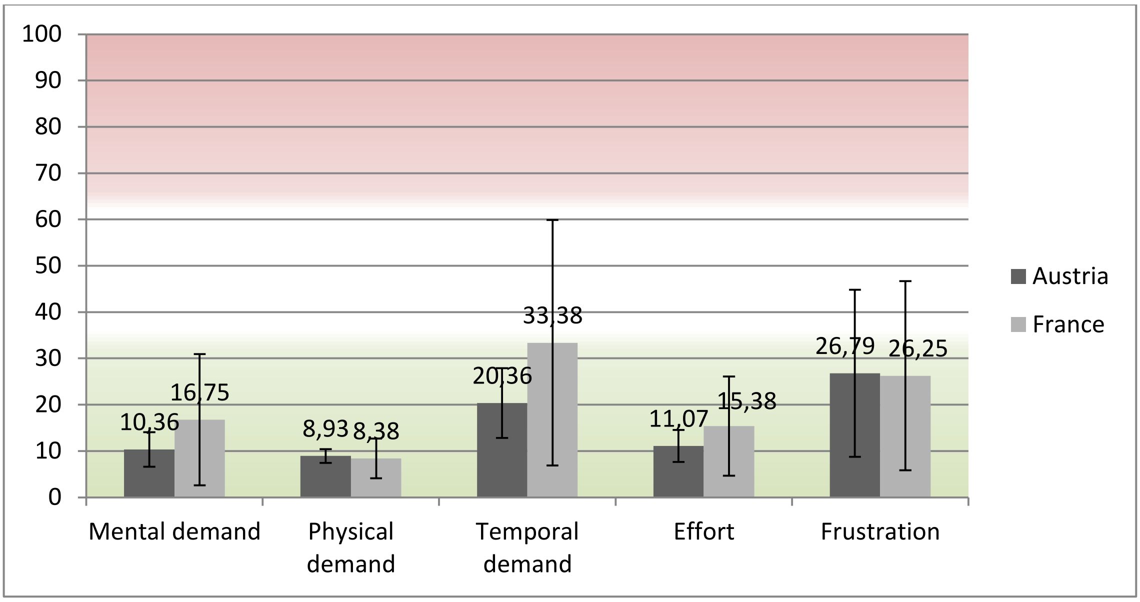


Fig. 5: Quantifying the user interaction performance with the NASA-TLX [3]

Overall, the task workload is judged as low indicating a good user interaction performance. Even though the French participants rated the temporal demand as considerable higher compared to the Austrian participants, this score hast to be interpreted carefully as great individual differences in the perception of the temporal demand are given (Mn: 33.38, Sd= 38.28). Also, the indicated frustration level varies but is perceived consistently as rather low (Austria: Mn= 26.79, Sd=25.83; French: Mn=26.25, Sd= 29.49).

Analyzing the task load over time (see Fig. 6), the results show a slight decrease after four weeks using the system.

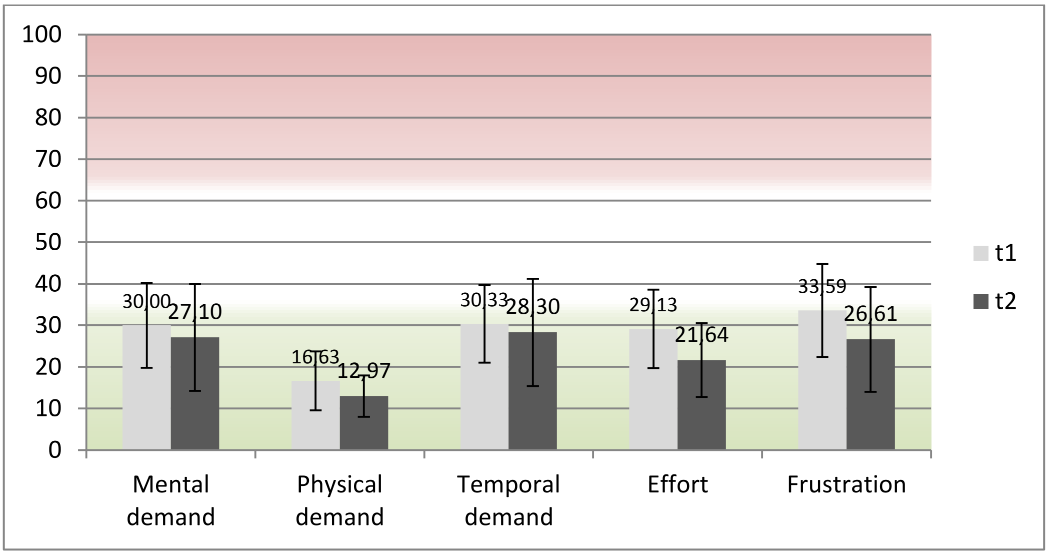


Fig. 6: Development of the task load over time

This slight decrease of the ratings can be explained with the immediate rather low task load. However, this trend is marginal and mainly concerns the scales effort (-7.49) and frustration (-6.98).

3.2.2 Quality of Experience metrics

To quantify the Quality of Experience, the SASSI [4], UEQ [6] and TAM3 [10] questionnaires were applied. In Italy, only reduced versions of the questionnaires were used in order to avoid overwhelming the participants.

The Subjective Assessment of Speech System Interfaces (SASSI) [4] serves as operationalization of the interaction quality, i.e. the perceived input and output quality and the system's cooperatively. Further, the results also indicate the ease of use [1]. In detail, a seven-point Likert scale is used to measure the response accuracy, likability, cognitive demand, annoyance, habitability, i.e. the pendant of "visibility" for speech-based systems, and speed.

Fig. 5 shows the average ratings and confidence intervals (p=0.95) for speech-based interaction in Austria and France.

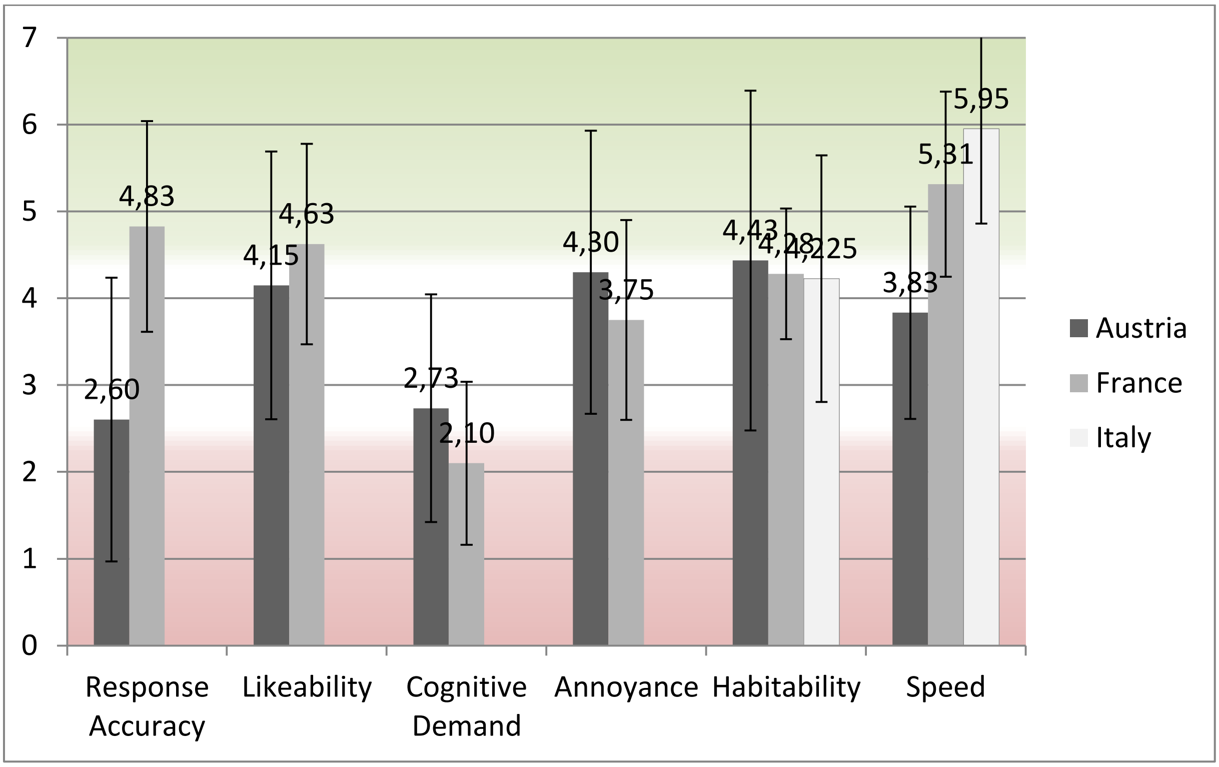


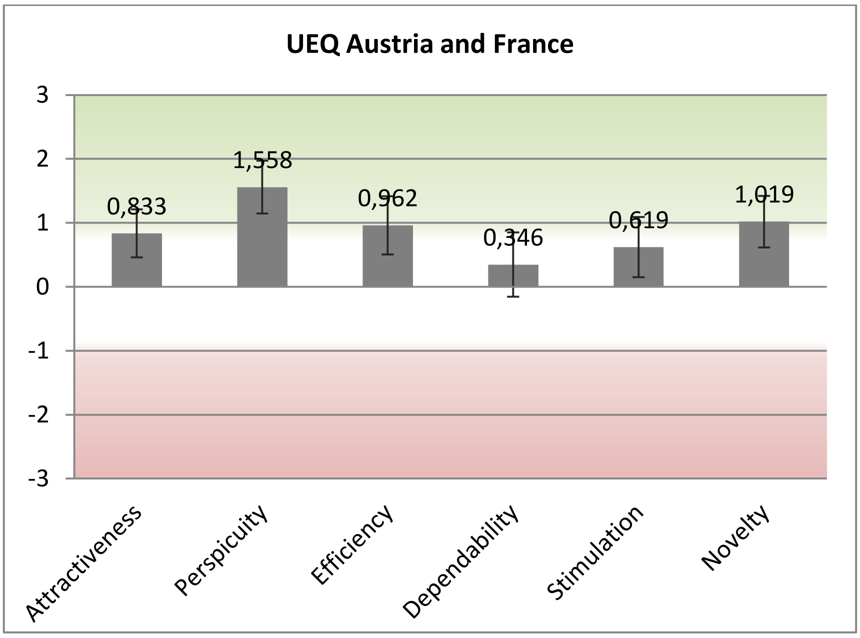
Fig. 7: SASSI [4] scores for speech-based interaction

While the French participants perceived the system response accuracy as rather good (Mn= 4.83, Sd=1.75), the Austrian ratings reveal potential improvements (Mn= 2.6, Sd=2.04). Also differences in the perception of the system's speed were found. The users in Italy and France rate the speed remarkably better (Italy: Mn=5.95, Sd=1.76; France: Mn= 5.31, Sd=1.54) compared to the Austrian participants (Mn=3.83, Sd=1.53).

In contrast, the likability (Austria: Mn=4.15, Sd=1.93; France: Mn=4.63, Sd=1.93) and habitability (Austria: Mn=4.43, Sd=2.45; France: Mn=4.28, Sd=1.08) are perceived as rather good by all participants and. In line with the NASA-TLX, the cognitive demand is rated as fair or rather low (Austria: Mn=2.73, Sd=1.64; France: Mn=2.1, Sd=1.35).

The average annoyance caused by the system usage is 4.3 (Sd=2.04) in Austria and 3.75 (Sd=1.66) in France and contradicts the other more positive rated scales. To get deeper insights, also the ratings of the User Experience Questionnaire (UEQ) [6] were analyzed. In detail, the UEQ was used to measure the Usability, Ease of Use and Joy of Use and the utility and usefulness of the system as defined by Wechsung et al. [1].

The analysis of the UEQ was conducted separately for Austria and France and Italy because the Italian participants were asked to fill in a reduced version of the questionnaire. Hence, the results for the Italian side offer insights for tendencies but have to be interpreted carefully.



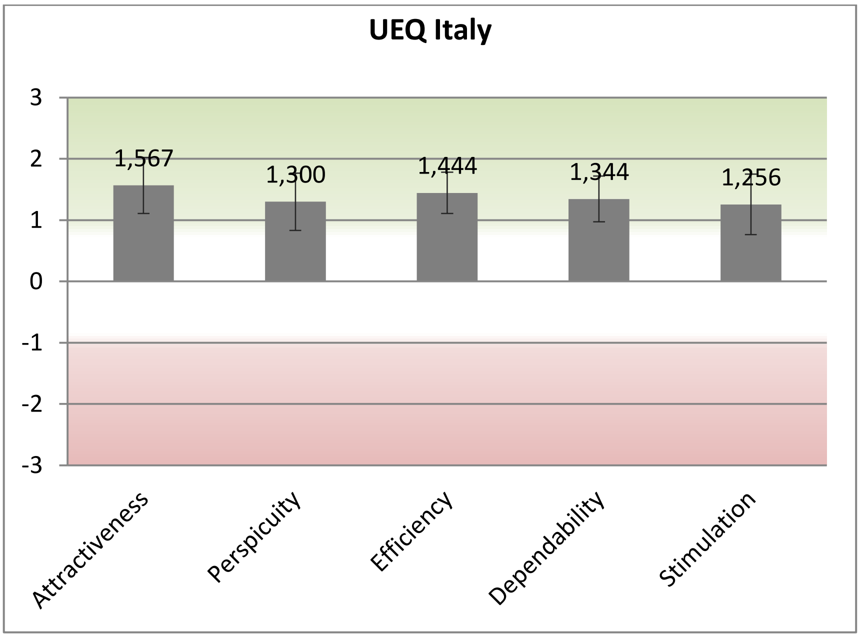


Fig. 8 The User Experience Questionnaire (UEQ) [6]

In general, the User Experience Questionnaires for all trial sites could be described as good or neutral. Attractiveness (Mn= 0,833, Sd=0.976), perspicuity (Mn= 1,558, Sd=1.075), efficiency (Mn= 0,962, Sd=1.187) and novelty (Mn= 1,019, Sd=1.051) are perceived as sufficient and positive. In contrast, the dependability (Mn= 0,346, Sd=1.308) and stimulation (Mn= 0,619, Sd=1.224) are rated as neutral, i.e. neither positive nor negative. Likewise, the analysis of the questionnaires collected in Italy indicates a positive User Experience.

The acceptability of a system is rather considered as the consequence of the Quality of Experience than a QoE factor itself [1]. Fig. 9 shows the results of the factors according to the Technology Acceptance Model 3 (TAM3) [10]. In detail, the average ratings on a seven-point Likert scale and the confidence intervals (p=0.95) are visualized.

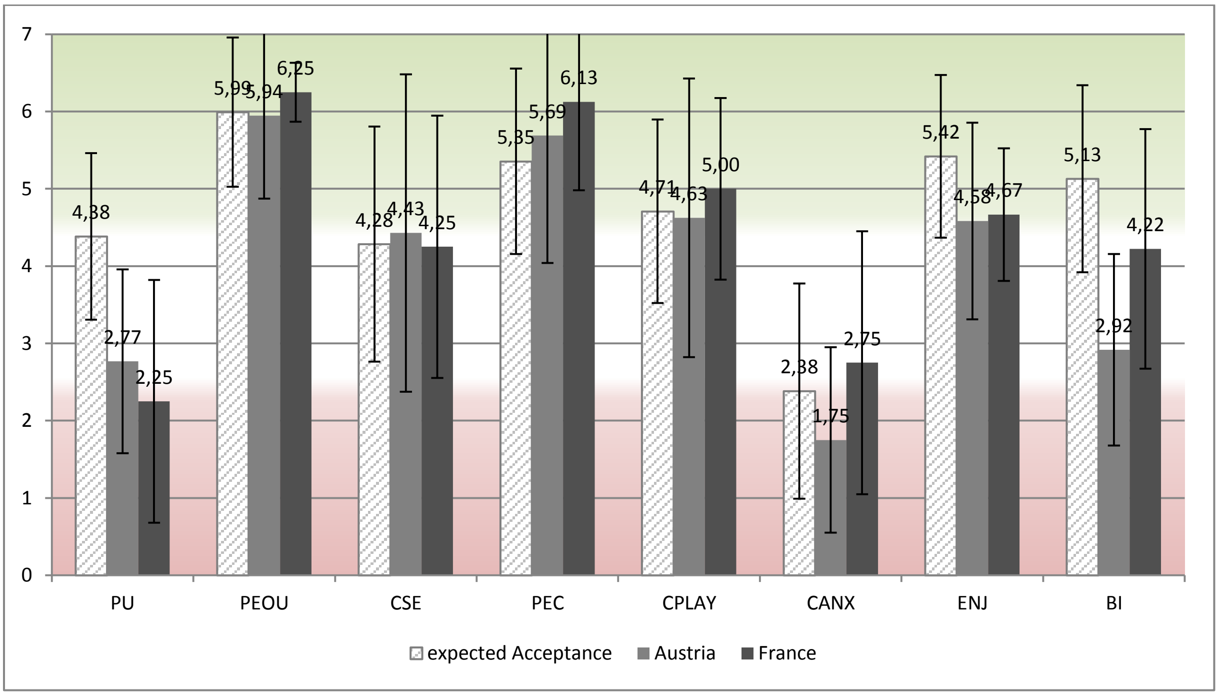


Fig. 9: Technology Acceptance Model (TAM3) [10]

In Austria and France, the perceived usefulness (PU) of the vAssist system is rated as below neutral (Austria: Mn=2.77, Sd=1.49; France: Mn=2.25, Sd=2.27) and compared to the expected perceived usefulness (asked at the beginning of the field study) it even decreased. As a result, the average behavioral intention (BI), which is directly influenced by the PU, is just 2.92 (Sd=1.55) in Austria. In France the behavioral intention is higher (Mn=4.22, Sd=2.24).

The perceived ease of use (PEOU) scores remarkable high at both trial sides (Austria: Mn=5.94, Sd=1.34; France: Mn=6.25, Sd=0.55). This is probably the result of the rather highly rated Computer self-efficacy (CSE) (Austria: Mn=4.43, Sd=2.57; France: Mn=4.25, Sd=2.45), the remarkable high rated perceptions of external control (PEC) (Austria: Mn=5.69, Sd=2.06; France: Mn=6.13, Sd=1.65) and the high computer playfulness (CPLAY) (Austria: Mn=4.63, Sd=2.25; France: Mn=5.00, Sd=1.69) and perceived enjoyment (ENJ) (Austria: Mn=4.58, Sd=1.59; France: Mn=4.67, Sd=1.24).

In other words, the participants perceived the vAssist system as easy to use because the subjects believe that they are able to perform the specific tasks with the system, the needed resources are given to use the system and they perceive the system as playful and think that it allowing spontaneity and an enjoyable experience.

The average computer anxiety (CANX), i.e. the apprehension or even fear when a person is faced with the possibility of using computers, is for both trial sides rather low (Austria: Mn=1.75, Sd=1.50; France: Mn=2.75, Sd=2.45)

3.2.3 Quality of Life metrics

The Quality of life was measured with the WHOQOL-BREF questionnaire [9] before the field study started and afterwards. Fig. 10 shows the mean values and the confidence intervals (p=0.95) of the participants:

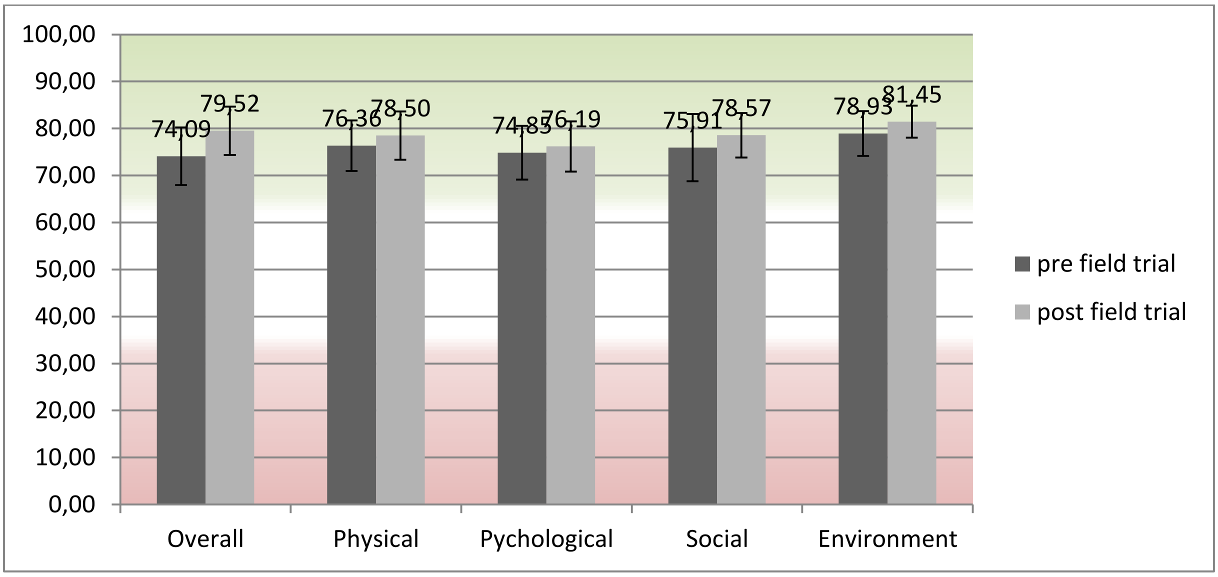


Fig. 10: Quality of Life (WHOQOL-BREF) time comparison [9]

The analysis of the mean value comparison reveals a significant increase of the rated overall Quality of Life (α = 0.025). In detail, the mean value regarding the overall QoL was at a level of 74.09 at first time of measurement (pre field trial) and after the system usage it increased to 79.52 (post field trial). Even though no significant increase for the single dimensions, i.e. the physical, psychological, social and environment-related Quality of Life, has been found, a slight trend similar to the overall score can be observed.

3.3 Qualitative Results

The qualitative results are reported separately for Austria, France and Italy revealing the common and divers issues that emerged during the evaluation of the vAssist services.

The analysis follows the instruction of Mayring [7] regarding inductive category development. We defined the content analytical units:

Coding unit: Clear semantic elements in the text/propositions

Recording unit: All CITs, DRMs, telephone interviews, problem-centered interviews, interviewer protocols and background materials

Context unit: CIT, DRM, telephone interviews, problem-centered interview, interviewer protocol and background material, questionnaires

The analysis follows the question which specific aspects regarding the Quality of the Service and the Quality of Experience can be found. Positive as well as negative aspects are relevant. Hence, this research question serves as the category definition. As the level of abstraction, we define aspects of the system. The categories serve as a basis for the development of main categories. The level of abstraction of the main categories is defined as general influencing factors which could be part of the taxonomy of Wechsung et al. [1] regarding QoS and QoE aspects of multimodal human–machine interaction. If during the development of the categories data with the abstraction level of the main categories is found, this data directly serves as input for the main categories and will not be ignored.

In addition, we want to investigate influencing factors on interaction modalities and general acceptance of multimodal interaction for health systems for older adults as multimodal interaction is a main part of the vAssist system. For this purpose, the same coding, recording and context units are applied. Categories are defined as influencing factors on the interaction modalities, e.g. contextual factors, and the acceptance and attitude of the participants regarding the interaction modalities. The level abstraction is set as for the QoS and QoE aspects.

Furthermore, we conduct quantifications for the categories for the Austrian trial site to offer insights regarding the importance of the categories. We determined the absolute frequencies of the categories, i.e. how often participants refer to a category.

The analysis of the categories allows us to identify system specific positive and negative aspects and further development. The analysis of the main categories serves on the one hand as a theoretical reference and allows on the other hand the comparison of vAssist with other systems or a theoretically perfect working system.

3.3.1 Austria

For the qualitative analysis of the Austrian trial site we coded 364 relevant propositions for the analysis of the system and 186 propositions for the analysis of the interaction modalities. The following sections show the results of the coding procedure according to Mayring [7] and interpretation.

3.3.1.1 Summary of categories and main categories

We identified ***user characteristics*** (24 propositions), such as *data protection and privacy concerns* or general *technology skepticism* but we could also assign propositions which indicate no concerns regarding the usage of systems like vAssist. Furthermore, ***usability*** related categories (104) were formed. For example, the *legibility* of the used text has to be improved *and triggering and closing the dialog* causes problems. Nevertheless, also indications for a *positive ease of use* were found. Regarding the ***utility*** of vAssist (82 propositions)*, several additional services* were suggested. A general *negative utility* was relatively rare compared to the perceived *utility of self-monitoring* and propositions indicating *positive utility* of vAssist in general. In additions, suggestions regarding the increase of the ***usefulness*** (84 propositions), the ***efficiency*** (20 propositions) and the ***interaction quality*** (50 propositions) were made. The following sections report these suggestions, the positive aspects of the system as well the overall satisfaction and general impressions of the system.

3.3.1.2 Overall satisfaction and general impressions

Some of the general impressions and statements of the participants can be rather attributed to ***user characteristics*** than to the system itself. For example, some participants have *privacy and data protection concerns* (5 propositions). These concerns were not mentioned with regard of the vAssist system but technological systems in general. Other participants are characterized by *technology skepticism* (6 propositions) in general. Some subjects also fear the *dependency on technology* (3 propositions) and in particular the risk that a system they rely on might be not available, e.g. because the battery is discharged. In contrast, other subjects explicitly do not have any concerns (10 propositions). They explain that advantages outweigh possible risks or that they do not see any privacy issues for the required data. For the majority, this is on the condition that the system is almost perfectly working.

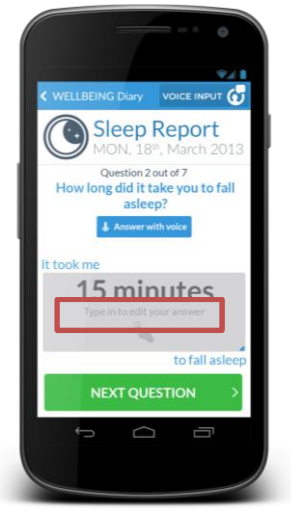
If the ***utility*** of vAssist is questioned (*negative utility*: 10 propositions), foremost the benefits of having a sleep report is not clear to the subjects. Other propositions are related to the interaction quality and just two participants query the utility of vAssist in general. In contrast, eleven propositions can be assigned to a *general positive utility* when participants state that they appreciate the idea of the vAssist system. Moreover, the *utility of self-monitoring* (14 propositions) is highly valued.

Nevertheless, participants also mentioned a desire for *additional services* (19 propositions). First, the graphical reports could be enhanced with additional information and interpretations which support the well-being and health-related behavior. For example, the system could actively suggest behavior changes or fitness activities etc. A calendar including reminders for medical appointments and the need for (re-)purchasing medication could be added. In addition, health data could be automatically added e.g. by connecting sensors. For the reports and diaries it is wished that they could be printed and sent. Further, the system could include a function for emergency calls and regular phone calls, i.e. communication services which are already implemented but not exhaustively tested during the field study. Furthermore, a strong *cooperation with physicians* is perceived as desirable (21 propositions) as the health as well as the well-being data give information about the way of life of patients. Hence, the reports can be shown or could be sent to the physician in charge and, if necessary, she or he could intervene. In addition, it was mentioned that the physician could recommend the vAssist system and this fact would improve the behavioral intend to use the system.

As a drawback of the system it was mentioned that the vAssist system has a too *long processing time* (2 propositions) indicating a negative impact on the ***efficiency***. Furthermore, the *stability and reliability of the speech interaction* (50 propositions), which can be assigned to the ***interaction quality***, was criticized. However, it should be noted that technical issues occurred during the Austrian field trial explaining the number of propositions. Server failures led to temporally not available speech interaction. Especially during the telephone interviews and in the CIT, participants reported these technical problems as, of course, they wanted the problems to be solved as fast as possible to continue testing the vAssist system via speech interaction resulting in an overrepresentation of that category from a quantitative point of view.

An important part of the system's ***usability*** is the ease of use. A rather *negative ease of use* is less prevalent (4 propositions) compared to a *positive ease of use* (8 propositions). Negative connoted statements indicate the need of slight improvement ("ein bisschen verbessern") whereas positive ones emphatically underline the ease of usage via touch and speech interaction. For example, one participant describes the usage of vAssist as being "as easy as teeth brushing" ("so einfach wie Zähneputzen").

3.3.1.3 Problematic aspects of the system



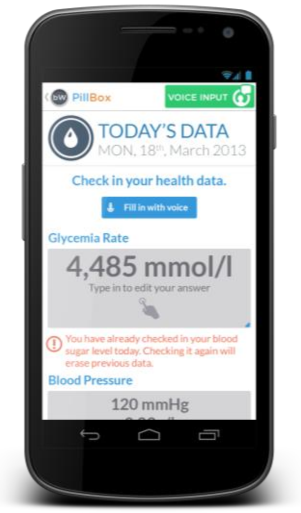
How long did it take you to fall asleep?

Type in to edit your answer

Fig. 11: Example for limited legibility

Participants also suggested improvements to gain a better ***usability***. The most important problem from a quantitative point of view is the *legibility* (19 propositions). The font size and contrast (e. g. grey front on white background) have to be increased (see Fig. 11 for an example) and reversed text, i.e. light font on dark background, should be avoided. This claim also concerns the legends and labels of the graphical reports. Even though the text size should be increased, *scrolling* should be avoided (10 propositions) as scrolling gestures require fine motor skills and the need for scrolling is not recognized by the subjects. It was suggested to display just one question per screen to overcome this problem. Regarding the general *navigation* (10 propositions) within the apps, a back and/or home button is missing. Also, the graphical reports cause problems as it is unclear to some subjects how to navigate.

Another problematic aspect is the *time input via clock* (5 propositions). The input via scrolling a time wheel as well as via the buttons (+/-) is perceived as hard and not likeable. Instead, participants prefer a numeric keyboard to enter the time. As the vAssist system automatically sets an example time to facilitate the understanding of which input is needed, overwriting should be supported. In general, keyboards (13 propositions) should be adjusted to the input. E.g. for numeric input, just numeric keyboards should be used. The regular QWERY keyboards cause problems regarding the legibility of the letters and the so-called fat finger problems as the size of the buttons is too small for older adults due to an age-related wider diameter of their fingers and fine moto impairments. The used *measuring unit* (4 propositions) of the blood pressure is not adaptable and does not take Austrian standards into account and thus, confuses some subjects.



Check in your health data.

Type in to edit your answer

You have already checked in your blood sugar level today. Checking it again will erase previous data.

Fig. 12 Inconsistency with the dialog

Furthermore, some *inconsistencies* (9 propositions) cause irritation (see Fig. 12 for an example). This problem primarily concerns the consistency between the GUI and VUI. For example, if the users want to enter the blood pressure value, the glycaemia rate is displayed on the top of the screen suggesting that this value should be entered despite the fact that system recognized that the users want to enter the blood pressure value.

The *status of the dialog* (15 propositions) also causes problems for some subjects. Some participants were observed to forget to tap the voice input button before they start talking. Others reported that the speech output started without intentionally triggering it. Further, participants were observed

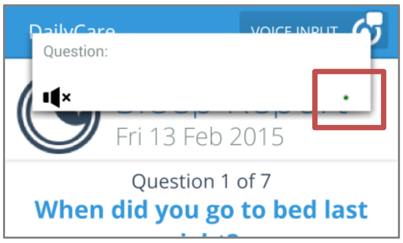
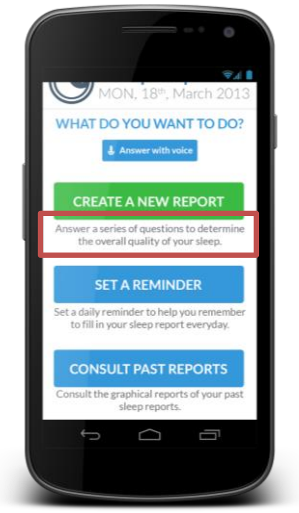


Fig. 13: Dialog of the vAssist system

unintentionally closing the dialog during the interaction because they touched the display. In addition to the triggering and closing the speech interaction problems, it might be unclear, whether the system is connected as the green or red dot which indicates the status of the connection is rather small (see Fig. 13).

Furthermore, the *status of the daily input* (6 propositions) could be improved (see Fig. 14). Participants describe this suggestion as a "detail" or "small thing" ("Kleinigkeit"). Nevertheless, a clear and more visible hint should indicate whether the daily reports were entered on the respective day. For example, a colored exclamation mark could show that a report has to be entered or a colored text could make clear that the specific report has been added.



Answer a series of questions to determine the overall quality of your sleep.

Set a daily reminder to help you remember to fill in your sleep report everyday.

Consult the graphical reports of your past sleep reports.

Fig. 14: Status of the daily input

To support a clear understanding, it is also important to choose the *wording* (9 propositions) of the GUI as well as the VUI carefully. Participants do not want English terms to be used as some of them neither speak English and nor are familiar with English technical terms. Translated terms and questions have to be chosen carefully to meet the correct and intended meaning. In addition, the wording should be consistent.

To support the ***usefulness*** of the system, which relates the ***utility*** of a system to its ***usability*** [1], ten propositions have been found regarding *concrete questions* of the sleep report. Foremost, the questions regarding the time and duration the user falls asleep are perceived as hard to answer and less useful. A 3-point Likert scale labeled with "long", "moderate" and "short" was suggested as users are not able to tell an exact duration. For example, participants explain that they were not looking at a clock the moment they fell asleep. This sarcastic statement indicates an emotional reaction or annoyance caused by this question. In addition, the 5-point Likert scale enhanced with emoticons and colors, which is used to rate the quality of the sleep, should also be replaced with a 3-point scale.

To accomplish full ***usefulness*** of the system, *additional input possibilities* (39 propositions) were suggested: Regarding the sleep report, the reason for waking up and the duration for each time, if multiple wake ups occurred, is required as assumed reasons could give information about physical or psychological causes. Blood pressure and glycaemia values should be entered multiple times per day. In general, the date of the reports should be modifiable to allow entries for passed days.

For the reports, also the missing *possibility to correct* (11 propositions) was mentioned. On the one hand, if a false input was made for the sleep report and the "next" button was pressed, i.e. the user has continued to the next questions on the next screen, there is no possibility to go back and correct the entry. On the other hand, this category refers to speech input. If an input is recognized incorrectly, the questionnaire has to be completed until the question if the input was recognized correctly comes up. In other words, corrections cannot be made immediately.

Furthermore, the *time input* is *limited* (5 propositions). Starting with 100 minutes the system forces the user to choose between hours and minutes as just two-digit input is possible. For example, if somebody's fitness activities took 105 minutes, the person has to round to the next full hour.

Regarding the *reminder* function (6 propositions), it is necessary that the reminders are perceptible on a visual and auditory level. However, this is mainly related to the smartphone and the operation software itself as the vAssist system uses the notifications of the Android operation system. Another category, which is related to the device used is the *battery drain* (18 propositions) affecting the perceived ***efficiency*** of the system. Participants complained about the frequent need to charge the smartphones.

3.3.1.4 Positive aspects of the system

First of all it should be noted, that the Austrian participants see their role rather as critics or critical reviewers of the tested system to allow further development by pointing out the drawbacks and negative aspects of the system. Hence, more negative statements and proposition do not necessarily indicate a low acceptance. Nevertheless we choose to report the quantifications of the coded proposition to give insights into the distributions of the positive aspects. However, the quantities have to be interpreted carefully because of the mentioned self-image of the participants.

As described in 3.3.1.2, several general aspects, such ease the *ease of use* (8 propositions) indicating a good ***usability,*** or the *utility of self-monitoring* (15 propositions), are seen as positive aspects. In particular for the health data (11 propositions), monitoring is appreciated and is described as “helpful” (“hilfreich”), “useful” (“sinnvoll”) and “interesting” (“interessant”), just to name a few statements. Especially the reports are appreciated. Further, a motivational effect of the fitness report (4 propositions) is expected by the participants.

The *graphical reports* (6 propositions) are perceived as “very nice“ (“sehr nett”). For example, the sleep report offers information on quantitative and qualitative aspects, i.e. the sleep quality and sleep durations, number of wake-ups etc. On the one hand, these propositions reveal the perceived ***utility*** of vAssist in general (32 propositions). On the other hand, the positive statements about the *graphical reports* also include references to well-designed ***aesthetics*** (3 propositions).

Furthermore, the *reminder* function is perceived as useful and “important” (“wichtig”) (10 propositions) with little drawbacks, e.g. the visibility of the notifications. Reminders include the notification for medication intake as well as support for remembering to fill in the vAssist diaries. Through the reminders, the cognitive effort of the respective task completion is reduced. Hence, reminders increase the ***usefulness*** of a system.

3.3.1.5 Interaction modalities

This section describes the analysis of influencing factors on interaction modalities and general acceptance of multimodal interaction for health systems for older adults. For this purpose in sum 16 categories which influence the interaction preferences and choices were identified on the basis of 185 relevant propositions collected during the field study.

80 propositions can be assigned to factors which target the ***characteristics of the user*** him-/herself. From a quantitative point of view, the most important one is *(fine) motor impairments* (24 propositions). Foremost, tremors and but also temporal impairments, e.g. after surgeries, and temporal or permanent injuries are mentioned. Also *visual impairments* were identified as a major factor (16 propositions). Almost all participants describe this condition as "poor vision" ("schlechtes Sehvermögen") indicating an extensive and broad understanding of this factor. If the visual or (fine) motor impairments are given, subjects consistently perceive the benefits of speech interaction.

Furthermore, *phonetic characteristics* are of great importance (13 propositions). Phonetics include dialect, differences in word stress, speech impediment, and temporal or permanent diseases, which also affect the voice, e.g. having a cold. In addition, potential differences between male and female voices were mentioned. On the one hand, these conditions and characteristics are described as a barrier for speech interaction. On the other hand, the subjects expressed the requirement of overcoming all these barriers, especially regarding cultural and regional specific terms and dialects. In contrast, if a person suffers from *hearing impairments* (3 propositions) including age related hearing problems, speech interaction is seen as not appropriate.

For *cognitive impairments* (3 propositions), one person stated that speech interaction would preferable compared to touch interaction. In contrast, two propositions were found indicating that speech interaction would be too complex for users with cognitive impairments. This conclusion is mainly related to the cognitive effort which is required to learn how to use new technologies. In general, previous *technology experience*, i.e. if the person is familiar with technological systems in general and speech interaction in particular, can be identified as another user characteristics related factor (8 propositions). Subjects describe speech interaction as a modality with a need of getting used to ("gewöhnungsbedürftig") or explain their general preference for speech interaction with the fact that they are using technological systems for a long time. It can be concluded that previous experiences dismantle barriers for speech interaction even more than for touch interaction as no propositions have been found related to touch input. However, once the first experiences have been gained, a major barrier has been overcome.

Another important factor for the general and specific decision on interaction modalities are *privacy concerns* (9 propositions). As an example, subjects often refer to public transportation where strangers could hear private information, i.e. private input and output. This concern is especially relevant when sensitive data like health or wellbeing data have to be entered. (Not) Using speech interaction when others are present is also related to another factor, which we call *respect* (3 propositions). Subjects hesitate to use speech interaction when speech input and output might "annoy" ("stören") others. This concern can be combined with *privacy concerns* but we also identified it as an independent factor existing within subjects who do not refer to privacy concerns. The argumentation for this factor is on a moral level: Subjects feel annoyed by others when they are e.g. on the phone in public. Hence, they would not use speech interaction. *Respect* can be seen as a *user characteristic*. But as it is strongly related to external circumstances, the content level of the factor as well as the subjects' argumentation for it, it can also be part of the user's current ***interaction context***.

In sum, 30 propositions can be attributed to the major factor *context*. The most prevalent category is *environmental noise* (12 propositions) and further underlining the drawbacks of speech interaction in public space or outside. This factor is also related to the *interaction quality* as environmental noise could impact the recognition quality. Other people talking and wind gusts are seen as the major problems in public space. In addition, participants refer to media with audio output like the radio or TV which also might cause false recognitions. Interestingly, the participants' explanations of this factor just refer to false recognitions but unintelligible output was never mentioned. A possible explanation of this observation is that the vAssist system offers visual support for the required input/its output which prevents misunderstandings.

However, *mobility* (6 propositions) and the need to interact *hands free* (4 propositions) are factors which influence the subjects in favor for speech interaction. On the one hand, speech interaction is seen as less distracting, e.g. while driving a car, allows being mobile and using the hands for other activities. On the other hand, going by bus or by car can be shaky or a person might wear gloves which cause problems interacting via touch. Another factor which is an argument for speech interaction is the *lightning condition* (2 propositions). In addition to *visual impairments*, for example bright sunlight might impact the visual perception of the screen which is needed for touch interaction. In contrast, subjects see the need for *connectivity* (3 propositions) stronger related to speech interaction compared to touch interaction. In detail, they refer to internet access and telephone networks. A likely explanation of this train of thoughts is that vAssist's speech interaction did not work without being connected to the internet but the interaction via touch was not visibly limited. Like *environmental noise*, *connectivity* is strongly related to the factor *interaction quality* (39 propositions) which is the most prevalent factor within the main category ***perceived characteristics of the system*** (78 propositions). *Interaction quality* includes the perceived stability and reliability of a system and is especially relevant for health services as subjects perceive the risk of false input as highly dangerous. For example, if a subject uses the medication reminder but the dosage of a prescription was recognized falsely, a serious health risk might be the consequence. Nevertheless, almost all coded propositions indicate that if a good interaction quality is given, the subjects prefer speech interaction but the trust in this interaction modality is limited. Hence, feedback on the recognized input and requests for confirmation are highly important. Furthermore, our results reveal that an instable system causes negative emotions, e.g. distress and annoyance.

Another important factor regarding the interaction modalities is the perceived *efficiency* (17 propositions). The most important resource which influences this factor is the temporal demand. Occasionally, efficiency is also related to effectivity (3 propositions). In other words, subjects take the needed resources, e.g. time, to successfully accomplish a task into account but focus on the accuracy and completeness of the goal achievement. Nevertheless, subjects perceive the efficiency of speech and touch interaction differently. About 30% of the propositions refer to speech interaction as more efficient. In contrast, more than half of the propositions which are assigned to the factor *ease of use* (in sum 16 propositions) are in favor of speech interaction. Subjects perceive speech interaction in general but also the absence of the touch keyboard as easier. Further, *joy of use* (1 proposition) can also influence the interaction modality choice. The identified proposition during our field study attributes joy of use to speech interaction.

Moreover, speech interaction can also enable a *para-social relationship* (5 propositions) which is related to the system personality and appeal and therefore also assigned to the *perceived characteristics of the system.* For example older and socially less included subjects can benefit from “talking to someone”, but the system as a “unknown being” (“unbekanntes Wesen”) causes also a need for getting used to it and subjects with a negative attitude towards technology might feel uncomfortable talking to a technological system. Nevertheless, no references between a para-social relationship and privacy concerns have been found.

To summarize, different factors which refer to the *characteristic of the user*, the *interaction context* and the *perceived characteristics of the system* were identified as major influencing factors on interaction modalities. Taking into account that several major factors are dynamic, the benefits of multimodal interaction like realized for the vAssist system are strongly given.

3.3.2 France

The following sections show the French results of the qualitative content analysis following the instructions of Mayring [7] and the respective interpretations.

3.3.2.1 Summary of categories and main categories

The main results of the French qualitative analysis show, that the perceived utility of the system (25 propositions) is dependent on the characteristics of the users (26 propositions). As a consequence, positive (9 propositions) as well as negative statements (11 propositions) regarding the utility of vAssist have been found. Further, the used device (14 propositions) negatively influenced the usability (31 propositions) as well as the general acceptance of the system (4 propositions). Moreover, the interaction quality (32 propositions) has to be improved.

3.3.2.2 Overall satisfaction and general impressions

Overall, the participants’ satisfaction is quite mixed: Interacting via touch, most of participants describe the vAssist apps as easy to use (7 propositions), quite clear and efficient but with speech interaction, vAssist services are perceived as unfinished and immature for a market release (16 propositions related to “stability and reliability”). Consequently, the main feature of vAssist has not convinced them as touch interaction, even if it provides a good user experience, is not perceived as innovative enough. In detail, comparing the vAssist speech recognition with Google voice recognition or Siri, the vAssist voice interaction does not offer any benefits (4 propositions).

Another aspect of trial has emerged from analysis as an important factor influencing the acceptance. A lot of proposition (14) refers to the smartphone characteristics (acer liquid, provided for trial). Some issues with speaker (4 propositions), calling behavior (2 propositions), battery draining (3 propositions) or screen size (4 propositions) have affected the perception of vAssist system without being related to vAssist system itself. In addition, an experimental bias has emerged despite efforts to avoid it. All participants kept their personal phone in addition of vAssist phone. The selected device offered a double sim feature, allowing them to use both numbers (personal and vAssist) on one phone. But all participants worried about learning of a new phone; so they kept their personal phone (feature phone or smartphone). This aspect of the trial leads to an unnatural use of vAssist phone: it was only used for vAssist's tasks (6 participants over 9). The others participants have taken the opportunity to "play" with a new smartphone, which is a more natural use.

From an experimental point of view, trial experience was described as too long or repetitive by several participants (22 propositions). The main issue was the lack of relevance of specific tasks by comparison with individual needs. For example the utility of the daily sleep report was perceived as low as participants have a good sleep quality (5 propositions).

3.3.2.3 Problematic aspects of the system

Main issue concerns the instability of speech interaction as several server failures occurred during the field trial (16 propositions). Sometimes, participants reported a lack of efficiency as the stated that the system is too slow for providing a good user experience (5 propositions). When speech interaction is more fluent, it was generally difficult to finish a dialog without issues as some bugs occurred (5 propositions). Two semantic categories are particularly concerned: recognition of numbers and report validation commands. Hence, also a not sufficient interaction quality was identified.

Regarding the legibility and thus usability, the screen size was perceived as too small for some text elements (6 propositions), but it was mainly related to the provided smartphone (4 inch). This result provides useful information about hardware recommendation. Further, some minor inconsistencies were highlighted (5 propositions). However, all of them are easily editable and did not affect overall perceived usability during the field trial.

The perceived utility of the vAssist was limited for some participants (11 propositions). In detail, persons without sleep issues have judged the task as very repetitive and useless (4 propositions). Nevertheless, others participants have described this feature very useful, allowing them to improve their sleep quality (4 propositions). Considering all propositions, the relevance of vAssist system is closely related to individual needs and the importance of a modular set of services is completely justified.

Most of problematic aspects of the Ambulatory Terminal concerned wearing issues (4 propositions). The clip of the device was not strong enough for keeping the AT well attached on the belt. In addition, 3 propositions are related to the size of the device: it is perceived as too big by comparison with similar devices. Two participants mentioned another device they tested during a study and described it as smaller and having a better clip.

Notification issues (3 propositions) were reported in both ways. One participant got a notification without using the AT, but two participants reported the lack of notification while wearing the device. Two participants also stated that they are not able to distinguish the plugs on the device. In other words, it is not clear which plug they have to use to charge the device and which one has to be used for the pulse rate sensor. Thus, the relevance of a wireless charging feature is confirmed.

3.3.2.4 Positive aspects of the system

As described previously, the usage of vAssist app via touch interaction has been perceived as quite easy and clear revealing a good ease of use and usability (5 propositions).

As described in section 3.3.2.3, the perceived utility is related to the user characteristics. If the person's characteristics entail a subjective need, e.g. a person suffers from sleeping problems, the subjective relevance of the services is strongly given and participants recognize the utility of the support vAssist is offering leading to a strong behavioral intention regarding future use of the system. As a consequence, participants appreciate the modular set of services is valued by the participants.

Even though several participants highlighted restrictions of the provided smartphones, after a first phase of familiarization the selected device seem more interesting than feature phones and participants were able use the smartphones without greater problems (5 propositions). Only two devices appeared as more efficient than French vAssist phone: IPhone (high-end class) and a Samsung Galaxy SII (former high-end class).

3.3.2.5 Interaction modalities

Regarding interaction modalities, user characteristics related factors have not been identified in France. Despite the little number of participants and heterogeneity of some factors (e.g. gender), modality choices appeared as highly individual and subjective: Age, gender or technology experience related factors were not found. However, phonetic characteristics were identified as an exception. While most of participants speak French without a significant accent, one of them was Luxembourgish with a little accent. She experienced speech interaction as difficult as a lot of misunderstandings and false recognitions occurred. We could a very poor efficacy of speech interaction as a much higher amount of commands were necessary. Further, the participant task completion often failed indication the influence on the interaction quality and effectivity.

The relationship between small screen size and small font was obvious during the field study. Most of participants have experienced issues to read smaller fonts (6 propositions). Their recommendations concerned bigger screen size and bigger fonts. The "speak" button for example (crucial feature of a speech recognition system) was not visible enough for some participants (3 propositions). Hence, the GUI is an important factor for speech interaction if users have to trigger the dialog via touch interaction.

From a general point of view, we could identify three kinds of users:

1) 1. "Slave of Habits" (3 participants):

They used to touch interaction and did not perceive interest of speech interaction. Even if they experienced a more stable interaction with another system (Google Now, Siri) they did not like it. This point of view could also influence the general usage of digital services, independently from interaction modality. For example, if they used to write their appointments in a paper calendar, a digital calendar is not relevant for them.

2) “Speech is Magic” (3 participants)

They have a great experience with speech recognition, even if problems with the vAssist solution occurred. They like text dictation (E-mails, SMS) and interaction with a device for example when their hands are occupied. Despite their speech interaction acceptance, they have high expectation regarding multimodal interaction and the respective systems. They are also skeptical about hypothesis of a full speech interaction system. For example, questions regarding back-up solutions in case of the loss of voice arose. For them, both modalities are still complementary.

3) “Ok when it works” group (3 participants):

These participants like the concept of speech interaction, but are not convinced by existing solutions. They think that speech interaction will be more satisfying in several years and indicate that they will observe the technological progress in this area.

3.3.3 Additional input from Italy

The Italian setup of the field study was reduced in order to avoid overwhelming the participants. Nevertheless, the involved persons expressed their opinion on the vAssist system during the field phase. These statements are valuable, serve as additional input on the vAssist system and should not be ignored. Hence, the following section reports these results even though the data basis was not collected in a strict methodological guided way. The structure of this section is oriented towards the sections 3.3.1 and 3.3.2. However, it should be noted, that categories were not formed according to Mayring [7] but give a first impression.

3.3.3.1 Summary of categories and main categories

In Italy, the most important observations include the following areas:

- User friendliness of the system
- Acceptance
- Medical visits frequency
- Familiarity with computing devices
- Sight
- Hearing
- Accent / dialect
- Motorial upper limbs skills

3.3.3.2 Overall satisfaction and general impressions

In Italy, the field trial was held in collaboration with the ASL4 of Prato. The medical staff of the ASL was highly interested in participating in the project because they expect the prototype to offer a new and original way to monitor and to assist patients, shortening distances and maintaining a constant and frequent contact with the primary users.

After some initial concerns all patients judged the interaction with vAssist as fun and user friendly and for some of them, the number of visits by their healthcare staff has decreased. However, patients affected by visual, auditory and motor disorders, unfamiliar with computing devices or with a heavy accent had experienced greater difficulties with interacting with the vAssist system.

3.3.3.3 Problematic aspects of the system

During the first period of the usage of the vAssist system, difficulties mainly related to unfamiliarity with the devices and with the "language" used occurred: Using touch screen devices and/or being confronted with words such as “workflows”, “templates”, “primary user” and “secondary user” did not result in immediate understanding.

For some patients, factors, such as bad hearing, bad sight, fine motor problems, heavy accents, have further complicated the system interaction.

Both, the medical staff and the patients, however, were constantly assisted by the I&S consultants and their doubts were promptly cleared.

3.3.3.4 Positive aspects of the system

After overcoming these initial difficulties, the vAssist applications was perceived as clearer, faster and more comprehensible and usable. As a consequence, the users described the vAssist system as very interesting and indicated their intention to start a process of home care through vAssist.

In some cases, patients, who had a rather high frequency of doctor visits, were able to decrease the number of appointments with their healthcare staff by constantly using vAssist.

Furthermore, the interaction with the vAssist system caused positive emotions, such as fun and pleasure.

3.4 Evaluation of the Service Delivery Models

Regarding business aspects of vAssist system in the participating countries, 25 primary users (Austria: 19; France: 6) and 5 secondary users (Italy) were questioned.

Table 3 sums up business related aspects per country. Fig. 15 shows the optimal price band of vAssist.

Table 3: Summary of the business model aspects per country

	Austria **Primary users**	**France** **Primary users**	**Italy** **Secondary users**
Participants	19	8	5
Usage of system similar to vAssist	1	1	0
Ownership of device needed for vAssist System	6	5	2
How important are costs for the usage of vAssist system	3.58 (5=very important)	3.75 (5=very important)	3.4 (5=very important)
Preference of using the own hardware	12	4	5

Analyzing the questionnaires, the following conclusions can be drawn:

Future users of vAssist

When it comes to the expected future users of the vAssist system, there are differences between Austria and France. While in Austria older adults who use the services on their own (14/19), persons with high affinity for new technologies (14/19) and seniors with marginal need for support (13/19) are seen as the future end users, in France four categories were mentioned by all participants (6/6): persons in need of care who use the system together with other persons, employees of institutional care providers, persons with physical impairments and persons with high affinity for new technologies. Two other user groups were identified in addition: family and medical stuff.

17/19 of the Austrian participants and all of French participants see themselves as part of the group of active seniors who use the services on their own. Ten of the Austrian participants described themselves as technology affine. In Italy, secondary users (5) describe themselves as part of the target group of employees of institutional care. In addition, active older adults (3/5), seniors with low need for support (4/5) and people with physical disabilities (4/5) were mentioned as potential users.

Only two of the Austrian and French participants are already using a system similar to vAssist: One person the iPhone blood pressure application and the other person an application similar to an online laboratory. No one of the Italian participants uses a system that is similar to the vAssist system.

Devices that are needed for the vAssist system are owned by 6/19 persons in Austria, while in France nearly all participants are equipped with that kind of devices (5/6). All different kinds of brands for smartphones are owned. Hence, no significant higher number for one model can be identified. Most of the participants in Austria and France would prefer to use their own hardware, only a few ones would like to rent it via the vAssist platform. The Italian participants stated that they would like to use their own hardware to avoid additional costs.

Financial aspects

In general, financial aspects are a very important factor for the decision whether the vAssist will be used or not. The average value regarding the importance of the costs is 3.7 (1=not important, 5=very important) for the primary users (Austria and France) and 3.4 for the secondary users (Italy)

The preferred payment model for all participants in France is “try and buy”: vAssist services would be delivered for a period of six month without any fee. Afterwards the single payment model should be offered. In Austria, 11/19 also would choose the “try and buy” in combinations with the pay per use (8/11) or flat rate model (6/11). In contrast, 7/19 persons are more interested in a flat rate without a free test period. Likewise, the primary users preferred the payment model “try and buy” (3/5). Two persons would prefer a flat rate.

Concerning the costs in detail, people were asked at what monthly rate the vAssist system is priced so low that the price would cause questioning the quality of vAssist. In Austria the average value is 13.22€ while in France it is a lot lower (3.5€). The service would be perceives as a bargain at 10.55€ in Austria and at 12.5€ in France. It gets expensive for Austrian users at 31.66 € and for French users at 21.6€. The system is too expensive for both groups at an average value of about 40€. The secondary users start from much lower costs: on average, too cheap would be at 3.10€, a monthly rate of 4.2€ would be perceives as a bargain, vAssist would get expensive at 10.60€ and it would be absolutely too expensive at 16.40€ per month. Resulting from this data basis, Fig. 15 shows the optimal prizing for the vAssist system.

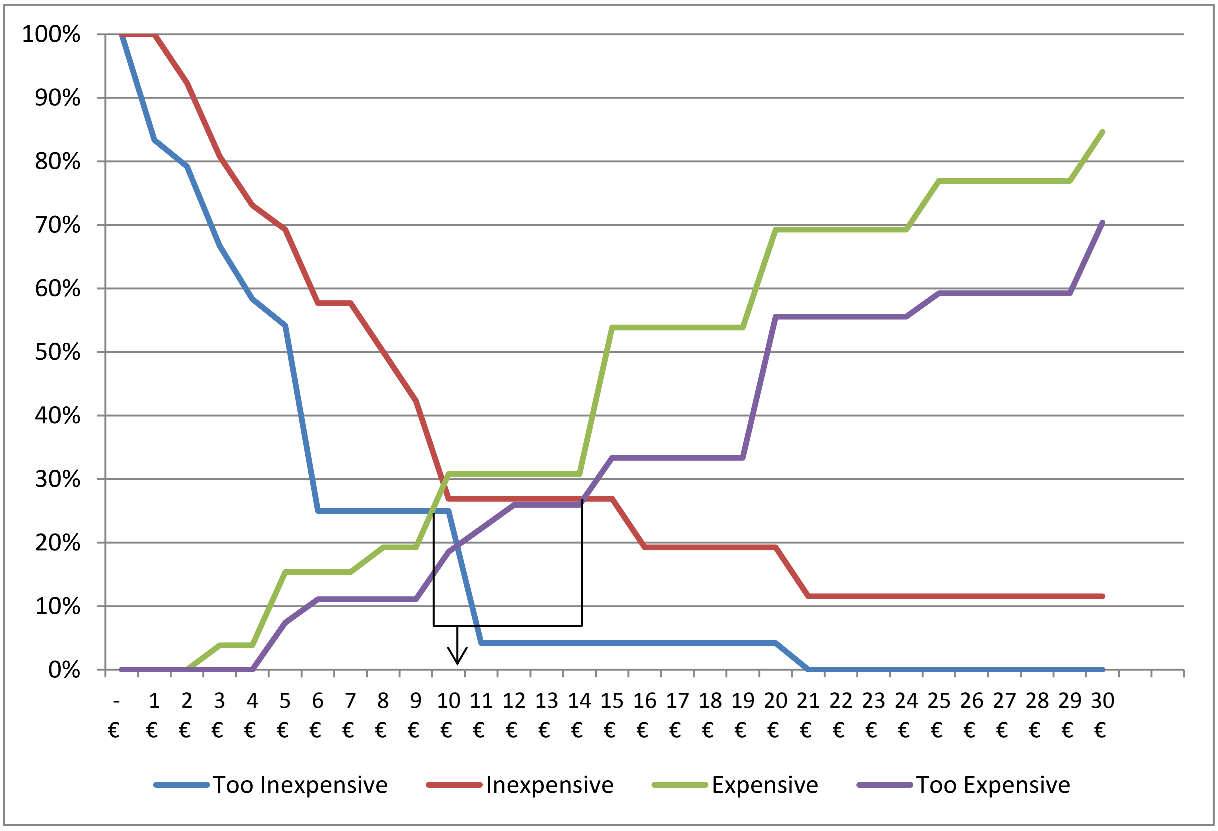


Fig. 15: Van Westendorp pricing model of vAssist

The range of an acceptable prizes starts at 9€ (Point of Marginal Cheapness) and ends with 14€ (Point of Marginal Expensiveness) per month. The Optimal Price Point is located at 10€ per month.

3.5 Conclusion and implications for overall system design

Conclusion and design implications are drawn from user statements and preferences which were gathered during the field evaluation of the system prototype.

Concerns regarding new technologies and dependency on such systems have to be taken seriously and underline the relevance of careful and exhaustive fulfilment of the demand regarding system related information. On the one hand, transparency regarding data storage and access has to be provided. On the other hand, more general aspects, like the benefits of technologically assisting services and the fact that in contrast to potential fears AAL technologies do not have the purpose to replace personal relationships, social inclusion etc., should be mentioned when suggesting the usage of such a system.

Furthermore, several suggestions regarding the improvement of the usability, utility and usefulness can be concluded: Even though the used font size should be increased to enable a better legibility, scrolling should be avoided as it might overstrain the fine motor skills of older adults and the need for scrolling might not be recognized. Hence, it was suggested to display just one question per screen to overcome this problem. Further, a back button should be added. For the time input, a numeric keyboard and overwriting should be offered and QWERTY keyboards should be avoided if possible. To ensure that the users do not forget the daily input, reminders have to be more intrusive and visual cues should be added to the GUI. Concerning the PillBox application, the measuring units, e.g. the blood pressure unit, should take local standards into account. To improve the DailyCare application, the duration of the wake ups should be replaced with a 3-point-Likert scale (“long” – “moderate” – “short”) as participants are not able to provide exact durations in minutes. In contrast, is should be possible to add an assumed reason/cause for waking up and in case of multiple wake ups, the durations should be reported separately. The input of minutes should be extended to three digits. Further, questions should be asked in the same order and with the same wording whether using touch or speech interaction. To support users to avoid closing the speech dialog unintentionally, a “close the dialog” button could be implemented. In contrast, closing the dialog by tapping somewhere on the screen should be disabled.

In addition, the interaction quality has to be improved by ensuring the stability and reliability of the speech interaction. The lack of the perceived interaction quality can be explained with server failures during the field study which led to temporally not available speech interaction and respective coded reports by the participants. Even though the quantity of the reports can be explained with the nature of a field study (reporting server failures to the contact persons), also the SASSI ratings of the response accuracy underline the importance of a stable as well as reliable working system. In addition, response time by the system and recognition of numbers have to be improved.

Even though participants want do purchase the needed hardware on their own, smartphone characteristics are identified as an important factor regarding the future acceptance of the system. For example, battery draining or screen size of the used smartphones affected the acceptance in a negative way. Hence, hardware recommendations should be provided to ensure that the used devices meet the requirements of the users.

Regarding the Ambulatory Terminal (AT), minor but important implications can be concluded. The clip of the device keeping the AT attached to the user’s belt has to be replaced with a stronger clip to ensure that users do not lose the device. In addition, the device is perceived as too big for everyday use.

In general, participants appreciate the utility of the vAssist system. In detail, the utility of self-monitoring including the (graphical) reports and the motivational effect of the fitness report are highly valued. The perceived usefulness of the system is rated relatively low compared to other acceptance factors according to the TAM3 analysis. However, the qualitative analysis reveals that this limitation is attributed to the desire for minor additional input possibilities and additional services and that the perceived relevance of the Assist system is closely related to individual needs justifying a modular set of services. Taking into account that the perceived usefulness links the perceived utility to usability aspects, the perceived usefulness can be improved by implementing the before mentioned suggestions, such as improving legibility and giving control over closing a speech dialog.

The quantitative analysis also reveals a low task load, i.e. mental, physical and temporal demand, effort and frustration, caused by the system interaction. In line, the ease of use is rated as good according to the TAM3 results and the UEQ analysis shows a highly rated perspicuity. Further, participants perceive the system as rather playful, allowing spontaneity and an enjoyable experience.

Moreover, the participants’ overall Quality of Life has increased after several weeks of system usage suggesting that the vAssist system could help to improve the users’ Quality of Life.

4 Summary

The D4.3 Field Trial Evaluation Report presents the results of the field evaluation conducted in Austria, France and Italy. The overall goal of the evaluations in the field was to ensure that issues related to quality of experience and quality of service factors are detected, to investigate the influence of the vAssist system on the users' life, e.g. their quality of life, and to evaluate the business model. Therefore, the taxonomy by Wechsung et al. [11] was operationalized and a mixed-method approach was applied.

Overall, the results reveal vAssist's potential in supporting the health monitoring and well-being of older adults. In detail, the utility of the system is recognized and indications for the increase of the participants' Quality of Life have been found. The ease of use and the perspicuity were rated as good and the system demands only low mental, physical and temporal resources. Participants perceived the system as rather playful, allowing spontaneity and an enjoyable experience. The analysis further revealed that limitations of the perceived usefulness are linked to the desire for minor additional input possibilities and additional services. Offering these additional services, a broad behavioural intention for the usage of the system can be anticipated. For that propose a “try and buy” in combination with either a flat rate fee or a charging per use payment model should be offered. Monthly costs should not extend 14€ and die optimal price is 10€ per month for the vAssist service excluding the required hardware. The monthly costs should not fall below 9€ as such a cheap prize would negatively influence the perception of the vAssist system. Further, as the perceived relevance of the Assist system is closely related to individual needs justifying a modular set of services.

5 References

[1] Bradley Margaret M., and Peter J. Lang. "Measuring emotion: the self-assessment manikin and the semantic differential." *Journal of behavior therapy and experimental psychiatry* 25.1 (1994): 49-59.

[2] Flanagan, J.C. (1954). The Critical Incident Technique. Psychological bulletin, 51(4), 327-358.

[3] Hart, S. G. & Staveland, L. E. (1988) Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati (Eds.) Human Mental Workload. Amsterdam: North Holland Press.

[4] Hone, Kate S., and Robert Graham. "Towards a tool for the subjective assessment of speech system interfaces (SASSI)." *Natural Language Engineering* 6.3&4 (2000): 287-303

[5] Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A survey method for characterizing daily life experience: The day reconstruction method. Science, 1776, 1776–1780

[6] Laugwitz, B.; Held, T. & Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. In: Holzinger, A. (Ed.): USAB 2008, LNCS 5298, pp. 63-76.

[7] Mayring, P. (2014). Qualitative Content Analysis. Theoretical Foundation, Basic Procedures and Software Solution. Klagenfurt, Austria. (free download via Social Science Open Access Repository SSOAR, URN: http://nbn-resolving.de/urn:nbn:de:0168-ssoar-395173).

[8] Scheibelhofer, E. (2008). Combining Narration-Based Interviews with Topical Interviews: Methodological Reflections on Research Practices. Int. J. Social Research Methodology. 11(5), 403-416.

[9] The WHOQOL Group. Development of the World Health Organization WHOQOL-BREF Quality of Life Assessment. Psychol Med. 1998;28:551–8.

[10] Venkatesh, V. and Bala, H. (2008), Technology acceptance model 3 and a research agenda on interventions. *Decision Science* 39(2), 273–315.

[11] Wechsung, Ina, et al. (2012)"Measuring the Quality of Service and Quality of Experience of multimodal human–machine interaction." *Journal on Multimodal User Interfaces* 6.1-2: 73-85.

Annex A: Template for the Content Analysis

Coding the text:

Case	Source	Reference (PillBox, DailyCare, DailyCom, vAssist in general)	Paraphrase (translated to English)	Reduction (category)

Forming main categories:

Case	Source	Reference (PillBox, DailyCare, DailyCom, vAssist in general)	Category	Abstraction and Reduction (main category)