Voice Interaction On Mobile Devices And Ways To Ensure Better User Experience.

Sara Ladner
Fachhochschule St. Pölten
it231509@fhstp.ac.at

ABSTRACT

This paper explores the use of voice interaction on mobile devices and other Voice User Interfaces (VUIs). It addresses challenges in current VUI designs, emphasizing the need for effective communication of system capabilities. The advantages and challenges of voice interaction on mobile devices are discussed, including various voice recognition methods. Guidelines for enhancing VUIs are introduced, focusing on tailoring conversations, addressing turn-taking challenges and prioritizing personalization.

1 INTRODUCTION

Interacting with a device via our voice has become common and Voice User Interfaces (VUIs) have been seamlessly integrated into our daily lives, driven by the widespread adoption of voice assistants like Google Home and Amazon Alexa. We tell phones, cars or even household devices, what we want them to do. As a result, voice assistants are experiencing rapid commercial growth in recent years [1].

If used right, voice interaction has the possibility to enhance user experience with different devices, this is also true for mobile applications. However, due to the complexity of designing accurately understanding flexible natural language commands it is not always easy to implement [2]. In what ways can we refine the voice interaction experience on mobile devices and Voice User Interfaces, overcoming current design obstacles and ensuring seamless communication of system capabilities?

2 VOICE USER INTERFACE (=VUI)

The general term for speech recognition technology that allows people to interact with a computer, smartphone or other device through voice is called Voice User Interface. In the last five years, users' attitudes towards VUIs have improved, with increasing recognition of their diverse applications, such as hands-free interaction and a natural communication with technology [3].

Currently, there is no established software architecture standard for voice assistants. Nonetheless, there are two fundamental operations involved in the processing of user utterances. First, the user's spoken words are transcribed into text using automated speech recognition (ASR). This step is crucial for converting spoken language into a format that can be further analysed and understood by the system. Then, the transcribed text undergoes Natural Language Understanding (NLU), which aims to extract the user's intent from the processed language. NLU systems analyse the text to understand the meaning, context, and specific commands or queries conveyed by the user [4].

3 VOICE INTERACTION ON MOBILE DEVICES

Voice interaction has several advantages, like expressive communication, easy learning, and hands-free usability. Employing voice commands to automate smartphone tasks enhances the interaction on mobile apps. However, creating voice interfaces for specific tasks is challenging due to the difficulty of accurately understanding flexible natural language commands. Developers often need extensive data, handwritten rules, or machine learning models, making efforts grow with the increasing number of tasks. This limitation results in only a few tasks having voice interfaces. Nevertheless, there is a strong desire, especially among non-programmers, to create voice interfaces for various tasks [2].

3.1 Voice Recognition

Nearly every mobile operating system includes access to at least one conversational user interface, typically in the form of an interactive personal agent [5]. Various acoustic signals are received by a microphone during the operation of the voice recognition system on the device, aiming to distinguish human voice from non-voice signals like background noise. Detecting voices accurately is important for the system to work well and understand emotions correctly. Among various methods to find voices, the energy-based one is chosen for mobile devices because it needs less computing power and gives reliable results. This method continuously checks the strength of sounds in fixed-length parts. If the strength of consecutive parts is louder than a certain level, it marks the beginning and end of voice signals [4].

3.2 Detection of emotions

Identified voice signals from the initial stage undergo processing to extract acoustic features that reveal emotional characteristics. Different types of acoustic features can illustrate human emotions, with certain spectral features like pitch, log-energy, and Mel-Frequency Cepstral Coefficient (MFCC) — a representation capturing the spectral content of the voice — being particularly valuable for distinguishing between emotion types. These features, extracted from each frame in the voice region, form a sequence of feature vectors [6].

4 SYSTEM FEEDBACK AND CHALLENGES

Voice User Interfaces are becoming increasingly popular in homes, yet users often underutilize available features. Researchers have identified that the learnability of VUIs poses a significant challenge for users, with the primary hurdle being errors in understanding and responding to natural language. This field involves the interaction between computers and humans using everyday language. Both new and frequent users face difficulties in comprehending and navigating through VUI features and commands, making it challenging to keep up with the introduction of new functionalities [7].

It seems that current VUI designs don't make it easy for users to adjust to how the machines work during a conversation. The designs lack proper feedback and information about what the system can do, making it challenging for users to try new tasks. Instead, users often use shorter sentences or simpler language with repeated phrases to make sure the VUIs understand them. These interactions can be frustrating, with users struggling to accomplish their intended tasks [8].

5 GUIDELINES AND IMPROVEMENTS

5.1 Task Domains and Context

Designing conversations for voice interfaces should be tailored to the task domain, ensuring comprehensive coverage of knowledge within that domain. Handling a wide range of topics in the task domain enhances usefulness and accommodates diverse user preferences. It is crucial for the system to possess sufficient knowledge in the specified task domains to satisfy user needs effectively. Immediate communication of the reason for task failure is essential, and considering user inferences based on the task domain can contribute to a more effective interaction [9].

Furthermore, a study about the transition from GUI (Graphical User Interface) to VUI conducted by Murad et al. (2021, S. 7), found that GUI experts had concerns about the inability of a lot of VUI devices to retain memory or context of previous requests, impacting interaction continuity. Devices lack memory of prior requests, hindering the ability to build on or refer to earlier queries. The absence of an undo capability was also highlighted, causing frustration due to the lack of contextual memory [10].

5.2 Turn-Taking

Phukon et al. (2022) discuss the significance of turn-taking in spoken interactions and the challenges of automating this process in human-computer conversational systems. Turn-taking involves the exchange of speech and silence between participants, and while researchers have developed solutions like silence-based models, which focus on recognizing and managing periods of silence in speech, and continuous models, designed to transcribe

speech seamlessly without explicit word segmentation, finding the right settings for these methods remains a challenge. Overall, automating turn-taking in conversational systems is an ongoing area of research with specific challenges, including avoiding awkward silences or overlapping speech [8].

5.3 Importance of personalizing

It is important to take users' background knowledge and inferences into consideration to tailor conversations according to users' needs and capabilities. There are various methods for achieving this, from learning and adjusting conversations based on user responses, to utilizing the task domain to establish a common ground. The need for adjustment arises from the potential lack of understanding or unpredictable remarks by users, such as questions for clarification, which the system may struggle to comprehend or answer [9].

Once a common ground is established, the guidelines recommend that voice systems should further adapt to each user. Personalization is key, involving the learning of user behaviors, language usage, and specific needs. The goal is to tailor the interaction to individual users, as highlighted by the suggestion to adapt the agent's style based on who the user is, how they speak, and even their emotional state. Creating user profiles is presented as one effective method to achieve this level of personalization [9].

5.4 Transparency

A VUI should distinctly communicate its capabilities and limitations to effectively manage user expectations. Providing this transparency is crucial for establishing accurate mental models and expectations during interactions. Presenting the system's abilities at the beginning and utilizing prompts helps convey the user's potential actions [9].

5.5 Clear Feedback

Clear feedback is paramount for user understanding, encompassing processing status and error notifications. Delivery should be swift and efficient to eliminate uncertainty about the system's comprehension and actions. Non-verbal feedback enhances user interaction. Visibility is key, ensuring the user is always aware of the system's status, especially during various functions. Comprehensive feedback on system failure, including reasons, is essential. User notification during processing or listening phases is vital, with progress indications being specific and unambiguous. The goal is to instill confidence in users regarding the accurate processing of conveyed information [9].

Building on the principles of clear feedback highlighted in usability guidelines and referring to the study from Murad et al. (2021, S. 7), GUI experts assessed two display-less devices, shedding light on usability concerns related to audio and visual feedback. These experts observed the use of colored indicators for actions such as volume adjustments and signaling when the interface was actively listening or processing user input. However, they encountered challenges with both visual and audio feedback, underscoring the imperative for enhanced clarity in the design. Some test participants expressed a preference for a design that minimizes the need for active observation of visual indicators, aligning with the goal of providing swift, efficient, and unambiguous feedback for optimal user confidence, as emphasized in established usability principles [10].

5.6 Make it “real”

Designing conversational interaction in VUIs that align with real-world conversational norms and dialogue patterns makes a VUI seem “realer”. Utilizing well-known conversational norms and allowing users to employ natural speech are essential. Key points include:

1. Speech: The importance of appropriate prosody, clear pronunciation, and user control over speech aspects, such as speed and tone. Non-verbal forms of auditory feedback are also recommended.
2. Realism: Matching interaction to real-world conversation models, user language structure, and incorporating familiar terms. However, realism should be balanced, considering the system's use case.
3. Conversation: Following natural turn-taking, brief responses, and using conversation to guide interaction. Repairing breakdowns collaboratively and

incorporating questions for guidance and common ground.

4. Dialogue: Designing dialogue trees, allowing for response variations, rephrasing prompts for better understanding, accepting a variety of user responses, and handling interruptions appropriately. The system should not interrupt users during input.

The overall goal is to create a VUI interaction that is natural, user-friendly, and aligned with users' expectations of conversation [9].

CONCLUSION

Several challenges with current VUI designs were identified, including difficulties in user adjustment during conversations, lack of proper feedback, and limited information about system capabilities. Users tend to simplify their language to ensure understanding, leading to frustration. This paper emphasizes the need for effective communication of the system's intelligence and capabilities to users. The advantages and challenges of voice interaction on mobile devices are discussed, highlighting the benefits of expressive communication and hands-free usability.

Guidelines for improving VUIs are presented, focusing on tailoring conversations to the task domain, addressing turn-taking challenges, and emphasizing the importance of personalization. Transparency in communicating system capabilities and limitations is underscored, along with the significance of clear and informative feedback to enhance user understanding. In conclusion, there is an ongoing need to persistently enhance VUIs by tackling current challenges and applying guidelines that emphasize user experience, transparency, and efficient communication.

REFERENCES

[1] Luger, E., Sellen, A. (2016). *"Like Having a Really bad PA": The Gulf between User Expectation and Experience of Conversational Agents* (p. 1). ACM Digital Library: https://dl.acm.org/doi/10.1145/2858036.2858288

[2] Pan, L., Yu, C., Li, J., Huang, T., Bi, X., & Shi, Y. (2022, May 6). *Automatically Generating and Improving Voice Command Interface from Operation Sequences on Smartphones* (p. 1-2). ACM Digital Library: https://dl-acm-org.ezproxy.fhstp.ac.at:2443/doi/10.1145/3491102.3517459

[3] Murad, C., Tasnim, H., & Munteanu, C. (2022, July 26-28). *"Voice-First Interfaces in a GUI-First Design World": Barriers and Opportunities to Supporting VUI Designers On-the-Job* (p. 2). ACM Digital Library: https://dl-acm-org.ezproxy.fhstp.ac.at:2443/doi/10.1145/3543829.3543842

[4] Tarakji, A. B., Xu, J., Colmenares, J. A., & Mohomed, I. (2018, June 10-15). *Voice enabling mobile applications with UIVoice* (p. 1-2). ACM Digital Library: https://dl-acm-org.ezproxy.fhstp.ac.at:2443/doi/10.1145/3213344.3213353

[5] Jaber, R., & McMillan, D. (2020, July 22-24). *Conversational User Interfaces on Mobile Devices: Survey* (p. 1). ACM Digital Library: https://dl-acm-org.ezproxy.fhstp.ac.at:2443/doi/10.1145/3405755.3406130

[6] Park, J.-S., & Jang, G.-J. (2015, October 21). *Implementation of Voice Emotion Recognition for Interaction with Mobile Agent* (p. 1). ACM Digital Library: https://dl-acm-org.ezproxy.fhstp.ac.at:2443/doi/10.1145/2814940.2815004

[7] Myers, C. M., Pardo, L. F., Acosta-Ruiz, A., Canossa, A., & Zhu, J. (2021, July 21). *"Try, Try, Try Again:" Sequence Analysis of User Interaction Data with a Voice User Interface* (p. 1). ACM Digital Library: https://dl-acm-org.ezproxy.fhstp.ac.at:2443/doi/10.1145/3469595.3469613

[8] Phukon, M., Shrivastava, A., & Balentine, B. (2022, November 9-11). *Can VUI Turn-Taking Entrain User Behaviours?* (p. 1). ACM Digital Library: https://dl-acm-org.ezproxy.fhstp.ac.at:2443/doi/10.1145/3570211.3570215

[9] Murad, C., Candello, H., & Munteanu, C. (2023, July 19-21). *What's The Talk on VUI Guidelines? A Meta-Analysis of Guidelines for Voice User Interface Design* (p. 1-11). ACM Digital Library: https://dl-acm-org.ezproxy.fhstp.ac.at:2443/doi/10.1145/3571884.3597129

[10] Murad, C., Munteanu, C., Cowan, B.R., Clark, L. (2021, July 27-29). *"Finding a New Voice: Transitioning Designers from GUI to VUI Design* (p. 7). ACM Digital Library: https://dl.acm.org/doi/10.1145/3469595.3469617