Development of a Heuristic Evaluation Tool for Voice User Interfaces

Martin Maguire(✉)

Design School, Loughborough University, Loughborough, Leicestershire LE11 3TU, UK
m.c.maguire@lboro.ac.uk

Abstract. Voice user interfaces (VUIs) are now a common means of interaction with IT systems. To assist in evaluating the usability of such interfaces, a set of evaluation heuristics assessing speech interfaces was developed by following an existing methodology for defining usability heuristics. Two groups of eight participants conducted an evaluation by inspection of three speech-based systems, a mobile phone assistant, a smart speaker and an in-car hands free phone system. One group used Nielsen and Molich's general heuristics for user interface design while the other group used the VUI heuristics. The second group found, on average, more problems than the first group. However, most heuristics from both sets were rated as useful during the study. This indicates that a mixture of both general and application specific heuristics are needed for a comprehensive evaluation to be performed. Experience from a pilot study, where a smart speaker was set up in a domestic setting, highlighted the need to also consider social and environmental issues to gain a complete picture of user experience when interacting with speech systems.

Keywords: Speech systems · Voice user interfaces · Heuristic evaluation · Smart speakers · Smart homes

1 Introduction

Voice user interfaces (VUIs) are having an increasing an effect in our daily lives. People are now appreciating the potential of a hands-free interface that uses natural language capability rather than requiring keyboard input. While speech-based telephone interfaces, in-vehicle voice recognition, and dictation systems have existed for a long time, the advent of voice-based assistants on a mobile phone or smart speaker has helped to make voice interaction a mainstream technology. This increases the need to provide tools and techniques to evaluate these systems.

This paper reports on a study to define and test whether usability heuristics developed specifically for VUIs, are better able to identify usability problems with VUIs, compared to using general-purpose usability heuristics.

A. Marcus and W. Wang (Eds.): HCII 2019, LNCS 11586, pp. 212–225, 2019.
https://doi.org/10.1007/978-3-030-23535-2_16

2 Literature Review

2.1 Development of Voice User Interface Technology

A voice-user interface (VUI) makes spoken human interaction with computers possible using speech recognition to understand spoken commands and questions and, typically, text-to-speech to play a reply. VUIs are not new, with the first elementary examples such as 'Radio Rex' being produced in the 1920s [1]. Rex was a small celluloid dog set into a wooden dog house. When the dog's name was called, it would jump out of the house to the owner. In the 1950s, Bell Labs built a system called 'Audrey' (Automatic Digit Recognizer) for single-speaker recognition of digits. This achieved a high degree of accuracy. These early systems had small vocabularies and were not much use outside of the lab. In the 1960s and 1970s, the research continued, expanding the number of words that could be understood and working toward "continuous" speech recognition (not having to pause between every word).

Organisations such as IBM and the U.S. Department of Defense experimented with speech recognition in the following decades, but it was only in the 1990s that it became a consumer product with Dragon releasing a consumer speech recognition product, Dragon Dictate, in 1990, and BellSouth launched the first consumer voice portal, VAL, in 1996. Testing by Xiong et al. [2] showed that automated systems performing a transcription task can reach parity or exceed the performance of human transcribers. In terms of recognition accuracy, machine errors are substantially the same as human ones, the main difference being that the machine was less able to identify backchannel utterances such as like "uh-huh" signalling that the speaker should keep talking, and hesitations sounds like "uh" which indicate that the current speaker has more to say and wants to keep his or her turn.

Speech recognition and voice commands also started to be built into operating systems such as Windows Vista and Mac OS X, as well as interactive voice response (IVR) systems for telephone callers. Voice interaction arrived on mobile devices for the first time in 2008 with the release of the Google Voice Search app for iPhones. This technology was later added to Google Search, Maps and the Chrome browser.

Voice recognition apps are now ubiquitous across mobile devices. Apple's Siri virtual assistant processed 1 billion queries per week in 2015, while 20 percent of Google searches on mobile devices are performed through voice recognition [3]. These services and devices depend on data and content assets acquired by these platforms to fulfil user requests. Thus, when a user asks Siri for directions, it can quickly leverage Apple Maps to provide a routing. When they ask Amazon Echo to play a song or read an Audible book, Alexa draws on those Amazon assets to play back the user's content [4].

In recent years, the benefits of speech technology have become more widely recognised since it enables systems to be commanded by voice without keyboard input and while the user is performing other tasks so that their eyes and hands may be busy e.g. when cooking or driving [5]. It uses conversational skills which most people have and apply naturally. A Stanford study showed that speaking (dictating) text messages was faster than typing, even for expert texters. Voice, which includes the characteristics of tone, volume, intonation, speed and emotion, conveys information that a textual

message generally does not. Speech interaction can have benefits for people with physical disabilities in controlling household devices such as TVs, lights, window blinds, heating controls, and security cameras, more easily than doing so directly or by using a remote handset.

2.2 Usability of Voice User Interfaces

As VUI systems have developed, researchers have gained experience in the design aspects that determine their usability. The authors Cohen et al. [6], in writing about interactive voice response (IVR) systems, describe many aspects of design e.g. persona, prosody (intonation, tone, stress and rhythm), error recovery, and prompt design, that are still relevant to today's VUIs. Harris [7] also describes a process for designing voice user interfaces including the voice or agent characteristics, dialogue design, scripting and iterative evaluation. In designing voice user interfaces, Harris emphasises the need to craft the interface for voice and not to try to match it to a visual user interface i.e. to create an auditory version of a GUI.

Cohen et al. [6] advises against making design decisions without consideration of the context or environment in which the system operates. In relation to this, Whitenton [8] emphasises the need for the system to be able to distinguish voice from interfering noise such as music or other sounds in the environment, and the ability to detect a voice input from a reasonable distance. Efficiency can be important for repetitive tasks so having to repeat multiple times, "please add milk to shopping list", "please add bread to shopping list", etc. can become laborious. Another principle described is the need to avoid more than 4 or 5 speech-based options as users must keep them in working memory in order to make the correct choice.

Asthana et al. [9] propose three dimensions for studying the usability of IVR design. The first is 'navigation' which is the time spent on announcements and selection of menu and submenu options which should be minimised while making the process clear. Secondly 'relevance' of information delivered to the user. This is determined by the ease with which users can accurately choose the option they want from the menu list. It was found that new callers tended not to select the wrong menu as they listened carefully to the options, while repeat callers tried to guess from previous usage which often led to an error especially if the order of the menu options, or the options themselves changed over time. This is an argument for maintaining consistency are far as possible when menus are updated. Thirdly, 'capacity' should be considered, which is the number of options in a menu balanced against the systems ability to correctly match user utterances to the options. If the number of options is too large, then the chance of an error increases.

Howell et al. [10] studied the use of spatial metaphors within a hierarchically structured mobile phone city guide service. They found that the use of spatial metaphors could lead to improved usability by capitalising on people's well-developed special abilities. The metaphors used were driving on a journey, managing a filing system, and a shopping journey. The study, which employed by first time users, showed that the office filing system metaphor borrowed from graphical user interfaces (GUIs) could be successfully transferred to a speech-based VUI.

Franzke et al. [11] compared a simulated speech recognition interface (using a 'wizard-of-oz' experiment) for a basic voice mail application, with functionally similar touch-tone and operator assisted versions. They found that subjects adjusted their behaviour when using the speech system compared to interacting with a human operator. Participants used less complex grammar when talking to a computer, less words per utterance, did not include the sentence subject as often, and tended to exclude the indirect objects from sentences, than when participants were talking to an operator. This may imply that a sophisticated natural language processing unit is not a necessity for a speech recognition application of the size and structure of a basic voice mail system. Speech was also regarded as generally more time efficient and subjectively easier to handle than key-command combinations since spoken commands are easier and faster to learn.

Damper and Gladstone [12] evaluated the IMAGINE speech-based interaction system to provide universal access to electronic services including disabled users. The system development initially concentrated on the application of an online shop. The system allowed basic shopping steps and speech specific steps. These included: logging on, setting speech output preferences, listing products by letter, browsing the catalogue, putting products in the basket, checking out, etc. Testing showed that users wanted to try out the system first. This idea of checking through the steps of an online process is as likely to be just as useful for a voice user interface as it is for a visual interface. The development and testing of the system identified some design rules that needed to be applied. These included removing the definite or indefinite article e.g. "a tin opener", keeping the list of products spoken below 6, and the need to recognise product codes as well as names. Also, when the user asks to browse all the browse options should be presented. This study shows that both general and specific requirements for a speech system will emerge as the application develops.

In studying IVR systems, Kim [13] states that user satisfaction with these systems is still low. Using a simulator to enable usability testing of speech systems, they identified four types of usability problem: (1) 'term ambiguity', where ambiguous terms and expressions can lead to delayed task completion, (2) 'phonetic deficiency' of speech output including pronunciation, volume and voice speed, (3) 'information navigation' where callers have to return to the root or previous menu to proceed with their task, and (4) 'cognitive overload' which can occur due to loss of concentration e.g. when listening to a list of menu options.

Portet et al. [14] reports a user evaluation study that assessed user acceptance and user concerns related to a smart home voice interface using a 'wizard-of-oz' technique. The study included scenarios for appliance control by voice, communication with the outside, responding to a system interruption to close a door or turn off an oven, and managing a shared calendar. The study included 18 people (8 older people, 7 relatives and 3 caregivers). They found that the issuing commands using keywords is well accepted compared with sentence-based commands. They also found that successful applications using speech recognition tend to have smaller vocabularies, and it is difficult to manage out-of-vocabulary and ill-formed commands uttered by the user. It was also discovered that people would tolerate having to repeat some commands when the VUI did not understand their first attempt, although this might diminish over time. Further findings were that natural voice outputs from the system was preferred to

synthetic outputs. Nine participants preferred the system to have a female voice, one preferred a male voice, while the remainder did not mind which gender it was.

Yankelovitch et al. [15] investigated the challenges of conversational interface design by development of a research prototype for voice command and interaction with email, a calendar, weather information and stock quotes. A study with 14 participants found several design challenges. To make the interaction feel conversational, prompting the user for input was avoided where possible, so allowing users to comfortably take the lead in formulating their input e.g. 'read the message' or 'skip this'. Other guidelines from the study were to ground the conversation, avoid repetition, and to handle interruptions. Immediate and informative feedback was also seen as essential so that users would know that the system has heard them and that their command had been recognised correctly. A long pause may result in the user trying to keep their conversational turn by using 'errs' and 'ums' which can result in recognition errors. Interpreting silence from the system is sometimes ambiguous as it may mean that the system is not working or simply that it did not hear the user input. The challenge of converting a visual interface (GUI) into a voice interface (VUI) was also addressed. Using voice, a user won't necessary use the correct menu command, e.g. 'tell me …' rather than 'what is …', and may use relative dates e.g. 'a week from tomorrow'. Numbering messages or tagging them with codes (g. old or new) may be a natural means of managing them visually but becomes awkward in a voice system. Also, when a dialog box is used to control flow with 'yes' and 'no' options, users may try other commands using speech e.g. 'send' or 'read the next message'.

Further aspects of voice interaction are provided by Bernholz [3]. He describes the importance of understanding users' expectations when they interact with a system and the scope of topics that the system covers. System feedback should make it clear what question it is answering. For example, as well as just providing a football score, it should name the teams and the date when the game was played which acts as useful confirmation. Apple Siri helps to solve these needs by listing ideas for possible questions when the users starts using the service and gives both visual and tactile feedback to show when it is listening.

Bernholz also mentions some practices to avoid when building a VUI as part of a mobile application. These include asking the user a question when the application expects a response, not making it clear about how the user should respond, giving the user too many choices, and being too verbose (e.g. "say 'football' for football. say 'basketball' for basketball…") and confirming the user's query too often. Confirmations should be reserved for important actions such as sending a message or making a purchase.

The results from these previous studies show that many findings from the past still apply to current VUIs. At the same time, the development of new voice input technological capability will open up new areas when design guidelines will be required.

2.3 Heuristic Evaluation

The method of heuristic evaluation developed by Nielsen and Molich [16] utilises experts who inspect an interface to identify usability problems. During the inspection, usability heuristics or 'rules of thumb' are used as a checklist to stimulate thinking and

to categorise the problems found. The results from all the experts are combined to create a comprehensive list of problems to be addressed by redesign. Nielsen and Molich's 10 usability heuristics are as follows:

1. Visibility of system status
2. Match between system and the real world
3. User control and freedom
4. Consistency and standards
5. Error prevention
6. Recognition rather than recall
7. Flexibility and efficiency of use
8. Aesthetic and minimalist design
9. Help users recognize, diagnose, and recover from errors

As speech-based systems become more common, the question arises as to whether speech as a form of user interaction is distinct enough from traditional screen-based interfaces that heuristics tailored to speech systems are needed to conduct usability inspections of them. A paper by Quiñones and Rusu [17] states that user interface heuristics such as Nielsen and Molich's, developed to evaluate traditional screen-based interfaces, are limited by not being able to evaluate the unique features of an application such as a voice user interface. Johnson and Coventry [18] studied the specific application of a VUI to a self-service automated teller machine (ATM). They question the use of traditional heuristics as their origin and general application concerns conventional screen-based, often desktop, interaction. Usability heuristics for use when evaluating VUIs have been produced by Cohen et al. [6] and Harris [7]. Both sources state the need for the adaptation of traditional heuristics when conducting heuristic evaluation due to their orientation towards screen-based user interfaces.

3 Pilot Study Work

To gain a more direct understanding of how users might interact with current voice interaction technology, a field study was conducted by installing an Amazon Echo/Alexa smart speaker in the kitchen of a student house for one week. The participants were four undergraduate students who were aware of speech-based assistants but stated that they had limited experience of using them. The intention was to see whether the participants used the smart speaker, what they used it for, and how they felt about the technology after one week of usage.

At the start of the week, the participants were shown how to interact with the device by using the wake-up command 'Alexa' and how to issue different commands such as play music or set a timer. They could interact with Alexa during the following week as they wished. The Alexa software app was available to access the user interactions and speaker responses made during the week. It was found that 27 interactions with the speaker were made which included requests to play music or the radio, set a timer (e.g. as a wake-up alarm or for cooking purposes), ask for information or a joke, download a skill or read the news headlines. Possibly more use of the device would have occurred

if it had been linked into home devices such as the control of lights or ordering groceries.

Despite the limited use, participants felt that they had interacted with the device effectively. There were comments about the limited accuracy/success rate of interactions, feeling self-conscious when speaking to the assistant with others present, and speculation about the microphone being 'always on' and listening in to their conversations. These results reflect the survey conducted by Milanesi [19] which showed that people's current use of consumer voice-based assistants may be at a basic level but as more services become reliant on them and they become integrated into homes, users will become more familiar with them and less self-conscious about using them.

A recent study by Adobe [20] indicates that smart speakers have led to a growing acceptance of voice interaction with systems. It reported that 72% of smart speaker owners are now comfortable with using voice assistants in front of others. Among people not owning smart speakers, only 29% are comfortable with doing so. Arguably, voice is becoming increasingly interwoven into our cultural fabric and will become a key element in how consumers engage with the world around them.

4 Development and Evaluation of VUI Heuristics

A study was conducted to define and validate a set of usability heuristics specifically for voice user interfaces (VUIs) and to see how effective they were in comparison with general purpose usability heuristics, when evaluating speech-based systems.

The VUI heuristics were developed following the method by Rusu et al. [21]. The method included the key stages of:

(1) Exploration – identify source material on problems related to speech interfaces.
(2) Description – group the problems by theme to create proto-heuristics.
(3) Correlation – refine the proto-heuristics by correlating them with well-established general-purpose sets of heuristics.
(4) Explication – specify the heuristics in a standard way and provide examples.
(5) Validation – comparison of new heuristics with benchmark set of heuristics.

The language used in the enhanced VUI heuristics was explicitly linked to the domain of speech interaction. Heuristics not seen as necessary for 'voice' were removed and new specific VUI heuristics were introduced. The results of each stage are as follows:

Exploration: A systematic review of literature relating to the usability of VUIs was conducted to build a relevant bibliography. Seventeen items of literature were identified including 12 journal papers, 2 books and 3 websites. From this sample, 72 usability related items were found comprising principles, guidelines, ideas and concepts related to usable VUIs.

Description: Athematic analysis was conducted on the items, resulting in 8 themes (see Table 1).

Table 1. VUI usability themes identified from the literature

Theme	Further detail
Cognitive load	• Limited capacity for short term memory • Recognition rather than recall
Speak the user's language	• Setting user expectations • Avoid ambiguity
Efficiency	• Avoid unnecessary words • Interface can be tedious when listing products or items
Feedback	• Visibility of system status • Usability improved with consistency of system-voice and feedback
Accuracy	• Network delays • Speech recognition more prone to error
Tolerant of errors	• Error prevention • Frustration from repetitious prompts adding no information
User control	• Flexibility and efficiency • Option to return to main menu at any time
Consistency	• Consistency in operation of voice system • VUI must be consistent with corresponding screen user interface

Correlation: Three widely used sets of 'traditional usability heuristics' were identified to correlate with the VUI heuristics. These included Nielsen [22] (see Sect. 2.3), Shneiderman and Plaisant [23] and Norman [24] (see Appendix). (The design principles of Shneiderman and Norman are listed in the Appendix. These were compared with the VUI usability themes to help validate them as a basis for final definition.

Explication: the heuristics were described using a standard template including an identifying number, name, definition, explanation and example (see Table 2). Cross-references with the literature are also shown in brackets.

The heuristics were evaluated to measure their effectiveness. Two groups of participants assumed the roles of usability experts, one group of eight was provided with Nielsen's general heuristics while the second group of eight used the VUI-specific heuristics. Participants were asked to complete usage scenarios with three speech-based systems to cover a range of usage contexts and identify as many usability problems with each as possible. The three systems included the Amazon Echo smart speaker, iPhone Siri voice assistant and the VW in-car hands-free kit for call acceptance, phone number selection by voice, and voice control of a media player. The participants then completed a debrief questionnaire asking them to state the advantages and disadvantages of using the heuristics they were allocated (traditional or enhanced) to identify which specific heuristics, in their set, that they found particularly useful.

Figure 1 shows that, on average, 13.5 usability problems per person were identified using Nielsen's general heuristics while 18.25 problems per person were identified with the VUI-specific heuristics.

An independent samples t-test showed that the difference in mean values was significant to the 1% level. This corresponds with the range of problems found by each

Table 2. Developed VUI Usability Heuristics

Number	Description
1	**Minimise short term memory load:** The user's short-term memory load should be kept to a minimum. In the absence of a companion screen display, listed information should be kept short and concise, containing only information necessary to the action being performed. The complexity of concepts the user must understand and the number of things they must learn to use the system must also be kept to a minimum. [12, 13, 22, 23] *Examples: Information that is listed shall be kept short and concise; Allow user to request repetition of previous system output*
2	**Accommodate conversational speech:** The system should speak in a natural way and adopt human-to-human speech conventions. This acts to increase the interaction flow and comprehension. [6, 15] *Examples: Ensure natural conversational flow including turn-taking, following conversation pragmatics and using a friendly tone and manner; Use terms that the user will understand; Be able to understand variations in dialect*
3	**Maximise efficiency:** Users want speed and efficiency. The fewer the number of steps that user-system dialog requires, the greater the perceived efficiency of the interaction with the system. [7, 15, 22] *Example: An action should not be broken into too many steps; Craft the interface for speech rather than try to create an auditory version of a GUI*
4	**Ensure adequate system feedback:** The system should always keep the user informed about what is going on through appropriate feedback within a reasonable time, providing, if necessary, confirmation of actions. [3, 15, 23, 24] *Examples: The system should avoid periods of silence during interaction and should provide confirmation of actions; In response to a user question, give feedback to confirm what question the system is answering; Allow processing of backchannel utterances and background noise*
5	**Ensure high accuracy to minimize input errors:** Recognition is important since errors degrade usability and lead to user frustration. [2, 6, 8, 9, 14] *Examples: The system should be enough to allow for natural speech with few requirements for the user to repeat utterances. Ensure noise in the environment does not interfere with the speech system; Users can tolerate a small amount of repetition of speech input if the system fails to understand it the first time*
6	**Recover from errors:** Users become confused and frustrated when errors occur. The system should enable easy recovery from errors and offer guidance to the user on how they can correct it. [22, 23] *Example: The system should provide error responses relevant to the error which has occurred and provide context to any error*
7	**Provide ability to control and interrupt:** The system should allow the user to interrupt if routed to a path they do not wish to follow. [15, 22, 23] *Example: The user can either interrupt with a new interaction or simply say 'stop'*
8	**Consistency and standards:** Users should be able to maintain their focus on one interface or the link to a second interface (e.g. screen display) should be clear and consistent in operation. [9, 22–24] *Example: The system could take a user's food order and then repeat it back or invite them to review it on screen*

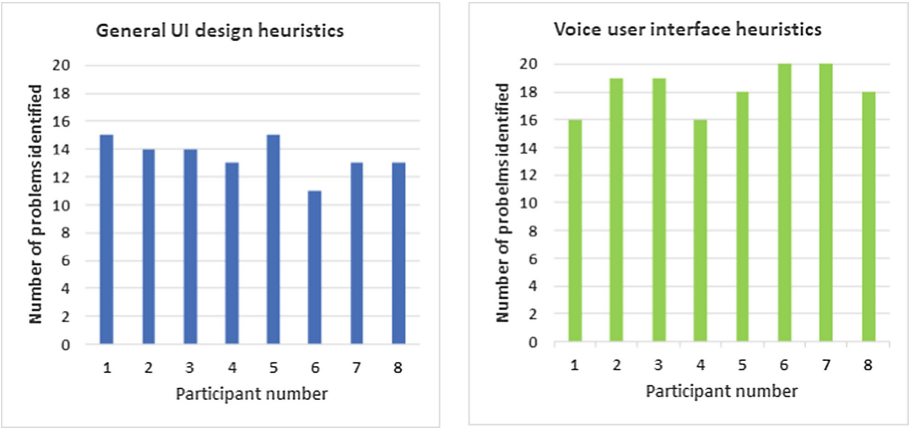

Fig. 1. Total number of problems identified per participant using either General or VUI heuristics

participant for each method: 11 to 15 for the general set of heuristics and 16 to 20 for the VUI-related set.

In a post-test questionnaire, participants were asked what they thought were the advantages and disadvantages for the set of heuristics they used. Interestingly, the benefits for both were similar: making the evaluation process easier and quicker, providing a structure or checklist to work to, and helping to identify problems that would otherwise not have been thought about. Fewer negative comments were received for either set. For the general heuristics, it was said that they were quite broad and could be better phrased to suit voice systems. For the VUI heuristics, the descriptions were thought to potentially limit or to narrow thinking during the evaluation and could include more general aspects of user interfaces.

Participants were asked to identify any heuristics that they had used during the evaluation that they found particularly useful. Figure 2 is a histogram showing the frequency with which each heuristic was selected. Most of the heuristics in both the general and VUI sets were regarded as especially useful by at least one participant. For the general set (Nielsen): (1) 'visibility of system status', and (9) 'help users recognise, diagnose, and recover from errors', were the most frequently cited. This may be an indication that people using VUIs still expect such applications to provide help to overcome errors although they might prefer that the help be integrated into the interaction dialogue and not be a separate task. This finding was also seen in a study by Bertini et al. [25] when assessing heuristics for use with mobile computing. Heuristic (8) 'aesthetic and minimalist design', was the only one not mentioned by any of the participants. While this might seem appropriate as it is normally applied to screen-based systems, it could be applicable to VUI in relation to the aesthetic qualities of the voice and adopting an efficient and effective conversational style.

Regarding the VUI heuristics, those referred to more frequently were: (2) 'adopt conversational speech', (5) 'achieve high accuracy', (7) 'support user control and interruption' and (8) 'consistency and standards'. Heuristic (4) 'ensure adequate system

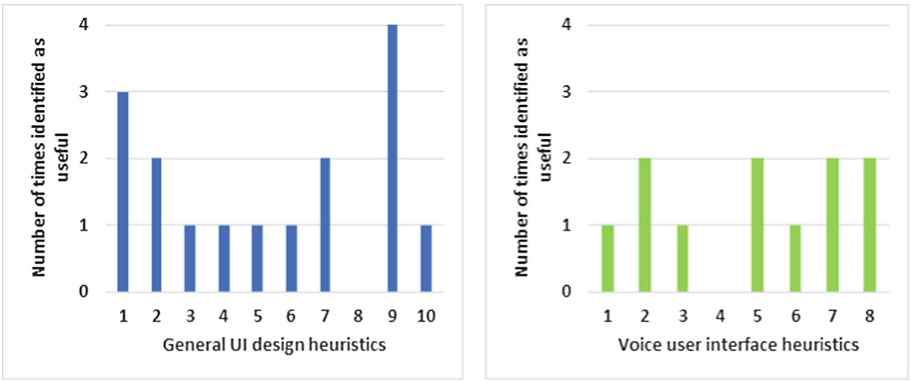

Fig. 2. Number of times each heuristic was chosen as being particularly useful

feedback' was not selected. This is surprising since this is central aspect of conversational interaction.

A limitation of the study was that while all participants did have knowledge of IT and usability, this may have been of variable level, so that some were better able to apply the heuristics evaluation method and identify problems than others.

5 Discussion

The finding that more usability problems were identified by participants who used VUI specific heuristics in comparison with general heuristics is consistent with similar studies looking into the effectiveness of application specific usability heuristics in other domains. For example, Inostroza et al. [26], who studied the use of heuristics for touch screen mobile devices, identified that evaluators using specific heuristics for these devices were able to identify more usability problems than the evaluators that used Nielsen's heuristics. However, the fact that many of the general heuristics were regarded as especially useful during the VUI evaluation study shows that there are other aspects of user interfaces, not necessarily speech specific, that should be included in an evaluation. This may mean that using a combination of the general and application specific heuristics is a more effective approach for conducting an evaluation of an application for a specific domain.

In can be said then that general usability heuristics clearly apply to VUIs e.g. minimising short term memory load and give suitable feedback after user input. However, there are other aspects for which heuristics can be generated e.g. following natural conversational conventions, that should also be covered by heuristics. Since voice interaction is also closely related to intelligent systems using natural language, this raises the issue of whether artificial intelligence related heuristics are also needed, such as whether system outputs show 'common sense' or reflect a knowledge of the real world. The Nielsen and Molistic heuristic 'match between the system and the real world' could be applicable here.

Broader aspects of VUIs that arose from the pilot study conducted in the student house relate to trust and social context. An extension of the heuristics may be needed to address them.

6 Conclusion

This study has investigated and evaluated usability heuristics specifically for auditory VUIs where interaction is conducted solely through voice. The usability heuristics were also generated following heuristic development method in a systematic way. It was anticipated at the start of the study that when conducting the usability evaluations, participants who used Nielsen's general heuristics would find many of them less relevant compared to participants who used the VUI-specific heuristics. However, the findings showed that participants using either set of heuristics found most of the heuristics available to them were useful. This may mean that the optimum design of an evaluation tool for usability inspection is one that combines both general and application specific heuristics. A modular approach could therefore be adopted where subsets of heuristics can be chosen to match a specific evaluation context.

Continuing from this study, further iterative development of the heuristics could be undertaken to reflect new developments in applying speech interfaces to intelligent systems and considering the broader contextual or environmental issues where VUI systems are implemented.

Acknowledgement. The author would like to acknowledge the work of the Loughborough Design School students, Simon Hughes and Daniel Essom, which this paper draws upon.

Appendix

Ben Shneiderman's eight golden rules for interface design:

1. Strive for consistency
2. Enable frequent users to use shortcuts
3. Offer informative feedback
4. Design dialog to yield closure
5. Offer simple error handling
6. Permit easy reversal of actions
7. Support internal locus of control
8. Reduce short-term memory load

Don Norman's principles of interaction design:

1. Visibility
2. Feedback
3. Constraints

4. Mapping
5. Consistency
6. Affordance

References

1. Jurafsky, D., Martin, J.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2nd edn. Pearson-Prentice Hall, Upper Saddle River (2009)
2. Xiong, W., et al.: Toward human parity in conversational speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(12), 2410–2423 (2017)
3. Bernholz, W.: VUI: Voice user interfaces and the future of voice in mobile apps (2017). https://www.dropsource.com/blog/vui-voice-user-interfaces-and-the-future-of-voice-in-mobile-apps/. Accessed 03 Mar 2019
4. Asher, M.: The history of user interfaces–and where they are heading (2017). https://www.cmo.com/features/articles/2017/7/20/a-brief-history-of-ui-and-whats-coming.html#gs.tcqMwFIo. Accessed 03 Mar 2019
5. Jones, D., Hapeshi, K., Frankish, C.: Design guidelines for speech recognition interfaces. Appl. Ergon. **20**(1), 47–52 (1989)
6. Cohen, M., Giangola, J., Balogh, J.: Voice User Interface Design, 1st edn. Addison-Wesley, Boston (2004)
7. Harris, R.: Voice Interaction Design, 1st edn. Morgan Kaufmann, San Francisco (2005)
8. Whitenton, K.: Voice interaction UX: brave new world…same old story. Nielsen Norman Group (2016). https://www.nngroup.com/articles/voice-interaction-ux/. Accessed 03 Mar 2019
9. Asthana, S., Singh, P., Singh, A.: Assessing designs of interactive voice response systems for better usability. In: Marcus, A. (ed.) DUXU 2013. LNCS, vol. 8012, pp. 183–192. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39229-0_21
10. Howell, M., Love, S., Turner, M.: Visualisation improves the usability of voice-operated mobile phone services. Int. J. Hum.-Comput. Stud. **64**(8), 754–769 (2006)
11. Franzke, M., Marx, A., Roberts, T., Engelbeck, G.: Is speech recognition usable? ACM SIGCHI Bull. **25**(3), 49–51 (1993)
12. Damper, R., Gladstone, K.: Experiences of usability evaluation of the IMAGINE speech-based interaction system. Int. J. Speech Technol. **9**(1–2), 41–50 (2006)
13. Kim, H.-C.: An experimental study to explore usability problems of interactive voice response systems. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ACIIDS 2012. LNCS (LNAI), vol. 7198, pp. 169–177. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28493-9_19
14. Portet, F., Vacher, M., Golanski, C., Roux, C., Meillon, B.: Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects. Pers. Ubiquitous Comput. **17**(1), 127–144 (2011)
15. Yankelovich, N., Levow, G., Marx, M.: Designing SpeechActs. In: CHI 1995 Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems, pp. 369–376. ACM Press, New York (1995)
16. Nielsen, J., Molich, R.: Heuristic evaluation of user interfaces. In: CHI 1990 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Seattle, Washington, pp. 249–256 (1990)

17. Quiñones, D., Rusu, C.: How to develop usability heuristics: a systematic literature review. Comput. Stand. Interfaces **53**, 89–122 (2017)
18. Johnson, G.I., Coventry, L.: "You talking to me?" Exploring voice in self-service user interfaces. Int. J. Hum.-Comput. Interact. **13**(2), 161–186 (2009)
19. Milanesi, C.: Voice Assistant Anyone? Yes please, but not in public! (2016). http://creativestrategies.com/voice-assistant-anyone-yes-please-but-not-in-public/. Accessed 03 Mar 2019
20. Adobe Digital Insights: State of voice assistants (2018). https://www.slideshare.net/adobe/adi-state-of-voice-assistants-113779956. Accessed 03 Mar 2019
21. Rusu, C., Roncagliolo, S., Rusu, V., Collazos, C.: A methodology to establish usability heuristics. In: ACHI 2011: The Fourth International Conference on Advances in Computer-Human Interactions, pp. 59–62 (2011)
22. Nielsen, J.: Usability Inspection Methods. Wiley, New York (1994)
23. Shneiderman, B., Plaisant, C.: Designing the User Interface, 1st edn. Addison-Wesley, Upper Saddle River (2010)
24. Norman, D.: The Design of Everyday Things. Basic Books, New York (2013)
25. Bertini, E., Gabrielli, S., Kimani, S.: Appropriating and assessing heuristics for mobile computing. In: Proceedings of the Working Conference on Advanced Visual Interfaces - AVI 2006 (2006)
26. Inostroza, R., Rusu, C., Roncagliolo, S., Jimenez, C., Rusu, V.: Usability heuristics for touchscreen-based mobile devices. In: ITNG 2012 Proceedings of the Ninth International Conference on Information Technology - New Generations (2012)