Voice and Graphical User Interfaces: Design Issues

Tomáš Nestorovič

University of West Bohemia in Pilsen
nestorov@kiv.zcu.cz

Abstract: This paper focuses user interfaces usability and design aspects. Above all, it discusses user interfaces as the top-level asset in speech signal processing. However, a short comparison of speech interfaces to traditional graphical interfaces is made. As such, the paper aims to provide a comprehensive overview of what the term "good user interface design" means. Our point of view consists of four stages that cover the interfaces from different points – logic, presentation, feedback, and consistency. Last but not least, the paper concerns psychological aspects of human-computer interaction, which are often neglected by speech user interface designers.

1 Introduction

Each interactive processing system offers the user an interface to engage in communication with it. Through that interface, mutual feedback is established. Its aim is obvious – getting both sides to understand each other. There is a wide range of *user interfaces* (UIs) and a particular decision is always a matter of application and conditions put on that application. In fact, the conditions may be divided into two groups: direct and indirect. Among the *direct conditions*, we can count in all explicit demands for the UI, e.g., structure of a dialogue (thus, defining and clustering functions into menus), or adaptability to the user (thus, his wishes, his skills). In contrast to direct, the *indirect conditions* dictate implicit demands, i.e., assumptions that an UI designer cannot affect and that always enter the UI design phase – e.g., focus group of users (thus, reflection of their interaction habits and skills) or environment wake influence (its variability and resulting *cocktail party effect*).

Unexpectedly fast spreading of computers was accompanied with evolution of UIs – from command terminals, through semi-character and graphical UIs (*Graphical UIs, GUIs*), to voice interfaces (*Voice UIs, VUIs*), usually accompanying GUIs to give an alternative in the interaction with the system. However, the VUIs seem to be a promising way of interaction between human and a computer (*Human-Computer Interaction, HCI*), especially due to increased naturalness from the human's point of view and higher information exchange effectiveness [1, 2]. Despite all its advantages, it cannot be doubted that the GUIs are more preferred in today's commercial applications [1]. The reasons for this paradox can be found in probable immaturity of speech recognition methods – reliability and performance [3, 4, 5, 6].

The rest of the paper is divided as follows. First, we will concern the graphical interfaces, their short history and design rules (Section 2). Providing GUIs pros and cons, we move to the voice interfaces, summarizing empirical knowledge about their design (Section 3). Finally, Psychological aspects of voice interaction are discussed (Section 4). The paper is concluded with summarization of this paper (Section 5).

2 Graphical User Interfaces (GUIs)

Semi-graphical and graphical interfaces emerged to solve them command terminals poor intuition and cumbersomeness problems. It is known [7] that learn quicker rather by recognizing visually known than by strict memorizing. An application equipped with graphical interface communicates with the user through *graphical controls*, laid out on its

window. This concept of representation using windows, icons, menus and pointers is abbreviated as the *WIMP concept* [8, 9].

2.1 Designing a GUI

As Hobart points out, the issue of today's GUIs is that only a low ratio of applications exhibit traces of good graphical interface design [7]. Nevertheless, as he in turn adds, it is extraordinary hard to delimit the notion of the term "good graphical interface design." However, there is a wide collection of recommendations and restrictions on how to get people to orientate better in an application. One of these is the *real world metaphor rule* [8] – a button with MasterCard logo most probably determines a place where to make an instant payment in the application.

In overall, graphical interface design is governed by the vague "look & feel" concept. References [7, 8] provide a comprehensive listing of recommendations that, if considered, help make the term more specific. They can be divided into the following categories.

2.1.1 Logics

This group accounts for properties like self-navigation skills (i.e., logics of the controls layout) and logics of operations carried out by the application back-end. Additionally, this category contains empirical rules regarding, for example, the position and relationship of controls (their significance gets lower "from left to right" and "top down" within a window), and the real world metaphor rule we mentioned above. Logic is also supported by the information being split into graphical dialogues. These dialogue windows are of twofold functionality – on one hand they catch user's attention [6], while on the other they put fragmented information into logically self-contained (and contextually dependent) blocks.

2.1.2 Presentation

This field regards visual properties of a GUI, e.g. controls layout *symmetry*, thus *optical lucidity* in general. Martínez summarizes peoples effort in putting vague terms like "handsome" or "beauty" into formal expressions. In his work [12], apart of the symmetry he also mentions the significance of the *golden section* (as one of the attributes of "handsome" and "beauty" is strict relation to the golden number ~ 0.618). Additionally, he writes about *uniformity* (breaking it, we can catch user's attention to a particular control) or *rhythm* (distance between controls). Purely graphical presentation can be extended with sound effects, however, ideally only in important situations [8]. Hobart in his work [7] shows an example of a wrong way designed GUI and presents solutions how to improve it.

2.1.3 Feedback

Through a feedback, an application expresses the acceptance of user's demand – that is why the feedback is of high importance. Considerable amount of users' frustration is caused by a bad or no feedback from an application [7]. Additionally as Hobart suggests, if a desired operation cannot be carried out within several seconds, it is mandatory to catch the user's attention (e.g. using a progress bar showing the computational state in percents). However, the number of seconds is an individual property of each user. Nevertheless, a general empirical rule dictates to prioritize processing of front-end operations at the expense of back-end.

2.1.4 Consistency

The consistency is important and covers a wide spectrum of interest – from consistency of used terminology, through consistency in functionality, to presentation consistency. Having

several sequel versions of a product, the consistency of the versions in the sequel is important also, since introducing a difference from any of the previous versions is generally not positively reckoned by the user's [8].

2.2 Common User Interface (CUA)

Currently, the design of graphical interfaces is highly affected by the *Common User Interface* (CUA) document, devised by IBM in 1987. The original aim of the document was to unify user, communication, and program interfaces within the IBM's *System Application Architecture* (SAA) platform. The document is defined in the following to publications:

- Systems Application Architecture Common User Access Guide to User Interface Design (SC34-4289-00)
- Systems Application Architecture Common User Access Advanced Interface Design Reference (SC34-4290-00)

The motivation to create the CUA document was the incongruity of user interfaces in the 1980s. The two publications above, therefore, not only cover all rules mentioned above, but moreover suggest rules to other design aspects, e.g. application key shortcuts or the order of actions in roll-down menus. Apart of it, the documents also define particular GUI elements describe their optimal usage and interaction. Finally, the CUA document is also adopted by several operation systems (OS/2, MacOS, and Windows), and is recommended to all applications that run under these systems. The goal of this step is to rise the users' prediction of each action results, thus speeding up users' familiarization with such applications [13]. The importance of the CUA document can be sensed if any of its recommendations is not met (e.g. swapped position of “Ok” and “Cancel” buttons in the Linux operation systems).

2.3 The drawbacks of GUIs

Graphical dialogue with an application is accompanied with a lot of advantages like high level of intuition. However despite that fact, having an application with GUI means to be unable to engage in communication with the computer for a considerable amount of people. These handicapped people are inhibited from using traditional *human interface devices* (HID) – mouse and keyboard. We can count in this group people with reduced finger motory skills [3, 11] or bad eyes [15]. Surely, GUIs can be extended with additional features to enable (partially) handicapped people to use them.[1] However, for a lot of people, these components seem to be insufficient, e.g., elder patients bound to a wheelchair [11].

3 Voice User Interfaces (VUIs)

Speech evolved through thousands of years seems to be the most effective and most natural means to share our thoughts [2, 11]. It therefore is not a surprise that voice user interfaces (VUIs), employing speech in communication with computers, seem to be a promising way of how to get the issue above solved.

3.1 Designin a VUI

Researches' attitudes on how to design a good VUI differ. While ones maintain the idea that their design *can* be governed by long-term experience with GUIs, the others [6] are of diametrically different opinion arguing that the resulting systems will be unusable. Finally,

1 For instance, commercial operation systems offer a set of functions usually referred to as accessibility, or a possibility to install a “reading agent,” thus a module transforming a text close to current position of a cursor into a speech output. Both extensions definitely make the communication through a GUI more effective, however, concern only people with enough motory skills, and enough eyes, respectively.

Bradburn [11] keeps relatively neutral saying that experience with the GUIs may be employed, however, it has to be a subject of a revision prior to application to any of speech system. Due to this diversity, the rules for creating a VUI are rather of fuzzy notion. Nevertheless, we still can divide them into the four groups like we did in case of the GUIs.

3.1.1 Logics

System response adequacy or dialogue flow logics are one of aspects of this group. A *dialogue system* (DS) accommodating a VUI is always inaccurate due to *automatic speech recognition* (ASR) imperfectness. From this point of view, the DS always needs to confirm each part of transferred information from the user to prevent the situation that a dialogue with it becomes unlogical (getting the user frustrated, see below in Section ???). Another factor accompanied with a successful VUI is the order of information elicited from the system, i.e., *dialogue management logics*. In majority of cases, the flow from more general concepts to the most specific concepts is recommended and preferred. This principle is in accordance with the general rule of information *constraining*. For example, in the car bazaar named Wheels [17] found the optimal order of car selection constraint (brand, type, color etc.) by questioning random users. Though the authors did not support the dialogue convergence to a successful end (presenting at most 5 cars to the user), they made the dialogue feel most logical to statistical majority of users.

Another of issues affecting the logics is the humans indirect meaning when speaking. The system, therefore, always should attempt to extract hidden intentions from user's utterance, instead of trying to fulfill the strict pragmatical intentions (e.g., “Do you know what time it is?” compared to “Say me what time it is.”). However, most of this theory adheres to the *plan-based dialogue management* which is considered as a highly vague and impractical means for implementing a human-computer speech communication channel. The reasons are that this kind of management lacks of enough formal grounds [18], and system's detection of human's plans is unreliable [19].

3.1.2 Presentation

The presentation criteria put more emphasis on *form* of system's utterances rather than the content, which determines aspects like proper speech timbre, intonation or accent, thus *prosody* in overall. Designer's first essential decision when creating a DS is whether synthetic or natural voice should be employed to produce system's responses. The current state of the technologies and methods is insufficient to produce naturally sounding utterances for the first case [6] (synthetic voice). On the other hand, the parameters of prosody can be adjusted easily, thus even an imperfect synthetic voice can be extended with simple emotive characteristics [20] (welcoming prompt can be said brightly, while informing the user that the stock is out of the desired item can be said sadly). The synthetic voice is recommended for DSs with high vocabulary variability that cannot be easily produced by a reasonable amount of natural speech records (e.g. business portals). As for the second option (natural speech), apparently, the issues with naturalness and emotion approaching are solved, however, new introduced – constant speech approaching, speech fluency, or segments sequent.

3.1.3 Feedback

Similarly as with GUI, the feedback plays a crucial role in VUI as well. It is mainly affected by the system's ability to react to the user's utterance in a reasonable time (ASR is the main issue), otherwise unnatural delays emerge [1, 6]. It seems to be practical to fill these gaps with another sound that catches the user's attention, informing her/him that the system is busy. The feedback does not, however, only consist of “some utterance.” In this case, the *content* is

essential. The appropriate content creation is described and discussed in [5] – e.g., one of them is the *incremental feedback* protecting an experienced user from long detailed utterances, while providing a novice with exhaustive prompts. Another suggestion is to avoid explicit lists of possibilities as they push a user into strict boundaries, suppressing his own initiative [6]. Additionally, the system should avoid repetitive expressions and replace them with naturally sounding equivalents [23] (i.e., instead of saying "The NASDAQ stock value raised 1.3% since yesterday. The Petroleum stock value raised 0.5% since yesterday. …" the system should say "The NASDAQ stock value raised 1.3% since yesterday, Petroleum 0.5, …").

One of another essential parts of dialogue management is the necessity to confirm each piece of information gained from the user. This was already mentioned earlier as a result of the ASR imperfectness. In fact, there are two ways how to implement a confirmation process in the DS – explicitly, or implicitly [24, 25]. However, in distinguished cases in which the user's new information does not affect the dialogue context in a critical way, the feedback can be omitted [6, 26]. In these cases, a direct system's response can substitute the confirmation.

3.1.4 Consistency

In case of voice interfaces, the consistency focuses mainly on the system's "person," i.e., its self-presentation and behaviour. For example,. we can count in the consistency of way of system voice production (synthetic and natural voice should not be mixed [27]), or consistency of voice prosody with utterance content [20]. The consistency also covers the system functionality – DSs usually offer a set of commands available throughout the entire session with an user (help or "go back" functions). These functions should always be accessible the same way to the user. For example, in [28] is demonstrated a simple multimodal navigation system where both modalities hold the same structure of menus to meet the demand of consistency (thus, the user does not need to learn two ways of interaction, as one is sufficient to manage both modalities).

4 Psychological Aspects in Human-Computer Interaction

The users interacting with DS voice interfaces are affected by individual psychological aspects, obviously arising when communicating with an artificial dialogue participant. From the past empirical research is clear that users communicate to the machines in a different way than they communicate to each other [29] – more structured way [1, 30]. The reason can be that people perceive the speech as its high and most natural means, and get the "speaking machine" as something unnatural and unobserved [22]. Another researchers extend this though with claiming that this is the reason why DS imitating humans natural habits are unreachable goal [10]. For example, user's of the voice controlled drawing application MacDraw [15] complained about the unnaturalness of the VUI. On the other hand, other researches stay in opposition to these conclusions, arguing that naturally acting DSs are reachable if they allow the users to employ the habits they learned during the natural human-human interaction [2, 4].

Several studies also prove that frustration emerges during the HCI for some users, especially as a result of improper structure of a dialogue [14], and apparent noncooperation of the system when fulfilling the user's demands [6]. The wrong design of an interaction is supported not only by its flow (e.g. redundancy of confirmations) but also formulation of utterances (unreasonably long and yet undescriptive or ambiguous utterances – a problem that share both VUI and GUI [7]). Restructuring or reformulating these utterances can reduce the frustration [14], though due to the ASR issues it is inevitable. As [6] shows, it is psychically exhaustive for an user to be unable to create a deterministic model of system's behaviour –

while in one session the system understands and correctly reacts to the user's request, in another session the identical user's utterance results into system's unexpected behaviour.

Some of the studies focused on expressive level of HCI. For example, people formulate their thoughts into more structured and shorter utterances when communicating with a machine [16]. If the users are exposed more modalities, their spoken utterances exhibit yet more reduction in comparison with pure VUIs, and contain 50% less disfluencies (i.e., false start, repetitions, or corrections) [2]. Gustavson [1] points out that humans interaction exhibits lower lexical variation, i.e., lower number of synonyms. It can be felt like people were talking in a command line way to machines – this is, however, something that machines cannot understand if they are build to behave in a natural way [22]. On the other hand, Zue and Glass [24] welcome the more structured way of interaction and argue with practical reasons they do not specify any closer. Nevertheless, the reasons most probably reflect the ASR issues. For example, authors of the voice controlled web browser [3] have designed the VUI as much structured as possible to go hand in hand with technologies capable to recognize just a minor portion of natural language. The mentioned VUI accommodates commands like "Go Back", "Follow Link <*number*>" etc., thus follows the well-known way of structured interaction the users' learned from interacting with GUIs. In fact, this application is also in accordance with another empirical rule stating that the phrases and terminology used by a system is a good guidance for the user to know how to express [1]. Therefore, the user attempts to adopt and imitate the partner's conversational style and habits [5]. If their utterance is not followed by expected system's response, they tend to change the intonation or reformulate their original sentence [21]. In case of multimodal systems, the system's improper reaction is usually a trigger for the users to change the modality [1].

Users are a priory unable to determine the level of complexity they may talk with to a DS. However, the quality of the DS self-presentation (no matter if verbal or multimodal) gives them a start guess of the system skills. In the final evaluation they either may the system overestimate or underestimate [5] – depending on how much the system is humanized by its designers. Seeing the problem from the opposite point of view, for a DS designer it is tough to predict and cover all possible utterances the users may make when interacting with a system.

From the above list of observations it is clear that people feel the extraordinaire situation when talking to a machine as an artificial conversational partner. These circumstances in which the system attempt to approach natural human behaviour and skills lead to change of their interaction style, as announced in the beginning of this section. We surely cannot generalize this attitude, however, we can delimit the most important two issues in the HCI:

- The users do not trust the uncommon (and speaking machine is uncommon).
- The users are unable to estimate the machine skills.

5 Conclusion

This paper aimed to put several essential design rules together and provide a clue how to create a good graphical and foremost voice user interfaces. It also presented the people's attitude to situation when machines' input and output is a natural speech. As a logically implying continuation, the research in affecting the humans' attitude by introducing an artificial avatar seems to be a reasonable future work.

Acknowledgement

This work was supported by grant no. 2C06009 Cot-Sewing.

Literatur

[1] Gustavson, J.: Developing Multimodal Spoken Dialogue Systems – Empirical Studies of Spoken Human-Computer Interaction. Ph.D. thesis, KTH, Department of Speech, Music and Hearing, 2002.

[2] Oviatt, S.: Multimodal Interfaces. In: Handbook of Human-Computer Interaction, Lawrence Earlbaum, New Jersey, 2002.

[3] Christian, K., et al.: A Comparison of Voice Controlled and Mouse Controlled Web Browsing. In: Proc. of Assets 2000, Arlington 2000, pp. 72-79.

[4] Hurtig, T., Jokinen, K.: On Multimodal Route Navigation in PDAs. In: Proc. of the 2nd Baltic Conference on Human Language Technologies. Tallin 2005, pp. 261-266.

[5] Yankelovich, N.: How do users know what to say? In: Interactions, Vol. 3, 1996, pp. 32-43.

[6] Yankelovich, N., Levow, G.-A., Marx, M.: Designing SpeechActs: Issues in Speech User Interfaces. In: SIGCHI`95, Human Factors in Computing Systems Proceedings. Denver 1995. pp. 369-376.

[7] Hobart, J.: Principals of Good GUI Design. Classic System Solutions. October 1995. http://www.classicsys.com/css06/cfm/article.cfm?articleid=20

[8] Dorey, S. J.: Common User Access – Principles of GUI Design. 2007. http://www.susandoreydesigns.com/software/CommonUserAccessGUIDesign.pdf

[9] Turk, M., Robertson, J.: Perceptual User Interfaces. In: Communications of the ACM, Vol. 43, 2000, pp. 32-34.

[10] Nils, D., Jonsson, A.: An empirically based computationally trackable dialogue model. In: Proc. of the 14th Annual Meeting of The Comgnitive Science Society, Bloomington, USA, 1992, pp. 783-780.

[11] Bradburn, K.: Issues and Guidelines for Designing Speech-Based User Interfaces. http://homepages.wmich.edu/~m0thiaga/kbradburn/kbradburn_issuesandguidelines.doc

[12] Martínez, J.S.: Lze popsat krásu matematickým výrazem? In: Umělá inteligence (5), Academia, Praha, 2007, ISBN 80-200-1445-4.

[13] Benson, C., et al: GNOME Human Interface Guidelines 2.0 – The GNOME Usability Project. 2004. http://developer.gnome.org/projects/gup/hig/2.0/hig-2.0.pdf

[14] Klemmer, S.R.: et al: SUEDE: A Wizard of Oz Prototyping Tool for Speech User Interfaces. In: Proc. of The 13th Annual ACM Symposium on User Interface Software and Technology (UIST), Vol. 2, 2000, pp. 1-10.

[15] Pausch, R., Leatherby, J.H.: An Empirical Study: Adding Voice Input to a Graphical Editor. In: Journal of the Americal Voice Input/Output Society, Vol. 9(2), 1991, pp. 55-66.

[16] Churcher, G.E., Atwell, E.S., Souter, C.: Dialogue management systems: A survey and overview. Report 97.06, University of Leeds, School of Computer Studies, Leeds, 1997.

[17] Meng, H.: Wheels: A Conversational System In The Automobile Classifieds Domain.“ In Proc. of ICSLP, 1996, pp. 542-545.

[18] Wilks, Y., Catizone, R., Turunen, M.: Dialogue management: State of the Art Papers, COMPANIONS Consortium, 2006.

[19] Bui, T.H.: Multimodal Dialogue Management – State of the art. Technical Report TR-CTIT-06-01 Centre for Telematics and Information Technology, University of Twente, Enschede 2006. ISSN 1381-3625.

[20] Nass, C., Ulla, G.F., Somoza, M.: The Effect of Emotion of Voice in Synthesized and Recorded Speech. http://www.stanford.edu/~nass/comm369/pdf/VoiceandEmotion.pdf

[21] Gustavson, J., Bell, L.: Speech technology on trial: experiences from the August system. Journal of Natural Language Engineering: Special Issue on Best Practice in Spoken Dialogue Systems, vol. 3. September 2003, pp. 273-286. ISSN:1351-3249.

[22] Edlund, J., et al: Two faces of spoken dialogue systems.“ In: Interspeech 2006 – ICSLP

Satellite Workshop Dialogue on Dialogues: Multidisciplinary Evaluation of Advanced Speech-based Interactive Systems, Interspeech 2006. Pittsburgh PA, USA, 2006.
[23] Nestorovič, T.: Grammar-Based Dialogue Management. In: Proc. of 9th International PhD Worskhop on Systems and Control. Izola 2008.
[24] Zue, V.W., Glass, J.R.: Conversational interfaces: Advantages and challenges. In: IEEE vol. 80, 2000, pp. 1166-1180.
[25] van Zanten, G.V.: Adaptive mixed-initiative dialogue management. In: Proc. of IVTTA, 1998.
[26] Matoušek, V., a Nestorovič, T.: Návrh hlasové komunikace s navigačním systémem automobilu a její implementace v jazyce VoiceXML. In: Proc. of NavAge, Prague, 2006.
[27] Nass, C., Simard, C., Takhteyev, Y.: Should Recorded and Synthesized Speech be Mixed?“ 2004. http://www.stanford.edu/~nass/comm369/pdf/MixingTTSandRecordedSpeech.pdf
[28] Nestorovič, T.: Navigation System: An Experiment. In: Proc. of NAG/DAGA, Rotterdam, 2009.
[29] Cohen, P.: Dialogue modeling, Survey of the State of the Art. In: Human Language Technology. Cambridge University Press, 1987, Cambridge.
[30] McGlashan, S.: Towards Multimodal Dialogue Management. In: Proc. of 11th Twente Workshop on Language Technology, Twente 1996. pp. 13-22.