Evaluating Personal Assistants on Mobile devices
Conceptual Paper

Julia Kiseleva
UserSat.com & University of Amsterdam
Amsterdam, the Netherlands
j.kiseleva@uva.nl

Maarten de Rijke
University of Amsterdam
Amsterdam, the Netherlands
derijke@uva.nl

ABSTRACT
The iPhone was introduced only a decade ago in 2007, but has fundamentally changed the way we interact with online information. Mobile devices differ radically from classic command-based and point-and-click user interfaces, now allowing for gesture-based interaction using fine-grained touch and swipe signals. Due to the rapid growth in the use of voice-controlled intelligent personal assistants on mobile devices, such as Microsoft’s Cortana, Google Now, and Apple’s Siri, mobile devices have become personal, allowing us to be online all the time, and assist us in any task, both in work and in our daily lives, making context a crucial factor to consider.

Mobile usage is now exceeding desktop usage, and is still growing at a rapid rate, yet our main ways of training and evaluating personal assistants are still based on (and framed in) classical desktop interactions, focusing on explicit queries, clicks, and dwell time spent. However, modern user interaction with mobile devices is radically different due to touch screens with gesture- and voice-based control and the varying context of use, e.g., in a car, by bike, often invalidating the assumptions underlying today’s user satisfaction evaluation.

There is an urgent need to understand voice- and gesture-based interaction, taking all interaction signals and context into account in appropriate ways. We propose a research agenda for developing methods to evaluate and improve context-aware user satisfaction with mobile interactions using gesture-based signals at scale.

CCS CONCEPTS
•Information systems → Users and interactive retrieval; Evaluation of retrieval results; •Human-centered computing → HCI design and evaluation methods;

KEYWORDS
Personal assistants, evaluation, conversational search

1 INTRODUCTION
Recent years have witnessed an explosive growth in the usage of gesture- and voice-controlled devices. The usage of mobile phones increased five-fold from 11.78% in October 2012 to 53.01% in December 2016, and it overtook the usage of desktops in October 2016. This spike happens due to availability of personal assistants. Spoken dialogue systems have been thoroughly studied in the literature [48, 58–60]. However, it has only been in recent years that a new generation of personal assistants, powered by voice, such as Apple’s Siri, Microsoft’s Cortana, Google Now, have become common and popular on mobile devices. One of the reasons for the increased adoption is the recent significant improvement in accuracy of automatic speech recognition [50].

Evaluation of effectiveness is an essential part of developing any interactive system such as web search and e-commerce applications. Modern evaluation methods, which were developed for desktops, heavily rely on interaction data, e.g., explicit queries and clicks that are massively logged [11–13, 69]. However, interaction signals on mobile devices are different due to the context of use and gesture- and voice-based control, like swipes, touch and voice conversations [4, 6, 38, 51, 63, 65]. As a consequence, there is an urgent need to develop new scalable techniques for understanding context-aware user satisfaction for gesture- and voice-controlled devices.

Our aim is to exploit voice- and gesture-based signals, trackable at large scale, for understanding context-aware user satisfaction with personal assistants. The overall aim leads to three specific research questions:

• RQ1: How to model interaction with gesture- and voice-controlled devices?
• RQ2: How to define context-aware user satisfaction with personal assistants in mobile environments?
• RQ3: How to predict context-aware user satisfaction with personal assistants using gesture-based signals on mobile devices?

2 SCIENTIFIC CHALLENGES
We list four central challenges that provide the background for the research questions listed above.

How and why to evaluate the effectiveness of an personal assistant? Previously, a common practice for evaluation was to create a ‘gold’ standard [55]. In modern personal assistants, there may be no general “correct” answers since the answers are highly personalized and contextualized, e.g., to a user’s location [3, 34, 66]
or a user’s past preferences [25, 36, 68]. User satisfaction is widely adopted as a subjective measure of the quality of the search experience [30]. Nowadays, it has become common practice to evaluate personal assistants on desktops by analysing interaction signals such as clicks (if users like they click) and dwell time (the actual length of time that a visitor spends on a page) [2, 11–13, 18, 23, 28, 29].

Currently, the research community is facing a challenge to evaluate user satisfaction at scale. Very large scale online controlled experiments, such as A/B testing and interleaving, have become a widely used technique for controlling and improving search quality based on data-driven decisions [26]. This methodology has been adopted by many leading companies [5, 14, 17, 57]. User behavior in voice- and gesture-controlled environment is very different when compared to desktops [38, 39, 42, 44, 64, 65], but our understanding of this difference is still fragmented at best. Unlike desktop computers with large displays and mouse-keyboard interactions [20–22, 47, 54], personal assistants come on mobile devices that have smaller displays and offer voice commands and a variety of gesture interactions, e.g., touch: swiping and zooming. Moreover, user behavior on mobiles is very context-dependent [71]. Therefore, traditional evaluation methods are not applicable for the growing mobile environment.

The fundamental problem limiting current progress in developing personal assistants for mobile environment is the lack of scalable methods to infer user satisfaction.

Why is context-awareness needed for evaluating user satisfaction? Kelly [30] proposes the following definition: “satisfaction can be understood as the fulfillment of a specified desire or goal.” Online user behavior is highly:

- context-dependent [1, 33–35, 37, 53, 67];
- sensitive to changes in the outside world [31, 32].

In a mobile environment, users are dealing with a much richer space of potential contextual situations, e.g., while driving, in the bus, on the way, a slow connection, compared to the relatively static desktop environment. These conditions have a great impact on ‘mobile’ user satisfaction. Similar experiences can be satisfying in one situation (Figure 1(A)'), e.g., a user is sitting in a hotel lobby with a fast wifi connection, and it can be totally frustrating in another situation (Figure 1(A)'''), e.g., when the same user is driving and having a slow data connection.

Therefore, situational context has to be studied in far greater details, allowing us to reason about how a user’s current environment impacts his satisfaction with personal assistants.

How to evaluate context-aware user satisfaction at scale? Eye-tracking techniques have been successfully used to gain an initial understanding of user interactions with mobile devices [43, 44] but they cannot applied at scale. In contrast, user gestures and voice commands can be collected and analysed at scale [65]. We suggest to model advanced voice- and gesture-based signals to predict context-aware user satisfaction for millions of users, which can be easily plugged-in into A/B testing platforms [5, 15, 16, 40].

Why analysing gestures is the way to infer context-aware user satisfaction? Analysing clicks heat maps, e.g., in Figure 1(B), is quite tricky. Because the screen size is small, it is difficult not to click inadvertently, or conversely, to have difficulty clicking an item. Analysing gesture-based patterns is a better way to infer user satisfaction, as it helps to decipher hidden behavioral aspects, e.g., swipes in Figure 1(C) clearly belong to left- and right-handed persons. Moreover, touch signals are extremely useful to predict user satisfaction for mobile search [38, 65]. Movements of the human body, e.g., gestures, reflect emotions [7, 8] that are closely connected with user satisfaction (Figure 1(A)). User emotions are used to evaluate voice-controlled systems [41, 52], e.g., changes in user intonation [56, 70]. We propose to exploit gesture-based and voice-based interaction to infer context-aware user satisfaction in mobile environment because they:

- are the primary ways to interact with mobile devices;
- are very sensitive to situational and behavioral aspects (Figure 1(C));
- reveal user emotions: satisfaction (Figure 1(A)'+') and frustration (Figure 1(A)'--');
- are highly scalable, both w.r.t. collection and analysis.

Figure 1: Illustrations of (A) how to define user satisfaction, (B) where users click on the mobile screen, and (C) how touches are tracked
We propose ventures into a new area of research, by moving beyond mobile interactions at the session-level. Then, one should consider movement events, e.g., the ‘shake’. We could use this rich set of gestures related users’ interaction with personal assistants. Capturing touch events (Figure 2(A)) is difficult in practice [27]; however, it is possible to infer touch-based interactions based on the mobile viewport [38, 65]. For instance, if an element is visible in the viewport at some point in time and then no longer visible, one can infer that a gesture must have taken place. To get a complete view of user gestures, we should capture that a gesture has taken place. To get a complete view of user satisfaction based on complex and subtle interaction patterns. Obtaining insight into this value is crucial if we want to evaluate user satisfaction based on complex and subtle interaction patterns. We need to encompass all gesture- and voice-based features to build an advanced representation of user interactions in a mobile setting.

RQ1: How to model interaction with gesture- and voice-controlled devices? We need to encompass all gesture- and voice-based features related users’ interaction with personal assistants. Capturing touch events (Figure 2(A)) is difficult in practice [27]; however, it is possible to infer touch-based interactions based on the mobile viewport [38, 65]. For instance, if an element is visible in the viewport at some point in time and then no longer visible, one can infer that a gesture must have taken place. To get a complete view of user gestures, we should capture (1) orientation and acceleration of a device in space (Figure 2(B)) that will allow us to model users’ hands position; (2) the GPS signal to infer changes in user locations; (3) movement events, e.g., the ‘shake.’ We could use this rich set of gesture-based features to build an advanced representation of interactions in a mobile setting.

RQ2: How to define context-aware user satisfaction with personal assistants in mobile environments? As a starting point, we can start from the approach presented in [38] to define user satisfaction with mobile interactions at the session-level. Then, one should extend it by introducing context-aware [33, 34, 37] and changing environments [31, 32]. In addition to unsupervised logs, dedicated user experiments should be conducted to gather rich, explicitly annotated data for further analysis and validation.

REFERENCES

1. Gediminas Adomavicius and Alexander Tuzhilin. 2010. Context-Aware Recommender Systems. CARS (2010).
2. Eugene Agichtein, Eric Brill, and Susan T. Dumais. 2006. Improving web search ranking by incorporating user behavior information. In SIGIR 19–26.
3. Ana O. Ables and Francisco C. Pereira. 2012. Making sense of location context. In ContextDD.
4. Myroslav Bachynskyi, Antti Oulasvirta, Gregorio Palmas, and Tino Weinkauf. 2014. Is motion capture-based biomechanical simulation valid for HCI studies? study and implications. In CHI 3215–3224.
5. Eytan Bakshy and Dean Eckles. 2013. Uncertainty in online experiments with dependent data: an evaluation of bootstrap methods. In KDD 1303–1311.
6. Joanna Bergstrom-Lehtovirta and Antti Oulasvirta. 2014. Modeling the functional area of the thumb on mobile touchscreen surfaces. In SIGCHI 1991–2000.
7. Daniel Bernhardt. 2010. Emotion inference from human body motion. Ph.D. Dissertation. University of Cambridge, UK.
8. Daniel Bernhardt and Peter Robinson. 2019. Detecting Emotions from Connected Action Sequences. In IJCV 1–11. https://doi.org/10.1007/s11263-005-0036-7.
9. Alexey Borisov, Ilya Markov, Maarten de Rijke, and Pavel Serdyukov. 2016. A Context-aware Time Model for Web Search. In SIGIR 205–214.
10. Alexey Borisov, Ilya Markov, Maarten de Rijke, and Pavel Serdyukov. 2016. A Neural Click Model for Web Search. In WWW International World Wide Web Conferences Steering Committee, 531–541.
11. Olivier Chapelle, Donald Metzler, Ya Zhang, and Pierre Crimpson. 2010. Expected reciprocal rank for graded relevance. In CIKM 621–630.
12. Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. 2015. Click Models for Web Search. Morgan & Claypool Publishers.
13. Aleksandr Chuklin, Pavel Serdyukov, and Maarten de Rijke. 2013. Click model-based information retrieval metrics. In SIGIR 493–502.
14. Alex Deng, Tianxi Li, and Yu Guo. 2014. Statistical inference in two-stage online controlled experiments with treatment selection and validation. In WWW 609–618.
15. Alex Deng, Jianfan Lu, and Jonathan Litz. 2017. Trustworthy analysis of online A/B tests: Pitfalls, challenges and solutions. In WSDM.
16. Alexey Drutsa, Gleb Gusev, and Pavel Serdyukov. 2015. Engagement Periodicity in Search Engine Usage: Analysis and its Application to Search Quality Evaluation. In WSDM. 27–36.
17. Alexey Drutsa, Gleb Gusev, and Pavel Serdyukov. 2015. Future User Engagement Prediction and Its Application to Improve the Sensitivity of Online Experiments. In WWW 256–266.
18. Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan T. Dumais, and Thomas White. 2005. Evaluating implicit measures to improve web search. TOIS 23, 2 (2005). 147–168.
19. David Grosf and Zoraide Callejas. 2016. A Neural Network Approach to Intention Modeling for User-Adapted Conversational Agents. Comp. Int. and Neurosc. 2016 (2016), 8402127:1–8402127:11.
20. Qi Guo and Eugene Agichtein. 2010. Ready to buy or just browsing? Detecting Web Searcher Goals from Interaction Data. In SIGIR 130–137.
21. Qi Guo and Eugene Agichtein. 2012. Beyond Dwell Time: Estimating Document Relevance from Cursor Movements and other Post-click Searcher Behavior. In WWW. ACM Press, New York, New York, USA, 569–578.
22. Qi Guo, Dmitry Lugan, and Eugene Agichtein. 2012. Predicting web search success with fine-grained interaction data. In CIKM 2050–2054.
