

## Zusammenfassung

Interaktive Dialogsysteme, die über gesprochene Sprache, graphische Schnittstellen (GUI) oder multimodale Nutzerschnittstellen bedient werden, finden immer weitere Verbreitung. Aktuelle Beispiele für solche multimodalen Dialogsysteme sind virtuelle persönliche Assistenten, wie Apples Siri, Google Now oder Microsoft Cortana.

Im Rahmen des Usability Engineering Lifecycles bedeutet eine kontinuierliche Usability-Evaluation, dass die Usability einer Nutzerschnittstelle auch nach kleinen Änderungen an der selbigen getestet wird. Mit klassischen Methoden des Usability Engineering, wie zum Beispiel einer Expertenevaluation oder einem Nutzertest, wäre dies nur unter hohem Aufwand bzw. gar nicht möglich. Eine Alternative bietet hier der Einsatz von automatischen Methoden zur Usability-Evaluation (AUE), insbesondere solchen, die auf der Simulation von Nutzerverhalten beruhen. Im Rahmen dieser Arbeit wird daher Folgendes untersucht:

- Wie kann Nutzungsverhalten simuliert werden?
- Wie können Parameter zu Beschreibung multimodaler Mensch-Computer Interaktionen automatisch erfasst werden?
- Wie können Nutzerverhaltenssimulationen bezüglich ihrer Ähnlichkeit zu empirischen Daten evaluiert und miteinander verglichen werden?

Zunächst wird dazu die Nutzerverhaltenssimulation (NVS) eines vorhandenen Werkzeugs zur AUE erweitert. Dazu werden technik- soziologische Methoden auf die Durchführung und Auswertung einer Nutzerstudie mit einem Smart Home System angewendet. Die resultierenden Simulationsergebnisse werden mit denen bestehender NVS und den Interaktionen der Nutzerstudie verglichen. Dabei wird gezeigt, dass die neu entwickelte Nutzersimulation den bestehenden Ansätzen hinsichtlich der Vorhersage von Maßen der Dialogeffizienz und -effektivität überlegen ist.

Des Weiteren werden ein parametrisches Datenmodell und ein darauf beruhendes Framework beschrieben. Diese dienen dazu, multimodale Mensch-Computer Interaktionen einheitlich zu beschreiben und für eine Usability-Untersuchung zur Verfügung zu stellen. Die Anwendbarkeit des entwickelten Frameworks wird anhand zweier empirischer Studien und eines Simulationsexperiments (mit demselben AUE Werkzeug wie oben) gezeigt.

Schließlich wird ein zweistufiges Verfahren zur Evaluation von NVS vorgeschlagen sowie angewendet und validiert. Das Verfahren beruht auf der Berechnung der Distanz zwischen jeweils zwei Dialogkorpora und dem anschließenden Vergleich der berechneten Distanzen. Im ersten Schritt wird ein dafür geeignetes Distanzmaß ermittelt. Dieses Distanzmaß wird dann im zweiten Schritt für die Bewertung der NVS verwendet.

Die Validität des Verfahrens zur Evaluation von Nutzerverhaltenssimulationen wird durch die Anwendung auf empirische und simulierte Dialoge geprüft. Die Anwendung auf Dialoge zwischen realen Nutzern und einem Sprachdialogsystem sowie zwischen zwei Varianten einer Nutzersimulation und demselben Sprachdialogsystem zeigt, dass das Verfahren korrekte Evaluationsergebnisse liefert

## Abstract

Interactive dialogue systems that are operated by spoken language, graphical user interfaces (GUI), or multimodal user interfaces propagate more and more. Currently, virtual personal assistants as Apple's Siri, Google Now, or Microsoft Cortana are well-known examples of such multimodal dialogue systems.

In the frame of the usability engineering life cycle, the user interface of a system should be tested after each change that is related to the human-machine interaction. Such a frequently testing is named continuous usability evaluation. A continuous usability evaluation causes high efforts or is even impossible, when using traditional evaluation methods as an expert evaluation or a user study. The usage of *automatic usability evaluation* (AUE) methods is an alternative to traditional methods, to enable the continuous usability evaluations of user interfaces. Here, AUE methods that base on the simulation of user behaviour are particularly auspicious.

Thus, the following three questions will be examined in the frame of this thesis:

- How can the behaviour of users be simulated?
- How can parameters describing multimodal human-computer interaction be captured automatically?
- How can user behaviour simulations be evaluated and be compared with each other about their similarity to empirically collected data?

To answer these questions, the *user behaviour simulation* (UBS) of an existing tool for AUE is extended. This extension is based upon a user study with a smart home system. Technical-sociological methods are used for the execution of the study and the analysis of the collected data. A comparison of the resulting UBS with former UBSs, as well as the empirical data, shows that the new simulation approach outperforms the former simulation. The improvement affects the prediction of dialogue metrics that are related to dialogue efficiency and dialogue effectiveness.

Furthermore, a parameter-based data model, as well as a related framework, is described. Both are used to uniformly describe multimodal human-computer interactions and to provide such descriptions for usability evaluations. The applicability of the framework is proofed by its application to two empirical studies and a simulation experiment. The latter is realised by the integration of the framework into the same AUE tool as mentioned in the previous paragraph.

Finally, a new two-stage method for the evaluation of UBSs is proposed, applied, and validated in this thesis. The method bases on the computation of a distance measures between two dialogue corpora and the pair-wise comparison of distances among several dialogue corpora (e. g. from two simulations and one empirical study). The first step of the developed method is the determination of an appropriate distance measure for the present data. In the second step, the determined distance measure is used to evaluate one or several UBSs.

The application of the UBS evaluation method to empirical and simulated dialogues and the analysis of the results show the validity of the method. In fact, its application to dialogues between human users and a spoken dialogue system, as well as between two variants of a UBS and the same dialogue system, shows that the method yields correct evaluation results.