Conversation is Multimodal - Thus Conversational User Interfaces should be as well

Stefan Schaffer
Norbert Reithinger
stefan.schaffer@dfki.de
norbert.reithinger@dfki.de
German Research Center for Artificial Intelligence (DFKI)
Berlin, Germany

ABSTRACT

In this paper we try to provoke by teasing the question "if conversational user interfaces should be multimodal?". Of course they should! In decades of research in multimodal HCI excellent arguments can be found. We substantiate our perspective with an example showing how conversational interaction becomes more robust and efficient through the use of multimodality.

CCS CONCEPTS

• **Human-centered computing → Natural language interfaces**.

KEYWORDS

conversational user interfaces, multimodal interaction, speech interfaces

ACM Reference Format:
Stefan Schaffer and Norbert Reithinger. 2019. Conversation is Multimodal - Thus Conversational User Interfaces should be as well. In *1st International Conference on Conversational User Interfaces (CUI 2019), August 22–23, 2019, Dublin, Ireland.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3342775.3342801

1 INTRODUCTION

Conversational User Interfaces (CUI) mimic a conversation with a real human. For decades, HCI research has been correctly arguing that natural language is the main interaction modality for conversational interfaces [3]. However, now that conversational interfaces are pop up everywhere (almost), the question as to whether the interaction with these interfaces should also be multimodal is treated a as new research topic on respective conferences. In this paper we argue why these interfaces should of course support a multimodal way of interaction.

2 FROM HUMAN TO HUMAN

Taking a step back, we first want to analyze what determines a conversation between humans. According to [17], no generally accepted definition of conversation exists, beyond the fact that a conversation involves at least two people talking together. [4] states that conversation is a joint activity in which two or more participants use linguistic forms and nonverbal signals to communicate interactively. Interpersonal communication is an exchange of information between two or more people [1]. Accordingly, communication is considered as the act of conveying meanings from one entity or group to another through the use of mutually understood signs, symbols, and semiotic rules. An important part of human communication is nonverbal communication, which includes body language, visual language, symbolism, touch, music and various forms of expressing oneself without words [6]. Humans also use means of communication in order to make it easier for the communication partner to understand them. Such communication tools range from simple signs illustrating what was said to conceptual models describing complex facts [7]. All in all it can be stated that human beings communicate multimodal between each other [10]. In this view we argue that also a conversation between humans must be considered as multimodal.

3 FROM HUMAN TO COMPUTER

Ever since the appearance of the "put-that-there" paradigm [2], multimodal user interfaces have been a subject of intensive scientific study in the HCI community. So called multimodal dialog systems have been developed with a wide range of research foci (e.g. [5, 16]) and for a variety of applications [9, 15]. It has been shown that multimodal interaction can offer advantages compared to unimodal interaction [12]. Multimodal interfaces allow humans to create inputs for a machine in a natural and concise form using the mode or mixture of modes that most precisely convey the intended meaning and to adjust this mix to reflect communication needs [11]. This style of interaction should also increase the intuitiveness of the user interface [8]. Symmetric multimodality means that all input modes (speech, gesture, facial expression) are also available for output, and vice versa. A dialogue system with symmetric multimodality must not only understand and represent the user's multimodal input, but also its own multimodal output [16].

Extensive knowledge about how, why and when multimodality can be effectively used is already available. Reasons why multimodality is still not very widespread today could be the higher development effort and the lack of experts in the field of multimodal HCI. The number of use cases in which supporting different interaction modalities for human computer interaction is not appropriate should be very small.

CUI 2019, August 22–23, 2019, Dublin, Ireland

ACM ISBN 978-1-4503-7187-2/19/08…$15.00
https://doi.org/10.1145/3342775.3342801

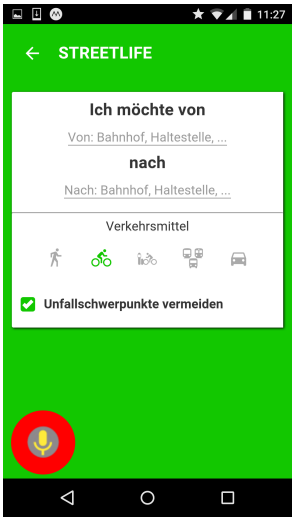

Figure 1: The STREETLIFE mobility app. The conversational view directly suggests the user a possible utterance ("I want to go from *origin* to *destination*.") which can be understood by the system

.

It can further be noted that in everyday usage the terms dialog and conversation are often used interchangeably and a dialog can be characterized as a conversation between two participants [4]. We therefore argue that the research findings in the area of multimodal dialog systems can be transferred to conversational HCI in large parts. Conversational user interfaces should make use of multimodality in order to increase e.g. robustness, efficiency, accessibility and intuitiveness.

4 FROM THEORY TO PRACTICE

So far we only argued on a theoretical basis. Using a few insights from our own research, we want to substantiate our demand for multimodal conversational interaction.

Our previous research showed that users tend to prefer input modalities with higher efficiency [14]. It was e.g. shown that users switch from GUI (graphical user interface) to VUI (voice user interface) usage if effort in terms of interaction steps can be saved. Based on these results we decided to make use of semi-conversational user interface in several projects (e.g. in [13]). By switching from a classical GUI to a conversational style of interacting with the user the efficiency of the interaction can be increased.

As an example Figure 1 shows the conversational view of the STREETLIFE app [13]. The app is a mobility application supporting the user to find itineraries from A to B. By tapping the microphone button the classical GUI (not shown) converts into the (shown) conversational view allowing the user to input origin and destination information via speech. In parallel to spoken input the graphical elements in the conversational view can be used to customize the modes of transport and to activate an "avoid accident hot-spots" feature. As a result of the conversational interaction the proposed itineraries are presented as a list in the GUI. This style of interaction exemplifies our demand for a reasonable use of multimodality as follows:

- for itinerary search a conversational spoken natural language input is more efficient then conversational text or classical GUI input.
- disambiguation for locations can be more efficiently resolved by means of the GUI, presenting possible options in a list
- the parallel use of the GUI while performing spoken input is a further advantage of multimodality
- as a part of the GUI, itinerary search results are presented in a list (not shown).

The semi-conversational user interface of the STREETLIFE app exemplifies how a conversational interaction style can be smoothly integrated into a graphical user interface. By using the strengths of both interfaces (GUI and CUI) the interaction gets more robust and efficient. According to results from multimodal interaction research this can also lead to higher user experience [18].

We are aware that the STREETLIFE App only partly supports our assumption, as it only supports the use of GUI and CUI for a well defined problem. The multimodal use of the app in different situation and context was not examined yet.

5 DISCUSSION AND CONCLUSION

Human human conversation is multimodal. A multimodal conversation between a human and a computer should therefore be more intuitive. Specific information is hard to efficiently integrate into a CUI, as e.g. a list or pictures can easier be presented within a GUI. Further, allowing a parallel combination of spoken input and GUI elements like buttons e.g. increases the input efficiency. Conversational user interfaces should be equipped with multimodal capabilities in order to utilize the advantages of multimodal interaction, and thus increase perceived user experience.

Our so far generally formulated position that conversational user interfaces should be multimodal leads to some open questions such as: what are the ethical, legal and social consequences if camera and microphone will be used as standard? How can the vast added complexity of adding multimodal support be handled from an engineering perspective? How can the context of the interaction be better taken into account? Can the context of interaction be used to define suitable multimodal support? As 98% of digital natives use the internet, will most of the "conversational natives" also use conversational interfaces, and will context adaptive multimodality be one of their requirements?

We leave these questions open to generate discussion and possible topics for future research. As an implication of our STREETLIFE app research, we would like to generate interest in defining a standard for different problem and context areas in which a particular combination of modalities should be used for multimodal CUIs in order to support the user in the best possible way.

ACKNOWLEDGMENTS

This research is part of the DAYSTREAM project of the research initiative "mFund", which is funded by the Federal Ministry of Transport and Digital Infrastructure (BMVI) of the Federal Republic of Germany under funding number 19F2031A.

REFERENCES

[1] Charles R Berger. 2008. Interpersonal communication. *The international encyclopedia of communication* (2008).

[2] Richard A Bolt. 1980. *"Put-that-there": Voice and gesture at the graphics interface*. Vol. 14. ACM.

[3] Susan E Brennan. 1991. Conversation with and through computers. *User modeling and user-adapted interaction* 1, 1 (1991), 67–86.

[4] Susan E Brennan. 2012. Conversation and dialogue. *Encyclopedia of the Mind. SAGE Publications* (2012).

[5] Philip R Cohen, Michael Johnston, David McGee, Sharon L Oviatt, Jay Pittman, Ira A Smith, Liang Chen, and Josh Clow. 1997. QuickSet: Multimodal Interaction for Distributed Applications. In *ACM Multimedia*, Vol. 97. 31–40.

[6] Mark L Knapp, Judith A Hall, and Terrence G Horgan. 2013. *Nonverbal communication in human interaction*. Cengage Learning.

[7] John Mylopoulos. 1992. Conceptual modelling and Telos. *Conceptual Modelling, Databases, and CASE: an Integrated View of Information System Development, John Wiley & Sons, New York, New York* (1992), 49–68.

[8] Anja B Naumann, Ina Wechsung, and Jörn Hurtienne. 2009. Multimodal Interaction: Intuitive, Robust, and Preferred?. In *IFIP Conference on Human-Computer Interaction*. Springer, 93–96.

[9] Dennis Perzanowski, Alan C Schultz, William Adams, Elaine Marsh, and Magda Bugajska. 2001. Building a multimodal human-robot interface. *IEEE intelligent systems* 16, 1 (2001), 16–21.

[10] Christina Regenbogen, Daniel A Schneider, Raquel E Gur, Frank Schneider, Ute Habel, and Thilo Kellermann. 2012. Multimodal human communication-targeting facial expressions, speech content and prosody. *Neuroimage* 60, 4 (2012), 2346–2356.

[11] Alexander I Rudnicky. 2005. Multimodal dialogue systems. In *Spoken multimodal human-computer dialogue in mobile environments*. Springer, 3–11.

[12] Stefan Schaffer, Benjamin Jöckel, Ina Wechsung, Robert Schleicher, and Sebastian Möller. 2011. Modality selection and perceived mental effort in a mobile application. In *Twelfth Annual Conference of the International Speech Communication Association*.

[13] Stefan Schaffer and Norbert Reithinger. 2014. Intermodal personalized Travel Assistance and Routing Interface. In *Mensch Computer 2014 - Tagungsband. Mensch & Computer, 14. Fachübergreifende Konferenz für Interaktive und Kooperative Medien - Interaktiv unterwegs - Freiräume gestalten, August 31-September 3, München, Germany*, Andreas Butz, Michael Koch, and Johann Schlichter (Eds.). De Gruyter Oldenbourg, 343–346.

[14] Stefan Schaffer, Robert Schleicher, and Sebastian Möller. 2015. Modeling input modality choice in mobile graphical and speech interfaces. *International Journal of Human-Computer Studies* 75 (2015), 21–34.

[15] Daniel Thalmann. 2000. The virtual human as a multimodal interface. In *Proceedings of the working conference on Advanced visual interfaces*. Citeseer, 14–20.

[16] Wolfgang Wahlster. 2006. Dialogue systems go multimodal: The SmartKom experience. In *SmartKom: foundations of multimodal dialogue systems*. Springer, 3–27.

[17] Martin Warren. 2006. *Features of naturalness in conversation*. Vol. 152. John Benjamins Publishing.

[18] Ina Wechsung and Anja B Naumann. 2009. Evaluating a multimodal remote control: The interplay between user experience and usability. In *2009 International Workshop on Quality of Multimedia Experience*. IEEE, 19–22.