WeldVUI: Establishing Speech-Based Interfaces in Industrial Applications

Mirjam Augstein, Thomas Neumayr, Sebastian Pimminger

WeldVUI: Establishing Speech-Based Interfaces in Industrial Applications

Mirjam Augstein[1 [0000−0002−7901−3765]], Thomas Neumayr[1 [0000−0003−3607−8873]], and Sebastian Pimminger[1]

University of Applied Sciences Upper Austria, Hagenberg, Austria
{mirjam.augstein, thomas.neumayr, sebastian.pimminger}@fh-hagenberg.at

Abstract. Voice User Interfaces (VUIs) and speech-based applications have recently gained increasing popularity. During the past years, they have been included in a wide range of mass-market devices (smart phones or technology installed in common car cockpits) and are thus available for many everyday interaction scenarios (e.g., making phone calls or switching the lights on and off). This popularity also led to a number of guidelines for VUI design, software libraries and devices for speech recognition becoming available for interface designers and developers. Although generally helpful, these resources are often broad and do not fully satisfy the specific requirements of certain industrial applications. First, grammar and vocabulary in such settings usually differ drastically from everyday scenarios. Second, common software libraries and devices are often not able to comply with the conditions in industrial environments (e.g. involving high levels of noise). This paper describes the iterative, user-centered design process for VUIs and functional speech-based interaction prototypes for the domain of industrial welding, including a two-stage Wizard of Oz procedure, rapid prototyping, speech recognition improvement and thorough user involvement. Our experiences throughout this process generalize to other industrial applications and so-called "niche applications" where grammar and vocabulary usually have to be established from scratch. They are intended to guide other researchers setting up a similar process for designing and prototyping domain-specific VUIs.

Keywords: Voice User Interface Design · User-Centered Design · Interaction Design · Speech-Based Interfaces · Industrial Applications.

1 Introduction

According to Cohen et al. [4, p. 5], a Voice User Interface (VUI) is "what a person interacts with when communicating with a spoken language application". Its elements include prompts ("all the recordings or synthesized speech played to the user during the dialog"), grammars (definition of "the possible things callers can say in response to each prompt") and dialog logic (the "actions taken by the system", e.g., "responding to what a caller has just said or reading out information retrieved from a database"). Cohen et al. further identify the following

advantages of speech systems compared with other access modes. They are *intuitive and efficient*, *ubiquitous*, *enjoyable*, and *hands-free, eyes-free*. Especially the latter is decisive in use cases or whole application domains that provoke functional impairment for the users. This is, for instance, the case during driving when the driver's hands should remain on the steering wheel and visual attention should be paid to what's happening on the road. Further, speech is generally considered to be intuitive and natural; already in 1960, John Licklider envisioned speech-based interaction as the "most natural means" of communication [17].

Nowadays, speech-based interaction has become omnipresent (e.g., in car cockpits, on smart phones and via smart speakers for home control). Most of the popular services like Amazon Alexa, Google Assistant, Apple Siri or Microsoft Cortana are cloud-based. They involve a wide spectrum of functionalities and a huge vocabulary, and many even offer a set of design guidelines. These systems often have in common that they allow for a *general, not further specified or restricted dialog about everyday things and services* (such as the weather or making phone calls). This makes them usable for a wide range of contexts and applications. Yet, it might also make them not very efficient if only a small, well-defined set of functionalities is needed. This is often the case in narrow, domain-specific scenarios such as certain industrial or other so-called "niche applications". Niche applications benefit from a VUI restricted to the most important functionalities needed in their context because i) the respectively small vocabulary can be memorized more easily, especially if in everyday use, and ii) they can be implemented for offline use more easily (which is often impossible for VUIs that rely on popular and high-performing cloud services). Both also enable short response times which contributes to the characteristics related to *ubiquitous use* and *efficiency* named in [4]. In contrast to VUIs for routine tasks, there are no gold standard VUI designs for most niche applications including those in the industrial domain. Thus, grammar, vocabulary and dialog structures must often be created from scratch and designed for the concrete needs of the particular target group.

This paper describes VUI design and application of speech-based interaction for industrial welding. Manual welding is a highly accurate process that requires the welder's hands to direct the welding torch or add welding rod. This leads to functional impairments during the process: neither can the hands be used for interaction with the machine (e.g., to change parameters), nor should the eyes be taken off the welding arc. There are numerous situations in which it would be beneficial regarding time and quality of the result if parameters could be changed on-the-fly, a *hands-free and eyes-free* [4] interaction method is required. The domain further involves additional restrictions: new solutions must be *low-cost*, *small-size*, *light-weight* and provide a *good User Experience (UX)* to be *integrable in the standard welding equipment* and *accepted by the market*.

We explain our methodology based on i) a two-stage Wizard-of-Oz (WoZ) test with welders including observation, open questions and standardized instruments (User Experience Questionnaire (UEQ) [16] and NASA TLX [9]), ii) field tests with the resulting functional speech interaction prototypes, and iii) systematic automated speech recognition tests and improvement. Further, we provide

an overview of related work on VUI design including (industrial) niche applications and evaluation of VUIs, and an introduction to the industrial welder's work environment. The paper's main contribution lies in the detailed description of the applied methodology. It is intended as a methodological guide for other researchers who face the challenge of conceptualizing and implementing VUIs and speech-based interaction solutions for industrial and other niche applications. Although the methods themselves (such as WoZ) are not novel, we applied them for problems where there exists limited reported experience. E.g., WoZ is commonly used in early prototyping phases but, to the best of our knowledge, it has not been applied for the elicitation of grammar and vocabulary for VUIs before.

2 Related Work

Here we describe related work on the design of VUIs in general and in specific niche applications, and on the evaluation of speech-based interactive systems.

2.1 Speech-Based Interaction and VUI Design

Turunen [31] distinguishes eight layers related to speech processing categorized into three groups: *acoustic layer, articulary layer and phonemic layer* (first group), *lexical layer and syntactic layer* (second group), and *semantic layer, pragmatic layer and discourse layer* (third group). Further, Turunen (based on [29]) summarizes the first two groups as "core speech technologies" and the second and third group as "speech applications". Our work focuses on the top layers in the third group as we mainly discuss VUI design and for the implementation itself rely on existing voice recognition technology (although we also report experiences with the optimization of voice recognition). Further, Turunen identifies several properties of speech recognition systems in the following categories: i) *vocabulary and language*, ii) *communication style* and iii) *usage conditions*. Our system can be described as follows, based on these categories: *vocabulary size* is *comparatively small* and uses *fixed phrases*. The *communication style* is *speaker-independent*[1] and the *speaking style* is mostly *discrete*. *Usage conditions* are *hostile* and *channel quality* might be *low*, due to the environmental conditions.

Cohen et al. [4] define five key principles for VUI design methodology that should especially help in projects with real-world constraints: i) *end-user input* ("inform design decisions with end-user input"), ii) *integrated business and user needs* ("find solutions that combine business goals and user goals"), iii) *thorough early work* ("avoid expensive downstream changes by focusing on thorough work in the early definition and design stages"), iv) *conversational design* ("move the design experience close to the user experience so that the designer can experience design elements in their appropriate conversational context"), and v) *context* ("make all design decisions with appropriate consideration of context"). These principles have been a basis for our VUI design as we thoroughly involved end

[1] This might be up to changes in a future version closer to a commercial product.

users but also other stakeholders and invested a high amount of time and effort in early design phases. Further we did exhaustive usage context research (using the Contextual Design methodology [12, 2]) prior to VUI design (see [1]).

An overview of specific requirements for VUIs is provided by Farinazzo et al. [8] based on [6, 7]. They argue that while most requirements that are common for graphical user interfaces apply also for VUIs (e.g., usability and feedback), there are additional criteria, especially related to the *transient* attribute of voice (in contrast to graphical interfaces which are *persistent*). This is also pointed out by Schnelle and Lyardet [30] who e.g., state that "speech is one-dimensional" (while the eye is active, the ear is passive and cannot browse a set of recordings the way eyes can scan a screen of text and figures), "speech is transient" (listening is controlled by the short term memory; speech is not ideal for delivering large amounts of data), "speech is invisible" (thus it is difficult to indicate to the users what actions they may perform), and "speech is asymmetric" (people can speak faster than they type but listen more slowly than they read).

Farinazzo et al. identify three categories of requirements: i) non-functional ones related to the *representation of the information*, ii) requirements related to *data input*, and iii) *technical issues*. Requirements in the first category "indicate the format that the interaction must assume in order to enable the system to deal with user inputs" and involve *consistency*, *feedback*, *support for all classes of users*, *minimization of the cognitive effort the user has to do in order to perform the tasks* and the *correctness, relevance and informativeness of system outputs*. Requirements in the second category include *appropriate recognition naturalness of speech and interaction*, *help mechanisms when the user is in difficult situations*, *error prevention* and *quick correction of inputs*. Finally, technical issues subsume *size of the vocabulary and the domain coverage and their effects on voice recognition*, *speaker dependence*, and *environmental influences such as noise*. The requirements defined by [8, 7, 6] guided our project along all phases.

Klemmer et al. [14] argue that building even simple speech-based interfaces requires technology expertise and takes considerable time and effort which is why many individuals are precluded from the design process. To allow for more rapid creation of speech interfaces, they introduce SUEDE, a prototyping tool for electronically supported WoZ. SUEDE's *design mode* enables designers to create dialogue examples including prompts and responses. The *test mode* enables testing with participants without the need for a functional speech backend. The wizard and participant are situated at different places, the wizard selects the appropriate dialogue elements and the system plays pre-recorded speech snippets to the participants according to the respective phase of a dialogue. In the *analysis mode*, SUEDE displays data collected in the test mode to the designer. The analysis interface is similar to the design interface but includes user transcripts from the test sessions and basic statistical information (e.g., the average time it took participants to respond to prompts). In our work described in this paper, we follow a similar path of steps (see Section 4). We first designed our VUIs using graphs and dialogue elements. During the two WoZ phases, we tested these dialogues with real users and recorded all potentially relevant information.

Next, we analyzed the data and used the findings for another iteration of VUI redesign. Our activities however exceeded the WoZ methodology as we actually implemented and tested fully functional prototypes. Yet, we fully agree with Klemmer et al. on the suitability of WoZ testing in early phases of VUI design.

2.2 Speech-Based Systems for Niche Applications

Rayner et al. [26, 25] describe Clarissa, a voice-based system that enables astronauts to navigate through complex procedures using only voice input and output. They argue that "the comparative success of the Clarissa project is perhaps more than anything due to its organization" that relied on a close cooperation of developers and the NASA [26]. They further state that the original conception of the system came from the astronauts themselves. This has impacted our decision to design a VUI for professional welders based on their natural behavior. The situation of astronauts is comparable to the situation of welders as part of the astronauts' tasks are also "frequently hands-busy and eyes-busy" [26]. Initially, Clarissa had a vocabulary of less than 50 words and involved a handful commands, the last version involves about 75 commands and 260 words. Its navigation concept involves commands like "next" or "previous" and can be used to adjust audio volume with commands like "increase volume" or "quieter". Our VUI involves similar concepts for changing parameters like welding current that can be altered either by setting it to concrete values or in- or decreasing it. Clarissa focuses on a user-initiated dialog (command and response) combined with TTS feedback (repeating the values that have been set). Rayner et al. use a grammar-based architecture instead of the more popular statistical one, arguing that i) no huge training data set was available and ii) the system was designed for experts who would have some time to learn its coverage. They refer to Knight et al. [15] who have found grammar-based approaches to work better for this kind of users. Our situation is similar and we use a grammar-based approach, too.

Another discussion of VUIs for a specific niche application is provided by Noyes and Haas [21] who describe speech-based systems for the military domain. Military environments are "often harsh and not amendable to the use of technology per se", military personnel are "often subject to extremes of temperature and humidity". Another primary consideration is ambient noise which degrades speech recognition performance as it interferes with a user's utterances. The environmental situation in industrial welding is similar: it is extreme regarding temperature and noise (the welding arc produces volume levels above 100 dB). Noyes and Haas describe several approaches to make speech recognition more robust in high-noise contexts, e.g., microphone-based (microphone arrays, noise-cancelling microphones) and algorithmic approaches.

Pires [24] describes experiments on commanding an industrial robot using the human voice. His work is relevant for us because of the concrete application examples which involve robotic welding. It is technically-focused and presents implementation details and concrete vocabulary used but does not describe the phase of VUI design. Our work does not tackle robotic welding but should improve efficiency, quality and UX during manual welding tasks. The welding tasks

described in [24] however could also be performed manually, thus the selection of parameters is interesting. Pires points out that several parameters like welding current have to be changed during the process and explains that the VUI must allow for numerical values to be commanded by users. Due to the noisy environment negatively influencing recognition, Pires suggests enabling speech input only when necessary. This was however not applicable for us due to the more variable nature of manual tasks. We thus had to consider approaches to enhance recognition during noisy phases of welding as discussed in [21]. A parallel of our work to Pires' is the use of a grammar-based approach (also see [26, 25]).

Rogowski [27] describes another industrial voice control system that facilitates voice control of robotized manufacturing cells. He notes the following requirements for industrial voice control systems: i) *correct recognition of all words of the voice command*, ii) *accurate recognition of numbers in commands* (as numbers are usually part of the language used in engineering), and iii) *instant reaction to the command*. Rogowski further argues that some requirements significant in other domains are not so restrictive in industrial voice control applications, e.g.: i) *users of such applications are usually qualified machine operators* and can be expected to *adapt themselves to some restrictions regarding voice command structure*, ii) *they can be expected to keep some discipline in speaking*, and iii) *the number of different actions to be performed by the machine is usually low*. These requirements and conditions apply also for our use case.

2.3 Evaluation of Speech-Based Systems

The evaluation of a VUI is a complex task that might differ from the evaluation of other interfaces. An article[2] published online by the Nielsen Norman Group e.g., points out that some classic usability criteria are severely constrained in VUIs, such as the *visibility of system status* or *supporting recognition over recall*.

Cordasco et al. [5] report the result of a lab trial evaluating vAssist, a voice-controlled care and communication service. It has been tested by 43 elderly in a WoZ setting, using several scenarios (e.g., recording and reporting users' medical data). The focus of the evaluation was on usability, learnability and intuitivity. The methodology involved interviews and questionnaires before, during and after user-system interaction, capturing qualitative and quantitative data. The Sinlge Ease Question (SEQ) [28] measured *how easy/difficult the interaction was* immediately after the respective scenario. After all scenarios, the AttrakDiff questionnaire [10] was used to assess *usability, effectiveness, efficiency, enjoyment and appeal of using*. The System Usability Scale (SUS) [3] measured *learnability* and the INTUI questionnaire [32] was used to evaluate *intuitivity*. Our procedure was similar (see Section 4), although we used different evaluation instruments.

A different approach to VUI evaluation is described by Farinazzo et al. [8]. They suggest an adoption of the methodology based on considerations around what is actually being evaluated (e.g., appropriate feedback) or the part of the system that is being evaluated (e.g., dialog management). These guidelines

[2] https://www.nngroup.com/articles/voice-interaction-ux/

helped us to design our evaluation methodology in different phases. Farinazzo et al. also suggest usability heuristics for VUIs around several categories: *suitable feedback*, *user diversity and perception*, *minimization of memorization efforts*, *appropriate output sentences*, *output voice quality*, *proper recognition*, *natural user speech*, *appropriate dialog start and adequate instruction*, *natural dialog structure*, *sufficiency of interface guidance*, *help*, *error prevention* and *handling errors*. The heuristics are used for an evaluation by VUI experts. Our procedure focused on user tests but most objectives can be assigned to the same categories.

Möller [20] further discusses assessment and evaluation of speech-based interactive systems. He distinguishes between a system's performance ("assessment") and its fitness for a specific purpose by the prospective user ("evaluation"). This classification is similar to the distinction of "performance evaluation" and "adequacy evaluation" used in [11]. Möller describes both as measurement processes and discusses along his taxonomy of quality aspects for speech-based interactive systems [19]. This taxonomy describes *quality factors* (e.g., usability, communication efficiency) and *quality aspects* (the factors' composing elements, e.g., speed for the factor communication efficiency). Our VUI design process concentrated on *evaluation*, gathering user feedback on grammar and vocabulary before implementation. Further, we analyzed a small set of indicators that could be assigned to *assessment* (e.g., time needed for a task). Unlike Möller's methodology ours mainly relied on qualitative data. The reasons for this are that i) we actually aimed at gaining insights in the users' subjective impressions, and ii) we had restricted access to professional welders in their work environment.

3 Activities of an Industrial Welder

Generally, welding is a process to permanently join two or more metallic parts. Our work only considers the process of arc welding, where a welding power supply creates an electric arc which melts the metals at the welding point. In some applications, welders have to follow a Welding Procedure Specification (WPS), i.e., a formal document that specifies parameters and settings for welding tasks. If no WPS is provided, welders either rely on their experience or have to do a number of test runs in a trial-and-error manner to identify optimal parameter settings. Thus, some welding power supplies provide standard characteristics with predefined parameter settings. The welder then only has to apply a minimal configuration (e.g., selection of material type and thickness). Often, welders work while standing, sometimes even while kneeling or crouching. A steady hand is required to produce high-quality weld joints. Thus, welders grasp the welding torch with both hands if possible and use all available utilities for additional stability. Further, even during the active welding process, welders need to adjust parameters like current or wire-feed speed. This may be done with a remote control integrated in the handle of the welding torch but requires a high level of experience in order to stay on the joint line and maintain a specific angle to the work surface. Often, altering parameters during the welding process is not even possible for experienced welders and the welding process needs to be interrupted.

This gave rise to the need for a hands-free (and eyes-free) communication with the welding power source. The welder's work environment can generally be described as hostile (according to the definition of [31]), involving high levels of noise, extreme temperatures, electromagnetic interference, and dust and dirt.

4 VUI Design Steps

During the Contextual Design [12, 2] process we followed [1], it became clear that speech would be a viable means to allow welders to manipulate parameters while still being able to use both hands for welding. Our iterative process involving two WoZ tests and analysis phases and several voice recognition improvement steps is sketched in Figure 1 (also see Sections 4.1 to 4.4). Our qualitative approach to user evaluation is conceptualized as a longitudinal lab experiment (with four points of data collection) that relies on observational data (audio, video and notes taken by two observers) and questionnaires answered by the participants but also quantitative data for comprehensive voice recognition tests.

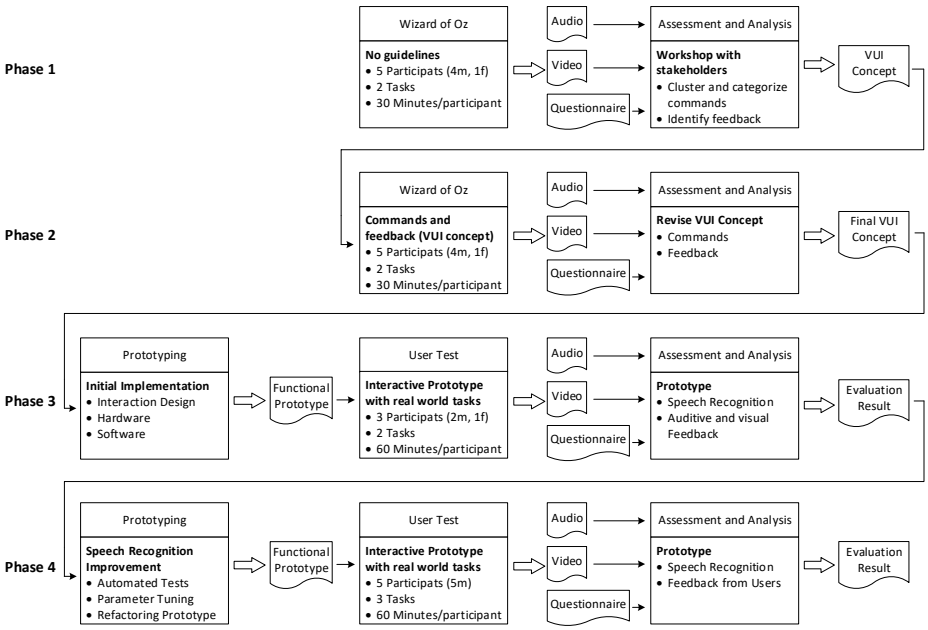

Fig. 1. Overview of the VUI design process.

4.1 Phase 1

Wizard of Oz. As we could not rely on any established VUI concepts in our specific domain, it was necessary to build the entire dialog systems from scratch. To

avoid constraining the welders by assumptions, we envisioned letting users decide how they wanted to "talk" to the machines. Because of the highly performance-oriented and time-critical nature any automated speech recognition would have had to provide for in the given context, we assumed that it was not possible to use a functional prototype at this early stage (e.g., due to the many ambiguities in natural language [23]). We thus used the WoZ paradigm [13] that was successfully employed by Cordasco et al. for their early VUI prototype study [5] and e.g., also recommended by Schnelle and Lyardet [30] or Klemmer et al. [14]. However, unlike earlier approaches, we applied WoZ in a novel way in this phase: not to test a first prototype but to elicit a first version of grammar and vocabulary in the given context. We were able to recruit five[3] professional welders as participants (four male, one female, aged 24 to 45) and divided them into groups of two using a rotation approach so that each of them first acted as welder, then as wizard (with the exception of the first wizard, who acted as the last welder). The welders had to perform predefined welding tasks and were instructed to articulate their needs (e.g., certain values for parameters) solely verbally (in form of instructions for the wizard), the wizards were briefly instructed to operate the welding machine's control panel and set the requested parameters or modes. In total, we analyzed five welder-wizard groups. By not specifying vocabulary or other guidelines in beforehand, we aimed at identifying the most important and frequent keywords, parameters and commands as well as groups of coherent concepts from the analysis of the dialogs. Similarly, we did not instruct the wizards to provide any specific kind of feedback and carefully chose two representative welding tasks that require numerous adjustments of parameters. Regarding the selection of the participants we strove to recruit at least one female welder to have a greater variety of different voices although this was not relevant for the WoZ tests. It was important for later tests with functional prototypes and we aimed at involving the same participants during all phases. During the WoZ test, a movable partition wall was used to separate the welder from the wizard to prevent visual cues (e.g., gestures or facial expressions) the participants might give or receive (see Figure 2). The tests lasted for about 30 minutes per group and were recorded with two cameras and four microphones, and observed by two observers. After each test, the welders answered a questionnaire consisting of open questions and two standardized instruments (NASA TLX [9] and UEQ[4] [16]). The open questions included perceived positive and negative aspects related to the voice-based interaction setting, whether the welders had to think about the commands to use, whether the voice-based interaction distracted from the welding task or whether the welder knew at any time what settings were currently active. Some of these questions were actually designed for the second WoZ test (we chose to use the same questionnaire to be able to compare the results). UEQ and NASA TLX were intended to gather participants' subjective judgments only as our small sample size does not allow to draw quantitative conclusions. After

[3] It was not possible to recruit more welders due to restricted availability at our project partner (many of them work in international support and are not always on-site).

[4] http://www.ueq-online.org/

the tests, audio and video material was synchronized and transcribed. The transcript was then placed into the synchronized audio and video files as subtitles to facilitate analysis of the videos that included a wide range of different noises.

Fig. 2. Setting during WoZ Phase 1.

Assessment and Analysis. We analyzed the obtained data to establish a list of the welders' most important and most frequently used concepts including the parameters that were named and changed during the test as well as the values and units thereof and responses the wizards provided as feedback. During this process, we categorized the commands and investigated how auditory feedback was obtained, resulting in a report. The report became basis for a discussion with stakeholders in a VUI design workshop where welders, welding technology specialists and engineers participated to check our assumptions made in the analysis process for plausibility. Example assumptions were that i) welders would like to finely tune the welding current starting from an initial value which includes the need to set values both absolutely and relatively, or ii) that only a handful of commands is sufficient to cover the most important needs during the welding process. These assumptions could be confirmed at the workshop. To get a more complete picture, we also invited product managers and scientific researchers in the area of welding technology. We presented a first categorization

of relevant commands and parameters and iteratively refined it in an interactive team process, leading to the following categories of welding scenarios:

- *Before the actual welding*: here, about 20 parameters are relevant, however only five (e.g., Job or Wire Feed Speed) are crucial for the VUI.
- *During the actual welding*: only one parameter (welding current) is needed that should, however, be fine-adjustable and configurable in several ways, e.g.: i) set current to a certain value , ii) continuous change (up and down) and iii) change by a delta (up and down). .
- *After the actual welding*: mostly concerning documentation, e.g., to record the quality of the result.
- *Feedback*: welders explicitly wished for confirmations of commands, clarifications in case of ambiguities or summaries of whole configuration processes.

After this step, we could answer the question *which* parameters had to be accessed and *which* ranges of values were possible, hence establishing a first version of the *vocabulary*. The next question was *how* the interactions that were formerly performed through touch screens, buttons and dials, can in future be done through speech, i.e., what the *grammar* should look like. We created an exhaustive list of sample dialogs based on our observations during the WoZ test and read the dialogs out loud to check if the envisioned commands also felt right after they were spoken, as recommended by Pearl [22]. The sample dialogs were again checked with welders to see if they made sense to them. Our observations yielded two very different approaches the welders adopted, depending on whether the welding process was currently active or not. *Before the welding process*, a *dialog-like free speech approach* was used to negotiate the necessary parameters with the wizard. As soon as the *actual welding process* was started, the interaction consisted of *short commands* regarding the necessary adjustments and short auditory responses as acknowledgements. Similarly, the vocabulary also was different for the two phases, with a much higher number of parameters relevant before the process was started. We therefore separated the use cases and created two different grammars. The third scope listed earlier (*after the actual welding*) was not further analyzed. According to the wishes of the stakeholders, the first VUI concepts were created for the German language. The VUI design also included acoustic feedback (repetition of the set value) complemented by visual feedback provided via different-color LEDs placed in the welding shield.

General Insights. Wizards often tried to *establish visual contact to the welders* (e.g., stood up to be able to look over the partition wall). This should be inhibited by test supervisors as it could enable implicit exchange of information through visual cues (e.g., mimics or gestures). The *wizards oftentimes acted "intelligently"* (e.g., one wizard asked whether the welder *really* wanted to use a certain setting, afterwards the welder corrected his command). Such intelligent behavior could be delusive for the welders as any kind of VUI that has not been designed as intelligent assistant will not be able to provide a comparable behavior. *Feedback is of tremendous importance* which became even more clear as the wizards differed strongly in the way and intensity they provided feedback.

Some repeated the parameters and values they had just set, others did not even indicate they had heard the command. In the latter case, welders asked whether the voice commands had i) been understood and ii) led to an adjustment of the parameters. While the participants whose wizards had provided prompt and conclusive feedback stated later that they always knew what state the system was in and felt in control of the process, those whose wizards did not provide feedback stated that they where not always aware of the parameters' values.

4.2 Phase 2

Wizard of Oz. The second WoZ phase was mainly intended to i) confirm the vocabulary and grammar defined in the first WoZ phase and ii) evaluate the adequateness of feedback. We planned on keeping most factors (participants, welding tasks, setup, recording equipment and questionnaires) constant to gain comparable results. One intentional change was that one fixed wizard was employed because this time his activity required a larger amount of training. He should not act intelligently and neither react to all kinds of natural language commands nor comment on the welder's instructions but strictly adhere to a predefined set of commands and reactions. Welders received an acoustic repetition of their command (i.e., the parameter and its new value) by the wizard after the parameters were set, and simple visual feedback. To make the wizard's task easier, he was provided a checklist of all relevant commands, his designated reactions and what he should do in case welders used commands that were not intended. Further, the wizard was granted a longer period of time to familiarize with his task. The welders received a quick training regarding the available commands. As the set of commands was comparatively small, all could memorize it easily. The WoZ test then involved the welding tasks already used for Phase 1 and identical questionnaires. We intended to involve the same welders that had taken part in Phase 1 but had to replace two due to a longer-term unavailability. We again had five participants (four male, one female, aged 24 to 45).

Assessment and Analysis. The second WoZ phase was intended to confirm or revise the vocabulary and grammar defined after the first phase. As we used identical questionnaires, we could compare the participants' answers after the second phase to see whether the results had changed due to restriction of grammar and vocabulary. Like for WoZ Phase 1, the results are not conclusive quantitatively due to the low number of participants. Thus, we considered the answers to NASA TLX and UEQ as weak indicators only (and do not report them in detail) while our main sources of information were the observers' notes, video and audio recordings and the participants' qualitative answers to the open questions.

General Insights. During the second WoZ test we could confirm that the *commands are easy to understand* and can be *learned and applied after a short time* of familiarization. Further, the *acoustic feedback* provided by the wizard led to an improvement in the *participants feeling in control* over the process and *knowing the state the parameters were in*. More generally, welders stated that *voice*

control had no negative impact on the welding tasks but is, to the contrary, rather supportive, especially because the process did not have to be interrupted. Further, we noted that the *restricted grammar and vocabulary* led to a *more efficient process* (the commands and reaction to them took less time compared to WoZ Phase 1) while the scope of the *vocabulary itself was sufficient*.

NASA TLX Comparison. The comparison of the answers to NASA TLX on an individual level of the three recurring participants seems to indicate that *mental demand* was increased in WoZ Phase 2. *Physical demand* tends to have decreased for two participants, although the welding tasks stayed the same. *Temporal demand* stayed on similar levels. *Performance* was estimated to be better by the participants in Phase 2. *Effort* was felt to have increased for one participant, while reduced for another. The perceived *frustration* has minimally increased for one participant but strongly decreased for another.

UEQ Comparison. The answers of the three recurring participants paint the following picture of the perceived UX in the two WoZ tests. *Attractiveness* (i.e., overall impression of the product) was slightly higher in Phase 2. *Perspicuity* (e.g., how easy it is to get familiar with a product) was notably lowered after Phase 2. The remaining categories *efficiency* (can users solve their tasks without unnecessary effort?), *dependability* (do users feel in control of the interaction?), *stimulation* (is it exciting and motivating to use the product?) and *novelty* (is the product innovative and creative?) were rated better after Phase 2.

4.3 Phase 3

Prototyping. After two WoZ iterations, we obtained a final VUI concept including vocabulary and grammar for welders for two different phases (*before* and *during the actual welding process*). Based on this approved concept, we started Phase 3 with prototyping a high-fidelity, interactive prototype. This allowed us to assess design details concerning interaction behavior, audio interface and speech recognition with our users in a real world setting. The Phase 3 prototype consisted of three main components: i) *speech recognition*, ii) *interaction controller*, and iii) *welding remote controller*. The system is designed in a modular way, thus each of the components can be easily adapted and replaced independently. The *speech recognition component* can use various open source and proprietary engines. We selected the platforms considering stakeholder requirements related to *offline service*, *weight*, *price* and *size*, and requirements obtained through our user tests. Depending on the features and functionality of the speech recognition engines, we also experimented with different language and acoustic models. This includes the use of grammars and statistical language models, such as N-gram models adjusted to our vocabulary. At this stage, no adaption or retraining of the acoustic model was done. We experimented with various prebuilt models in German (and few in English). The *speech recognition component* recognizes the spoken commands based on the vocabulary in the VUI concept and returns tokens to the *interaction controller*. The interaction controller maintains

system state, controls feedback mechanisms (acoustic and visual feedback) and interfaces with the welding equipment through a *remote controller component*. Acoustic feedback is given through a TTS system with small speakers placed inside the welding shield. Multicolor LEDs provide visual feedback through ambient light. The use of a soft light is crucial to not affect the welder's vision.

User Tests. After the initial implementation and tests in our lab environment, a first user test with the functional VUI was conducted. The same measuring instruments as in the preceding WoZ phases were employed: audio and video recordings, observation and the same questionnaires. Further, we could now compare the spoken voice commands with the detected commands. As in the WoZ phases, five participants sequentially worked on the same tasks. Unfortunately, voice recognition in this phase was strongly affected by the noise level of the welding arc so that primarily in the higher current areas (with a sound pressure level of up to 110 dB resulting from a wire feed of around 8 m/min), efficient work was impossible. As the tasks involved in the user test all required the welding current to be in these areas, the experiments were canceled after three runs. Yet, important insights could be gained during this test (see below).

Assessment and Analysis. According to the observers' impression, the vocabulary and grammar could be internalized fast by all three participants (two of them already took part in a preceding WoZ test) and consequently the VUI was usable with only few errors on the participants' side. The analysis of audio and video material along with observer notes, however, confirmed our initial impression that this prototype's bad recognition rate prevented an efficient way of working although it was successfully tested in the lab before. This also led to participants using different pitches and going from screaming to whispering in order to achieve voice input in the higher current areas but to no avail. These problems were also reflected in the results of our questionnaires. To prevent a lasting frustration, it was necessary to explain that this early prototype had not been tested under realistic conditions before but it worked reasonably well in the absence of excessive environment noise, and that our next steps comprise a systematic improvement of the recognition rate. Further insights were gained regarding feedback: acoustic feedback was found to be too quiet and visual feedback was not clear enough to be easily recognized while the welding arc was active. Also, participants perceived a short delay before provision of feedback.

4.4 Phase 4

Prototyping. The prototype underwent several iterations of refactoring and improvements based on the findings of Phase 3. This included acoustic and visual feedback and the interface to the welding machine for more robustness and faster responses during welding parameter changes. As analyzed in Phase 3, the biggest drawback of our prototype was the bad recognition rate in the loud

working environment. Thus, Phase 4 mainly focused on speech recognition improvement. We set up an infrastructure for systematic, repeatable tests and automated analyses. Our experiments considered different *microphone types* (with different directional characteristic, e.g. omnidirectional, cardioid and shotgun), *noise levels* and *pre-processing of the audio signal* (normalization and filtering). Our *audio database* consisted of samples from 7 persons (4m, 3f) with 70 different commands and 3 repetitions per person. The samples were recorded with 7 microphones (in price ranges between 2€ and about 500€) in parallel, with and without welding noise, resulting in more than 17.500 recorded and transcribed audio samples. The raw samples (with a base welding background noise and further pre-processing) showed recognition rates ranging from 34.9% to 69.4% for the various microphones. By automatically adjusting the gain on some microphones, the recognition rate could be further improved. Generally, we observed that the pronunciation and dialect is important as we noted significantly higher recognition rates for our best speakers. Additionally, experiments with audio noise reduction filtering brought another improvement of 3.1% on average across all microphone types. Among others, we applied a Wiener filter [18, Ch. 6] in our noise reduction approach. It should be noted that parameterization is essential. Some filters and their parameter setting respectively, worsened the recognition rate significantly. All (even little) improvements of the recognition rates however came with considerable implementation effort and lots of additional testing. In some cases, potential improvement approaches even led to a step back and a decrease in the recognition rate, and changes had to be reverted.

User Tests. After extensive effort to improve the recognition rate in the systematic tests with recorded welding noise it was now necessary to test again under real world conditions. Therefore, the same tasks and instruments as in the previous user tests were used while focusing on i) the correctness and timeliness of the speech detection and ii) the participants' subjective feedback on the other. Only after a sufficient recognition rate could be achieved, is it possible to assess if the VUI concept was designed and implemented successfully.

Assessment and Analysis. Firstly, the focus of this assessment was to show whether there was a significant improvement in the recognition rate. This could be confirmed because even at the noise levels where recognition had badly failed earlier, many voice commands were now correctly recognized. Still, we could not get close to a perfect or human-like recognition when environment noise increased. Especially short commands were hard to detect and were therefore replaced by longer synonyms. Secondly, we received predominantly positive feedback by our five participants. Because only one participant had taken part in Phase 3 where the recognition had not worked sufficiently, a high *contrast effect* could be ruled out. Nevertheless, his results are reported separately here. The participant liked the acoustic and visual feedback, although the TTS output was still too quiet for him. Thus, he was not always aware of the currently active parameters. He suggested extending the range of possible synonyms for

commands and to personalize certain effects of commands. He also stated that training would help to become better at working with the VUI ("practice makes perfect"). The other participants mainly liked the short reaction times and all of them stated that the voice commands did not distract them from working. All participants also explained that there was not much thinking involved to memorize and use the commands, at least after an initial familiarization.

5 Discussion

In this section we reflect on our VUI design process, summarize our findings and discuss their implications for a broader spectrum of contexts. The paper's main contribution lies in the detailed description of our applied methodology during the process of designing VUIs, consisting of two WoZ phases and two stages of prototyping. It is important to note that we used WoZ in a rather uncommon way: not for experiments with early prototypes but for establishing first concepts for VUIs in a context-constrained domain. WoZ enabled us to elicit appropriate grammar and vocabulary for our use case which would hardly have been possible without observing potential end users' natural behavior and dialogs in their working environment. Further, it helped us in the second phase to confirm our first VUI design under realistic conditions. We strongly encourage other researchers (independent of the domain, VUIs should be designed for) to apply WoZ with real users and under realistic conditions (i.e., giving them real-world tasks and their familiar equipment). Further, it was extremely helpful not to constrain the welders' natural dialog in the first phase. Regarding the WoZ process, we observed that *visual contact* between two persons talking to each other is *very important* to them. Most participants tried to avoid the partition wall, e.g., by getting on their tiptoes. The natural visual cues that aid our face-to-face communication are usually lost in VUIs. Thus, many currently available VUIs (e.g., those integrated with Amazon Alexa[5]) use different colors or LED blinking patterns (e.g., the Amazon Echo Dot) or even more enhanced graphical user interfaces (e.g., the Amazon Echo Show). To account for this basic human need, our prototypes also use *auxiliary visual feedback*. Further, we recommend other VUI designers using WoZ to consequently prevent any visual cues to be exchanged even if this comes at high effort. Insufficient spatial separation of user and wizard might lead to unreliable results regarding the identified dialog behavior. We also highly recommend applying a two-phase WoZ process *and* involving all stakeholders before finalizing a VUI in an industrial or other niche application context. The second phase helped us to test whether the identified grammar and vocabulary also worked without further verbal information exchange. The stakeholder workshop revealed additional requirements (e.g., the need to introduce command synonyms welders don't use in their colloquial dialogs but are present in the control panels of the welding machines, for reasons of consistency).

In summary, our *experiences with WoZ* during early phases of VUI design are predominantly *positive* due to its flexibility (not being constrained by technical

[5] https://developer.amazon.com/de/alexa

limitations). This was particularly important in the first WoZ phase because it would hardly have been possible to implement a functional prototype that allows for an unlimited, natural dialog around the welding domain. Prior to the WoZ tests we did experiments with popular cloud services but noted high error rates for domain-specific vocabulary (while recognition was excellent for more general dialogs). A *possible drawback of using WoZ* while establishing VUI concepts for manual welding and comparable applications lies in the *time-critical nature of the tasks*. Responding fast enough and still sticking to the predefined set of commands and responses is challenging for the wizard who has to i) interpret the voice command solely according to a given protocol, ii) set the parameter and iii) give correct and timely feedback. *Thorough training of the wizard* is necessary.

Generally, the findings of our tests have shown that the niche application of *manual industrial welding can strongly benefit from the use of voice interaction*. Especially, welders pointed out i) *the time-saving nature of voice interaction*, ii) that they *do not have to interrupt the welding process* in case parameters need to be changed and iii) that the *installation of this kind of new technology neither requires a major change regarding their equipment nor requires additional equipment*. The tests have also shown that although the *vocabulary needs to be sufficiently small* (to be easily memorized), *a short training period is acceptable* for the target group. In case the range of functions utilized by end users differs strongly for different usage contexts (e.g., before or during the welding process), we can recommend designing different VUIs (and related grammars etc.). This reduced the complexity of the individual VUIs and the error rate on the users' side. Yet, all related VUIs should be designed on the basis of consistent principles (e.g., related to feedback and reaction to errors). We also noted that for the welders it is of tremendous importance that the *vocabulary matches their natural usage of language in the domain* (e.g., in case there were two or more designations for the same command, they rigorously preferred the one that was closer to their colloquial term). As the related preferences might differ among persons but potentially also among languages, we chose to *introduce synonyms for parameters for which there exist different names*. This was appreciated by the welders.

Our *utilization of the two standardized instruments UEQ and Nasa TLX* suffered from the well-known *problem of longitudinal study designs* that it might be *difficult to get the same participants* for all evaluations. While this would have at least allowed us to compare the two questionnaires' results on an individual level, only three recurring participants took part in Phases 1 and 2 and only one was present during all phases. Still, in these early phases it allowed us some important insights. For example, the comparison of the two NASA TLX results shows that the perceived *mental demand* had increased for the three participants. This might be connected to the well-known challenge in VUIs that users must memorize the set of allowed commands, and acknowledges the aforementioned infringement of Nielsen's heuristics of *visibility of system status* or *supporting recognition over recall*. Equally, the UEQ category *perspicuity* that was also rated lower after WoZ Phase 2 has a strong connection to *learnability*. All other categories of the UEQ were rated better after the second run.

Regarding implementation, we recommend early testing of functional prototypes under realistic conditions, even if first prototypes might terribly fail. Early tests allowed us to identify i) those kinds of environmental noise where recognition rates go down drastically and ii) the conditions where recognition is only marginally affected. Our automated test environment then allowed us to perform a great number of systematic tests. These tests enabled us to identify the best-performing settings for different conditions. If a VUI should be speaker-independent, we recommend thorough tests with different speakers who should be highly diverse regarding their timbre, volume, pronunciation and dialect. Our first tests revealed severe differences in recognition rates among different speakers of up to 25% although all of them tried to speak clearly and loudly.

6 Conclusion

VUI designers can usually resort to guidelines, gold standards or best practices (e.g., the Amazon VUI design guidelines[6]), or experiences with similar systems. Yet, we identified a lack thereof, when a VUI has to be established in a domain off well-trodden paths. Thus, we presented a four-phase approach of creating a VUI concept and functional prototype in a context-constrained application domain from scratch. By first observing the natural conversation between welders in a WoZ test during manual welding tasks, we could establish the basic vocabulary and grammar that was later discussed with domain experts and finally validated with the help of a second WoZ test. During the authoring of the sample dialogs that followed the initial observations, specific focus was placed on the requirements identified by Farinazzo et al. [8] such as *consistency* (we ensured that the same commands have the same effects in different situations), *feedback* (acoustic and visual), or the *minimization of cognitive effort* (we used a small set of commands and allowed synonyms where more than one identifier was observed in WoZ Phase 1). The analysis of data we gathered during the different phases indicates that the VUI concept we established was well usable by our participants. Phase 3 has shown the specific challenges related to voice recognition in noisy environments. However, we could reach significant and promising improvements during our extensive test and voice recognition enhancement activities in Phase 4. The prototypes have shown sufficient potential for application in real world settings in the domain of industrial welding and are currently taken one step further towards a commercial product by our industrial project partner.

Acknowledgements

The work described in this paper has been conducted within the scope of the project *Welding Interaction in Future Industry* funded through the BRIDGE 1 program, managed by the Austrian Research Promotion Agency (FFG). Project partners are the University of Applied Sciences Upper Austria, LIFEtool gemeinnützige GmbH and Fronius International GmbH.

[6] https://developer.amazon.com/designing-for-voice/

References

1. Augstein, M., Neumayr, T., Pimminger, S., Ebner, C., Altmann, J., Kurschl, W.: Contextual design in industrial settings: Experienes and recommendations. In: Proceedings of the 20th International Conference on Enterprise Information Systems. Funchal, Madeira, Portugal (2018)
2. Beyer, H., Holtzblatt, K.: Contextual design. Interactions pp. 32–42 (1999)
3. Brooke, J.: Sus - a quick and dirty usability scale. Usability Evaluation in Industry **189**(194), 4–7 (1996)
4. Cohen, M., Giangola, J., Balogh, J.: Voice User Interface Design. Addison-Wesley (2004)
5. Cordasco, G., Esposito, M., Masucci, F., Riviello, M.T., Esposito, A., Chollet, G., Schlögl, S., Milhorat, P., Pelosi, G.: Assessing voice user interfaces: The vassist system prototype. In: Proceedings of the 5th IEEE Conference on Cognitive Infocommunications. Vietri sul Mare, Italy (2014)
6. Dybkjaer, L., Bernsen, N.O.: Usability evaluation in spoken language dialogue systems. In: Proceedings of the ACL 2001 Workshop on Evaluation Methodologies for Language and Dialogue Systems (2001)
7. Farinazzo, V., Salvador, M., De Oliveira Neto, J.a.S., Kawamoto, A.S.: Requirement engineering contributions to voice user interface. In: Proceedings of the First International Conference on Advances in Human-Computer Interaction. Sainte-Luce, France (2008)
8. Farinazzo, V., Salvador, M., Kawamoto, A.L., De Oliveira Neto, J.a.S.: An empirical approach for the evaluation of voice user interfaces. In: Matrai, R. (ed.) User Interfaces. InTech (2010)
9. Hart, S.G., Staveland, L.E.: Development of nasa-tlx (task load index): Results of empirical and theoretical research. Advances in Psychology **52**, 139–183 (1988)
10. Hassenzahl, M.: Hedonic, emotional and experiential perspectives on product quality. In: Ghaoui, C. (ed.) Encyclopedia of Human Computer Interaction. Idea Group Reference (2006)
11. Hirschman, L., Thompson, H.S.: Overview of evaluation in speech and natural language processing. In: Survey of the State of the Art in Human Language Technology. Oxford University Press (1997)
12. Holtzblatt, K., Jones, S.: Contextual inquiry: A participatory technique for system design. In: Schuler, D., Namioka, A. (eds.) Participatory Design. Principles and Practices, chap. 9. Lawrence Erlbaum Associates (1993)
13. Kelley, J.F.: An iterative design methodology for user-friendly natural language office information applications. ACM Trans. Inf. Syst. **2**(1), 26–41 (Jan 1984). https://doi.org/10.1145/357417.357420, http://doi.acm.org/10.1145/357417.357420
14. Klemmer, S., Sinha, Anoop K. and Chen, J., Landay, James A. and Aboobaker, N., Wang, A.: Suede: A wizard of oz prototyping tool for speech user interfaces. In: Proceedings of the 13th Annual Symposium on User Interface Software and Technology. pp. 1–10. San Diego, California, USA (2000)
15. Knight, S., Gorrell, G., Rayner, M., Milward, D., Koeling, R., Lewin, I.: Comparing grammar-based and robust approaches to speech understanding: A case study. In: Proceedings of Eurospeech 2001. Aalborg, Denmark (2001)
16. Laugwitz, B., Schrepp, M., Held, T.: Construction and evaluation of a user experience questionnaire. In: Proceedings of USAB 2008. pp. 63–76. Graz, Austria (2008)

17. Licklider, J.: Man-computer symbiosis. IRE Transactions on Human Factors in Electronics **HFE-1**(1) (1960)
18. Loizou, P.C.: Speech Enhancement: Theory and Practice. CRC Press, Inc., Boca Raton, FL, USA, 2nd edn. (2013)
19. Möller, S.: A new taxonomy for the quality of telephone services based on spoken dialogue systems. In: Proceedings of the 3rd SIGdial Workshop on Discourse and Dialogue. Philadelphia, PA, USA (2002)
20. Möller, S.: Assessment and evaluation of speech-based interactive systems: From manual annotation to automatic usability evaluation. In: Chen, F., Jokinen, K. (eds.) Speech Technology. Theory and Applications. Springer (2010)
21. Noyes, J., Haas, E.: Military applications: Human factors aspects of speech-based systems. In: Chen, F., Jokinen, K. (eds.) Speech Technology. Theory and Applications. Springer (2010)
22. Pearl, C.: Designing Voice User Interfaces: Principles of Conversational Experiences. " O'Reilly Media, Inc." (2016)
23. Pieraccini, R., Suendermann, D., Dayanidhi, K., Liscombe, J.: Are we there yet? research in commercial spoken dialog systems. In: Proceedings of the International Conference on Text, Speech and Dialogue. pp. 3–13. Pilsen, Czech Republic (2009)
24. Pires, N.: Robot-by-voice: Experiments on commanding an industrial robot using the human voice. Industrial Robot: The International Journal of Robotics Research and Application **32**(6), 505–511 (2005)
25. Rayner, M., Hockey, B.A., Renders, J.M., Chatzichrisafis, N., Farrell, K.: Spoken language processing in the clarissa procedure browser. Natural Language Engineering **1**(1) (2005)
26. Rayner, M., Hockey, B.A., Renders, J.M., Chatzichrisafis, N., Farrell, K.: Spoken dialogue application in space: The clarissa procedure browser. In: Chen, F., Jokinen, K. (eds.) Speech Technology. Theory and Applications. Springer (2010)
27. Rogowski, A.: Industrially oriented voice control system. Robotics and Computer-Integrated Manufacturing **28**(3), 303–315 (2012)
28. Sauro, J., Dumas, J.: Comparison of three one-question, post-task usability questionnaires. In: Proceedings of the 27th International Conference on Human Factors in Computing Systems. Boston, MA, USA (2009)
29. Schmandt, C.: Voice Communication with Computers: Conversational Systems. Van Nostrand Reinhold Co, New York (1994)
30. Schnelle, D., Lyardet: Voice user interface design patterns. In: Proceedings of the 11th European Conference on Pattern Languages of Programs. Irrsee, Germany (2006)
31. Turunen, M.: Jaspis – A Spoken Dialogue Architecture and its Applications. Ph.D. thesis, University of Tampere, Department of Information Studies, Tampere, Finland (2004)
32. Ullrich, D., Diefenbach, S.: Intui. exploring the facets of intuitive interaction. In: Tagungsband Mensch & Computer 2010: Interaktive Kulturen. Duisburg, Germany (2010)